UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

LÊNIN FERREIRA BARBOSA

**Caracterização automática de grupos acadêmicos utilizando bibliometria e
análise de redes sociais**

São Paulo

2017

LÊNIN FERREIRA BARBOSA

**Caracterização automática de grupos acadêmicos utilizando bibliometria e
análise de redes sociais**

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 23 de outubro de 2017. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2017

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Barbosa, Lênin Ferreira

Caracterização automática de grupos acadêmicos utilizando bibliometria e análise de redes sociais / Lênin Ferreira Barbosa ; orientador Luciano Antonio Digiampietri. – 2017.

64 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida

1. Redes sociais - Análise. 2. Bibliometria - Ferramentas - Automação. 3. Ciência da computação. I. Digiampietri, Luciano Antonio. II. Título.

CDD 22.ed.– 303.4833

CRB-8 4936

Dissertação de autoria de LÊNIN FERREIRA BARBOSA, sob o título “**Caracterização automática de grupos acadêmicos utilizando bibliometria e análise de redes sociais**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 23 de outubro de 2017 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Universidade de São Paulo

Presidente

Prof. Dr. Jesús Pascual Mena Chalco

Universidade Federal do ABC

Prof. Dr. Roberto Hirata Junior

Universidade de São Paulo

Prof. Dr. Daniel de Angelis Cordeiro

Universidade de São Paulo

Dedico esse trabalho à amada mãe, Marli do Nascimento Ferreira, a qual sempre quis ter um filho mestre.

Agradecimentos

Para você que está lendo essa dissertação, eu lhe deixo um aviso. Caso pretenda iniciar um projeto dessa magnitude, eu recomendo a priori que seu estado emocional, profissional e pessoal estejam bem resolvidos. Esse trabalho foi feito durante um período difícil em minha vida. Fiz muito questionamentos e duvidei muito de minha capacidade durante a escrita.

Eu agradeço em primeiro lugar ao meu orientador e amigo Luciano Digiampietri, por ter sempre me ouvido durante as crises, por ter me apoiado e ter me estimulado a sempre seguir em frente. A maioria das pessoas teria se cansado, principalmente quando eu mesmo havia desistido de mim. Ele confiou e me respeitou no meu tempo, com a calma e serenidade que tanto me faltava.

Agradeço a minha mãe por sempre ter se dedicado de corpo e alma para que eu pudesse alcançar esse objetivo. Que cuidou de minhas roupas, casa e comida para que eu pudesse me dedicar ao meu projeto. Um conforto essencial e muito carinho para ver seu filhote crescer.

Agradeço ao meu amigo e irmão William Takahiro Maruyama, por ter me ouvido e compartilhado suas experiências e conquistas comigo. Somos amigos desde adolescência e estamos crescendo juntos ainda. Obrigado por esses anos.

Agradeço ao meu namorado Adriano Rodrigues dos Santos por estar comigo, por rir e chorar, por cuidar de mim e me ajudar a dar novos passos na minha independência.

Reconheço que esse trabalho está longe de ser meu máximo potencial, o que é um ótimo sinal, afinal serve como um estímulo para continuar a aprender. Hoje eu aprendi um novo caminho e terminei o mestrado melhor do que comecei. Concertei alguns aspectos da minha vida e tantos mais ainda tenho que melhorar. Agora caminho um passo de cada vez, com novos horizontes em mente.

Resumo

BARBOSA, Lênin Ferreira. **Caracterização automática de grupos acadêmicos utilizando bibliometria e análise de redes sociais**. 2017. 65 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017.

A avaliação acadêmica (de docentes, departamentos, programas de pós-graduação ou instituições) é uma atividade rotineira e utilizada para, por exemplo, a concessão de recursos para um projeto, definição da verba que será liberada a um dado programa de pós-graduação ou mesmo para a autorização do oferecimento de um curso de doutorado para um programa de pós-graduação. Contudo, determinar a qualidade do ensino não é uma tarefa trivial, com metodologias comparativas ainda em estudo. Dois destes critérios são informações bibliométricas e análise das redes de coautorias dos programas. Assim, o desenvolvimento de métodos e ferramentas para automatizar parte da avaliação ou mesmo para identificar a importância de diferentes métricas podem tanto auxiliar no processo de avaliação como servir para auxiliar as pessoas ou os grupos que serão avaliados. Neste trabalho foram desenvolvidas ferramentas para a realização automática de análise bibliométrica e de redes sociais para a caracterização e ranqueamento de grupos acadêmicos, bem como investigar a relação entre diferentes medidas e alguns ranqueamentos acadêmicos existentes. As ferramentas foram avaliadas considerando departamentos internacionais e programas de pós-graduação da área de Ciência da Computação. Os resultados de nosso estudo indicaram quais métricas contribuíram positivamente e negativamente para a posição dos ranqueamentos dos programas.

Palavras-chave: Análise de grupos. Análise de redes sociais. Bibliometria.

Abstract

BARBOSA, Lênin Ferreira. **Automatic characterization of academic groups using bibliometrics and social networks analysis**. 2017. 65 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2017.

The academic evaluation (of researchers, departments, graduate programs or institutions) is a routine activity and used for the granting of resources for a project, definition of the funds that will be released to a given graduate program or even for the authorization of the offering of a doctorate course. However, this activity is complex with comparative methodologies still in evaluation. Two criteria used are bibliometrics information and scientific social network analysis. Thus, the development of methods and tools to automate part of the evaluation or even to identify the importance of different metrics for the evaluation can both aid the evaluation process and serve as a guide for the researchers or groups that will be evaluated. In this work, tools were developed for performing automatically bibliometric and social networks analysis for the characterization and ranking of academic groups, as well as to investigate the relationship between different measures and some existing academic rankings. The tools were evaluated considering international departments and graduate programs in the Computer Science area. Our results showed which metrics contributed positively and negatively in the rank positioning of the programs.

Keywords: Group analysis. Social networks analysis. Bibliometrics.

Lista de figuras

Figura 1 – Bases utilizadas na literatura	27
Figura 2 – Correlação entre características. As linhas e as colunas são as características apresentadas na tabela 3.	41
Figura 3 – Rede de coautorias - departamentos	46
Figura 4 – Legenda de cores	46
Figura 5 – Rede de coautorias - todas as arestas	46
Figura 6 – Rede de coautorias - apenas arestas entre pesquisadores do mesmo departamento	46
Figura 7 – Ranqueamento dos programas de acordo com as métricas utilizadas . .	47
Figura 8 – Correlações entre as métricas analisadas	48
Figura 9 – Plotagem das duas componentes principais	51
Figura 10 – Correlações entre as métricas dos programas nacionais	52
Figura 11 – Correlações das métricas dos programas nacionais com a nota Capes .	53
Figura 12 – Plotagem das duas componentes principais dos programas nacionais . .	54
Figura 13 – Resultado da regressão linear aplicada a cada métrica analisada	56
Figura 14 – Estimativa da nota CAPES considerando a regressão linear para métricas individuais	57

Lista de tabelas

Tabela 1 – Métodos de avaliação utilizadas de acordo com os objetivos do estudo .	28
Tabela 2 – Departamentos ou Programas de Computação Analisados	34
Tabela 3 – Características extraídas das citações.	37
Tabela 4 – Medidas Bibliométricas Utilizadas	38
Tabela 5 – Métricas da Análise de Redes Sociais Utilizadas	38
Tabela 6 – Classificação dos programas de pós-graduação brasileiros	39
Tabela 7 – Características ranqueadas por seletores de atributos	42
Tabela 8 – Desempenho da solução proposta.	43
Tabela 9 – Matriz de confusão	44
Tabela 10 – Resultado da classificação utilizando Naive-Bayes	58
Tabela 11 – Matrix de confusão da classificação dos programas nacionais	58

Sumário

1	Introdução	12
1.1	<i>Objetivos</i>	13
1.1.1	Objetivo Geral	13
1.1.2	Objetivos específicos	13
1.1.3	Limitações	13
1.1.4	Resumo da Metodologia	14
1.1.5	Organização do Texto	15
2	Conceitos Básicos	16
2.1	<i>Desambiguação</i>	17
2.2	<i>Bibliometria</i>	19
2.3	<i>Indicadores para análise de Redes Sociais</i>	20
2.4	<i>Bases de dados bibliográficas</i>	21
3	Revisão da Literatura	22
3.1	<i>Planejamento</i>	22
3.2	<i>Condução</i>	23
3.3	<i>Extração/Documentação</i>	25
3.3.1	Documentação	26
3.4	<i>Trabalhos Complementares</i>	28
3.5	<i>Considerações Finais sobre a Revisão</i>	30
4	Materiais e Métodos	32
4.1	<i>Materiais</i>	32
4.2	<i>Métodos</i>	35
4.2.1	Desambiguação	35
4.2.2	Análises Bibliométricas	37
5	Resultados	40
5.1	<i>Desambiguação</i>	40
5.2	<i>Caracterização dos grupos</i>	45
5.2.1	Programas e departamentos internacionais	45

5.2.2	Programas nacionais	51
6	Conclusão	59
	Referências ¹	62

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

De acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), no censo de 2011, havia 2.365 instituições superiores em funcionamento. Em 2012, último registro atualizado, foram 2.416. O aumento de cursos superiores propiciou a uma parcela maior da população o acesso a esse nível de escolaridade. O próximo passo na vida acadêmica consiste na pós-graduação, *lato e stricto sensu*. Com o maior ingresso de pessoas, maior a importância em demonstrar a qualidade do ensino ante a sociedade. Contudo, determinar a qualidade do ensino não é uma tarefa trivial, com metodologias comparativas ainda em estudo (LAENDER et al., 2008).

Limitando nosso foco de estudo aos programas de pós-graduação, mais especificamente os de *stricto sensu* em ciências da computação, pode-se encontrar sistemas de ranqueamentos que combinam diferentes parâmetros em suas análises para categorizar os programas. Alguns ranqueamentos internacionais (por exemplo, The Guardian¹, Shanghai², Top Universities³) divulgam os critérios estabelecidos para a classificação. Um destes critérios são informações bibliométricas dos programas. Porém, de um modo geral, os dados utilizados e a metodologia para o cálculo do ranqueamento não são divulgados publicamente.

Ao optar-se na avaliação bibliométrica, a tarefa de obter-se os dados brutos a serem utilizados pode ser não trivial. De acordo com a tipo de informação a ser levantada pode-se ter problemas na integração dos dados durante o cálculo. Por exemplo, ao avaliar-se a produção bibliográfica, é necessário que os autores dos artigos sejam identificados univocamente. Esse processo, nomeado desambiguação de entidades na literatura, investiga as maneiras mais eficientes de identificar autores em bases de dados científicas.

Em relação à produção intelectual, há também a necessidade de uma metodologia para determinar qual o impacto dos artigos produzidos pelas instituições. O uso de análise bibliométrica permite avaliar quantitativamente a difusão da informação, o que pode ser usado como base para o estudo do impacto (DIODATO, 1994).

A lacuna encontrada na literatura consiste no caminho reverso. É possível, a partir de uma análise bibliométrica automática, caracterizar um departamento ou programa de

¹ The Guardian: <http://www.theguardian.com/education/universityguide>

² Academic Ranking of World Universities: <http://www.shanghairanking.com/>

³ Worldwide university rankings: <http://www.topuniversities.com/>

pós-graduação? Considerando sistemas internacionais de avaliação, pode-se identificar as características das universidades brasileiras e compará-las ante as universidades melhores posicionadas nesses ranqueamentos?

1.1 *Objetivos*

1.1.1 Objetivo Geral

Objetivo deste trabalho foi desenvolver e integrar ferramentas que permitam a realização automática de análise bibliométrica e de redes sociais para a caracterização e o ranqueamento de grupos acadêmicos, bem como investigar a relação entre diferentes medidas e quatro ranqueamentos acadêmicos existentes (três internacionais e um nacional).

1.1.2 Objetivos específicos

O objetivo principal foi dividido em cinco objetivos específicos.

- Identificar o estado da arte na análise bibliométrica de grupos sociais acadêmicos;
- Desenvolver ferramentas para identificar corretamente (desambiguar) os dados dos participantes dos grupos de pesquisas e suas produções;
- Caracterizar os grupos acadêmicos de acordo com as medidas bibliométricas e sociométricas;
- Analisar a correlação entre as métricas e os sistemas de ranqueamento.
- Utilizar uma abordagem baseada em inteligência artificial para o posicionamento de cada grupo acadêmico nos ranqueamentos de acordo com as métricas estudadas.

1.1.3 Limitações

Os ranqueamentos utilizados nesta dissertação foram TOP Universities, The Guardian e Shanghai. Todos esses ranqueamentos abrangem ao menos 100 universidades. A apresentação geral do método de avaliação entre eles é similar quanto aos critérios avaliados, porém a não divulgação dos detalhes da metodologia de cálculo prejudica a comparação com outros resultados.

A base escolhida para identificar a produção acadêmica na área de Ciência da Computação (área do estudo de caso realizado) foi a Digital Bibliography & Library Project⁴ (DBLP). Essa base coleta as referências bibliográficas contidas nas principais revistas e eventos científicos da área de computação e correlatos, identifica os autores e os indexa numa base de dados.

1.1.4 Resumo da Metodologia

A metodologia desta dissertação de mestrado envolveu a realização de uma revisão sistemática, no intuito de avaliar as atuais aplicações de análises de redes sociais na análise de importância ou influência de indivíduos ou grupos e quais os principais métodos de validação de resultados. Essa literatura foi complementada com estudos de análise de grupos. Considera-se um grupo nessa dissertação o conjunto de pesquisadores que pertençam a um programa de pós-graduação ou a um departamento.

A rede social acadêmica foi construída com base em coautorias. Inicialmente, foram selecionados 20 departamentos internacionais e oito programas de pós-graduação em Ciência da Computação com notas 6 ou 7 atribuídas pela CAPES. Adicionalmente, um estudo adicional foi realizado considerando os 61 programas de pós-graduação brasileiros na área de Ciência da Computação que possuíam mestrado acadêmico ou doutorado durante o triênio 2010-2012.

Apesar dos registros da DBLP serem indexados e terem passado por um processo de desambiguação, alguns desses ainda apresentam incoerências quanto à identificação do autor. Em particular foram identificados casos em que duas chaves referenciavam um único pesquisador. Para resolver esse tipo de problema, foi realizado um estudo sobre desambiguação, com o desenvolvimento de um algoritmo para a identificação automática dos registros em duplicidade. A validação dos resultados foi feita com o auxílio de dados da Plataforma Lattes.

Para esse trabalho calculamos métricas bibliométricas referentes a participação em coautoria, número de publicações entre outras referenciadas no 4. Estes cálculos foram realizados após a resolução dos autores e de suas produções. Os dados da DBLP em complemento com informações do Microsoft Academic Search⁵ foram suficientes para o

⁴ <http://dblp.uni-trier.de/db/>

⁵ <http://academic.research.microsoft.com/>

cálculo das métricas bibliométricas utilizadas neste trabalho. Calculamos também algumas métricas de rede social, como densidade, assortatividade entre outras também referenciadas no capítulo 4.

A última etapa consistiu na categorização dos departamentos ou programas de acordo com as métricas e na apuração da influência dessas quanto ao posicionamento dos programas no ranqueamento.

1.1.5 Organização do Texto

Essa dissertação está dividida em seis capítulos, incluindo este primeiro capítulo introdutório. O capítulo 2 apresenta os conceitos básicos. O capítulo 3 detalha o processo de revisão sistemática e estudos complementares. O capítulo 4 contém os métodos e materiais utilizados. O capítulo 5 apresenta e discute os resultados obtidos. Por fim, o capítulo 6 apresenta as conclusões e trabalhos futuros.

2 Conceitos Básicos

A representação de uma rede social é comumente realizada por meio de grafos, nos quais os vértices representam os indivíduos (pessoas, grupos ou instituições) e as arestas representam as relações entre indivíduos. De acordo com os objetivos de estudo acerca da rede social, opta-se pela utilização de determinadas métricas (DIGIAMPIETRI, 2015), assim como qual tipo de grafo, de acordo com as definições a seguir.

De acordo com Cormen et al. (2001), um grafo G é um conjunto composto por dois outros conjuntos $G = (V, E)$. Isto é, um grafo G é composto por um conjunto de vértices ou nós V e um conjunto de arestas E , sendo que cada aresta representa uma conexão entre dois elementos do conjunto V . De acordo com a esparsidade do grafo, pode-se optar por duas representações: lista de adjacências, recomendada para grafos muito esparsos, ou matriz de adjacências, para densos (CORMEN et al., 2001). A matriz de adjacência tem a vantagem de permitir a consulta imediata sobre a existência de uma aresta entre dois vértices. Além disso, dependendo da métrica a ser calculada, a utilização da matriz pode ser mais recomendada do que a lista de adjacências.

Grafos podem ser direcionados ou não. Em grafos não direcionados, caso o vértice u esteja conectado com v , v também está conectado a u . Em grafos direcionados, u pode estar conectado a v , sem que v esteja conectado a u . Grafos podem ser simples ou ponderados. Um grafo ponderado é aquele cujas arestas recebem pesos para indicar diferentes níveis de relacionamento (CORMEN et al., 2001).

O *caminho* de um grafo consiste no subconjunto de arestas que ligam um vértice u a outro v , sendo que, para um grafo não ponderado, o caminho mínimo ou geodésico consiste naquele com o menor número de arestas entre os vértices u e v (DIESTEL, 2006). O componente conexo de um grafo é de um subgrafo no qual todos os vértices estão conectados por ao menos um caminho. Denomina-se componente gigante o componente conexo que possui a maior quantidade de vértices. Muitas análises de redes sociais (e algumas de suas métricas) são realizadas considerando-se apenas este componente (WASSERMAN; FAUST, 2009).

Um grafo é considerado completo se existirem arestas ligando cada um dos seus pares de vértices, o que implica a existência de $V * (V - 1)/2$ arestas. Uma **clique** de um grafo é um subgrafo completo desse grafo (SZWARCFITER, 1986).

O restante deste capítulo descreve quatro assuntos necessários para o entendimento do presente trabalho: Desambiguação de nomes; Bibliometria; Análise de redes sociais; e Bases de dados acadêmicas.

2.1 Desambiguação

A identificação correta de autores é essencial ao trabalho bibliométrico. Contudo, a ambiguidade é um problema, pois pode fazer com que dois autores sejam identificados como apenas um, assim como o caso contrário, no qual a produção de um autor está fragmentada em diferentes registros.

