UM ESTUDO SOBRE A GENEALOGIA ACADÊMICA BRASILEIRA

WELLINGTON JOSÉ DAS DÔRES

UM ESTUDO SOBRE A GENEALOGIA ACADÊMICA BRASILEIRA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação. como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto Henrique Frade Laender
Coorientador: Fabrício Benevenuto de Souza

Belo Horizonte
Dezembro de 2017

Dôres, Wellington José das

D695e Um estudo sobre a genealogia acadêmica brasileira / Wellington José das Dôres. – Belo Horizonte, 2017
xxii, 64 f. : il. ; 29cm

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Alberto Henrique Frade Laender
Coorientador: Fabrício Benevenuto de Souza

1. Computação – Teses. 2. Redes Complexas 3. Árvores genealógicas acadêmicas. 4. Plataforma Lattes. I. Orientador. II. Coorientador. III. Título

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Um estudo sobre a genealogia acadêmica brasileira

WELLINGTON JOSÉ DAS DÔRES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ESTEVAM BARBOSA DE LAS CASAS
Departamento de Engenharia de Estruturas - UFMG

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 21 de Dezembro de 2017.

Agradecimentos

Os meus mais sinceros agradecimentos a todos que me apoiaram durante esses dois anos de amplas vivências e experiências que enriqueceram a minha vida pessoal, profissional e acadêmica. Gostaria de agradecer especialmente

- À minha família que sempre me apoiou seja de longe ou próxima de mim;
- À Carolina Barone que divide comigo as lutas e as glórias de viver e que me apoia sempre;
- Aos amigos Amir, Jedson, Rodrigo, Guilherme e Manu;
- Aos amigos de sempre, que independente da distância estarão sempre comigo;
- Aos amigos do LDB - Laboratório de Bancos de Dados, principalmente aqueles com quem compartilhei bons papos na hora do almoço;
- Ao David e ao Sadraque que, além de exemplos, me deram todo o apoio necessário;
- Ao Elias pela parceria e contribuição neste trabalho;
- Ao Prof. Alberto H. F. Laender por estar sempre à disposição, pelo apoio e confiança para que pudesse seguir em frente;
- Ao Prof. Fabrício Benevenuto de Souza que esteve presente desde o início da minha carreira acadêmica;
- Ao Thiago Magela Rodrigues Dias pela cessão da coleção de dados sobre os doutores obtida da Plataforma Lattes;
- Aos colegas do LoCuS - Laboratório de Computação Social;
- Ao DCC/ICEx que sempre ofereceu um ambiente propício à pesquisa;
- À CAPES, CNPq e Fapemig por financiarem parcialmente este trabalho.

"De tudo, ficaram três coisas: a certeza de que ele estava sempre começando, a certeza de que era preciso continuar e a certeza de que seria interrompido antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança, do medo uma escada, do sono uma ponte, da procura um encontro."

(Fernando Sabino)

Resumo

Ao longo da história, muitos pesquisadores contribuíram de maneira notável para a ciência, não apenas no avanço do conhecimento, mas também na mentoria de novos pesquisadores. Atualmente, identificar e estudar a formação de novos pesquisadores ao longo dos anos é uma tarefa desafiadora, uma vez que os repositórios atuais contendo dados sobre as orientações acadêmicas estão estruturadas de forma descentralizada em diversos sítios espalhados pela Web. Nesta dissertação, foi construída uma coleção de árvores genealógicas acadêmicas que mostram as relações orientador-orientando tanto no mestrado quanto no doutorado. As árvores foram construídas a partir de dados extraídos dos currículos de todos os doutores cadastrados na Plataforma Lattes até abril de 2017. Para isso, foi desenvolvido um algoritmo capaz de processar os dados de cada currículo, desambiguar os nomes dos pesquisadores e encontrar as relações tanto explícitas quanto implícitas de orientação entre cada um dos pesquisadores presentes nos currículos coletados. Este trabalho também inclui uma análise das árvores genealógicas acadêmicas geradas considerando as diferentes grandes áreas do conhecimento conforme definidas pelo CNPq. Para tal, foram definidas métricas que auxiliam o entendimento da estrutura e a evolução das diferentes árvores construídas. Os resultados apresentados mostram como algumas árvores se destacam das demais no contexto da ciência brasileira. Foram também detectadas diferenças entre as árvores das grandes áreas do conhecimento. Mais importante, as árvores construídas podem ser acessadas por um portal aberto à comunidade científica, possibilitando entender um pouco mais o avanço e as contribuições em termos de mentoria de novos pesquisadores.

Palavras-chave: Redes Complexas, Árvores Genealógicas Acadêmicas, Plataforma Lattes.

Abstract

Along the history, many researchers provided remarkable contributions to science, not only advancing knowledge but also in terms of mentoring new scientists. Currently, identifying and studying the formation of new researchers over the years is a challenging task, since current repositories of theses and dissertations are organized in a decentralized way through many digital libraries spread across the Web. In this dissertation, we built a collection of academic genealogy trees that show the relationships between advisors and advisees at both Master's and PhD levels. These trees were build from data extracted from curricula of all researchers with a PhD degree registered at the Lattes Platform until April 2017. To do that, we developed an algorithm for processing data from each curricula, disambiguating researchers' names and finding both explicit and implicit advising relationships involving all researchers present in each collected curriculum. Our work also includes an analysis of the academic genealogy trees built, considering the different knowledge areas defined by CNPq. For such a purpose, we defined specific metrics to help understanding the structure and the evolution of each genealogy tree built. Our results show that some of the trees are more remarkable than others in the context of the Brazilian science. We also detected differences in the trees from different knowledge areas. More important, the genealogy trees built can be accessed through a web portal open to the public, making it possible to understand better the advances and contributions in terms of mentoring new researchers.