Para exemplificar, ao se procurar pela produção em uma ferramenta de busca, tipicamente a informação básica que se possui é o nome completo e/ou o nome usado em citações, o que pode não ser o suficiente. Por exemplo, ao buscar GUPTA, R. no DBLP¹, encontramos 113 registros de autores diferentes, o que ilustra a homonímia (afinal, se temos apenas a identificação de GUPTA R., qual deles é o correto?). Por outro lado, se buscarmos no mesmo DBLP o nome RAJESH GUPTA, verificamos que existem dois possíveis registros para seu nome (Rajesh K. Gupta e Rajesh Gupta). Os trabalhos descritos a seguir foram utilizados como embasamento para a solução de desambiguação desenvolvida neste projeto.

Uma das propostas para identificar corretamente os autores consiste na utilização de um registro central de identificação. O papel desse cadastro é armazenar e identificar de forma única os autores, como o realizado pela ORCID, uma instituição não lucrativa². Smalheiser e Torvik (2009) 2009 explicam porque não podemos confiar em um registro central. O primeiro ponto para que seja possível uma identificação correta depende da participação maciça da comunidade acadêmica em registrar-se no órgão. As informações preexistentes e as já inseridas obrigatoriamente devem estar corretas. Caso não ocorra esta participação, os dados não serão completos e corretos. Ademais, Smalheiser e Torvik (2009) colocam que a desambiguação ser efetuada por observação humana não é eficiente. Os resultados são diferentes de acordo com os realizadores. Para completar nomes comuns, como Lee e Smith, por exemplo, são mais difíceis de serem identificados corretamente por um observador humano. Ele terá de recorrer a outras informações adicionais gastando um tempo superior ao processamento computacional.

¹ dblp.uni-trier.de

² <https://orcid.org/> acessado em 23/03/2017

Milojevic (2013) relata que a utilização apenas do primeiro nome (mesmo abreviado) e do último sobrenome apresenta resultados bastante satisfatórios para grupos pequenos e fechados de autores. Porém, isto não é aplicável a nomes asiáticos, sobretudo para a comunidade Chinesa e Coreana, nas quais há grande frequência de repetição devido ao pequeno conjunto de sobrenomes. Para esses casos, técnicas avançadas de desambiguação são necessárias para separar os autores agrupados erroneamente (STROTMANN; ZHAO, 2012).

Uma das estratégias para identificar autores é verificar o grupo de coautores dos trabalhos. Em geral, se há semelhança entre os grupos de coautores dos artigos publicados pelos dois autores em verificação, aumenta-se a chance dos mesmos serem a mesma pessoa.

Usar somente modelos com base nos nomes e na rede de coautoria também pode não ser suficiente. Para comunidades pequenas, onde grupos praticamente mantêm-se isolados, os resultados são satisfatórios (WHITE; MCCAIN, 1988). Porém para muitos estudos reais, nos quais os pesquisadores podem atuar em diferentes áreas e colaborar com inúmeros grupos, a abordagem perde eficiência.

Milojevic (2013) relata que, de acordo com o enfoque da pesquisa, a complexidade dos algoritmos pode ser ajustada. Para pesquisas onde não existe uma grande quantidade de grupos, ou não há muita similaridade entre as entidades pode-se optar por utilizar um algoritmo simplificado. Por exemplo, pode-se usar somente o nome dos autores como critério de desambiguação. Quando esse critério não é suficiente para identificação, aumenta-se a complexidade do algoritmo. De acordo com os trabalhos mais relevantes na área, as técnicas de distância de *strings* (como distância de edição e distância cosseno), mineração de dados e técnicas de agrupamento têm obtido resultados satisfatórios (DIGIAMPIETRI; BARBOSA; LINDEN, 2015). Outras abordagens aplicadas para a verificação dos registros envolvem diretamente o uso de inteligência artificial (DIAZ-VALENZUELA; MARTÍN-BAUTISTA; VILA, 2014), com a aplicação de lógica *fuzzy* sobre os registros. Deve-se ter cuidado com o número de variáveis e combinação de técnicas, pois o algoritmo pode tornar-se altamente complexo. Contudo, não existe um consenso quanto a complexidade necessária para que o algoritmo possa ser preciso.

Uma técnica em dois passos explorada por Ferreira, Machado e Goncalves (2012) consiste em utilizar primeiro um método não supervisionado para agrupar os artigos. O segundo é outro supervisionado, para derivar as regras. O usuário então fornece um *feedback* ao algoritmo, para que esse se aprimore. A apresentação de um *feedback* sob uma

pequena amostragem dos casos ambíguos durante a primeira fase permitiu um aumento na precisão em 10%.

Fica evidente a necessidade de desambiguação, sobretudo em aplicações de nível mais alto, como no caso de análise de produtividade do autor e, em especial, ao considerar grupos de autores de diferentes instituições Lima (2003), Ferreira, Goncalves e Laender (2012).

2.2 Bibliometria

A bibliometria consiste em estudos quantitativos sobre a produção científica. Existem ao menos quatro termos ligados a análise bibliométrica, sendo eles bibliometria, cientometria, informetria e webometria. Para esse mestrado, apenas os conceitos de bibliometria foram utilizados. Bibliometria é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação (TAGUE-SUTCLIFFE, 1992).

Digiampietri (2015) descreve as quatro principais medidas utilizadas nos trabalhos de produção acadêmica (DIGIAMPIETRI, 2015): (a) Medidas baseadas no **número de artigos**, que contabilizam o número de artigos por pesquisador e por grupos, assim como a média em ambos os casos. (b) Medidas baseadas no **número de autores por artigos**, englobam contabilização e média de autores por artigo, média num determinado grupo, etc. (c) Medidas para a **qualificação dos veículos de publicação** consideram algum critério qualificatório dos eventos ou veículos nos quais são efetuadas as publicações. (iv) Medidas baseadas no **número de citações**, são aquelas que utilizam a soma do número de citações de artigos de pesquisadores e grupos de pesquisa, médias, medianas ou medidas derivadas como o índice-h e o índice-g.

O índice-h é determinado pela quantidade mínima de citações que o conjunto de artigos produzidos por um pesquisador. Caso um pesquisador tenha um artigo com 30 citações e outro com 3, seu índice-h será 3. O índice-g é calculado com base na média quadrática das citações dos artigos onde o número de citações dos artigos de cada um dos artigos que compuseram a média seja menor ou igual a própria.

O uso dessas métricas deve ser usado com parcimônia no ato de inferir qualidade a uma pesquisa científica. Fatores como área da ciência, precisão em aferir a quantidade de

citações, critério de seleção de idiomas para formação da base de dados influenciam essas medidas citeLeiden2003.

2.3 Indicadores para análise de Redes Sociais

A análise de redes sociais pode ser feita utilizando diferentes metodologias de construção de grafos. Contudo, a maioria dos estudos envolve o cálculo da densidade, coeficiente de agrupamento ou clusterização, assortatividade, centralidade e centralização. Essas medidas são comumente utilizadas para realizar análises de grupos, determinação de influência e outros comportamentos.

A **densidade** é a razão entre a quantidade de arestas de um grafo e o número máximo possível de arestas (TAGUE-SUTCLIFFE, 1992). Um grafo que possua a densidade muito baixa é dito esperso.

O **coeficiente de agrupamento** ou clusterização, mede a transitividade das relações de um grafo. Divide-se o total de cliques de tamanho três pela quantidade de trios de vértices com ao menos um caminho de tamanho dois os conectando. Essa métrica é tipicamente associada a estabilidade (ou maturidade) da rede social (TAGUE-SUTCLIFFE, 1992).

A **assortatividade** mede a tendência de nós com as mesmas características se conectarem, sendo possível aplicar essa medida, no âmbito de análise de redes sociais, em relação a instituições, nacionalidades, gênero, grau, etc (DIGIAMPIETRI, 2015).

A **centralidade** indica a importância de um vértice na rede de acordo com uma determinada característica. As centralidades mais utilizadas são: a de grau ou *degree* (calculada considerando o grau de um vértice em relação ao grau dos demais vértices da rede); a de proximidade ou *closeness* (distância do vértice em relação aos demais); intermediação *betweenness*, que indica a quantidade de vezes que o nó aparece nos caminhos mínimos do grafo; autovalor (eigenvalue), para verificar a importância do nó em relação a importância dos demais nós aos quais está conectado.

Por fim a **centralização** é uma medida global (isto é, da rede e não individual de cada nó, assim como a densidade) que utiliza da centralidade para verificar a importância do nó mais central da rede em relação a toda a rede. Quanto maior a centralização em uma rede, maior a influência desse nó na rede e, potencialmente, mais dependente a rede é desse nó (FREEMAN, 1979).

2.4 Bases de dados bibliográficas

Obter os dados bibliográficos dos pesquisadores e estabelecer os relacionamentos entre os pesquisadores costumam ser as atividades mais desafiadoras da análise de redes sociais. Atualmente, existem bases indexadoras, como a do DBLP (que indexa artigos internacionais da área de computação) e a sessão bibliográfica da Plataforma Lattes, que concentra em seu banco o currículo dos pesquisadores, centralizando informações de projetos de pesquisa, atuação profissional, áreas de interesse e publicações. Por meio desses dados, é possível desenvolver a rede social acadêmica nacional e identificar grupos de trabalho.

A estrutura e o método de inserção de novos registros é distinto em cada base. Na base DBLP os registros são organizados por publicação, assim a lista de coautorias de um único artigo pode ser extraída de maneira simples a partir do registro de uma publicação, mas, apesar de haver uma tentativa de padronizar os dados dos autores, uma mesma pessoa pode possuir mais de um identificador dentro desta base. Assim, é necessário utilizar estratégias de resolução de nome de autores.

Já na base de dados da Plataforma Lattes, os currículos são a unidade base centralizadora de informações. Desta forma, para uma dada pessoa é fácil encontrar todas as publicações cadastradas em seu currículo (lembrando-se que os dados são preenchidos pelos donos dos currículos), porém a identificação das relações de coautoria é mais complexa, pois cada artigo pode aparecer de diferentes formas nos currículos de seus coautores. Sendo assim, faz-se necessária a utilização de estratégias de resolução de publicações.

A base DBLP conta atualmente com registros de mais de 3,8 milhões de publicações e mais de 2 milhões de registros de autores (neste último caso é possível encontrar duplicações: dois ou mais registros que se referem ao mesmo autor). Já a Plataforma Lattes contém o currículo de mais de 5 milhões de pessoas e mais de 10 milhões de registros de artigos completos publicados em periódicos ou anais de eventos (contando-se a replicação de artigos entre os currículos dos coautores).

Estas duas bases foram utilizadas neste projeto com diferentes objetivos, conforme será apresentado no restante deste texto.

3 Revisão da Literatura

Para ter uma ampla visão da literatura e dos trabalhos correlatos, optou-se por realizar uma revisão seguindo os passos de uma revisão sistemática sobre bibliometria e, em específico, de como a bibliometria é utilizada para a análise de grupos acadêmicos. A revisão sistemática identifica e discute os trabalhos acadêmicos mais relevantes dentro de uma *string* de busca. Como essa metodologia tem passos bem definidos, ela permite a auditoria posterior dos resultados bem como a atualização da revisão. Portanto, definiu-se o objetivo dessa revisão sistemática em sumarizar as principais técnicas utilizadas para avaliar redes acadêmicas, com enfoque sobre a colaboração de trabalhos. Uma revisão sistemática divide-se em três etapas: planejamento, condução e documentação. A etapa de planejamento consiste no estabelecimento dos dados a serem levantados, quais bases serão utilizadas e quais as palavras chaves a serem pesquisadas. Durante a condução, estabelece-se quais serão os critérios de inclusão e exclusão dos trabalhos resultantes da busca, prosseguindo com a leitura dos trabalhos aceitos para a fase de documentação. Durante a documentação, analisa-se os artigos obtidos, sendo o resultado a resposta a pergunta inicial formulada.

3.1 Planejamento

A etapa de planejamento sumariza-se com o protocolo de pesquisa, onde determinam-se as questões as quais a revisão sistemática visa responder. Nessa etapa também são escolhidas as bases e as palavras de busca a serem utilizadas. Um protocolo é dividido em:

- Questão: Objetivo da Revisão Sistemática
- População: Artigos que serão avaliados
- Controle: Artigos que servem como base para comparação a pesquisa
- Resultado: O que se espera encontrar ao concluir a revisão
- Palavras Chaves: Palavras utilizadas para definir os critérios de pesquisa
- Fontes: Bases a serem consultadas e critérios de seleção

Neste protocolo, a questão principal a qual pretendeu-se responder foi: Quais as técnicas e análises de grupo mais utilizadas?

O critério para formar a população foram artigos que aplicaram uma análise estrutural para estudar a rede social acadêmica.

O grupo controle foi composto por artigos previamente selecionados em máquinas de busca, além de outros artigos recomendados pelo orientador.

O resultado esperado da revisão sistemática foi o de obter-se um sólido conhecimento nas métricas e premissas utilizadas. As técnicas e metodologias de construção e análise de redes sociais assemelham-se de acordo com o objetivo do estudo.

A aplicação pretendida foi a de servir como base do conhecimento para futuros trabalhos.

As palavras-chaves para essa Revisão sistemática foram: *Academic Assessment, Academic Group, Academic Network, Academic social network, Group Analyses, Group Methodology, Link Analysis, Collaborative Network, Scientometrics, co author, coauthorship, research group*.

O critério de seleção de veículos restringiu àquelas que possuíam algum enfoque relacionado à bibliometria, disponibilizando artigos completos. A linguagem escolhida foi inglês, por ser a mais utilizada em artigos acadêmicos internacionais. Como resultado, foram selecionadas os artigos das máquinas de busca da ACM, IEEE, PLOS One, Scientometrics e Science Direct. Por não serem fontes que publicam apenas artigos relacionados ao assunto da revisão, diferentes combinações das palavras-chaves foram aplicadas, as quais estão detalhadas durante a seção 3.2, para maximizar a chance dos resultados das buscas serem aderentes aos objetivos dessa revisão.

3.2 Condução

Para selecionar os artigos, os seguintes critérios foram utilizados:

- **Inclusão**
- Calcular métricas bibliométricas e de análise de redes;
- Apresentar uma metodologia de validação;
- Ser um artigo completo.
- **Exclusão**
- Não apresentar um método quantitativo de validação;
- Não detalhar as métricas utilizada;

- Não ser focado em grupos acadêmicos;
- Não estar disponível para leitura.

Para ser aceito na seleção, o artigo precisa satisfazer a todos os critérios de inclusão e não apresentar nenhum critério de exclusão. A avaliação desses critérios foi realizada por uma avaliação preliminar, restrita ao título e ao *abstract*. Caso nenhum critério de exclusão fosse verificado nessa etapa, prossegue-se com a leitura completa.

ACM: Duas pesquisas resultaram em resultados significantes. A primeira, com a string “(*academic and group*) AND (*keywords:scientometrics*)”, retornaram 10 artigos. Desde grupo, seis foram aprovados na avaliação preliminar. Contudo, apenas um foi aprovado na leitura completa.

A segunda pesquisa feita com a string “(*academic*) AND (*network or group*) AND (*not sentiment*) AND (*keywords: link analysis*)” retornou 18 artigos, sendo seis aprovados na leitura preliminar e três aprovados na leitura completa. “*Not sentimento*” foi acrescentado devido ao fato das buscas preliminares retornarem muitos artigos relacionados a análise de sentimentos.

IEEE: Nessa base de dados, duas pesquisas foram relevantes. A primeira com a string “(*link analysis*) AND (*collaborative*) AND (*Academic Network*)”, resultando em dois artigos, sendo um descartado na leitura preliminar. A segunda busca foi “(*research group*) AND (*co author*) AND (*analysis*)”, resultando em 28 artigos. 11 foram aprovados na leitura preliminar e desses três foram aprovados na leitura completa.

PLOS One: Nessa base a string com melhor resultado foi “*everything:academic social network*”, com retorno de três artigos, todos aprovados na leitura preliminar, e com um aprovado na leitura completa.

Scientometrics: Por ser uma base especializada em bibliometria, essa foi a base com maior quantidade de artigos retornados. A string básica de busca foi “*link analysis AND collaboration AND network AND coauthorship*”. O resultado gerou centenas de artigos. Limitamos o escopo a artigos ligados a ciência da computação e com acesso livre. A quantidade de artigos reduziu-se para 10, com quatro selecionados na leitura preliminar e três selecionados na leitura completa.

Science Direct: A melhor string de busca aplicada a essa base foi “(*Coauthorship Network*) AND (*Collaboration*)”. Por termos mais de centenas de resultados, limitamos aos artigos publicados entre os anos de 2011 e 2015, restritos aos periódicos: Social Networks,

Journal of Infometrics e Information Processing and Managment. Restringiu-se ainda aos tópicos *Network OR Collaboration*. A busca retornou 14 artigos, sendo 11 aceitos na leitura preliminar e oito selecionados na leitura completa.

3.3 Extração/Documentação

Durante a etapa de extração buscou-se encontrar as informações bases consideradas relevantes para a revisão. A primeira etapa consistiu em avaliar o critério empregado pelos artigos selecionados quanto a escolha da base de dados.

Para estudos bibliométricos, as bases precisam conter sumariamente o título do artigo, o nome dos autores e o veículo de publicação. Preferencialmente, a base deve conter um mecanismo de indexação de autores, para agilizar o trabalho de identificação. Dentre as mais utilizadas nos artigos selecionados estão a DBLP, MAS, Plataforma Lattes, Information Science & Library Science. Nesta seção, nos organizamos e analisamos as métricas e algoritmos desenvolvidos para criar e analisar redes sociais acadêmicas, agrupando-os de acordo com os objetivos.

Predição de Colaboração: A produção de artigos acadêmicos, majoritariamente, envolve a dedicação de um grupo de pesquisadores. Dada a globalização e a difusão do conhecimento, aumenta-se a possibilidade de estabelecer relacionamentos com outros pesquisadores a fim de colaboração. Contudo, dado o volume de publicações, encontrar pesquisadores potenciais para novos trabalhos pode não ser trivial.

Uma solução a esse problema consiste no reconhecimento de padrões de pesquisa dos pesquisadores e sua agregação por afinidade. Esse padrão foi fundamentado a partir de uma análise de rede social. Nesse caso, dois modelos de redes foram utilizados. O primeiro usou áreas de interesse (arestas), enquanto o segundo usou como arestas a colaboração dos pesquisadores nos artigos. Em ambos os casos, optou-se por grafos ponderados, cujos pesos podem ser calculados com base na evolução temporal dos interesse/atuação do autor (TYLENDÁ; ANGELOVA; BEDATHUR, 2009) (CHEN et al., 2011) (DATTA et al., 2013), (MENA-CHALCO; CESAR, 2011) ou de acordo com os trabalhos onde ambos pesquisadores foram coautores (CHEN et al., 2011), (YANG; HONG; DAVISON, 2010), (HUYNH et al., 2014).

O uso de redes sociais encontraram taxa de correlação positiva entre a métrica de interesse e a coautoria efetiva. Ao simular as possíveis correlações os algoritmos obteve predições similares e melhores do que os trabalhos correlatos. Contudo, todos observaram que os resultados de suas trabalhos apenas aplicam-se ao conjunto de dados selecionado, e não puderam afirmar se obteriam o mesmo sucesso caso aplicassem-no em bases com um maior volume de publicações e áreas de interesse.

Influência da rede: A influência de um pesquisador indica qual sua importância no direcionamento das pesquisas efetuadas por seu grupo. Os artigos selecionados apresentaram dois objetivos. O primeiro foi estabelecer o perfil dos pesquisadores mais influentes, de acordo com seus colaboradores (SARIGOL et al., 2014) (LUNA; REVOREDO; COZMAN, 2013). O segundo, avaliar como o trabalho de um determinado pesquisador afeta os trabalhos dos correlatos (FREIRE; FIGUEIREDO, 2011) (CIMENLER; REEVES; SKVORETZ, 2015) (ORTEGA, 2014). Dentre os resultados obtidos, não apenas a qualidade do pesquisador, mas dos institutos também pode ser aferida e comparada (YAN; ZHAI; FAN, 2013) (ORTEGA; AGUILLO, 2013) (ABBASI; HOSSAIN; LEYDESDORFF, 2012) buscando encontrar padrões entre os melhores qualificados e determinar quais características em comum são compartilhadas entre esses.

Predição de Citação: Outra aplicação do estudo de redes sociais é identificar qual o possível número de citações que um artigo receberá. A rede social foi estabelecida em torno de colaboração e influência de rede. Isto é, quando um autor ou grupo de autores produz um trabalho, eles serão citados de acordo com sua influência na rede social (ARIF; ALI; ASGER, 2014) (DIGIAMPIETRI et al., 2014). O uso de classificadores serviu para tentar definir padrões na citação de acordo com regionalidade, tempo de publicação e instituições de origem.

3.3.1 Documentação

Ao analisar as bases utilizadas nos estudos, nota-se que as mais utilizadas foram a Microsoft Academic Search, a Plataforma Lattes e a DBLP, conforme figura 1. Essas bases contem o maior número de referências e trabalham com ferramentas de indexação, o que facilita o trabalho de identificação de autores.

Figura 1 – Bases utilizadas na literatura

Fonte: Lênin Ferreira Barbosa, 2017

O grafo da rede social, de acordo com os objetivos do estudo, tem suas regras para definição das arestas de maneiras distintas, conforme demonstrado nos trabalhos acima. Na maioria dos casos, opta-se pelo grafo ponderado, pois tão importante quanto defini-las é comparar a intensidade dos relacionamentos entre os grafos. Um dos aspectos comuns aos estudos de redes sociais, independentemente do objetivo, consiste na identificação e categorização de grupos. Esse procedimento pode ser denominado análise de grupos, cuja metodologia envolve combinações e variações dos procedimentos definidos por Spearman's (vide capítulo 2). Dentre as análises mais utilizadas estão *Betweenness* e *Centrality*. Em conjunto a essas técnicas, costuma-se utilizar classificadores como SVM, Naive-Bayes e Random Forest. O critério para utilização depende do quão esparsa for a rede e do nível de influência entre os vértices. A Tabela 1 indica as principais métricas, tanto bibliométricas quanto de análise de rede social, de acordo com os objetivos dos estudos.

Spearman também definiu o método de validação para indicar se existe de fato uma correlação entre as métricas utilizadas para a formação da rede social. Esse método é conhecido como Spearman's Correlation Rank. A validação, sobretudo quando se aplicam novas métricas de rede, é necessária para provar de que o resultado obtido não se deve a uma variável aleatória.

Tabela 1 – Métodos de avaliação utilizadas de acordo com os objetivos do estudo

Objetivo	Análises
Influência do Pesquisador	Ego Network
	Principal Components Analysis
	Social Network Analysis
Influência de pesquisa	Poisson Regression Modeling
Recomendação de Coautoria	Vertex Similarity
	Lexical Similarity
Caracterização de programas	Social Network Relation
Predição de Citação	Hindcasting approach
Desambiguação	K-Metrics
	K-Means
Predição de Coautoria	Space-Temporal Index
	Logistic Regression
	Ego Network
	Bayesian Network
	BaseLine Models
	C-index
	Exponential Random Graphs

Fonte: Lênin Ferreira Barbosa, 2017

3.4 Trabalhos Complementares

Dada a natureza generalista da revisão, alguns artigos foram pesquisados para complementar o estudo em relação a ranqueamentos. Dentre as pesquisas encontradas, destaca-se dois artigos que envolvem o ranqueamento de pesquisadores quanto ao nível de influência que eles possuem dentro do grupo de pesquisa ao qual pertencem.

Han et al. (2013) analisaram a qualidade de um pesquisador colocando como critério qual a aptidão do pesquisador em fazer parte de um comitê de congresso. A métrica desenvolvida teve como base estudos de influência. A premissa do autor é de que os pesquisadores mais influentes são os mais indicados para compor o comitê. Quanto maior o número de conexões representadas pelas coautorias de um determinado autor, maior será a chance desse ser um especialista.

Para determinar qual a área de interesse dos especialistas, foi utilizada a análise KL-Divergence sob os títulos dos artigos. Os resultados foram suavizados com *Dirichlet smooth*. Segundo os autores, este trabalho apresentou uma boa acurácia para encontrar os especialistas, porém, o perfil dos mais indicados não corresponderam as expectativas iniciais.

Li e Gillet (2013) optaram por determinar a influência de autores analisando a citação de seus trabalhos. A base escolhida de dados foi a Mendeley, que possui artigos em diferentes áreas. Esse estudo teve como premissa que: (i) os pesquisadores com maior influência na rede e impacto acadêmico fossem os pesquisadores seniores, com diversos coautores; (ii) O uso de três métricas (R-Index, número de citações e número de leitores) deve ser suficiente para determinar o fator de influência social do pesquisador sobre a rede; e (iii) as influências sociais e acadêmicas medem diferentes dimensões, portanto não devem estar altamente correlacionadas.

Para a verificação da influência acadêmica obtida, ao invés de considerar métricas tradicionais, como índice h ou PageRank, utilizou-se a quantidade de vezes que os artigos de cada autor foram lidos, informação disponível na base escolhida (LI; GILLET, 2013). Com isso, as métricas sociais foram: número máximo de leitores por trabalho, total de leitores e índice-R (variação do índice-h desenvolvido pelos autores) que verifica o impacto e a produtividade de um trabalho em um determinado grupo R de leitores. Para analisar a influência social, foram analisadas as correlações de Spearman entre três características: transitividade, grau e proximidade em relação ao número de leituras dos artigos. Os resultados obtidos indicaram que houve correlação positiva entre as duas primeiras métricas sociais. O índice R correlacionou-se positivamente com o total de número de leituras, mas não com o número máximo de leituras por artigo. Dentre os resultados obtidos, conforme maior é o *rank* de influência, maior a quantidade de coautores que o pesquisador possui. Verificou-se também que as influências sociais e acadêmicas não se correlacionaram. Com isso, comprovou-se a diferente dimensionalidade das influências.

Digiampietri et al. (2014), partindo da produção científica dos diferentes programas de pós-graduação brasileiros, verificaram quais os aspectos comuns aos programas melhor classificados no país. O objetivo principal foi o de identificar como os programas relacionavam-se entre si. A base de dados escolhida foi a Plataforma Lattes. A rede de pesquisadores (vértices) foi construída considerando como relacionamento (aresta) a coautoria dos artigos. Os autores também estudaram quais as correlações existentes entre as métricas empregadas. A investigação da rede baseou-se em medidas de centralidade e de agrupamento. O coeficiente de correlação de Spearman foi empregado para verificar a correlação da qualidade do programa com métricas de citação como *Scholar h-index*, *Scholar CC*, *MS g-index*, Qualis entre outras. Os resultados obtidos indicaram correlação positiva das métricas de citação com o ranqueamento, conforme previsto. Após analisar as

arestas e seus comportamentos intragrupos, Digiampietri et al. (2014) aplicaram Análise de Componentes Principais (PCA) para reduzir a dimensionalidade das componentes e assim, verificar se é possível separar os principais programas dos demais, além de verificar como cada componente está relacionado ao ranqueamento do programa.

A conclusão do trabalho indicou que os programas melhor classificados possuem alta colaboração entre si e tendem a manter esses vínculos ao longo do tempo, enquanto programas novos ou com piores classificações, tendem a ter baixa colaboração com outros programas, apresentando poucas e esporádicas colaborações.

3.5 Considerações Finais sobre a Revisão

Os artigos da área de bibliometria com maior relevância estão concentrados na base da Scientometrics, em seguida pela Science Direct. Ao total foram 83 artigos avaliados, sendo 50 os aprovados após leitura a completa, publicados majoritariamente entre os anos de 2013 e 2014.

Uma das lacunas encontradas durante a revisão realizada foi quanto aos grupos utilizados para a formação das redes sociais. Os artigos selecionados nessa revisão utilizaram pequenos grupos de pesquisa para identificar padrões dentro das redes sociais, sobretudo nos estudos de influência. As dificuldades de se utilizar bases maiores estão relacionadas a identificação correta de autores e de suas produções bibliográficas.

Nos estudos de área de interesse, a utilização de grupos pequenos tornou a identificação da área de professores interdisciplinares mais imprecisa. Portanto, mesmo categorizando os estudos pelos seus objetivos, não existe um consenso estabelecido sobre um modelo de rede social a ser aplicado. Por ser uma pesquisa de caráter exploratório quanto às técnicas utilizadas para determinados objetivos, não foram considerados trabalhos de investigação quanto a relação entre características de redes sociais e ranqueamentos internacionais. Essa lacuna foi explorada por essa dissertação.

Para auxiliar na identificação dos autores, no contexto da ciência da computação, a DBLP e o Microsoft Academic Search (MAS) aplicam esforços nessa área, indexando autores em seus bancos de dados. Contudo a desambiguação de autores ainda é necessária para garantir a correta identificação de autores e artigos.

A coautoria em produções científicas foi utilizada em todos os estudos, seja como peso para indicar influências ou como resultado, nos trabalhos de recomendação, indicando a relevância dessa informação na análise de redes sociais acadêmicas. As principais técnicas de avaliação dos resultados nos estudos empregam análise de correlação de métricas, independentemente da estrutura de rede social e do objetivo do estudo.

4 Materiais e Métodos

Neste capítulo são detalhados os dados utilizados e os métodos aplicados nesta dissertação. Conforme detalhado na Seção 2.1, desambiguar os registros dos autores é uma atividade essencial para estudos bibliométricos. Portanto, o primeiro objetivo consistiu em identificar de modo automatizado os registros ambíguos de nossa base de dados. Com base no conhecimento adquirido após a revisão sistemática, detalhada no Capítulo 3, foram selecionadas as principais características relacionadas à identificação de autores em nossa base de dados. Com base nesses dados, foi desenvolvido um algoritmo para a automatização desse processo. O processo de desambiguação empregado bem como seus resultados são apresentados na seção 4.2.1.

Após a desambiguação dos autores, foram extraídas as características bibliométricas dos programas de pós-graduação. Nosso objetivo foi caracterizar os diferentes programas ou departamentos e comparar os programas nacionais melhores avaliadas pela CAPES com alguns dos principais departamentos internacionais (avaliadas de acordo com alguns ranqueamentos internacionais). Os programas nacionais considerados de níveis internacionais são aquelas cujas notas emitidas pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) alcançaram as notas 6 ou 7. A metodologia desse trabalho foi detalhada na seção 4.2.2.

4.1 Materiais

Em novembro de 2014 foi copiada toda a base DBLP. Desta base foram extraídas cerca de 3 milhões de publicações, atribuídas a um total de 1,5 milhões de autores. Em março de 2016, fez-se o *download* da lista presente na CAPES de professores permanentes em programas de pós-graduação da área de Ciência da Computação. Esses dados foram utilizados para desenvolver um método de desambiguação, que foi aplicada para a realização dos estudos de análise de grupos apresentados na seção 4.2.2.

Especificamente para o estudo de caso sobre desambiguação, utilizou-se como amostra o conjunto de professores permanentes de um programa brasileiro de pós-graduação em Ciência da Computação. Do conjunto de programas, o programa da Unicamp, com

48 professores registrados foi selecionado aleatoriamente para a realização dos testes e validação do algoritmo de desambiguação.

Selecionada a amostra de registros para a desambiguação, iniciou-se a identificação e coleta de dados das universidades melhor ranqueadas na área de ciências da computação no âmbito mundial.

Uma particularidade dos ranqueamentos internacionais é que departamentos, unidades ou universidades são, em geral, avaliados com a mesma metodologia. O sistema de avaliação nacional tipicamente avalia unicamente os programas de pós-graduação (que podem, inclusive, conter docentes de diferentes universidades). Por essa diversidade, nesta dissertação a palavra departamento será utilizada para tratar tanto um departamento (internacional) quanto um programa de pós-graduação (nacional).

Para a seleção dos vinte departamentos internacionais foram utilizados três ranqueamentos internacionais. Shanghai¹, Top Universities Top Universities² e The Guardian³. A tabela 2 apresenta os programas selecionados. Foram selecionados os departamentos que mais apareceram entre as 15 primeiras posições dos diferentes ranqueamentos. Adicionalmente outros cinco departamentos com ranqueamento similar aos dos programas brasileiros melhor ranqueados foram selecionados. Essa seleção teve como base identificar as diferentes características presentes nos programas internacionais. Por fim, foram selecionados os oito programas brasileiros de pós-graduação em Ciência da Computação considerados pela CAPES de nível internacional (isto é, programas nota 6 ou 7). A Tabela 2 contém a lista desses departamentos e programas. A ordenação da tabela decorre da média obtida pelos programas nos três ranqueamentos. Essa regra aplica-se aos 18 primeiros programas que possuem uma posição clara no *rank*. Esse método não pode ser aplicado aos demais programas pelos seguintes motivos. Há casos onde o programa não apareceu em um ou mais ranqueamentos, portanto, sua média seria injusta em relação aos demais. Outros programas estão numa faixa de posição devido a estarem empatados com outros programas naquele ranqueamento, o que impossibilita o cálculo da média.

Selecionados os programas ou departamentos, a lista de seus pesquisadores foi obtida a partir do site dos respectivos departamentos ou do site da CAPES (no caso dos programas brasileiros de pós-graduação). A obtenção dessa lista foi manual para os

¹ Academic Ranking of World Universities: <http://www.shanghairanking.com/>

² Worldwide university rankings: <http://www.topuniversities.com/>

³ The Guardian: <http://www.theguardian.com/education/universityguide>

Tabela 2 – Departamentos ou Programas de Computação Analisados

Universidade	Abrev. usada	País	Shanghai	Top Universities	The Guardian
MIT - Electrical Engineering and Computer Science Department	MIT EECS	Estados Unidos	2	1	1
Stanford University	Stanford	Estados Unidos	1	2	2
Harvard - School of Engineering and Applied Sciences	Harvard SEAS	Estados Unidos	4	4	4
Carnegie Mellon University	CMU	Estados Unidos	6	5	4
University of California - Berkeley	UCBerkeley	Estados Unidos	3	7	6
Princeton University	Princeton	Estados Unidos	5	10	9
Swiss Federal Institute of Technology in Zurich	ETH Zurich	Suíça	17	8	9
University of Cambridge	Cam	Reino Unido	23	6	8
University of Toronto	Toronto	Canada	11	11	16
University of Oxford	Oxford	Reino Unido	34	3	3
University of California - Los Angeles	UCLA	Estados Unidos	9	16	17
Hong Kong University of Science and Technology	HKUST	China	39	14	8
The University of Texas - Austin	UTexas	Estados Unidos	7	34	22
Cornell University	Cornell	Estados Unidos	8	26	31
California Institute of Technology	Caltech	Estados Unidos	11	28	27
University of Illinois	Illinois	Estados Unidos	13	30	30
Chinese University of Hong Kong	CUHK	China	37	22	18
Ecole Polytechnique Federale de Lausanne	EPFL	Suíça	35	18	29
University of Washington	UW	Estados Unidos	51-75	28	20
Universidade Estadual de Campinas	UNICAMP	Brasil	101-150	51-100	51-100
Universidade de São Paulo - São Carlos	USP São Carlos	Brasil	151-200	51-100	51-100
Universidade de São Paulo - São Paulo	USP São Paulo	Brasil			
Boston University	BU	Estados Unidos	101-150	101-150	
Universidade Federal do Rio Grande do Sul	UFRGS	Brasil		151-200	
Universidade Federal do Rio de Janeiro	UFRJ	Brasil		151-200	
Universidade Federal de Pernambuco	UFPE	Brasil		201-250	
Pontifícia Universidade Católica do Rio de Janeiro	PUC-Rio	Brasil		201-250	

Fonte: Digiampietri, Linden e Barbosa (2016)

departamentos internacionais, pesquisando as páginas bibliográficas dos pesquisadores a partir dos sites dos departamentos. Para os programas nacionais, os dados foram obtidos do sistema Sucupira. A identificação dos registros dos pesquisadores na base DBLP foi feita de maneira semiautomática. Isto é, fez-se o emprego do algoritmo de desambiguação desenvolvido neste projeto, mas os dados foram validados manualmente.

Listados os pesquisadores, os dados de suas publicações foram obtidos automaticamente da base DBLP. Com base no título e ano da publicação foram obtidas as citações dos artigos, extraídas automaticamente do site *Microsoft Academic Search* (MAS). Os índices G e H foram calculados com base nesses dados sobre citações.

4.2 Métodos

Esta seção apresenta os métodos aplicados nos estudos de desambiguação e de análises bibliométricas e de redes sociais. Para ambos os estudos fez-se a extração de características, seguida da classificação, para identificar quais delas possuem maior relevância para alcançar os respectivos objetivos. No caso da desambiguação, a relevância está ligada à probabilidade de identificação correta se duas referências (perfis do DBLP) se referirem ou não a um mesmo autor. Para o caso da análise de grupos, a classificação estava ligada à relevância da métrica em relação ao posicionamento dos departamentos (ou programas) no ranqueamento.

As métricas de redes sociais foram calculadas utilizando a biblioteca *igraph* do arcabouço R⁴ e as demais métricas foram calculadas utilizando ferramentas desenvolvidas durante este projeto de mestrado ou pelo grupo de pesquisa no qual este projeto está contextualizado.

Os resultados desses estudos foram publicados em dois artigos e um terceiro foi aceito para publicação. O primeiro refere-se a desambiguação e os dois últimos às avaliações ou caracterizações dos departamentos internacionais e nacionais. Estas são as referências para estes artigos: Digiampietri, Barbosa e Linden (2015), Digiampietri, Linden e Barbosa (2016) e Linden, Digiampietri e Barbosa (2017). A subseção 4.2.1 descreve os métodos utilizados para o estudo de desambiguação, enquanto que a subseção 4.2.2 apresenta os métodos para as análises de grupos.