Keywords: Complex Networks, Academic Genealogy Trees, Lattes Platform.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A ciência vive em constante evolução, de modo que novas teorias, tecnologias e áreas vão surgindo a cada dia e outras são substituídas com o passar do tempo. No início, a ciência estava em grande parte diretamente associada à religião e à filosofia. Com o passar do tempo, a ciência evoluiu para diferentes ambientes e em diferentes ritmos, que permitiram responder inúmeros desafios e construir diversos pilares da sociedade atual. Recontar essa história, através dos laços de orientação entre pesquisadores, é entender como foram criados os alicerces que permeiam a nossa sociedade.

Em termos do Brasil, as primeiras universidades datam do início do século XX bem distante do surgimento das primeiras universidades do mundo, tendo os primeiros professores das instituições brasileiras vindo de diversos países da Europa [Schwartzman, 2006]. Isso e o fato de diversos acadêmicos brasileiros terem buscado suas titulações no exterior liga a ciência brasileira diretamente às raízes da ciência mundial. As origens do doutorado aqui no Brasil encontram-se definidas na Lei de Diretrizes e Bases da Educação Brasileira (LDB 4024/61) que entrou em vigor em 1961, incluindo formalmente os cursos de pós-graduação como parte integrante da estrutura da educação brasileira.

Apesar de relativamente recente, o ensino superior no Brasil vem se consolidando cada vez mais, com o surgimento de novos programas de pós-graduação espalhados por todo o país. De acordo com os resultados da última avaliação da CAPES[1], no quadriênio 2013-2016 houve um aumento de 25% no número de programas de pós-graduação reconhecidos no Brasil. Do mesmo modo, o número de doutores cresce a cada ano, chegando a centenas de milhares de pesquisadores com o grau de doutor, de acordo com dados da Plataforma Lattes [Dias, 2016]. Esses números começaram a atrair mais e mais a atenção para trabalhos envolvendo a produção científica e as redes de cola-

[1] http://avaliacaoquadrienal.capes.gov.br

boração acadêmica dos mais diversos grupos de pesquisa do país [Canto & Hannah, 2001; Delgado-Garcia et al., 2014; Laender et al., 2008; Mena-Chalco et al., 2014; Silva et al., 2017]. Mais recentemente, um tipo particular de rede, as árvores genealógicas acadêmicas [Dores et al., 2016; Damaceno et al., 2017], começaram a chamar a atenção. Apesar de ainda não serem amplamente estudadas, são tão importantes quanto as demais. Uma das principais funções de um professor, além da pesquisa, é a formação de recursos humanos. Como essas redes se desenvolvem em relação à formação de novos doutores e também como elas têm evoluído é o tema desta dissertação. Árvores genealógicas são uma forma de representar graficamente as gerações de grupos de pessoas como, por exemplo, no caso de uma família. Portanto, nesta dissertação, para representar as relações de tutoria entre professores e estudantes de pós-graduação, adotou-se o termo Árvore Genealógica Acadêmica que nada mais é do que um grafo direcionado acíclico [Bang-Jensen & Gutin, 2008] que descreve as relações de formação acadêmica nos programas de pós-graduação das diversas áreas do conhecimento.

A relação de tutoria pode ser decisiva na carreira de um indivíduo, uma vez que este pode seguir os passos do tutor e tender a copiar as suas melhores habilidades. Nas árvores genealógicas acadêmicas, o elo entre os indivíduos se dá através da relação entre o orientador e seus estudantes. Em algumas áreas já existem esforços para se construir as árvores genealógicas acadêmicas como, por exemplo, os casos da Matemática [Jackson, 2007], da Física [Chang, 2003] e da Neurociência [David & Hayden, 2012], entre outros. Esse tipo de rede guarda em suas relações toda a estrutura que envolve uma ou mais áreas do conhecimento. Através delas pode-se visualizar as relações entre diferentes áreas como, por exemplo, Ciência da Computação e Engenharia Elétrica, ou Biologia e Medicina. Essas árvores também registram quem são os grandes formadores de recursos humanos nas diferentes áreas do conhecimento.

Ainda segundo a última avaliação quadrienal da CAPES, o Brasil conta com 4.175 programas de pós-graduação, divididos entre mestrado profissional, mestrado acadêmico e doutorado. Todos os anos, são concedidos dezenas de milhares de títulos de mestre e doutor aos estudantes desses programas espalhados pelo país. Com as árvores genealógicas acadêmicas pode-se observar melhor como se dá a formação desses mestres e doutores, e como tem ocorrido o crescimento da pós-graduação no país. Apesar disso, ainda não há nenhum estudo que ofereça uma análise sobre como se dá a formação de novos pesquisadores ou mesmo responda a questões envolvendo a origem da formação desses pesquisadores.

Uma das grandes dificuldades para se construir tais árvores é a falta de dados em formato digital e também a forma como esses dados estão organizados. Grande parte desses dados está distribuída por repositórios descentralizados, muitas vezes, até

mesmo privados. Outro problema é que, apesar de haver uma padronização sugerida para o armazenamento desses dados em bibliotecas digitais [Lagoze & Van de Sompel, 2001], não há um cuidado por parte dos administradores em seguir as padronizações sugeridas. Apesar de suas próprias restrições, a Plataforma Lattes do CNPq[2] ainda é uma das melhores fontes para tal aplicação, sendo considerada modelo para vários países [Lane, 2010].

Uma das estratégias mais comuns para a coleta de grandes volumes de dados hoje em dia é o chamado *crowdsourcing* [Howe, 2006], onde o esforço de uma multidão é empregado para se obter dados que, sem tal esforço, seria quase impossível obtê-los. Grandes exemplos de sistemas de *crowdsourcing* são a própria Plataforma Lattes, a Wikipédia[3] (escrita e correção de artigos sobre diversos temas) e o aplicativo Waze[4] (informação sobre trânsito em tempo real), entre muitos outros. Assim, com a ajuda do processo de *crowdsourcing* é possível atingir patamares de coleta de dados que outros métodos, por melhores que sejam, não são capazes de conseguir. Um bom exemplo de um esforço baseado em *crowdsourcing* para a geração de árvores genealógicas acadêmicas é o Mathematics Genealogy Project [Jackson, 2007] que registra a formação acadêmica de mais de 221.000 matemáticos de todo o mundo.

Entretanto, o processo de *crowdsourcing* muitas vezes apresenta uma taxa de crescimento muito baixa ou por vezes acaba por se extinguir por falta de contribuidores. Um exemplo é o projeto Theoretical Computer Science Genealogy [Johnson, 1984] que tentou reunir a genealogia de todos os pesquisadores que trabalham na área de Teoria da Computação. O projeto contou com o apoio do SIGACT[5], mas aparentemente encontra-se inativo. Esse exemplo mostra que apesar de promissor o projeto necessitava do apoio de toda a comunidade envolvida, através do preenchimento de um formulário na WWW, para alcançar seus objetivos. Entretanto, de modo geral, grande parte dos dados sobre pesquisadores encontra-se espalhada pela WWW, seja em repositórios acadêmicos, bibliotecas digitais ou mesmo em bancos de dados de projetos como o da Matemática, muitas vezes de acesso restrito.

1.1 Objetivos da Dissertação

Os objetivos desta dissertação são construir e analisar um tipo específico de rede, as árvores genealógicas acadêmicas, que, assim como as redes de coautoria, colaboração

[2] http://lattes.cnpq.br
[3] https://www.wikipedia.org
[4] https://www.waze.com/pt-BR
[5] http://sigact.acm.org/genealogy

científica, relações entre pessoas e outras, também possui propriedades e características próprias. Identificar essas características e propriedades contribui de diversas formas para desenvolver e aprofundar o conhecimento das relações descritas pela rede.

São diversos os desafios que envolvem a construção de uma árvore genealógica acadêmica. O foco principal desta dissertação é uma análise ampla das árvores genealógicas acadêmicas de uma grande parcela dos pesquisadores brasileiros. Para a construção dessas árvores, utilizamos dados dos currículos de todos os doutores cadastrados na Plataforma Lattes. Apesar de centralizar todos os dados, a forma como esses dados são apresentados na Plataforma Lattes não permite visualizar a origem desses pesquisadores, a não ser pelo seu orientador e instituição onde se formou e realizou seus estudos de pós-graduação. Assim, o principal objetivo deste trabalho é gerar, através dos dados disponíveis na Plataforma Lattes, as árvores genealógicas acadêmicas de todos os pesquisadores atuantes no Brasil. Para isso o trabalho foi dividido nas seguintes etapas:

1. Extração dos dados relevantes presentes em uma coleção de currículos previamente coletada da Plataforma Lattes;
2. Preprocessamento dos registros gerados para limpeza e padronização dos dados;
3. Identificação e desambiguação das entidades (pesquisadores) mencionadas nos currículos coletados;
4. Construção das árvores genealógicas acadêmicas a partir dos laços de orientação identificados entre os pesquisadores;
5. Caracterização e análise das árvores construídas.

1.2 Motivação

Segundo a última avaliação quadrienal da Capes, o Brasil obteve um crescimento de 94% no número de doutores formados no país em relação ao período 2010-2012[6]. Explorar como se deu o crescimento da rede de orientações acadêmicas e possibilitar a visualização dessa rede, além do reconhecimento de esforços individuais e das instituições para o alcance de tais resultados, são algumas das principais contribuições deste trabalho. Mais ainda, construir a rede de orientações nos permite navegar até as origens da ciência no país. A partir dessa rede, é possível recontar a história da ciência

[6]http://avaliacaoquadrienal.capes.gov.br/home/sai-o-resultado-da-1a-etapa-da-avaliacao-quadrienal-2017

brasileira, mostrando como se deu o seu desenvolvimento e o surgimento de novas áreas, e identificando os seus pioneiros, ou seja, os grandes responsáveis pelo desenvolvimento de cada área.

A chamada rede de orientações acadêmicas representa as relações de orientação dentro da academia. Nesse caso, a rede a ser explorada é a rede de formação de mestres e doutores. Tal rede mostra a relação temporal entre doutores, mestres e seus respectivos orientadores. A origem do título de "doutor" é proveniente da era medieval, sendo a Alemanha um dos precursores do modelo de doutorado utilizado até hoje no mundo [Park, 2005]. Devido ao seu sucesso, tal modelo foi copiado por instituições espalhadas pelo mundo e até hoje é o modelo utilizado para conceder o grau máximo da academia.

Um repositório de dados sobre relacionamentos acadêmicos fornece diversos benefícios adicionais à comunidade acadêmica, permitindo aos novos membros descobrir suas raízes e também se alinhar com o contexto de seu campo, além de servir como inspiração para novos pesquisadores. Estudos sobre a genealogia acadêmica têm permitido a compreensão sobre quais ambientes de treinamento têm produzido os pesquisadores mais produtivos ao fim de suas carreiras [Ali & Panther, 2008; Malmgren et al., 2010; Tuesta et al., 2015].

Disponibilizar o acesso a essa rede para que outros pesquisadores possam utilizá-la como base em suas pesquisas ou apenas pela curiosidade de identificar quem são os outros doutores em sua mesma linhagem é da maior importância para traçar como se deu a formação de grupos de pesquisa e o próprio desenvolvimento das diversas áreas de pesquisa. Explorar tal rede, por meio de uma interface gráfica, torna essas descobertas muito mais simples e intuitivas que o processo atual de se pesquisar manualmente os diversos repositórios disponíveis na Web.

Por fim, a genealogia acadêmica pode ser utilizada como meio para documentar e organizar, através de uma rede, as relações de orientação ou supervisão acadêmica de cada pesquisador. Em escala mundial, a identificação dos pesquisadores ancestrais é uma tarefa desafiadora pois atualmente não existem muitos repositórios que permitam o registro de informações da linhagem acadêmica de pesquisadores associados a diferentes áreas de atuação acadêmica.