4.2.1 Desambiguação

A busca pelo primeiro e último nome dos 48 pesquisadores do programa de pós-graduação em Ciência da Computação da Unicamp resultou na identificação de 82 registros do DBLP. Dos 48 pesquisadores, 29 possuíam um único registro. O processo de busca não identificou nenhum registro DBLP para dois professores. Para estes casos, pesquisou-se manualmente variações possíveis na base. Estas pesquisas indicaram que estes professores não possuíam registro na base DBLP.

Para os 17 professores restantes (para os quais mais de um registro DBLP foi encontrado como potencial registro) adotou-se a seguinte estratégia de resolução de

⁴ <https://www.r-project.org/>

ambiguidade. Os registros que potencialmente se referenciam a um mesmo professor (isto é, que foram recuperados com base na busca de um mesmo primeiro e último nomes) foram combinados dois a dois e cada par foi rotulado manualmente como se referissem (ou não) a uma mesma pessoa (independentemente dessa pessoa ser ou não o professor permanente que estava sendo procurado). Esse processo resultou na identificação de 102 pares.

De acordo com Ferreira, Machado e Gonçalves (2012), há quatro fontes de dados a serem consideradas no processo de desambiguação de nomes: lista de autores, citação, rede de coautorias e informações adicionais.

A partir dos registros bibliométricos contidos nas publicações é possível identificar a lista de autores. Com isso, verifica-se se o registro autoral é compatível com seu nome. Contudo, essa informação não é suficiente para desambiguar, tendo em vista que os nomes utilizados nas referências dos artigos, muitas vezes, não correspondem ao nome completo dos autores, e mesmo que correspondesse, os casos de homonímia não permitiriam a correta atribuição da autoria. Avaliar dados de citação consiste em comparar as referências de citação desse autor, a distância temporal das publicações, como o ano de publicação, título do evento/periódico e título do artigo. Por exemplo, quanto maior for a distância entre anos de publicação entre dois artigos, menores são as chances dos autores serem a mesma pessoa.

Ao avaliar redes de coautoria, pretende-se identificar se há vizinhos em comum entre os autores. Se entre os artigos publicados esses vizinhos são comuns, é provável que os diferentes registros de autores se referem a mesma pessoa. Por fim, em alguns casos é necessário buscar na web o nome do autor para capturar dados adicionais, como instituição na qual trabalha, orientações etc.

Após o tratamento preliminar dos dados, extraiu-se 14 características dos pares de registros potencialmente ambíguos. A Tabela 3 apresenta todas as características agrupadas de acordo com o tipo de informação utilizada para cálculo/extração. Ressalta-se que as características dos grupos de mineração de texto e período de publicação usaram como base uma amostra de 1.828 artigos relacionados aos registros dos autores ambíguos.

Tabela 3 – Características extraídas das citações.

Tipo de característica	Característica	Descrição
características da rede social de coautorias	vizinhos em comum	número de vizinhos em comum na rede social acadêmica de todos os autores da DBLP
	são vizinhos	indica se o par de autores é ou não vizinho na rede social
características extraídas dos nomes	distância de edição	distância de edição entre os nomes dos dois autores
	distância relativa	proporção entre a distância de edição e o tamanho do menor nome dentre os autores
	primeiro nome diferente	indica se os autores têm seus primeiros nomes diferentes (um do outro)
	último nome diferente	indica se os autores têm seus últimos nomes diferentes (um do outro)
	proporção de diferentes nomes do meio	proporção (em relação ao número total de nomes) de nomes diferentes entre os autores, contabilizando nomes adicionais como diferentes
	proporção de diferentes abreviações	proporção (em relação ao número total de nomes) de nomes abreviados diferentes entre os autores, contabilizando abreviações adicionais como diferentes
	nomes invertidos	indica se há inversão da posição das partes dos nomes entre os autores
	nomes ou abreviações diferentes	indica a proporção de nomes efetivamente diferentes entre os autores (sem contar a presença de nomes adicionais)
características baseadas na mineração de texto	mineração de texto dos títulos dos artigos	métrica baseada em TFIDF que compara a frequência das palavras dos títulos entre os dois autores e entre o corpus formado pelos títulos de todos os autores avaliados
	log(MT)	logaritmo do valor resultante da mineração de texto
características baseadas nos anos de publicação dos artigos	intersecção do período de publicação	intersecção entre os períodos de publicação dos dois autores
	distância em anos entre publicações	distância mínima em anos entre as publicações dos dois autores (apenas se a intersecção for igual a zero)

Fonte: Digiampietri, Barbosa e Linden (2015)

4.2.2 Análises Bibliométricas

A partir dos dados de publicações é relativamente simples determinar a rede de coautoria, pois nos registros do DBLP de cada publicação há a lista dos identificadores dos autores. Para o estudo que analisa os 20 departamentos internacionais e os 8 programas nacionais, foram calculadas métricas de coautoria de 29 redes sociais.

Para cada artigo identificado, buscou-se o número de citações no site *Microsoft Academic Search*. A busca é baseada na comparação do título e do ano de publicação. Para confirmar se o artigo encontrado na busca é efetivamente o artigo procurado, desenvolveu-se um algoritmo que utilizava casamento aproximado de *strings*, permitindo uma distância de edição de no máximo 10% do tamanho do título do artigo. Ao todo, 73.232 artigos receberam 1.197.416 citações (658.721 para artigos em periódicos e 538.695 para artigos em eventos).

Tabela 4 – Medidas Bibliométricas Utilizadas

Abrev. Utilizada	Descrição
Pubs	Quantidade total de publicações em periódicos ou anais de eventos dos pesquisadores de um dado departamento.
PubsCo	Quantidade total de publicações em periódicos ou anais de eventos em coautoria com pesquisadores de outro departamento do estudo.
Cits	Quantidade total das citações às publicações dos artigos do departamento.
CitsCo	Quantidade total das citações às publicações dos artigos do departamento envolvidas em coautoria com pesquisadores de outro departamento.
PubsPP	Taxa de publicações por pesquisador.
PubsCoPP	Taxa de publicações em coautoria por pesquisador.
CitsPP	Taxa de citações por pesquisador.
CitsCoPP	Taxa de citações por pesquisador envolvendo as publicações em coautoria com pesquisadores de outro departamento.

Fonte: Digiampietri, Linden e Barbosa (2016)

Tabela 5 – Métricas da Análise de Redes Sociais Utilizadas

Abrev. Utilizada	Descrição
Medidas Locais	
Betweenness _{<i>l</i>}	Centralidade de intermediação (betweeness) - baseada na frequência em que um dado vértice aparece entre todos os caminhos mínimos da rede.
Closeness _{<i>l</i>}	Centralidade de proximidade (closeness)
Degree _{<i>l</i>}	Centralidade de grau - baseada no grau do vértice.
Eigen _{<i>l</i>}	Centralidade de autovalor (eigenvalue) - baseada na importância dos vértices que estão ligados ao vértice em análise.
Page-Rank _{<i>l</i>}	Centralidade PageRank - baseada na quantidade e qualidade das ligações existentes envolvendo o vértice em análise.
Medidas Globais	
Clique _{<i>g</i>}	Tamanho da clique máxima - tamanho da maior clique na rede.
MCC _{<i>g</i>}	Média dos caminhos mínimos - valor médio do caminho mínimo entre todos os pares de vértices da rede.
C.Clusterização _{<i>g</i>}	Coefficiente de clusterização - medida da transitividade das relações da rede.
Degree _{<i>g</i>}	Centralização de grau - medida do quão importante o vértice mais importante da rede é em relação aos demais (em relação ao grau).
Closeness _{<i>g</i>}	Centralização de proximidade - medida do quão importante o vértice mais importante da rede é em relação aos demais (em relação à proximidade).
Eigen _{<i>g</i>}	Centralização de autovalor - medida do quão importante o vértice mais importante da rede é em relação aos demais (em relação ao autovalor).
Assort.Grau	Assortatividade de grau - medida da tendência de nós de mesmo grau estarem ligados.
Diâmetro _{<i>g</i>}	Diâmetro - diâmetro da rede (distância mínima entre os dois nós mais afastados na rede).
Densidade _{<i>g</i>}	Densidade da rede - relação entre o número de arestas da rede e o número máximo de arestas possíveis na rede.

Fonte: Digiampietri, Linden e Barbosa (2016)

Um estudo adicional foi realizado considerando os 59 programas nacionais em Ciência da Computação que possuíam mestrado acadêmico ou doutorado no ano de 2010. Os métodos aplicados para o cálculo de métricas foram os mesmos dos programas internacionais. Adicionalmente, executou-se um algoritmo de classificação com o objetivo de tentar identificar a nota CAPES de cada programa de acordo com as métricas/características calculadas.

O sistema de avaliação dos programas de pós-graduação no Brasil usa cinco critérios, conforme estabelecido pela CAPES. Os pontos observados são: Objetivos do programa; Faculdade; Estudantes; Produção intelectual e Inserção social (políticas de extensão). Os programas considerados de qualidade recebem notas na escala de 3 a 7. A tabela 6 apresenta a última (triênio 2010-2012) classificação dos 59 programas de pós-graduação em Ciências da Computação analisados.

Tabela 6 – Classificação dos programas de pós-graduação brasileiros

Nota	Número de Programas
3	25
4	21
5	5
6	3
7	5

Fonte: Lênin Ferreira Barbosa, 2017

De acordo com a metodologia do CNPq, a atribuição do estrato QUALIS⁵ para a área Ciência da Computação utiliza como base o índice *JCR* para a classificação dos periódicos e o índice *h* para a classificação dos artigos em eventos. Cada veículo de publicação está classificado em um estrato entre oito possíveis, sendo estes A1, A2, B1 a B5 e C. Os artigos recebem uma pontuação específica dependendo do estrato do veículo em que foram publicados. A nota final do quesito produção intelectual de cada programa está relacionado à soma das publicações realizadas pelos pesquisadores do respectivo programa.

Os resultados dos três estudos realizados são apresentados no capítulo 5.

⁵ Disponível no endereço eletrônico <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacao>

5 Resultados

O presente capítulo apresenta os resultados tanto da desambiguação de nomes de autores quanto da caracterização dos programas nacionais em ciência da computação e departamentos internacionais com base na análise bibliométrica e de redes sociais. A primeira seção apresenta os resultados do estudo de desambiguação. Estes resultados foram utilizados antes da realização dos estudos das análises bibliométricas e de redes sociais, apresentados na segunda seção deste capítulo.

5.1 Desambiguação

Conforme apresentado no capítulo anterior, o processo de desambiguação proposto foi testado considerando os docentes permanentes do programa de Pós-Graduação em Ciência da Computação da UNICAMP. Dos 48 pesquisadores do programa, 82 apresentaram registros compatíveis com a DBLP. Detectou-se ambiguidades em 17 desses registros. Ao combinar os possíveis registros formaram-se ao todo 102 pares de referências que precisaram ser identificadas como referentes ou não a um mesmo pesquisador. Parte dos resultados apresentados nesta seção já foram publicados (DIGIAMPIETRI; BARBOSA; LINDEN, 2015).

Dentre o conjunto de características extraídas (vide tabela 3), a característica “são vizinhos” não foi verdadeira para nenhum dos pares. O valor verdadeiro indica que um par de referências a autores está em posições diferentes nas relações de coautoria de uma mesma publicação. Para os casos verdadeiros haveria uma menor probabilidade dos registros daquele par referirem-se a mesma pessoa, pois implicaria que um autor apareceu duas vezes na lista de autores de uma única publicação. Esta característica não se mostrou verdadeira para nenhum dos dados da amostra e por isso não será apresentada nas discussões a seguir.

O atributo *classe*, criado após verificação manual dos pares, indica se o par de registros do DBLP corresponde ou não a um mesmo autor. O valor *um* foi atribuído ao caso em que os dois registros pertencem ao mesmo autor, caso contrário, o valor *zero* foi atribuído. Com esse dado, calculou-se a correlação de Pearson entre o atributo *classe* e as características selecionadas. Valores positivos indicam que a característica está

correlacionada positivamente com o fato dos dois registros pertencerem ao mesmo autor. Correlações negativas indicam que quanto menor o valor da característica, maior a chance do valor da *classe* ser *um*.

Foram também calculadas as correlações entre todas as características para verificar quais são ou não correlacionadas (figura 2). Identificou-se que *proporção de diferentes nomes do meio* e *nomes ou abreviações diferentes* possuem correlação de 0,669 e *mineração de texto* e *log(MT)* possuem correlação de 0,741. Esse é um indicativo de que estes pares de características apresentam informações semelhantes (e, de certa forma, redundantes).

Figura 2 – Correlação entre características. As linhas e as colunas são as características apresentadas na tabela 3.

	Classe	vizinhos em comum	distância de edição	distância relativa	primeiro nome diferente	último nome diferente	proporção de diferentes nomes do meio	proporção de diferentes abreviações	nomes invertidos	nomes ou abreviações diferentes	mineração de texto	log(MT)	intersecção do período de publicação	distância em anos entre publicações
classe	1.000	0.320	-0.244	-0.304	-0.163	-0.464	-0.622	0.054	-0.066	-0.560	-0.033	0.173	0.159	-0.035
vizinhos em comum	0.320	1.000	-0.152	-0.166	-0.068	-0.192	-0.331	-0.088	0.133	-0.232	-0.019	0.096	0.100	-0.087
distância de edição	-0.244	-0.152	1.000	0.945	0.220	0.258	0.382	0.012	0.041	0.312	0.021	-0.007	-0.225	0.143
distância relativa	-0.304	-0.166	0.945	1.000	0.263	0.268	0.378	0.048	0.087	0.322	0.073	-0.002	-0.195	0.135
primeiro nome diferente	-0.163	-0.068	0.220	0.263	1.000	0.081	0.317	-0.142	0.652	0.273	-0.083	-0.131	0.013	0.005
último nome diferente	-0.464	-0.192	0.258	0.268	0.081	1.000	0.416	0.022	0.235	0.418	0.045	-0.159	-0.086	0.019
proporção de diferentes nomes do meio	-0.622	-0.331	0.382	0.378	0.317	0.416	1.000	-0.357	0.130	0.669	0.096	-0.057	-0.186	0.116
proporção de diferentes abreviações	0.054	-0.088	0.012	0.048	-0.142	0.022	-0.357	1.000	-0.142	0.092	-0.036	-0.041	0.085	-0.127
nomes invertidos	-0.066	0.133	0.041	0.087	0.652	0.235	0.130	-0.142	1.000	0.020	-0.065	-0.101	-0.075	-0.002
nomes ou abreviações diferentes	-0.560	-0.232	0.312	0.322	0.273	0.418	0.669	0.092	0.020	1.000	0.099	-0.008	-0.142	0.097
mineração de texto	-0.033	-0.019	0.021	0.073	-0.083	0.045	0.096	-0.036	-0.065	0.099	1.000	0.741	0.172	-0.110
log(MT)	0.173	0.096	-0.007	-0.002	-0.131	-0.159	-0.057	-0.041	-0.101	-0.008	0.741	1.000	0.341	-0.227
intersecção do período de publicação	0.159	0.100	-0.225	-0.195	0.013	-0.086	-0.186	0.085	-0.075	-0.142	0.172	0.341	1.000	-0.668
distância em anos entre publicações	-0.035	-0.087	0.143	0.135	0.005	0.019	0.116	-0.127	-0.002	0.097	-0.110	-0.227	-0.668	1.000

Fonte: Lênin Ferreira Barbosa, 2017

As três maiores correlações encontradas entre a *classe* e as demais características e atributos, conforme apresentado na figura 2, são *nomes e abreviações diferentes*, *proporção de diferentes nome do meio* e *último nome diferente*.

Essas correlações são negativas, indicando que maiores diferenças entre os nomes das referências dos autores estão correlacionadas com uma menor chance das referências serem de um mesmo autor. Este resultado era esperado, tendo em vista que os maiores desafios da desambiguação estão nos casos nos quais a comparação entre nomes por si só não consegue evidenciar se estes pertencem ou não a uma única pessoa.

Dentre as correlações não relacionadas aos nomes dos autores, a que obteve o maior valor positivo foi a característica *vizinhos em comum* (correlação igual a 0,32). Também obtiveram correlações positivas, porém menos intensas, a correlação com *log(MT)* e com

intersecção do período de publicação. No caso da característica *vizinhos em comum*, quanto maior o número de coautores em comum, maiores as chances das referências serem de um mesmo autor. As características relacionadas à Mineração de Texto objetivam verificar se há semelhanças entre os títulos das publicações dos autores. A correlação positiva indica que quanto maior esta semelhança, maior será a chance das publicações serem de um mesmo autor. A correlação positiva com *intersecção do período de publicação* indica que se os dois registros de autores do DBLP publicarem artigos em uma janela semelhante de tempo haverá maior probabilidade dos registros referenciarem a uma mesma pessoa.

Um outro modo de identificar quais características são mais influentes no processo de desambiguação é pelo uso de seletores. Neste projeto foram utilizados os seletores de atributos disponíveis na plataforma Weka, de forma a ranquear as características mais relevantes de acordo com a *classe*. A tabela 7 apresenta o nome do seletor utilizado e a ordem em que cada atributo/característica foi colocado no ranqueamento produzido pelo seletor.

Tabela 7 – Características ranqueadas por seletores de atributos

característica / seletor de atributos	GainRatioAttributeEval	CfsSubsetEval	ChiSquaredAttributeEval	FilteredAttributeEval	InfoGainAttributeEval	SymmetricalUncertAttributeEval	ReliefAttributeEval	SVMAttributeEval
vizinhos em comum							11	10
distância de edição							7	11
distância relativa							9	7
primeiro nome diferente	5		5	5	5	5	5	9
último nome diferente	3	1	3	3	3	3	1	3
proporção de diferentes nomes do meio	1	2	1	1	1	1	3	1
proporção de diferentes abreviações							8	8
nomes invertidos							13	12
nomes ou abreviações diferentes	2	3	2	2	2	2	2	2
mineração de texto	4	4	4	4	4	4	12	13
log(MT)	4		4	4	4	4	6	4
intersecção do período de publicação							4	5
distância em anos entre publicações							10	6

Fonte: Digiampietri, Barbosa e Linden (2015)

Observa-se pela tabela 7 que seis características foram definidas como as mais importantes pelos seletores. Elas foram ranqueadas de acordo com a informação que agregam no processo de resolução de entidades. Em confluência com os resultados das correlações, as características relacionados aos nomes *proporção de diferentes nomes do meio*, *nomes ou abreviações diferentes*, *último nome diferente* e *primeiro nome diferente* foram selecionadas. Entre as não vinculadas aos nomes, *mineração de texto* e $\log(MT)$ foram selecionadas pela maioria dos seletores. Assim, dentre as principais características selecionadas, estão apenas às relacionadas aos nomes dos autores e às métricas de mineração de textos utilizadas.