1.3 Trabalhos Relacionados

Entre os principais trabalhos sobre redes complexas estão os trabalhos pioneiros sobre redes de colaboração científica realizados por Newman [2001a,b]. Utilizando dados re-

ferentes a artigos científicos das áreas de Física, Biomedicina e Ciência da Computação, Newman gerou e estudou as redes de colaboração entre os autores desses artigos. Esses estudos aumentaram ainda mais o interesse em analisar os mais diversos tipos de rede.

Diversos trabalhos têm utilizado dados disponíveis em bibliotecas digitais, principalmente de artigos científicos, para compreender a estrutura das redes de colaboração científica formadas pelos dados agregados e disponíveis nessas bibliotecas. Por exemplo, os trabalhos de Cunningham [2001], Dawson et al. [2014], Liu et al. [2005], Menezes et al. [2009] e Sarigöl et al. [2014] utilizam dados de conferências internacionais para construir tais redes. Outros trabalhos empregam métricas de análise de grafos para caracterizar os diversos padrões de colaboração entre os autores de artigos científicos [Glänzel, 2001; Newman, 2004; Uddin et al., 2012].

Mais especificamente, alguns trabalhos buscam estudar a evolução temporal das redes de colaboração acadêmica e propõem modelos que capturam os mecanismos que afetam essa evolução, como os estudos de Barabási et al. [2002] e Perc [2010]. Já Alves et al. [2013] e Yan & Ding [2009] utilizam as redes de colaboração científica para detectar seus principais líderes e entender qual o papel desses lideres nas diferentes comunidades científicas. Demirkan & Demirkan [2012] mostram que empresas de biotecnologia dependem bastante das redes sociais envolvendo pesquisadores para a troca e produção de conhecimento. Já Kumar & Jan [2013] examinaram o tamanho do componente gigante da rede de coautoria em quatro disciplinas da área de engenharia. Seus resultados apontam que, das quatro disciplinas, duas já possuem o componente gigante bem formado e as outras duas ainda estão em estágio de desenvolvimento.

Da mesma forma, há também alguns esforços que visam documentar, analisar e classificar as redes de orientação acadêmica. Chang [2003] apresenta uma retrospectiva acadêmica de importantes físicos da American Physical Society e descreve as suas respectivas árvores genealógicas acadêmicas. Apesar de não deixar claro como os dados foram obtidos, o artigo descreve detalhadamente a carreira desses físicos e ao fim apresenta as árvores genealógicas acadêmicas em formato de infograma.

A partir de cartas enviadas a todos os programas de doutorado em Matemática dos EUA, solicitando o nome, título da tese e orientador de todos os seus alunos, Coonce iniciou o Mathematics Genealogy Project [Jackson, 2007]. De acordo com Coonce, apenas 25% a 30% dos programas responderam ao seu pedido, porém, os dados recebidos foram suficientes para dar início ao projeto[7] que hoje conta com mais de 200 mil registros de matemáticos do mundo todo.

Não muito distante, David & Hayden [2012] criaram um repositório para arma-

[7]http://genealogy.math.ndsu.nodak.edu

zenar as árvores genealógicas acadêmicas de todos os pesquisadores da área de Neurociência. O projeto[8], que iniciou-se apenas no papel, deu origem a um banco de dados relacional que, devido ao interesse de diversos pesquisadores, foi posteriormente disponibilizado para o público na WWW. Hoje, o projeto, que se expandiu, conta com um acervo de diversas áreas da ciência.

Em comum, esses projetos coletam dados sobre pesquisadores que trabalham em diferentes áreas visando estabelecer suas genealogias acadêmicas. Um outro projeto, o Academic Genealogy procurou documentar as famílias acadêmicas de pesquisadores do mundo todo, compartilhando a informação gerada por meio de um Wiki[9], recentemente desativado. Vale ressaltar que grande parte dos dados desses projetos foram obtidos e são mantidos por meio de um esforço de *Crowdsourcing*.

Também há trabalhos que buscam apenas analisar, compreender e modelar as estruturas das árvores genealógicas acadêmicas de pessoas ou áreas específicas. Tuesta et al. [2015] analisaram as árvores genealógicas acadêmicas de pesquisadores brasileiros que compõem a grande área de Ciências Exatas e da Terra. Em seu trabalho, eles exploram a correlação entre o tempo de orientação e a produtividade dos pesquisadores analisados, apresentando evidências que mostram que o desenvolvimento e o aprendizado do aluno não fica limitado apenas ao tempo de atuação com seu orientador, indo muito além dessa relação acadêmica.

Em outro trabalho, Malmgren et al. [2010] investigaram o desempenho dos discípulos na relação mestre-discípulo. Para isso, eles analisaram dados das árvores genealógicas acadêmicas dos matemáticos. Em seu artigo, eles demonstraram que a fecundidade, número de alunos orientados, dos pesquisadores da Matemática é correlacionada com outras métricas de sucesso. Seus resultados mostram que orientadores com fecundidade baixa, orientam alunos com fecundidade 37% maior que o esperado, enquanto que orientadores com alta fecundidade obtêm sucesso apenas nos primeiros dois terços de sua carreira.

Rossi & Mena-Chalco [2014] introduziram métricas topológicas para analisar a estrutura de uma árvore genealógica. Como estudo de caso, eles utilizaram a árvore genealógica acadêmica do matemático J. Bernoulli. As métricas introduzidas procuram avaliar de maneira quantitativa e qualitativa determinados aspectos dos pesquisadores presentes nas árvores, contribuindo para um melhor entendimento da estrutura dessas árvores e o enriquecimento das análises realizadas.

Em um outro esforço, utilizando dados da Networked Digital Library of Theses

[8]http://neurotree.org
[9]http://phdtree.org

and Dissertations, NDLTD[10], foram construídas e analisadas as árvores genealógicas acadêmicas referentes a parte das teses e dissertações armazenadas naquele repositório [Dores et al., 2016]. Devido à característica mais genérica da NDLTD, as árvores construídas eram menores e a rede derivada dessas árvores muito mais esparsa do que a que é analisada nesta dissertação.