Na base de artigos levantados na revisão sistemática (vide capítulo 3), não se identificou o uso de mineração de textos aplicada a títulos na resolução de nomes de autores. Essa é uma informação relevante, tendo em vista que a mineração de textos auxiliou no processo de desambiguação. Uma segunda informação que vale destaque é que a *distância de edição*, tanto em seu valor absoluto quanto relativo (*distância relativa*), não se destacou como elemento agregador de informação. Esse resultado foi inesperado, tendo em vista que essa é uma medida amplamente citada nos trabalhos correlatos.

Com base nas características calculadas, optou-se por tratar o problema de resolução de nomes como um problema de classificação binária, no qual, cada par de referências a autores (registros de autores diferentes do DBLP) será classificado como pertencente ou não ao mesmo autor.

Com base em alguns resultados empíricos, optou por utilizar o classificador *Random Forest* (RODRIGUEZ; KUNCHEVA; ALONSO, 2006). O método escolhido para fazer a validação dos resultados foi a validação cruzada de 10 subconjuntos (*10 fold cross-validation*). As medidas de desempenho utilizadas foram: Taxa de Verdadeiro-Positivos (VP); Taxa de Falso-Positivos (FP); Precisão; Revocação; Medida-F¹; e Área sob a curva ROC. Estas medidas são apresentadas na tabela 8.

Tabela 8 – Desempenho da solução proposta.

classe	VP	FP	Precisão	Revocação	Medida-F	Área ROC
F	1	0,333	0,957	1	0,978	0,977
T	0,667	0	1	0,667	0,8	0,977
Média Ponderada	0,961	0,294	0,962	0,961	0,957	0,977

Fonte: Digiampietri, Barbosa e Linden (2015)

¹ Média harmônica da precisão e revocação

Pode-se notar que 100% dos casos nos quais de fato os registros eram de diferentes pesquisadores (Verdadeiro da Classe F [verdadeiro-negativos]) foram identificados corretamente, enquanto 66% dos casos em que eram iguais foram corretamente identificados, resultando numa média ponderada de acertos de 96,1%. Esses resultados foram considerados promissores, sendo superiores aos trabalhos correlatos que utilizaram a base DLBP (medida-F geral em torno de 0,93 (CULOTTA et al., 2007)). Porém, não se pode comparar diretamente os resultados porque não se sabe ao certo qual a amostra foi utilizada nos outros estudos.

A Tabela 9 apresenta a matriz de confusão resultante do processo de identificação dos pares. Como mencionado, observa-se que nenhum par negativo (isto é, um par de nomes referentes a pessoas diferentes) foi classificado como positivo. Os quatro erros de classificação ocorreram quando pares de nomes referentes à mesma pessoa foram identificados como sendo de pessoas diferentes. Nestes quatro casos um dos nomes do par a ser avaliado era composto apenas por dois nomes (primeiro nome e último sobrenome) e, além disso, nenhuma característica utilizada indicou evidências de que se referenciavam às mesmas pessoas.

Tabela 9 – Matriz de confusão

Classificado como			
F	V		
90	0	F	Classe correta
4	8	V	

Como uma etapa final do processo de desambiguação, não relacionada diretamente à solução automática proposta para esta atividade, foi realizada uma análise da quantidade de artigos científicos que seriam descartados em caso da não realização de um processo de desambiguação (informação relevante para estudos bibliométricos).

Para isto, foram analisados os registros DBLP dos 1.219 pesquisadores dos 59 programas brasileiros de pós-graduação em Ciência da Computação que possuíam doutorado ou mestrado acadêmico durante o triênio 2010 a 2012.

Dos 1.219 pesquisadores, 239 possuem de dois a quatro registros na base DBLP (19,6%). A análise apresentada a seguir identificou a quantidade de artigos que seriam ignorados por um sistema qualquer que considerasse apenas as produções do pesquisador mais produtivo no caso de registros ambíguos (isto é, um sistema ou uma análise que não realizasse desambiguação de nomes).

Os 239 pesquisadores com registros ambíguos possuem 5.659 artigos registrados na base do DBLP, porém, se para cada um deles considerarmos apenas o registro DBLP que publicou a maior quantidade de artigos, teremos 4.908 artigos. Isto é, 751 artigos seriam desconsiderados, ou seja, 13,27% da produção total destes pesquisadores. Na média, cada um desses pesquisadores possui 23,7 artigos na base do DBLP e, sem o processo de desambiguação, estariam sendo ignorados cerca de 3,14 artigos de cada um destes pesquisadores. Para alguns destes pesquisadores a soma das publicações de seu segundo (e, se houver, terceiro e quarto) registros no DBLP corresponde a menos de 1,2% de sua produção total cadastrada no DBLP. Porém, para outros, esta porcentagem chega a 60%.

Destaca-se que, ao todo, os 1.219 pesquisadores publicaram 33.553, assim, os 751 artigos, que não seriam considerados caso não houvesse um processo de desambiguação de nomes, correspondem à apenas 2,24% do total.

5.2 *Caracterização dos grupos*

Assim como na etapa de desambiguação, o processo de caracterização dos grupos iniciou-se com o cálculo ou a extração das características. Um dos objetivos do projeto foi identificar quais eram as características mais relevantes ou mais correlacionadas à posição de todos os departamentos em relação aos ranqueamentos internacionais. Uma análise adicional foi realizada considerando os programas nacionais de pós-graduação em Ciência da Computação.

5.2.1 Programas e departamentos internacionais

A figura 3 apresenta a primeira rede, indicando as coautorias interdepartamentais. A legenda (figura 4) indica qual cor corresponde a qual departamento. Nota-se que todos os programas nacionais estão localizados à esquerda da figura, com pouca colaboração internacional. Os departamentos internacionais, ao contrário, apresentam maior número de arestas entre si, portanto maior colaboração.

As figuras 5 e 6 apresentam a rede de pesquisadores e suas coautorias. O grafo da figura 5 indica todas as relações de coautorias existentes, enquanto o grafo da figura 6 apresenta apenas as coautorias existentes dentro de cada departamento.

Figura 3 – Rede de coautorias - departamentos

Figura 4 – Legenda de cores

Fonte: Digiampietri, Linden e Barbosa (2016)

Figura 5 – Rede de coautorias - todas as arestas

Figura 6 – Rede de coautorias - apenas arestas entre pesquisadores do mesmo departamento

Fonte: Digiampietri, Linden e Barbosa (2016)

Cinco métricas de centralidade individuais foram aplicadas exclusivamente ao grafo da figura 3, no qual cada departamento é um nó. Para cada departamento, foi criado um grafo em que cada pesquisador corresponde a um nó e cada aresta representa a relação de

coautoria entre dois pesquisadores. Assim, foram criados 28 grafos adicionais. Para esses, nove métricas globais de rede foram calculadas. As métricas encontram-se nas tabelas 4 e 5.

Assim, os 28 programas ou departamentos foram caracterizados por 34 métricas. De modo sintético, são elas o número de pesquisadores em cada departamento, as 16 medidas bibliométricas apresentadas, as cinco medidas de centralidade de cada departamento (na rede na qual cada departamento é um nó), nove medidas globais extraídas das redes de cada departamento (em que cada docente é um nó), e as posições desses departamentos nos três ranqueamentos utilizados.

Figura 7 – Ranqueamento dos programas de acordo com as métricas utilizadas

Departamento	Pesq	Pubs	PubsCo	Cits	CitsCo	PubsPP	PubsCoPP	CitsPP	CitsCoPP	Pubs_5anos	PubsCo_5anos	Cits_5anos	CitsCo_5anos	PubsPP_5anos	PubsCoPP_5anos	CitsPP_5anos	CitsCoPP_5anos	Betweenness_g	Closeness_g	Degree_g	Page-Rank_g	Eigen_g	Clique_g	MCC_g	C.Clusterizacao_g	Degree_g	Closeness_g	Eigen_g	Assort. Grau	Diametro_g	Densidade_g	Shanghai	TopUniversities	TheGuardian	
MIT EECS	17	11	6	5	5	4	4	2	3	13	13	8	19	6	11	5	18	17	11	11	13	10	10	13	27	18	24	9	18	12	24	2	1	1	
Stanford	11	12	3	6	3	9	3	4	1	12	6	5	5	9	7	6	7	5	3	3	4	3	4	7	18	6	19	6	20	3	25	1	2	2	
Oxford	19	13	19	13	14	7	21	13	14	9	24	4	6	4	21	2	5	26	18	18	20	17	10	11	11	26	20	11	7	6	15	15	3	3	
Harvard SEAS	24	20	11	16	8	11	2	10	4	21	10	21	10	8	2	15	6	21	11	11	12	11	23	20	20	14	12	20	23	22	14	4	4	4	
CMU	2	5	2	4	4	13	7	9	7	6	2	2	18	13	8	8	19	3	3	3	3	4	10	5	26	16	23	4	14	3	28	6	5	4	
UCBerkeley	4	1	1	1	1	5	1	1	2	3	1	6	4	7	3	9	9	2	1	1	2	1	1	14	9	15	5	15	6	6	23	3	7	6	
Cam	27	25	27	18	26	15	26	17	22	27	27	20	8	17	24	10	2	13	18	18	18	18	23	26	23	2	14	14	24	26	7	14	6	7	
ETH Zurich	4	9	16	11	13	16	22	18	17	8	15	11	25	15	19	14	26	10	7	7	7	8	10	12	24	7	13	12	25	12	18	13	8	9	
Princeton	1	10	4	3	2	21	16	12	5	10	5	19	20	27	15	25	23	6	5	5	6	5	23	2	25	25	26	8	21	1	26	5	9	9	
HKUST	12	8	17	15	18	6	18	19	19	5	21	15	17	3	22	19	17	22	14	14	15	16	10	10	14	20	15	13	1	6	13	18	11	7	
Toronto	17	15	14	14	15	17	13	15	15	18	14	9	22	19	12	7	21	16	9	9	10	9	10	1	17	12	16	7	22	1	16	11	10	11	
UCLA	20	2	9	9	10	1	6	3	10	1	3	7	11	1	1	3	10	4	5	5	5	7	10	25	16	4	27	2	13	25	20	9	12	12	
CUHK	23	26	24	23	20	25	19	22	20	25	26	27	26	26	27	26	25	27	22	22	27	20	23	27	14	11	21	16	27	26	1	17	14	13	
Cornell	6	6	5	2	6	12	10	7	9	7	8	10	9	14	13	13	15	11	7	7	8	6	10	3	22	24	25	3	5	3	27	8	15	19	
UTexas	9	7	7	7	7	8	9	6	8	11	4	16	13	12	4	20	13	8	11	11	11	15	4	6	12	22	11	10	15	6	22	7	19	15	
EPFL	13	4	12	8	12	2	11	5	12	4	16	13	15	2	18	16	12	7	9	9	9	13	23	4	19	13	18	5	17	6	19	16	13	17	
UW	21	16	10	12	9	10	5	8	6	17	11	23	21	10	5	22	20	25	14	14	17	12	10	15	21	8	22	19	26	22	17	19	16	14	
Caltech	28	28	28	19	24	19	25	16	18	28	28	24	27	16	23	11	27	28	28	28	28	21	28	28	28	28	28	28	1	28	26	5	10	16	16
Illinois	8	3	8	10	11	3	12	14	13	2	12	14	12	5	17	21	16	1	1	1	1	2	10	8	13	23	17	18	9	6	21	12	18	18	
UNICAMP	10	18	15	22	19	26	15	24	21	16	8	25	14	22	9	24	14	23	26	22	23	26	1	21	2	9	3	25	3	19	8	20	20	20	
USP São Paulo	14	27	20	26	21	28	23	25	25	24	20	28	23	28	20	28	24	15	22	22	24	25	4	19	1	17	10	26	12	12	3	22	20	20	
USP São Carlos	21	24	21	24	22	25	23	24	20	23	24	22	17	1	3	18	16	1	3	14	21	21	21	24	4	23	3	10	6	27	19	22	20	20	
BU	25	23	22	17	17	18	17	11	11	26	23	18	24	23	14	12	22	19	14	14	16	14	10	22	10	5	9	23	4	19	6	20	23	23	
UFMG	15	21	13	20	16	23	14	21	16	19	7	3	1	21	6	4	1	12	20	20	19	22	3	16	5	3	2	21	11	12	11	24	24	23	
UFRGS	7	14	25	24	25	20	27	26	26	14	19	12	2	20	26	18	4	9	17	17	14	19	4	18	7	21	8	17	2	12	9	24	24	23	
UFRJ	15	19	21	27	28	22	24	27	28	20	25	26	27	25	25	27	27	18	22	22	25	23	10	9	8	27	7	24	10	12	12	24	24	23	
UFPE	2	17	26	28	27	27	28	28	27	15	21	17	7	24	28	23	11	24	27	27	26	28	10	17	6	19	4	28	8	12	10	24	27	23	
PUC-Rio	25	22	18	21	22	14	8	20	23	23	17	22	16	11	10	17	8	20	22	22	22	22	27	4	24	4	1	1	22	16	22	2	24	27	23

Fonte: Digiampietri, Linden e Barbosa (2016)

Para cada uma destas medidas, os departamentos foram ranqueados da 1ª posição (maior valor para a métrica) até a 28ª (menor valor para a métrica), conforme apresentado na Figura 7. Os fundos das células estão coloridos de acordo com o valor do ranqueamento: quanto mais próximo de 1 a célula tem seu fundo com um verde mais intenso, por outro lado, células com valores próximos a 28 são coloridas com fundo vermelho. Destaca-se que o departamento com maior valor para uma dada métrica é ranqueado em 1º lugar, porém um valor maior para uma métrica não necessariamente representa uma característica positiva (por exemplo, alguns dos programas pior ranqueados possuem valores altos de centralização de proximidade ($Closeness_g$)).

Figura 8 – Correlações entre as métricas analisadas

	Pesq	Pubs	PubsCo	Cits	CitsCo	PubsPP	PubsCoPP	CitsPP	CitsCoPP	Pubs_5anos	PubsCo_5anos	Cits_5anos	CitsCo_5anos	PubsPP_5anos	PubsCoPP_5anos	CitsPP_5anos	CitsCoPP_5anos	Betweenness_I	Closeness_I	Degree_I	Page-Rank_I	Eigen_I	Clique_g	MCC_g	C.Clusterizacao_g	Degree_g	Closeness_g	Eigen_g	Assort.Grau	Diametro_g	Densidade_g	Shanghai	TopUniversities	TheGuardian
Pesq	1.00	0.63	0.48	0.35	0.36	0.02	0.10	0.01	0.09	0.66	0.54	0.37	0.26	-0.11	-0.01	-0.20	0.01	0.47	0.33	0.41	0.48	0.31	0.47	0.63	0.08	-0.24	0.13	-0.01	0.35	0.73	-0.53	0.14	-0.05	-0.00
Pubs	0.63	1.00	0.74	0.77	0.68	0.74	0.56	0.61	0.52	0.97	0.68	0.65	0.32	0.54	0.32	0.25	0.17	0.57	0.69	0.79	0.83	0.72	0.33	0.64	-0.24	-0.09	-0.27	0.43	0.24	0.63	-0.79	0.49	0.33	0.33
PubsCo	0.48	0.74	1.00	0.86	0.94	0.57	0.87	0.75	0.84	0.67	0.90	0.54	0.23	0.38	0.69	0.27	0.11	0.56	0.74	0.83	0.81	0.79	0.32	0.63	-0.01	0.05	-0.30	0.03	-0.01	0.57	-0.86	0.71	0.47	0.40
Cits	0.35	0.77	0.86	1.00	0.92	0.74	0.74	0.89	0.87	0.68	0.70	0.61	0.16	0.48	0.49	0.07	0.49	0.75	0.84	0.84	0.84	0.89	0.08	0.58	-0.61	0.02	-0.56	0.65	-0.14	0.53	-0.84	0.83	0.65	0.60
CitsCo	0.36	0.68	0.94	0.92	1.00	0.62	0.87	0.85	0.94	0.60	0.82	0.54	0.18	0.41	0.66	0.35	0.09	0.45	0.74	0.82	0.80	0.84	0.18	0.54	-0.45	0.08	-0.43	0.46	-0.15	0.49	-0.81	0.81	0.61	0.55
PubsPP	0.02	0.74	0.57	0.74	0.62	1.00	0.65	0.82	0.65	0.69	0.40	0.54	0.26	0.81	0.35	0.49	0.28	0.31	0.59	0.67	0.68	0.69	0.35	0.35	-0.44	0.04	-0.46	0.31	-0.03	0.28	-0.59	0.59	0.57	0.55
PubsCoPP	0.10	0.56	0.87	0.74	0.87	0.65	1.00	0.84	0.89	0.42	0.79	0.44	0.22	0.48	0.37	0.57	0.18	0.38	0.64	0.72	0.70	0.79	0.02	0.34	-0.56	0.17	-0.54	0.60	-0.22	0.28	-0.64	0.70	0.50	0.57
CitsPP	0.01	0.61	0.75	0.89	0.85	0.82	0.84	1.00	0.92	0.50	0.60	0.52	0.16	0.57	0.59	0.55	0.16	0.34	0.64	0.72	0.70	0.79	0.02	0.34	-0.56	0.17	-0.54	0.60	-0.22	0.28	-0.64	0.70	0.50	0.57
CitsCoPP	0.09	0.52	0.84	0.87	0.94	0.65	0.89	0.92	1.00	0.42	0.70	0.48	0.13	0.53	0.68	0.74	0.48	0.13	0.33	0.68	0.70	0.73	0.63	0.45	0.46	-0.15	-0.12	0.25	0.04	0.45	-0.71	0.57	0.27	0.22
Pubs_5anos	0.66	0.97	0.67	0.68	0.60	0.69	0.46	0.50	0.42	1.00	0.62	0.64	0.35	0.55	0.24	0.22	0.20	0.20	0.62	0.72	0.75	0.73	0.63	0.45	0.46	-0.15	-0.12	0.25	0.04	0.45	-0.71	0.57	0.27	0.22
PubsCo_5anos	0.54	0.68	0.90	0.70	0.82	0.40	0.79	0.60	0.70	0.62	1.00	0.52	0.37	0.22	0.74	0.24	0.24	0.54	0.59	0.70	0.73	0.63	0.45	0.46	-0.15	-0.12	0.25	0.04	0.45	-0.71	0.57	0.27	0.22	0.22
Cits_5anos	0.37	0.65	0.54	0.61	0.54	0.54	0.44	0.52	0.48	0.64	0.52	1.00	0.55	0.44	0.32	0.80	0.48	0.45	0.53	0.58	0.65	0.57	0.39	0.46	-0.21	0.16	-0.17	0.46	0.19	0.56	-0.61	0.42	0.43	0.40
CitsCo_5anos	0.26	0.32	0.23	0.16	0.18	0.26	0.21	0.16	0.16	0.35	0.37	0.55	1.00	0.29	0.23	0.02	0.04	0.94	0.32	0.17	0.19	0.29	0.11	0.37	0.06	0.26	0.14	0.28	-0.03	0.40	0.22	-0.23	0.09	0.08
PubsPP_5anos	0.11	0.54	0.38	0.48	0.41	0.81	0.48	0.57	0.46	0.55	0.22	0.44	0.29	1.00	0.45	0.47	0.34	0.33	0.57	0.51	0.46	0.45	-0.06	0.11	-0.23	-0.04	-0.24	0.50	0.09	0.03	-0.52	0.38	0.46	0.44
PubsCoPP_5anos	0.32	0.69	0.49	0.66	0.35	0.78	0.59	0.70	0.24	0.74	0.32	0.23	0.45	1.00	0.32	0.25	0.45	0.49	0.46	0.57	0.55	0.49	0.46	0.25	0.10	-0.04	0.30	0.02	0.23	0.00	0.00	-0.33	0.45	0.53
CitsPP_5anos	0.01	0.25	0.27	0.45	0.35	0.49	0.37	0.55	0.46	0.22	0.24	0.80	0.40	0.47	0.37	1.00	0.67	0.26	0.19	0.32	0.34	0.38	0.41	0.10	0.04	-0.34	-0.31	-0.26	0.53	-0.06	0.10	-0.33	0.45	0.53
CitsCoPP_5anos	0.47	0.57	0.55	0.49	0.45	0.31	0.34	0.33	0.52	0.54	0.32	0.33	0.43	0.19	0.57	0.57	0.32	0.09	0.80	0.75	0.75	0.58	0.18	0.36	-0.02	0.05	0.04	0.36	0.18	0.36	-0.65	0.33	0.12	0.05
Betweenness_I	0.33	0.69	0.74	0.75	0.74	0.59	0.61	0.64	0.68	0.62	0.59	0.53	0.17	0.57	0.57	0.32	0.09	0.22	1.00	0.80	0.80	0.80	0.88	0.06	0.48	-0.32	0.05	-0.25	0.51	0.01	0.40	-0.83	0.60	0.44
Closeness_I	0.41	0.79	0.83	0.84	0.82	0.67	0.70	0.72	0.74	0.70	0.70	0.58	0.19	0.51	0.55	0.34	0.10	0.75	0.96	1.00	0.97	0.95	0.16	0.56	-0.38	0.12	-0.31	0.49	-0.01	0.50	-0.83	0.68	0.53	0.44
Degree_I	0.48	0.83	0.81	0.84	0.80	0.68	0.66	0.70	0.73	0.73	0.68	0.57	0.29	0.46	0.49	0.38	0.19	0.75	0.90	0.97	1.00	0.93	0.21	0.58	-0.31	0.14	-0.30	0.48	0.01	0.56	-0.61	0.67	0.49	0.40
Page-Rank_I	0.31	0.72	0.79	0.89	0.84	0.68	0.71	0.79	0.81	0.63	0.63	0.57	0.11	0.45	0.46	0.41	0.02	0.58	0.88	0.95	0.93	1.00	0.05	-0.52	-0.56	0.09	-0.50	0.56	-0.18	0.46	-0.78	0.79	0.62	0.51
Eigen_I	0.47	0.33	0.32	0.08	0.18	0.06	0.23	0.02	0.03	0.35	0.45	0.39	0.37	-0.06	0.25	0.10	0.23	0.18	0.06	0.16	0.21	0.05	1.00	0.26	0.51	0.17	0.41	-0.33	0.55	0.42	-0.15	-0.11	-0.20	-0.11
Clique_g	0.63	0.64	0.63	0.58	0.54	0.35	0.34	0.34	0.37	0.60	0.46	0.46	0.06	0.11	0.10	0.04	-0.07	0.36	0.48	0.56	0.58	0.52	0.26	1.00	-0.23	-0.29	-0.20	0.20	0.08	0.92	-0.69	0.34	0.42	-0.61
MCC_g	0.08	-0.24	-0.31	-0.61	-0.45	-0.44	-0.32	-0.56	-0.52	-0.15	-0.15	-0.21	0.26	-0.23	-0.04	-0.34	0.22	-0.02	-0.32	-0.38	-0.39	-0.56	0.51	-0.23	1.00	-0.01	0.80	-0.67	0.66	-0.07	0.07	-0.42	-0.64	-0.28
C.Clusterizacao_g	-0.09	0.05	0.02	0.08	0.04	0.27	0.17	0.15	-0.12	0.21	0.16	0.14	-0.04	0.30	0.31	0.28	0.05	0.05	0.12	0.14	0.09	0.17	-0.29	-0.01	1.00	0.14	-0.14	-0.14	-0.25	-0.30	0.17	-0.02	0.13	0.11
Degree_g	0.13	-0.27	-0.30	-0.56	-0.43	-0.46	-0.30	-0.54	-0.46	-0.25	-0.12	-0.17	0.28	-0.24	0.02	-0.26	0.28	0.04	-0.25	-0.31	-0.30	-0.50	0.41	-0.20	0.80	1.00	0.01	-0.72	0.50	-0.09	0.20	0.25	1.00	-0.58
Closeness_g	0.01	0.43	0.38	0.65	0.46	0.51	0.31	0.60	0.48	0.37	0.25	0.46	-0.03	0.50	0.22	0.53	-0.03	0.36	0.51	0.01	-0.01	0.48	0.46	0.42	0.92	-0.07	-0.30	-0.09	0.20	0.25	1.00	-0.88	0.31	0.21
Eigen_g	0.24	-0.01	-0.14	-0.15	-0.03	-0.14	-0.22	-0.25	0.30	0.04	0.19	0.40	0.09	0.00	-0.06	-0.06	0.28	0.18	0.01	-0.01	0.01	-0.18	0.55	0.08	0.66	-0.25	0.50	-0.30	1.00	0.25	-0.03	-0.38	-0.42	-0.44
Assort.Grau	0.35	0.24	-0.01	-0.14	-0.15	-0.03	-0.14	-0.22	-0.25	0.30	0.04	0.19	0.40	0.09	0.00	-0.06	-0.06	0.28	0.18	0.01	-0.01	0.01	-0.18	0.55	0.08	0.66	-0.25	0.50	-0.30	1.00	0.25	-0.03	-0.38	-0.42
Diametro_g	0.73	0.63	0.57	0.53	0.49	0.29	0.26	0.26	0.30	0.61	0.45	0.56	0.22	0.03	0.00	0.10	0.04	0.36	0.40	0.50	0.56	0.46	0.42	0.92	-0.07	-0.30	-0.09	0.20	0.25	1.00	-0.88	0.31	0.21	0.21
Densidade_g	-0.53	-0.79	-0.86	-0.84	-0.81	-0.59	-0.64	-0.68	-0.72	-0.72	-0.71	-0.61	-0.23	-0.52	-0.55	-0.33	-0.10	-0.65	-0.83	-0.83	-0.81	-0.78	-0.42	-0.15	-0.69	0.42	0.17	0.35	-0.57	0.63	0.55	0.44	-0.44	-0.44
Shanghai	0.14	0.49	0.71	0.83	0.81	0.59	0.70	0.81	0.83	0.39	0.57	0.42	0.09	0.38	0.46	0.45	0.05	0.33	0.60	0.68	0.67	0.79	-0.11	0.34	-0.67	-0.02	-0.57	0.63	-0.38	0.31	-0.67	1.00	0.78	0.77
TopUniversities	-0.05	0.33	0.47	0.65	0.61	0.57	0.60	0.63	0.64	0.30	0.27	0.43	0.06	0.46	0.25	0.53	0.13	0.12	0.48	0.53	0.49	0.62	-0.20	0.24	-0.66	0.13	-0.54	0.55	-0.42	0.21	-0.45	0.78	0.97	1.00
TheGuardian	-0.02	0.33	0.46	0.60	0.59	0.55	0.50	0.58	0.62	0.30	0.28	0.42	0.08	0.45	0.25	0.49	0.14	0.06	0.43	0.48	0.45	0.57	-0.13	0.21	-0.60	0.11	-0.48	0.47	-0.40	0.21	-0.45	0.97	1.00	0.97

Fonte: Digiampietri, Linden e Barbosa (2016)

A Figura 8 contém a correlação de postos de Spearman calculada para todos os pares de métricas utilizados. As três últimas linhas e colunas contêm as correlações entre cada medida e os três ranqueamentos internacionais utilizados. Destaca-se que os três ranqueamentos possuem uma correlação maior ou igual a 0,76 entre si. De acordo com as próprias fontes, os ranqueamentos Shanghai e Top Universities possuem metodologias similares e portanto, esse não foi um resultado inesperado.

Valores positivos nessa tabela (nas últimas três linhas ou colunas) indicam que há uma correlação positiva entre a respectiva métrica e o ranqueamento do programa. Isto é, quanto maior for o valor da respectiva métrica, melhor ranqueado o programa será.

Observa-se que seis das sete medidas com maiores correlações positivas com os ranqueamentos são bibliométricas. Estas sete métricas são listadas a seguir, ordenadas de maneira decrescente pela média do valor da correlação com os três ranqueamentos: CitsCoPP, Cits, CitsPP, CitsCo, Page-Rank_l, PubsPP e PubsCoPP. Observa-se que a quantidade de citações em coautoria (com pesquisadores de outro departamento) por pesquisador foi a medida mais correlacionada com os ranqueamentos, seguida por outras três medidas relacionadas a citações. Isto é compatível com o conceito de que os melhores departamentos publicam artigos que recebem um número grande de citações (o que pode indicar uma pesquisa de boa qualidade e/ou publicada em veículos mais conhecidos e citados).

Analisando a média de citações por artigo dos programas ou departamentos analisados obtemos o valor 16,6. No caso dos dez melhor ranqueados, o valor sobe para cerca de 20 e, no caso dos programas brasileiros, esta média cai para perto de 5. Este indicador ilustra a grande diferença de algumas medidas entre os principais programas nacionais e os principais departamentos internacionais.

Das medidas de análise de redes sociais, a que apresentou a maior correlação positiva com os ranqueamentos foi a centralidade PageRank_l (0,66). Todas as medidas de centralidade analisadas (considerando a rede em que cada departamento foi representado como um nó) possuem correlações positivas com o ranqueamento.

Ao se analisar as correlações negativas com os ranqueamentos, destacam-se três medidas de redes (com correlações, na média, abaixo de -0,5). A primeira é o coeficiente de clusterização (C.Clusterização_g) com correlações entre 0,6 e 0,66. Isto indica que quanto maior a transitividade entre as relações de coautoria em um departamento pior o seu ranqueamento, sugerindo que o aumento de transitividade nas relações de coautoria dentro

de um departamento é prejudicial ao programa (o que poderia, por exemplo, ser um indício de um grupo muito fechado).

A segunda medida de destaque é a densidade da rede de cada um dos departamentos ($Densidade_g$). A correlação indica que quanto mais densa a rede é, pior será a posição do departamento, indicando que departamentos que têm muitas arestas internas tendem a não ter um ranqueamento muito bom.

Por fim, destaca-se a correlação negativa entre os ranqueamentos e a centralização de proximidade ($Closeness_g$), indicando que uma maior importância do pesquisador mais central de um dos departamentos em relação aos demais (considerando-se a proximidade como medida de importância) leva a uma tendência do departamento ser pior ranqueado. Isto indica que um departamento no qual não há apenas um elemento central (mas sim, que tem sua centralização dividida entre seus pesquisadores) tende a ser melhor ranqueado.

Para estes três índices, os programas brasileiros encontram-se na zona laranja-vermelho, isto é, perto das dez últimas posições. Na métrica $C.Clusterização_g$ os brasileiros estão nas últimas oito posições. Estas medidas indicam um relacionamento mais intenso entre os pesquisadores brasileiros de um mesmo programa e que cada programa é mais dependente de seu pesquisador mais central, diferentemente do que ocorre nos melhores programas internacionais.

Foi utilizada a técnica de Análise de Componentes Principais para mapear as 31 métricas dos departamentos (todas as métricas menos as classificações nos ranqueamentos internacionais) em um espaço bidimensional. A figura 9 apresenta os 28 departamentos mapeados de acordo com as duas componentes principais.

A primeira componente (eixo das abscissas na Figura 9) representa 51,4% da variância dos dados originais e está mais relacionada a medidas de “importância” dos departamentos (citações recebidas e as centralidades $PageRank_i$, $Degree_i$ e $Closeness_i$). Já a segunda componente (eixo das ordenadas) representa cerca de 13% da variância dos dados e está mais relacionada ao modo de “interação” dos pesquisadores de um departamento. As três medidas mais intensamente representadas na segunda componente são Assortatividade de Grau, Coeficiente de Clusterização e Tamanho da Clique Máxima. A componente principal teve correlação maior ou igual a 0,59 com todos os ranqueamentos internacionais utilizados, destacando-se a correlação de 0,81 com o ranqueamento Shanghai.

Destaca-se que nove dos dez departamentos melhor ranqueados encontram-se no lado direito da figura (valores positivos no eixo das abscissas). Por outro lado, dos dez

Figura 9 – Plotagem das duas componentes principais

Fonte: Digiampietri, Linden e Barbosa (2016)

pior ranqueados, oito se encontram do lado esquerdo da figura. Chama a atenção na figura que os oito programas de pós-graduação brasileiros analisados encontram-se na região esquerda inferior da figura (formando um grupo no qual não há nenhum departamento internacional entre os analisados), sugerindo que eles possuem características parecidas entre si e bastante diferentes dos demais departamentos.

5.2.2 Programas nacionais

A análise dos programas nacionais foi realizada de maneira similar a apresentada para os departamentos internacionais. Destaca-se duas características adicionais. A primeira é que diversos programas nacionais não estão presentes nos ranqueamentos internacionais, assim, analisou-se a relação entre as métricas calculadas e a classificação dada pela CAPES no triênio 2010-2012, chamado neste trabalho de *CAPES atual* e no triênio 2008-2010, chamado neste trabalho de *CAPES anterior*. Adicionalmente, duas outras métricas foram calculadas para cada programa: os índices h e g (calculados com base nas citações aos artigos extraídas do site Microsoft Academic Search).

A figura 10 apresenta a correlação entre todas as métricas. A última linha e a última coluna contêm as correlações entre cada uma das métricas e o posicionamento do programa de acordo com sua nota CAPES. Programas que possuem nota 7, ficaram com

a 1ª posição de ranqueamento de acordo com a nota, assim como programas com nota três ficaram empatados na última posição do ranqueamento.