Em um recente trabalho, Damaceno et al. [2017] também utilizaram dados da Plataforma Lattes para gerar e analisar árvores genealógicas acadêmicas de pesquisadores brasileiros. Nesse trabalho, os autores apresentam um algoritmo para a geração dessas árvores considerando, entretanto, apenas a formação de doutores e procurando caracterizar a capacidade individual dos pesquisadores nesse tipo de formação acadêmica. Eles também apresentam uma análise dessas árvores com base nas grandes áreas do conhecimento do CNPq indicadas pelos pesquisadores como sendo aquelas de sua atuação.

Assim, o foco principal desta dissertação é realizar uma análise das árvores genealógicas acadêmicas dos pesquisadores brasileiros construídas a partir de dados coletados da Plataforma Lattes. Esta dissertação busca assim não só aprimorar alguns dos trabalhos citados como também complementar outros que analisam essas árvores.

1.4 Contribuições

As principais contribuições desta dissertação são:

- Criação de um repositório contendo os dados de todas as orientações de mestrado e doutorado registradas na Plataforma Lattes até abril de 2017;
- Construção e caracterização das árvores genealógicas acadêmicas a partir dos dados coletados, apresentando detalhes sobre a estrutura da respectiva rede e destacando os principais atores nela presentes [Dores et al., 2017];
- Análise das árvores construídas a partir de métricas específicas que permitem uma melhor compreensão de como ocorreu o processo de formação de nossos pesquisadores e grupos de pesquisa.

1.5 Organização da Dissertação

Os demais capítulos desta dissertação estão organizados da seguinte forma. O Capítulo 2 apresenta uma visão geral sobre redes complexas e suas características, incluindo

[10]http://ndltd.org

uma descrição dos principais conceitos e métricas utilizados ao longo da dissertação. O Capítulo 3 descreve o conjunto de dados utilizado para construir as árvores genealógicas acadêmicas, bem como as várias etapas envolvidas nessa tarefa. A seguir, o Capítulo 4 apresenta uma caraterização geral dessas árvores e uma análise topológica de acordo com as grandes áreas do conhecimento conforme definidas nos currículos dos respectivos pesquisadores. Finalmente, o Capítulo 5 apresenta as principais conclusões desta dissertação e algumas direções para trabalhos futuros.

Capítulo 2

Redes Complexas

Este capítulo apresenta uma visão geral dos conceitos, perspectivas e aplicações de redes complexas, bem como dos fundamentos teóricos necessários para analisar as árvores genealógicas acadêmicas consideradas nesta dissertação.

2.1 Introdução

O estudo de como o mundo se relaciona vem sendo foco de diversos trabalhos e livros, nos mais variados contextos. A disciplina de redes complexas, que permite estudar os padrões de inter-relacionamento de elementos do mundo real, consolida-se cada vez mais como um campo de estudos interdisciplinar, influenciando diversas áreas, como, Ciência da Computação, Biologia e Física [Strogatz, 2001].

Diversos aspectos do mundo real podem ser representados através das chamadas redes complexas [Easley & Kleinberg, 2010]. Assim, desde a década de 1930 sociólogos têm utilizado essas redes, modeladas matematicamente como grafos, com a finalidade de estudar o comportamento da sociedade e as relações entre indivíduos dentro de diversos contextos [Granovetter, 1973]. Com o advento dos computadores e o desenvolvimento de algoritmos e técnicas de análise de dados na Ciência da Computação, os estudos realizados pelos sociólogos foram gradativamente incorporados a essa área criando um novo campo de estudo chamado Ciência dos Dados. Além disso, com a crescente evolução e expansão da Web, junto à imensa quantidade de dados compartilhados nos últimos anos, principalmente por meio das chamadas redes sociais, esse campo vem se desenvolvendo cada vez mais. Assim, redes dos mais diversos tipos e tamanho podem agora ser estudadas, fazendo com que surjam, cada vez mais, novas métricas e que análises mais sofisticadas possam ser realizadas sobre elas [Scott, 2017].

Matematicamente, uma rede G pode ser definida como um grafo [Bondy & Murty, 1976] $\boldsymbol{G(V, E)}$ tal que, V é um conjunto não vazio de objetos denominados nodos (vértices) e E é um subconjunto de pares não ordenados de nodos contidos em V, denominados arestas.

Figura 2.1. Exemplo de uma rede complexa.

Mais especificamente, uma rede complexa pode ser definida como sendo uma coleção de objetos (nodos) na qual cada objeto se relaciona com outros por meio de conexões (arestas). A rede da Figura 2.1 ilustra as relações de amizades entre um grupo de pessoas cadastradas na rede social Facebook[1]. Nessa rede, as pessoas cadastradas são representadas pelos nodos e pessoas que possuem uma relação de amizade são interligadas através de uma aresta.

Na rede ilustrada, é possível perceber por meio da posição e também pelas cores dos nodos que ela está dividida em diversos grupos ou comunidades, cada um deles formado devido a alguma característica semelhante entre os seus nodos. No caso específico da rede da Figura 2.1, temos grupos de amigos que compartilham em comum a mesma etnia, o local de trabalho, a instituição em que estudam ou mesmo a região em que residem. Em um outro exemplo, pesquisadores poderiam ser agrupados de

[1]https://pt-br.facebook.com

acordo com os tópicos de seus artigos científicos, as conferências em que apresentam esses artigos ou mesmo por tempo de carreira.

Diversos fenômenos ou situações que ocorrem no mundo real possuem características que permitem a sua representação por meio de uma rede complexa, por exemplo, a Internet, sistemas biológicos, rotas de aviões, coautorias, amizades, rodovias, receitas e muitos outros. Chen et al. [2015] apresentam um estudo sobre como essa área vem se desenvolvendo ao longo dos anos, seu futuro e os resultados mais significativos.