Figura 10 – Correlações entre as métricas dos programas nacionais

C.Clusterização_g	Pesq	1.00	0.85	0.77	0.70	0.68	0.36	0.69	0.51	0.81	0.75	0.83	0.71	0.79	0.72	0.53	0.35	0.67	0.54	0.72	0.72	0.53	0.22	-0.25	0.04	0.31	0.39	0.79	-0.85	0.71	0.67	
	Pubs	0.85	1.00	0.88	0.89	0.74	0.91	0.80	0.71	0.81	0.92	0.73	0.97	0.83	0.89	0.81	0.85	0.38	0.84	0.77	0.82	0.86	0.87	0.87	0.81	0.73	0.81	0.80	-0.77	0.79	0.75	
	PubsCo	0.77	0.88	1.00	0.82	0.89	0.78	0.84	0.72	0.81	0.68	0.76	0.84	0.84	0.80	0.87	0.72	0.77	0.77	0.82	0.84	0.77	0.85	0.85	0.82	0.67	0.74	0.15	-0.06	0.20	0.17	
	Cits	0.79	0.89	0.88	1.00	0.81	0.79	0.66	0.96	0.70	0.95	0.88	0.84	0.74	0.90	0.79	0.49	0.86	0.75	0.78	0.82	0.84	0.84	0.84	0.83	0.69	0.66	-0.26	-0.19	0.13	0.18	
	CitsCo	0.70	0.74	0.89	0.81	1.00	0.64	0.77	0.74	0.66	0.87	0.66	0.79	0.71	0.84	0.54	0.63	0.84	0.54	0.63	0.81	0.72	0.70	0.74	0.73	0.59	0.62	0.14	-0.04	0.10	0.20	
	PubsPP	0.58	0.91	0.78	0.79	0.64	1.00	0.69	0.77	0.59	0.83	0.72	0.88	0.76	0.78	0.74	0.93	0.66	0.79	0.74	0.74	0.74	0.76	0.76	0.75	0.64	0.70	0.18	-0.03	0.23	0.02	
	PubsCoPP	0.36	0.60	0.84	0.56	0.77	0.69	1.00	0.55	0.68	0.81	0.51	0.56	0.81	0.54	0.59	0.63	0.90	0.57	0.71	0.64	0.54	0.60	0.43	0.47	0.03	0.16	0.25	0.03	0.09	0.43	
	CitsPP	0.63	0.79	0.72	0.96	0.74	0.67	0.56	1.00	0.88	0.88	0.87	0.95	0.75	0.65	0.84	0.72	0.68	0.47	0.85	0.70	0.71	0.76	0.75	0.75	0.62	0.52	-0.15	0.13	0.09	0.13	
	CitsCoPP	0.51	0.61	0.81	0.70	0.96	0.50	0.83	0.68	1.00	0.72	0.68	0.54	0.72	0.60	0.70	0.50	0.62	0.60	0.78	0.62	0.57	0.62	0.63	0.62	0.48	0.47	0.11	0.05	0.24	0.14	
	C.Clusterização_g	IndiceH	0.81	0.92	0.86	0.96	0.83	0.83	0.88	0.72	1.00	0.80	0.90	0.87	0.87	0.80	0.70	0.52	0.82	0.70	0.79	0.86	0.86	0.86	0.72	0.27	0.13	0.15	0.15	0.27	0.69	-0.73
IndiceG		0.75	0.83	0.76	0.98	0.77	0.72	0.51	0.97	0.68	0.90	1.00	0.78	0.69	0.87	0.75	0.66	0.45	0.84	0.71	0.74	0.78	0.80	0.80	0.79	0.68	0.57	0.29	-0.22	0.08	0.18	
Pubs_5anos		0.83	0.97	0.84	0.67	0.88	0.56	0.77	0.54	0.86	0.77	0.88	0.78	0.90	0.82	0.83	0.74	0.78	0.81	0.83	0.83	0.82	0.77	0.24	-0.14	0.17	0.23	0.14	0.13	0.28	0.71	
PubsCo_5anos		0.71	0.83	0.94	0.74	0.79	0.78	0.81	0.65	0.72	0.67	0.69	0.86	1.00	0.78	0.82	0.77	0.88	0.73	0.78	0.80	0.77	0.82	0.82	0.81	0.64	0.68	0.11	-0.03	0.14	0.28	
Cits_5anos		0.79	0.89	0.80	0.90	0.71	0.78	0.54	0.84	0.60	0.87	0.67	0.88	0.78	1.00	0.82	0.76	0.56	0.98	0.80	0.74	0.80	0.81	0.81	0.80	0.66	0.67	0.19	-0.16	0.10	0.21	
CitsCo_5anos		0.72	0.81	0.87	0.79	0.64	0.74	0.69	0.72	0.78	0.80	0.75	0.78	0.82	0.82	1.00	0.69	0.64	0.77	0.79	0.99	0.73	0.70	0.75	0.74	0.72	0.60	0.68	0.15	-0.18		
PubsPP_5anos		0.53	0.85	0.72	0.70	0.54	0.83	0.63	0.69	0.50	0.70	0.66	0.50	0.77	0.76	0.69	1.00	0.70	0.77	0.79	0.68	0.68	0.70	0.70	0.70	0.62	0.63	0.16	0.00	0.29	0.08	
PubsCoPP_5anos		0.35	0.58	0.72	0.49	0.63	0.66	0.80	0.47	0.67	0.52	0.45	0.62	0.88	0.56	0.64	0.70	1.00	0.57	0.66	0.62	0.55	0.50	0.50	0.40	0.42	-0.02	0.12	0.18	0.13		
CitsPP_5anos		0.67	0.94	0.74	0.96	0.67	0.79	0.57	0.65	0.80	0.82	0.84	0.83	0.73	0.98	0.77	0.57	1.00	0.80	0.70	0.75	0.76	0.75	0.75	0.58	0.58	0.16	-0.12	0.12	0.15		
CitsCoPP_5anos		0.64	0.77	0.82	0.75	0.81	0.74	0.71	0.70	0.78	0.76	0.71	0.74	0.78	0.80	0.99	0.70	0.66	0.80	1.00	0.69	0.65	0.69	0.67	0.55	0.62	0.14	-0.15	0.27	0.17		
C.Clusterização_g	Betweenness_I	0.72	0.82	0.84	0.78	0.72	0.74	0.64	0.71	0.82	0.79	0.74	0.78	0.80	0.74	0.73	0.68	0.62	0.70	0.68	0.62	0.70	0.68	0.62	0.70	0.68	0.55	0.75	0.69	0.95	0.67	
	Closeness_I	0.79	0.85	0.79	0.82	0.70	0.74	0.64	0.74	0.57	0.86	0.78	0.81	0.77	0.80	0.70	0.68	0.55	0.75	0.69	0.68	0.95	0.97	1.00	0.98	0.63	0.66	0.12	-0.15	0.04	0.29	
	Degree_I	0.81	0.87	0.85	0.84	0.74	0.78	0.60	0.78	0.63	0.86	0.80	0.83	0.82	0.81	0.75	0.59	0.76	0.69	0.95	0.97	1.00	1.00	0.98	0.62	0.63	0.61	0.10	0.10	0.08	0.29	
	Page-Rank_I	0.80	0.87	0.85	0.84	0.74	0.78	0.61	0.78	0.63	0.86	0.80	0.83	0.82	0.81	0.74	0.70	0.60	0.75	0.69	0.96	0.97	1.00	1.00	0.98	0.62	0.63	0.11	-0.16	0.03	0.29	
	Eigen_I	0.80	0.85	0.82	0.83	0.73	0.75	0.58	0.75	0.62	0.86	0.79	0.82	0.81	0.80	0.72	0.70	0.59	0.75	0.69	0.89	0.98	1.00	1.00	0.93	0.63	0.65	0.12	-0.10	0.03	0.24	
	Clique_g	0.71	0.76	0.67	0.69	0.59	0.64	0.43	0.62	0.48	0.72	0.68	0.77	0.64	0.66	0.60	0.62	0.40	0.58	0.55	0.61	0.63	0.62	0.63	0.63	0.55	0.61	-0.10	0.15	0.01	0.40	
	MCC_g	0.75	0.80	0.74	0.66	0.62	0.70	0.47	0.52	0.47	0.70	0.57	0.77	0.68	0.67	0.68	0.63	0.43	0.58	0.52	0.62	0.62	0.66	0.65	0.63	0.65	1.00	0.02	0.00	0.44	-0.03	
	C.Clusterização_g	0.22	0.25	0.15	0.28	0.14	0.18	0.03	0.25	0.11	0.27	0.29	0.24	0.11	0.19	0.15	0.16	-0.02	0.16	0.14	0.04	0.11	0.12	0.11	0.12	0.61	0.02	-1.00	-0.30	-0.10	-0.24	
	Degree_g	-0.25	-0.15	-0.06	-0.19	-0.04	-0.03	0.16	-0.15	0.05	-0.13	-0.22	-0.14	-0.03	-0.16	-0.18	0.00	0.12	-0.12	-0.15	-0.18	-0.15	-0.14	-0.15	-0.10	-0.10	-0.30	1.00	0.32	-0.10	-0.41	
	Closeness_g	0.04	0.20	0.10	0.13	0.19	0.23	0.25	0.13	0.24	0.15	0.24	0.17	0.14	0.11	0.05	0.29	0.18	0.12	0.27	0.08	0.10	0.15	0.04	-0.01	0.10	0.32	1.00	-0.52	-0.09	0.34	
Eigen_g	0.31	0.17	0.16	0.20	0.22	0.03	0.09	0.14	0.18	0.23	0.21	0.19	0.18	0.23	0.22	0.11	0.19	0.08	0.13	0.15	0.17	0.24	0.31	0.22	0.15	-0.52	1.00	-0.02	0.00	-0.55		
Assort.Grau	0.39	0.31	0.28	0.20	0.18	0.16	0.09	0.13	0.12	0.27	0.19	0.31	0.28	0.22	0.24	0.10	0.10	0.17	0.21	0.21	0.24	0.26	0.29	0.24	0.40	0.10	0.60	0.40	-0.09	-0.02		
Diâmetro_g	0.79	0.80	0.73	0.65	0.60	0.67	0.43	0.50	0.43	0.69	0.57	0.78	0.66	0.67	0.67	0.60	0.40	0.56	0.50	0.61	0.62	0.65	0.65	0.62	0.68	0.96	0.10	0.10	0.14	0.24		
Densidade_g	-0.85	-0.77	-0.77	-0.72	-0.65	-0.59	-0.48	-0.58	-0.50	-0.73	-0.67	-0.78	-0.76	-0.78	-0.66	-0.57	-0.51	-0.64	-0.67	-0.64	-0.67	-0.68	-0.69	-0.69	-0.66	-0.69	1.00	-0.69	0.62	0.61		
CapesAnterior	0.71	0.79	0.66	0.78	0.59	0.74	0.43	0.73	0.48	0.77	0.73	0.78	0.82	0.76	0.82	0.76	0.65	0.69	0.61	0.74	0.73	0.78	0.76	0.60	0.61	0.11	0.11	0.00	-0.55	-0.14		
CapesAtual	0.67	0.75	0.63	0.73	0.59	0.69	0.43	0.68	0.51	0.72	0.71	0.75	0.81	0.73	0.83	0.67	0.43	0.69	0.61	0.67	0.67	0.67	0.67	0.66	0.56	0.57	0.27	0.16	0.07	0.11		
C.Clusterização_g	Pesq	1.00	0.85	0.77	0.70	0.68	0.36	0.69	0.51	0.81	0.75	0.83	0.71	0.79	0.72	0.53	0.35	0.67	0.54	0.72	0.72	0.53	0.22	-0.25	0.04	0.31	0.39	0.79	-0.85	0.71	0.67	
	Pubs	0.85	1.00	0.88	0.89	0.74	0.91	0.80	0.71	0.81	0.92	0.73	0.97	0.83	0.89	0.81	0.85	0.38	0.84	0.77	0.82	0.86	0.87	0.87	0.81	0.73	0.81	0.80	-0.77	0.79	0.75	
	PubsCo	0.77	0.88	1.00	0.82	0.89	0.78	0.84	0.72	0.81	0.68	0.76	0.84	0.84	0.80	0.87	0.72	0.77	0.77	0.82	0.84	0.77	0.85	0.85	0.82	0.67	0.74	0.15	-0.06	0.20	0.17	
	Cits	0.79	0.89	0.88	1.00	0.81	0.79	0.66	0.96	0.70	0.95	0.88	0.84	0.74	0.90	0.79	0.49	0.86	0.75	0.78	0.82	0.84	0.84	0.84	0.83	0.69	0.66	-0.26	-0.19	0.13	0.18	
	CitsCo	0.70	0.74	0.89	0.81	1.00	0.64	0.77	0.74	0.66	0.87	0.66	0.79	0.71	0.84	0.54	0.63	0.84	0.54	0.63	0.81	0.72	0.70	0.74	0.73	0.59	0.62	0.14	-0.04	0.10	0.20	
	PubsPP	0.58	0.91	0.78	0.79	0.64	1.00	0.69	0.77	0.59	0.83	0.72	0.88	0.76	0.78	0.74	0.93	0.66	0.79	0.74	0.74	0.74	0.76	0.76	0.75	0.64	0.70	0.18	-0.03	0.23	0.02	
	PubsCoPP	0.36	0.60	0.84	0.56	0.77	0.69	1.00	0.55	0.68	0.81	0.51	0.56	0.81	0.54	0.59	0.63	0.90	0.57	0.71	0.64	0.54	0.60	0.43	0.47	0.03	0.16	0.25	0.03	0.09	0.43	
	CitsPP	0.63	0.79	0.72	0.96	0.74	0.67	0.56	1.00	0.88	0.88	0.87	0.95	0.75	0.65	0.84	0.72	0.68	0.47	0.85	0.70	0.71	0.76	0.75	0.75	0.62	0.52	-0.15	0.13	0.09	0.13	
	CitsCoPP	0.51	0.61	0.81	0.70	0.96	0.50	0.83	0.68	1.00	0.72	0.68	0.54	0.72	0.60	0.70	0.50	0.62	0.60	0.78	0.62	0.57	0.62	0.63	0.62	0.48						

Fonte: Linden, Digiampietri e Barbosa (2017)

Observa-se na figura 10 uma correlação bastante alta entre as cinco medidas de centralidade calculadas. Destaca-se também que a métrica *Densidade_g* é a que possui as maiores correlações negativas com as demais métricas.

A figura 11 corresponde a um recorte da figura 10 com a correlação das métricas com a nota CAPES atual de cada programa. A alta correlação positiva com a nota CAPES do triênio anterior (*CapesAnterior*) é esperada, pois a maioria dos programas não tem sua nota CAPES alterada de um ano para o outro e, para aqueles que sofrem essa alteração, ela corresponde à mudança de apenas um ponto de mudança na classificação.

Figura 11 – Correlações das métricas dos programas nacionais com a nota Capes

Métrica	Correlação com CAPES
CapesAnterior	0.85
Pubs	0.75
Pubs_5anos	0.75
Cits_5anos	0.73
Cits	0.73
IndiceH	0.72
IndiceG	0.71
Closeness_l	0.70
Eigen_l	0.69
CitsPP_5anos	0.69
PubsPP	0.69
CitsPP	0.68
Pesq	0.67
Degree_l	0.67
Page-Rank_l	0.67
PubsPP_5anos	0.67
Clique_g	0.65
CitsCo_5anos	0.63
PubsCo	0.63
CitsCoPP_5anos	0.61
Diâmetro_g	0.61
PubsCo_5anos	0.61
Betweenness_l	0.60
CitsCo	0.59
MCC_g	0.57
CitsCoPP	0.51
PubsCoPP	0.43
PubsCoPP_5anos	0.43
C.Clusterização_g	0.27
Assort.Grau	0.20
Eigen_g	0.11
Closeness_g	0.07
Degree_g	-0.16
Densidade_g	-0.59

Fonte: Lênin Ferreira Barbosa, 2017

Assim como para os departamentos internacionais, as principais métricas correlacionadas positivamente com a avaliação (no caso, a nota CAPES) são bibliométricas. Porém há diferenças entre quais são essas métricas. Para os programas nacionais, as duas principais métricas foram número de publicações e número de publicações nos últimos cinco anos, seguidas por duas métricas relacionadas a citações (citações nos últimos cinco anos

e número total de citações). Estas quatro métricas são seguidas por mais duas medidas bibliométricas derivadas das citações: o índice h e o índice g.

Em seguida, aparecem duas medidas de centralidade da rede na qual cada programa corresponde a um nó: centralidade de proximidade ($Closeness_i$) e de autovalor ($Eigen_i$).

Destaca-se ainda que há mais 19 métricas com correlações positivas acima de 0,5 com a classificação CAPES.

Diferentemente da análise das correlações negativas envolvendo os departamentos internacionais, para os programas nacionais apenas uma métrica se destacou com um valor absoluto alto para sua correlação negativa com a classificação. Mesmo assim, a mesma métrica obteve o menor valor de correlação em ambos os casos, a densidade das redes que representam cada departamento ($Densidade_g$), servindo como evidência adicional de que um programa é pior ranqueado se houver uma quantidade relativa maior de arestas entre os seus membros. Duas hipóteses levantadas são: que programas cujas colaborações ficam centradas apenas dentro do próprio programa potencialmente produzirão uma menor quantidade de artigos e/ou eventualmente ficarão mais restritos aos mesmos assuntos. Adicionalmente, um esforço em colaboração interna pode implicar em menos tempo/disposição para colaborações externas.

Conforme apresentado, a análise de componentes principais dos programas nacionais e internacionais (figura 9) identificou o agrupamento dos programas nacionais num dos quadrantes. Realizou-se uma análise similar, utilizando apenas os programas nacionais e o resultado da visualização dos programas nas duas principais componentes pode ser observado na figura 12.

Figura 12 – Plotagem das duas componentes principais dos programas nacionais

Fonte: Digiampietri, Linden e Barbosa (2016)

Observa-se que a classificação dos programas está mais relacionada ao eixo das abscissas enquanto, novamente, o eixo das ordenadas está mais relacionado à forma na qual os programas se relacionam. Os programas com notas 7 encontram-se a direita da figura. No extremo oposto encontra-se a maioria dos programas nota 3. Em torno do valor zero do eixo das abscissas estão a maioria dos programas nota 4. Todos os programas com nota 5 e 6 encontram-se do lado direito da figura.

Segundo essa figura, observa-se uma mistura considerável entre programas nota 3 e 4. Duas hipóteses foram levantadas para esta característica. A primeira é que as métricas utilizadas não são suficientes para uma boa separação entre estes programas. A segunda, é que existe uma pressão muito grande para a mudança de notas entre programas nota 3 e 4, tornando essa faixa de avaliação muito competitiva, com diversos programas no limite para subir de nota e alguns no limite para cair.

Para cada uma das métricas estudadas, calculou-se uma função utilizando regressão linear para estimar o valor da avaliação CAPES atual em relação a estas métricas.

A figura 13 apresenta as 35 funções calculadas e os respectivos valores de R^2 . Apesar das funções mapearem métricas de diferentes naturezas em apenas cinco valores possíveis da avaliação CAPES, algumas funções obtiveram valores de R^2 relativamente altos. Treze destas funções apresentaram R^2 maior do que 0,7, destacando-se o valor $R^2 = 0,82526$ para a função que utiliza o número total de publicações (*Pubs*) e corresponde à $CAPES_Atual = 3,034 + 0,002 * Pubs$.

Com base nas funções calculadas é possível estimar uma nota CAPES para um dado programa existente ou para um grupo de pesquisadores que pretenda propor a criação de um programa de pós-graduação. A figura 14 apresenta os valores estimados de nota CAPES para um dos programas que atualmente possui nota 3.

Observa-se pela figura 14 que a nota estimada para este programa (durante o período analisado), pela maioria das métricas, está próxima de 3,5. A média dessas estimativas foi de 3,64, sendo a menor estimativa igual a 3,06 (estimada utilizando o índice h) e máxima de 4,06 considerando o coeficiente de clusterização.

Por fim, utilizou-se uma combinação de seleção de atributos e classificação para se realizar uma verificação experimental da capacidade dos atributos utilizados para a inferência da nota CAPES dos programas. Para isto foram utilizadas as implementações dos algoritmos disponíveis no arcabouço Weka.

Figura 13 – Resultado da regressão linear aplicada a cada métrica analisada

CAPES_Atual =	2.274	+	0.084 *Pesq	R ² =	0.58904
CAPES_Atual =	3.034	+	0.002 *Pubs	R ² =	0.82526
CAPES_Atual =	3.024	+	0.007 *PubsCo	R ² =	0.75322
CAPES_Atual =	3.419	+	0.000 *Cits	R ² =	0.72154
CAPES_Atual =	3.567	+	0.001 *CitsCo	R ² =	0.53513
CAPES_Atual =	2.464	+	0.087 *PubsPP	R ² =	0.72071
CAPES_Atual =	2.755	+	0.200 *PubsCoPP	R ² =	0.46455
CAPES_Atual =	3.311	+	0.015 *CitsPP	R ² =	0.65373
CAPES_Atual =	3.445	+	0.033 *CitsCoPP	R ² =	0.48680
CAPES_Atual =	2.621	+	0.110 *IndiceH	R ² =	0.77624
CAPES_Atual =	2.812	+	0.052 *IndiceG	R ² =	0.75089
CAPES_Atual =	3.009	+	0.005 *Pubs_5anos	R ² =	0.78365
CAPES_Atual =	3.056	+	0.016 *PubsCo_5anos	R ² =	0.72349
CAPES_Atual =	3.805	+	0.001 *Cits_5anos	R ² =	0.30559
CAPES_Atual =	3.705	+	0.003 *CitsCo_5anos	R ² =	0.46268
CAPES_Atual =	2.249	+	0.221 *PubsPP_5anos	R ² =	0.68437
CAPES_Atual =	2.746	+	0.488 *PubsCoPP_5anos	R ² =	0.45474
CAPES_Atual =	3.742	+	0.056 *CitsPP_5anos	R ² =	0.36151
CAPES_Atual =	3.655	+	0.147 *CitsCoPP_5anos	R ² =	0.48577
CAPES_Atual =	3.270	+	0.288 *IndiceH_5anos	R ² =	0.48487
CAPES_Atual =	3.467	+	0.079 *IndiceG_5anos	R ² =	0.53567
CAPES_Atual =	3.412	+	0.015 *Betweenness_l	R ² =	0.71843
CAPES_Atual =	-4.328	+	1236.8 *Closeness_l	R ² =	0.72313
CAPES_Atual =	2.413	+	0.094 *Degree_l	R ² =	0.76835
CAPES_Atual =	2.166	+	150.3 *Page-Rank_l	R ² =	0.76370
CAPES_Atual =	2.229	+	4.022 *Eigen_l	R ² =	0.73778
CAPES_Atual =	1.548	+	0.725 *Cliq_ug	R ² =	0.49647
CAPES_Atual =	2.197	+	0.696 *MCC_g	R ² =	0.32199
CAPES_Atual =	3.804	+	1.342 *C.Clusterizacao_g	R ² =	0.02064
CAPES_Atual =	4.404	+	-1.734 *Degree_g	R ² =	0.01953
CAPES_Atual =	3.650	+	4.518 *Closeness_g	R ² =	0.15903
CAPES_Atual =	1.238	+	3.238 *Eigen_g	R ² =	0.03122
CAPES_Atual =	4.230	+	0.522 *Assort.Grau	R ² =	0.00747
CAPES_Atual =	2.659	+	0.259 *Diametro_g	R ² =	0.31130
CAPES_Atual =	5.361	+	-18.64 *Densidade_g	R ² =	0.34779

Fonte: Lênin Ferreira Barbosa, 2017

O seletor de atributos *CfsSubsetEval* identificou seis atributos como os mais relevantes em relação à classe: *Pubs*, *Cits*, *PubsPP*, *PubsCo_5anos*, *Closeness_l* e *C.Clusterizacao_g*.