2.2 Conceitos Básicos sobre Redes Complexas

Dentre as redes complexas destacam-se as redes sociais. Tais redes formam-se por meio das mais diversas formas de interação social. Nesta dissertação, focamos em redes sociais na quais as relações ocorrem no meio acadêmico como resultado de orientações de mestrado e doutorado. Assim, acadêmicos são representados por meio dos nodos de uma rede e as orientações entre eles são representadas por arestas nessa mesma rede. Outra característica dessas redes é a presença de arestas dirigidas, isto é, todas as arestas da rede possuem uma direção no sentido da orientação, ou seja, do orientador para o orientado. Também é possível haver, entre dois nodos, mais de uma aresta com mesma origem e destino. Assim, uma rede contendo arestas dirigidas e sem ciclos é modelada como um tipo específico de grafo também denominado multigrafo dirigido ou, simplesmente, multidígrafo [Bollobas, 1998].

Normalmente, as redes complexas são analisadas a partir de características de sua estrutura. Muitos dos algoritmos utilizados para a análise dessas redes foram herdados da área de teoria dos grafos. Na literatura, há diversos problemas modelados como redes, existindo também diversas métricas que auxiliam na extração de informação topológica seja dos nodos, por exemplo, grau e conectividade, como também das arestas, como peso e força da aresta na rede [Granovetter, 1973]. Por fim, várias propriedades gerais da estrutura de uma rede podem ser consideradas, como grau médio dos nodos, diâmetro e número de componentes, entre outras [Bondy & Murty, 1976]. Algumas dessas propriedades são definidas a seguir.

2.2.1 Grau de um Nodo

O grau de um nodo é definido pelo número de vizinhos conectados a ele, sendo que dois nodos são considerados vizinhos quando possuem uma aresta em comum. Assim, o grau de um nodo é dado pelo número de arestas que ligam esse nodo a outros. Formalmente,

o grau de um nodo v em uma rede G é igual ao número de arestas que são incidentes a v. E o grau total de uma rede G é medido pela soma dos graus de todos os seus nodos.

A Figura 2.2 mostra dois nodos com diferentes valores de grau, o primeiro conectado a cinco vizinhos, portanto de grau 5, e o outro conectado a dois vizinhos, ou seja, de grau 2. Quanto mais conexões um nodo possui maior o seu grau e, consequentemente, maior a sua importância na rede [Havel, 1955].

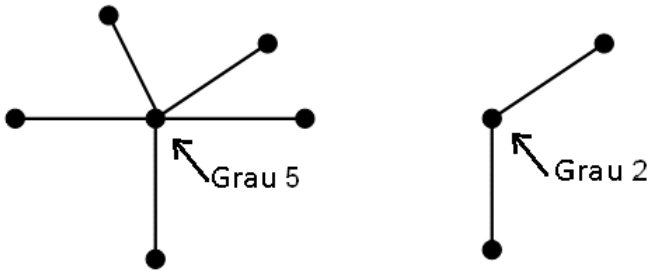

Figura 2.2. Nodos com diferentes valores de grau.

Em redes com arestas dirigidas, podemos considerar dois graus distintos para um mesmo nodo: *o grau de saída*, que corresponde ao total de arestas partindo do nodo e chegando aos nodos vizinhos, e *o grau de entrada*, que corresponde ao total de arestas que saem dos nodos vizinhos e chegam a esse nodo. Por exemplo, no Twitter[2] todas as conexões entre os usuários podem ser divididas em seguidores e seguidos. Essa conexão pode ser modelada como uma aresta direcionada partindo de um nodo v_j para um nodo v_i, se v_j é um seguidor de v_i. Então podemos calcular o grau de entrada de um nodo v_i como o número de nodos com arestas direcionadas para v_i (seguidores), enquanto o seu grau de saída corresponde ao número de arestas que partem de v_i para outros nodos (seguidos).

2.2.2 Caminho e Diâmetro

Em uma rede, um caminho é definido como uma sequência finita ou infinita de nodos conectados por uma sequência de arestas onde os nodos são todos diferentes uns dos outros [Bollobas, 1998]. Em uma rede direcionada, um caminho é uma sequência de arestas dirigidas que se conectam a uma sequência de nodos seguindo o sentido das arestas. O tamanho de um caminho pode ser definido pelo número de arestas do mesmo ou pelo número de nodos no caminho menos um. O diâmetro (comprimento) de uma rede é definido pelo tamanho do maior caminho nela existente. A Figura 2.3

[2]http://twitter.com

mostra, destacado em vermelho, o caminho cujo tamanho corresponde ao diâmetro de uma rede.

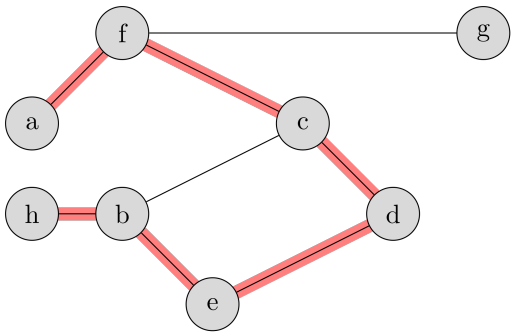

Figura 2.3. Exemplo de um caminho em uma rede, cujo tamanho corresponde ao seu diâmetro.

2.2.3 Componente Conectado

Algumas vezes, uma rede não está completamente conectada e isso faz com que a mesma seja fragmentada em diferentes partes. Dizemos que um componente conectado de uma rede corresponde a um subconjunto de nodos tal que:

(i) Todo nodo nesse subconjunto possui um caminho para todos os demais;

(ii) Esse subconjunto não é parte de um conjunto maior com a propriedade que todo nodo pode alcançar todos os demais.

Em uma rede direcionada, um componente é denominado fortemente conectado se e somente se, para qualquer par de seus nodos (a, b), houver um caminho de a para b e também um caminho de b para a. Já um componente é fracamente conectado se, ignorando a direção dos nodos, houver pelo menos um caminho entre todos os pares de nodos. O componente que engloba o maior número de nodos da rede é denominado componente gigante. Devido às características das árvores genealógicas acadêmicas, nesta dissertação denominamos componentes apenas aqueles fracamente conectados [Bollobas, 1998], já que devido à falta de ciclos uma árvore genealógica não possui caminhos entre todos os seus pares de nodos.