Utilizou-se a validação cruzada em 10 subconjuntos para testar a classificação utilizando o um classificador *Naive-Bayes*. Dos 59 programas, 41 foram classificadas corretamente, o que equivale à 69,5% do total. A classificação dos 18 programas que foram classificados incorretamente (30,5%) ocorreu por apenas um ponto na nota CAPES. A tabela 10 apresenta o sumário dos resultados da classificação.

Ao se analisar a matriz de confusão (tabela 11), observa-se que os programas notas 3 e 7 apresentaram a maior taxa de acerto na classificação, respectivamente 88% e 100%. Todos os erros de classificação qualificaram a nota do programa uma classe acima ou abaixo da correta. Os programas com nota 6 e 4 foram aqueles que apresentaram os piores resultados na classificação (33% e 43%, respectivamente).

Figura 14 – Estimativa da nota CAPES considerando a regressão linear para métricas individuais

Fonte: Lênin Ferreira Barbosa, 2017

Tabela 10 – Resultado da classificação utilizando Naive-Bayes

Medida	(Valor)
Instâncias classificadas corretamente	41 (69,5 %)
Instâncias classificadas incorretamente	18 (30,5 %)
Estatística Kappa	0,55
Erro médio absoluto	0,13
Erro médio quadrático	0,31

Fonte: Lênin Ferreira Barbosa, 2017

Tabela 11 – Matrix de confusão da classificação dos programas nacionais

Matrix de confusão					
Classificado como					
a	b	c	d	e	Classe correta
5	0	0	0	0	a = 7
1	1	1	0	0	b = 6
0	0	4	1	0	c = 5
0	0	2	9	10	d = 4
0	0	0	3	22	e = 3

Fonte: Lênin Ferreira Barbosa, 2017

Dos programas com nota 6, um foi classificado corretamente, um foi classificado como 7 e outro como 5. Dos programas nota 4, nove foram classificados corretamente, dois foram classificados como nota 5 e 10 como nota 3.

6 Conclusão

Conforme apresentado ao longo desta dissertação, a tarefa de quantificar a qualidade dos programas de pós-graduação não é trivial. As metodologias dos ranqueamentos internacionais citados, assim como a realizada pela CAPES, englobam aspectos de produção acadêmica, inclusão social, reconhecimento entre outros.

Considerando apenas dados oriundos da produção acadêmica, buscou-se nesta dissertação encontrar quais as medidas bibliométricas e de análise de redes sociais acadêmicas de maior influência sobre o posicionamento dos programas ou departamentos. Com o uso de uma revisão sistemática (capítulo 3), fez-se o levantamento das bases de dados bibliográficas ou acadêmicas, de quais os procedimentos necessários para tratar esses registros e quais as principais métricas utilizadas em trabalhos correlatos.

Após a escolha da base de dados, resolveu-se a ambiguidade autoral dos registros considerando quatorze características extraídas dos dados da DBLP, conforme subseção 4.2.1. As importâncias das características sobre os autores foram analisadas considerando-se a correlação de Pearson e diferentes seletores de atributos. Tratou-se a desambiguação como um problema de classificação binária, no qual um metaclassificador foi utilizado para combinar as características extraídas para pares de referências a autores. O metaclassificador separou os pares em *pertencentes à mesma pessoa* ou *não pertencentes à mesma pessoa*.

Para a amostra selecionada, atingiu-se uma precisão média de 96% e uma medida-F superior a 0,95. Esse resultado foi positivo e superior aos trabalhos correlatos, contudo algumas considerações devem ser destacadas. A amostra utilizada neste estudo possuía poucos casos em que o par representava a mesma pessoa (apenas 12 instâncias). Outro ponto é que os trabalhos correlatos coletaram as informações bibliométricas de outras bases além do DBLP ou mesmo usando outros registros da base do DBLP, o que limita a comparação de resultados.

Considerando os 1.219 pesquisadores permanentes dos programas nacionais de pós-graduação em Ciência da Computação, observou-se que para os pesquisadores da amostra, 19% deles possuem mais de um registro na base DBLP. Se apenas um registro fosse considerado em uma análise bibliométrica, o estudo desconsideraria, na média, mais de três artigos de cada um desses pesquisadores.

A base do DBLP continuamente desambigua os registros autorais. Ao se avaliar quantos artigos estariam sendo ignorados caso o estudo realizado neste trabalho não realizasse um processo desambiguação, observou-se que apenas 2% dos artigos seriam desconsiderados. Para essa base específica, a não utilização de métodos de desambiguação não prejudicaria significativamente estudos bibliométricos sobre o conjunto total.

Após a desambiguação da base, realizou-se o estudo bibliométrico e de análise de redes sociais para identificar quais as métricas têm maior relevância em relação aos ranqueamentos internacionais utilizados (Top Universities, Shangai e The Guardian). Selecionou-se 20 departamentos internacionais e 8 programas de pós-graduação nacionais (considerados de nível internacional), dos quais 34 métricas foram extraídas ou calculadas. Um segundo estudo bibliométrico, com as mesmas métricas, foi efetuado considerando apenas os programas nacionais de pós-graduação em Ciência da Computação avaliados pela CAPES com notas entre 3 e 7. De modo similar ao estudo de desambiguação, a relevância das métricas foi avaliada em relação à correlação de Pearson com os ranqueamentos internacionais para os 28 departamentos e com o ranqueamento nacional para todos os programas avaliados pela CAPES. Complementou-se o resultado com o uso de diferentes seletores de atributos, conforme seções 5.2.1 e 5.2.2.

A análise dos departamentos internacionais indicou que as métricas que influenciam positivamente um departamento estão ligadas a publicações, citações e medidas de centralidade. As posições dos programas são consistentes dentro do modelo de valoração de pesquisa usados pelos mesmos (e que aproximadamente coincidem nos três ranqueamentos internacionais apresentados).

A análise de componentes principais considerando os 28 departamentos indicou que os programas nacionais estão próximos, mas suas características são distintas àquelas dos melhores programas mundiais. Assim, é possível levantar uma hipótese de que existe uma “ciência à brasileira”, cujos critérios de qualidade parecem ser exatamente os mesmos adotados mundialmente. Evidenciou-se que os programas brasileiros têm um nível de cooperação entre si e com departamentos internacionais distintos do que é comum entre os melhores programas mundiais.

A segunda análise de componentes principais, contendo os programas de pós-graduação nacionais, indicou um agrupamento dos mesmos em relação as suas notas e métricas. Notou-se que os programas com melhores avaliações possuíam menores métricas

de transitividade e de centralização, embora a correlação dessas não tenha sido forte com o sistema de ranqueamento nacional.

Fez-se uma estimativa/inferência das notas dos programas brasileiros utilizando as métricas extraídas. A primeira estimativa, baseada em regressões lineares, indicou que o número total de publicações possui uma melhor aproximação com a nota CAPES.

O uso de um classificador Naive Bayes para a estratégia utilizando aprendizado de máquina e seleção de atributos mais relevantes na definição de notas obteve maior precisão para os programas com maior e menor notas, respectivamente. A classificação dos programas nota 4 obteve índices baixos de acerto. Esta informação, em conjunto com a análise de componentes principais, permite observar que há uma alta mistura dos programas nota 4 com aqueles nota 3, ao menos em relação às métricas avaliadas.

A conclusão dessa dissertação é de que é possível estimar a avaliação de um programa com base em métricas bibliométricas e de análise de redes sociais. Mesmo sem considerar os demais critérios utilizados para a definição final da nota e levando em consideração o sistema discreto de notas aplicado pela CAPES, pode-se obter um resultado com o intervalo máximo de uma classe de incerteza.

Como esperado, tanto nos ranqueamentos internacionais quanto no nacional, o número de publicações e citações é essencial para garantir um bom posicionamento. Em contrapartida, a métrica de densidade foi considerada um critério negativo em ambos os sistemas. Em relação às outras métricas, foi possível observar que há diferenças quanto ao impacto de cada uma entre os sistemas. Contudo, futuras investigações precisam ser feitas para corroborar e detalhar estas considerações.

Um dos caminhos para aumentar o nível de reconhecimento dos programas nacionais pode ser o incremento da colaboração com outros departamentos internacionais. Um maior número de conexões entre departamentos pode auxiliar o desenvolvimento de novas ideias e fortalecer os programas em conjunto.

Como trabalhos futuros, pode-se desenvolver um projeto focado exclusivamente na classificação dos programas realizada pela CAPES, considerando informações adicionais como as orientações realizadas.

Referências¹

- ABBASI, A.; HOSSAIN, L.; LEYDESDORFF, L. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, v. 6, n. 3, p. 403 – 412, 2012. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S175115771200003X>. Citado na página 26.
- ARIF, T.; ALI, R.; ASGER, M. Author name disambiguation using vector space model and hybrid similarity measures. In: *Contemporary Computing (IC3), 2014 Seventh International Conference on*. [S.l.: s.n.], 2014. p. 135–140. Citado na página 26.
- CHEN, H.-H. et al. Collabseer: A search engine for collaboration discovery. In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2011. (JCDL '11), p. 231–240. ISBN 978-1-4503-0744-4. Disponível em: <http://doi.acm.org/10.1145/1998076.1998121>. Citado na página 25.
- CIMENLER, O.; REEVES, K. A.; SKVORETZ, J. An evaluation of collaborative research in a college of engineering. *Journal of Informetrics*, v. 9, n. 3, p. 577 – 590, 2015. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157715000516>. Citado na página 26.
- CORMEN, T. H. et al. *Introduction to Algorithms*. Cambridge, MA: MIT Press, 2001. Citado na página 16.
- CULOTTA, A. et al. Author disambiguation using error-driven machine learning with a ranking loss function. In: *Sixth International Workshop on Information Integration on the Web (IIWeb2007)*. [S.l.: s.n.], 2007. Citado na página 44.
- DATTA, S. et al. Factors influencing research contributions and researcher interactions in software engineering: An empirical study. In: *Software Engineering Conference (APSEC), 2013 20th Asia-Pacific*. [S.l.: s.n.], 2013. v. 1, p. 34–41. ISSN 1530-1362. Citado na página 25.
- DIAZ-VALENZUELA, I.; MARTÍN-BAUTISTA, M. J.; VILA, M. A. A fuzzy semisupervised clustering method: Application to the classification of scientific publications. In: A., L. et al. (Ed.). *Information Processing and management of uncertainty in knowledge-based systems - 15th International Conference, IPMU 2014*. [S.l.: s.n.], 2014. p. 179–188. Citado na página 18.
- DIESTEL, R. *Graph Theory. Graduate Texts in Mathematics*. 3. ed. New York: Springer-Verlag, 2006. Citado na página 16.
- DIGIAMPIETRI, L.; LINDEN, R.; BARBOSA, L. Caracterizando departamentos e programas de computação utilizando análise de redes sociais e bibliometria. In: *V Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2016)*. [S.l.: s.n.], 2016. Citado 8 vezes nas páginas 34, 35, 38, 46, 47, 48, 51 e 54.
- DIGIAMPIETRI, L. A. *Análise da Rede Social Acadêmica Brasileira*. [S.l.], 2015. Tese de Livre Docência. Citado 3 vezes nas páginas 16, 19 e 20.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

DIGIAMPIETRI, L. A.; BARBOSA, L. F.; LINDEN, R. Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP. In: *IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015) - Anais do XXXV Congresso da Sociedade Brasileira de Computação (CSBC2015)*. [S.l.: s.n.], 2015. Citado 6 vezes nas páginas 18, 35, 37, 40, 42 e 43.

DIGIAMPIETRI, L. A. et al. Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLoS ONE*, Public Library of Science, v. 9, n. 4, p. e94541, 04 2014. Citado 3 vezes nas páginas 26, 29 e 30.

DIODATO, V. P. *Dictionary of Bibliometrics*. [S.l.: s.n.], 1994. Citado na página 12.

FERREIRA, A. A.; GONCALVES, M. A.; LAENDER, A. H. F. A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, v. 41, n. 2, p. 15–26, 2012. Citado na página 19.

FERREIRA, A. A.; MACHADO, T. M.; GONCALVES, M. A. Improving author name disambiguation with user relevance feedback. *Journal of Information and Data Management*, v. 3, n. 3, p. 332–347, 2012. Citado 2 vezes nas páginas 18 e 36.

FREEMAN, L. C. Centrality in social networks: Conceptual clarification. *Social Networks*, v. 1, p. 215–239, 1979. Citado na página 20.

FREIRE, V.; FIGUEIREDO, D. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, Springer-Verlag, v. 17, n. 4, p. 255–266, 2011. ISSN 0104-6500. Disponível em: <http://dx.doi.org/10.1007/s13173-011-0041-7>. Citado na página 26.

HAN, S. et al. Recommending program committee candidates for academic conferences. In: *Proceedings of the 2013 Workshop on Computational Scientometrics: Theory & Applications*. New York, NY, USA: ACM, 2013. (CompSci '13), p. 1–6. ISBN 978-1-4503-2414-4. Disponível em: <http://doi.acm.org/10.1145/2508497.2508498>. Citado na página 28.

HUYNH, T. et al. Collaborator recommendation for isolated researchers. In: *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on*. [S.l.: s.n.], 2014. p. 639–644. Citado na página 25.

LAENDER, A. et al. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM Special Interest Group on Computer Science Education*, ACM, v. 40, n. 2, p. 135–145, 2008. Citado na página 12.

LI, N.; GILLET, D. Identifying influential scholars in academic social media platforms. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA: ACM, 2013. (ASONAM '13), p. 608–614. ISBN 978-1-4503-2240-9. Disponível em: <http://doi.acm.org/10.1145/2492517.2492614>. Citado na página 29.

LIMA, J. G. S. *Gerenciamento de Dados Climatológicos Heterogêneos para Aplicações em Agricultura*. Dissertação (Mestrado) — IC-UNICAMP, October 2003. Orientação C. B. Medeiros. Citado na página 19.

LINDEN, R.; DIGIAMPIETRI, L.; BARBOSA, L. “Brazilian Style Science” - An Analysis of the Difference between Brazilian and International Computer Science Departments and Graduate Programs Using Social Networks Analysis and Bibliometrics. *Social Network Analysis and Mining*, v. 7, n. 44, 2017. Citado 2 vezes nas páginas 35 e 52.

LUNA, J.; REVOREDO, K.; COZMAN, F. Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, Springer London, v. 19, n. 4, p. 397–409, 2013. ISSN 0104-6500. Disponível em: <http://dx.doi.org/10.1007/s13173-013-0108-8>. Citado na página 26.

MENA-CHALCO, J.; CESAR, R. Towards automatic discovery of co-authorship networks in the brazilian academic areas. In: *e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*. [S.l.: s.n.], 2011. p. 53–60. Citado na página 25.

MILOJEVIC, S. Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, v. 7, n. 2, p. 767–773, 2013. Citado na página 18.

ORTEGA, J. L. Influence of co-authorship networks in the research impact: Ego network analyses from microsoft academic search. *Journal of Informetrics*, v. 8, n. 3, p. 728 – 737, 2014. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157714000613>. Citado na página 26.

ORTEGA, J. L.; AGUILLO, I. F. Institutional and country collaboration in an online service of scientific profiles: Google scholar citations. *Journal of Informetrics*, v. 7, n. 2, p. 394 – 403, 2013. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157713000023>. Citado na página 26.

RODRIGUEZ, J.; KUNCHEVA, L.; ALONSO, C. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 28, n. 10, p. 1619–1630, Oct 2006. ISSN 0162-8828. Citado na página 43.

SARIGOL, E. et al. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, Springer Berlin Heidelberg, v. 3, n. 1, 2014. Disponível em: <http://dx.doi.org/10.1140/epjds/s13688-014-0009-x>. Citado na página 26.

SMALHEISER, N. R.; TORVIK, V. I. Author name disambiguation. *Annual Review of Information Science and Technology*, Wiley Subscription Services, Inc., A Wiley Company, v. 43, n. 1, p. 1–43, 2009. ISSN 1550-8382. Disponível em: <http://dx.doi.org/10.1002/aris.2009.1440430113>. Citado na página 17.

STROTMANN, A.; ZHAO, D. Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, v. 63, n. 9, p. 1820–1833, 2012. Citado na página 18.

SZWARCFITER, J. L. *Grafos e algoritmos computacionais*. [S.l.]: Campus, 1986. 216 p. Citado na página 16.

TAGUE-SUTCLIFFE, J. An introduction to informetrics. *Information Processing & Management*, v. 28, n. 1, p. 1 – 3, 1992. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/030645739290087G>. Citado 2 vezes nas páginas 19 e 20.

TYLEND, T.; ANGELOVA, R.; BEDATHUR, S. Towards time-aware link prediction in evolving social networks. In: *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*. New York, NY, USA: ACM, 2009. (SNA-KDD '09), p. 9:1–9:10. ISBN 978-1-60558-676-2. Disponível em: <http://doi.acm.org/10.1145/1731011.1731020>. Citado na página 25.

WASSERMAN, S.; FAUST, K. *Social network analysis: methods and applications*. 19. ed. [S.l.]: Social network analysis: methods and applications, 2009. Citado na página 16.

WHITE, H. D.; MCCAIN, K. W. Visualizing a discipline: An author co-citation analysis of information science 1972-1995. *Journal of the American Society for Information Science*, v. 49, n. 4, p. 327–355, 1988. Citado na página 18.

YAN, X.; ZHAI, L.; FAN, W. C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics*, v. 7, n. 1, p. 223 – 239, 2013. ISSN 1751-1577. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1751157712000946>. Citado na página 26.

YANG, Z.; HONG, L.; DAVISON, B. D. Topic-driven multi-type citation network analysis. In: *Adaptivity, Personalization and Fusion of Heterogeneous Information*. Paris, France, France: Le centre de hautes etudes internationales d'informatique documentaire, 2010. (RIAO '10), p. 24–31. Disponível em: <http://dl.acm.org/citation.cfm?id=1937055.1937062>. Citado na página 25.