2.2.4 Florestas, Árvores, Folhas e Raízes

Na teoria dos grafos, árvores são grafos que não contêm ciclos, representando visualmente uma estrutura hierárquica. A importância das árvores é evidente quando olhamos para a sua ampla aplicação em várias situações que vão desde a representação

de árvores familiares na Biologia, fonte de diversas terminologias, até complexas estruturas de dados na Ciência da Computação [Mehlhorn & Sanders, 2008]. Assim sendo, em uma rede composta por componentes conectados sem a presença de ciclos, podemos considerar esses componentes como árvores. Além disso, caso a rede seja composta por vários componentes, podemos considerá-la como sendo uma floresta.

Em uma árvore, alguns nodos possuem uma terminologia própria dependendo de determinadas propriedades. Assim, os nodos de uma árvore com grau de saída 0 são denominados *folhas*. Do mesmo modo, os nodos que possuem apenas arestas de saída são denominados *raízes*.

2.2.5 Árvores Genealógicas Acadêmicas

De acordo com o dicionário Collins[3], uma árvore genealógica é definida por um gráfico que mostra toda a linhagem de uma família por meio de suas diversas gerações e dos relacionamentos entre elas. A partir desta definição, consideramos, nesta dissertação, uma árvore genealógica acadêmica como sendo a representação da linhagem de um acadêmico, onde os nodos correspondem a todos os pesquisadores direta ou indiretamente ligados a ele e os relacionamentos representam as orientações realizadas por ele e todos os seus descendentes. Matematicamente, uma árvore genealógica acadêmica é representada por um grafo dirigido acíclico (GDA[4]).

2.2.6 Descendência

Descendência pode ser definida como sendo o conjunto de todos os indivíduos (pesquisadores) de uma árvore genealógica acadêmica que possuem um ancestral (orientador) em comum. Por exemplo, se consideramos a árvore genealógica acadêmica do físico Albert Einstein, todos os alunos que direta ou indiretamente têm uma relação acadêmica com ele, são considerados seus descendentes. Como métrica, estamos interessados em saber o número de descendentes presentes em uma árvore, mais especificamente, o número de nodos abaixo da raiz.

2.2.7 Linhagem

A linhagem de uma família é representada pelo número de gerações abaixo do primeiro ancestral. Uma linhagem acadêmica seria igual ao número de gerações de orientandos presentes na árvore. Matematicamente, a linhagem pode ser definida pela profundidade

[3]https://www.collinsdictionary.com/dictionary/english/genealogical-tree
[4]Do inglês DAG - Directed Acyclic Graph.

da árvore, a partir do seu nodo raiz. Uma forma de medir a profundidade de uma árvore seria calculando o tamanho de seu maior caminho mínimo.

2.2.8 Fecundidade

Fecundidade na biologia é definida como sendo a capacidade de um indivíduo gerar outros. Assim, a fecundidade acadêmica está associada ao número de indivíduos orientados por um acadêmico durante a sua carreira [Malmgren et al., 2010]. Portanto, definir a fecundidade de uma árvore é algo próximo a definir a sua capacidade de propagação, ou seja, corresponde ao número médio de descendentes acadêmicos gerados a partir de orientações de pesquisadores que passaram a orientar novos pesquisadores (Equação 2.1).

$$f(arvore) = \frac{(|(nodos)|)}{(|grau_s(nodos) > 0|)} \tag{2.1}$$

Nesta dissertação, a métrica fecundidade mede a média de filhos de todos os nodos que não sejam folhas. Por exemplo, uma árvore com uma raiz, dois filhos e quatro netos tem fecundidade média igual a dois. Ou seja, cada nodo "adulto" (nodo que não seja folha) dessa árvore gerou em média dois filhos. Assim, para cada árvore temos a fecundidade média dos nodos envolvidos. Com essa métrica espera-se entender melhor como se dá a evolução das árvores. Estabelecer a fecundidade média de uma árvore genealógica acadêmica auxilia a compreensão de sua formação, capacidade de propagação e diferenciação em relação a outras árvores.

2.2.9 Densidade

Em um grafo, a densidade é definida como sendo a razão entre o número de arestas existentes e o número de arestas possíveis [Coleman & Moré, 1983]. Assim um grafo simples, sem arestas, possui densidade zero enquanto um grafo simples, mas completo, possui densidade igual a um. A Equação 2.2 mostra como é calculada a densidade para uma árvore genealógica acadêmica, ou seja, a densidade é a razão entre o número total de orientações realizadas pelo pesquisador e seus descendentes, e duas vezes o número de seus descendentes, já que, para cada descendente, espera-se que ele tenha tido duas orientações, uma de mestrado e outra de doutorado.

$$d(arvore) = \frac{(|arestas|)}{(2 \times |nodos|)} \tag{2.2}$$

Com a métrica densidade é possível medir o quão densa é uma árvore, ou em outras palavras, qual é a reincidência de orientações nessa árvore. Uma árvore com uma densidade alta indica um grupo mais fechado, no qual seus membros normalmente buscam novas orientações dentro do mesmo grupo do seu orientador anterior. Por exemplo, um pesquisador que foi orientado no mestrado por um determinado pesquisador tende a ser orientado no doutorado por outro pesquisador ligado ao seu orientador anterior, tornando assim a árvore mais densa.

Capítulo 3

Materiais e Métodos

Este capítulo descreve inicialmente a coleção de dados utilizada para a construção das árvores genealógicas acadêmicas, apresentando algumas estatísticas gerais sobre os dados considerados. A seguir, é descrito o processo de tratamento dos dados, incluindo as etapas de extração e limpeza dos dados, o algoritmo de construção das árvores genealógicas acadêmicas e a estratégia de desambiguação de nomes adotada para o casamento dos pesquisadores identificados nos currículos. Por fim, é apresentada uma descrição de como são armazenados os dados necessários para a visualização das árvores.

3.1 Coleção de Dados

Nesta dissertação foram utilizados dados extraídos diretamente de currículos presentes na Plataforma Lattes por meio de uma ferramenta de coleta e extração de dados, denominada LattesDataXplorer, desenvolvida especificamente para essa finalidade [Dias, 2016]. A Plataforma Lattes surgiu como um esforço do CNPq na integração de repositórios de dados de pesquisadores, grupos de pesquisa e instituições em um único ambiente. Sua principal função é facilitar o acesso e gerenciamento de dados utilizados por instituições, pesquisadores e agências de ciência e tecnologia de todo o país.

A Plataforma Lattes tornou-se um padrão nacional para o registro acadêmico de estudantes e pesquisadores do país, sendo hoje o sistema adotado por quase todas as instituições de ensino e pelas agências de pesquisa do país. Devido à sua constante atualização, crescente confiança e abrangência, a plataforma tornou-se uma ferramenta indispensável para a análise de mérito das solicitações de bolsas e projetos de pesquisa submetidos às diversas agências de fomento do país.

Hoje a Plataforma Lattes conta com mais de cinco milhões de currículos de indivíduos envolvidos em atividades de pesquisa em todo país. Cada currículo apresenta

diversos dados sobre a vida acadêmica de cada indivíduo, incluindo dados pessoais, como endereço e a afiliação, graus obtidos durante a sua vida acadêmica e dados sobre atividades acadêmicas como projetos e disciplinas ministradas, publicações científicas e orientações acadêmicas. Para realizar o processo de construção das árvores genealógicas acadêmicas, utilizamos os currículos de todos os doutores cadastrados na plataforma até abril de 2017, perfazendo um total de 256.845 currículos.

Figura 3.1. Seções Identificação, Endereço e Formação acadêmica/titulação do currículo Lattes do Prof. Marcos André Gonçalves do DCC/UFMG.

Identificação

Nome Marcos André Gonçalves

Nome em citações bibliográficas GONÇALVES, Marcos André;Gonçcalves, Marcos André;Marcos A. Gonçalves;MARCOS GONÇALVES;GONÇALVES, MARCOS;GONÇALVES, MARCOS A.;GONÇALVES, MARCOS;GONCALVES, MARCOS

Endereço

Endereço Profissional Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Av. Antônio Carlos, 6627 - Departamento de Ciência da Computação - Prédio do ICEx, Sala 4010, Pampulha, Belo Horizonte, Minas Gerais, Brasil
Pampulha
31270901 - Belo Horizonte, MG - Brasil
Telefone: (31) 34095860
URL da Homepage: http://www.dcc.ufmg.br/~mgoncalv

Formação acadêmica/titulação

1999 - 2004 Doutorado em Computer Science.
Virginia Tech, VIRGINIA TECH, Estados Unidos.
Título: Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications, Ano de obtenção: 2004.
Orientador: Edward A Fox.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.
Palavras-chave: digital libraries; ontology; semantic modeling; log standard; quality; theory.
Grande área: Ciências Exatas e da Terra

1996 - 1997 Mestrado em Ciência da Computação (Conceito CAPES 7).
Universidade Estadual de Campinas, UNICAMP, Brasil.
Título: Uso de modelos hipermidia em bibliotecas digitais para dados geograficos,Ano de Obtenção: 1997.
Orientador: Claudia Bauzer Medeiros.
Bolsista do(a): Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, Brasil.

1992 - 1995 Graduação em Bacharelado em Ciência da Computação.
Universidade Federal do Ceará, UFC, Brasil.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.

Após a obtenção dos currículos no formato XML (eXtensible Markup Language), identificamos os marcadores que correspondem às seções que contêm os dados necessários para a construção das árvores. Esses dados encontram-se em duas partes específicas dos currículos.

A primeira parte, formada pelas seções Identificação, Endereço e Formação acadêmica/titulação do pesquisador, inclui dados pessoais do pesquisador, como nome e endereço profissional, e detalhes de sua formação acadêmica, como título obtido e a instituição que o concedeu, nome do orientador, título da dissertação ou tese, e ano de conclusão para cada uma das titulações obtidas. Essa parte inclui, portanto, os dados referentes aos ancestrais acadêmicos do pesquisador. A Figura 3.1 mostra as seções de Identificação, Endereço e Formação acadêmica/titulação do currículo do Prof. Marcos

André Gonçalves do Departamento de Ciência da Computação da UFMG. Na figura podemos perceber que há dados sobre o seu doutorado e mestrado, e também sobre a sua graduação. Porém, apenas nos casos do doutorado e mestrado encontramos dados adicionais sobre seus orientadores, Prof. Edward Fox da Virginia Tech e Profa. Claudia Bauzer Medeiros da Unicamp, respectivamente. O preenchimento prévio dos campos referentes aos orientadores é imprescindível para possibilitar a descrição completa da genealogia acadêmica de um pesquisador.

A outra parte do currículo que também contribui para a construção da árvore genealógica de um pesquisador é a que inclui a seção Orientações e supervisões concluídas na qual estão listadas todas as atividades em que o pesquisador exerceu o papel de orientador ou coorientador. Essa seção possui duas subseções principais. A primeira, Dissertação de mestrado, lista as orientações e coorientações de mestrado realizadas pelo pesquisador, enquanto a segunda, Tese de doutorado, lista as suas orientações e coorientações de doutorado. As listas de orientações dessas duas subseções incluem, além do nome do aluno orientado, o título do trabalho realizado, a instituição que concedeu o título por ele obtido e o período de estudo, bem como o papel do pesquisador (orientador ou coorientador).

Figura 3.2. Parte da seção Orientações do currículo Lattes do Prof. Marcos André Gonçalves do DCC/UFMG.

Dissertação de mestrado

1. Clebson Cardoso Alves de Sá. Optmizing emsembles of boosted additive bagged tress for learning-to-rank. 2016. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
2. Masoumeh Nezhadbiglari. ScholarTrendLeaner: Predicting Scholar Popularity as Early and Accurate as Possible. 2016. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
3. Alan Filipe. Heurísticas para Desambiguação de Nomes de Autores em Referências Bibliográficas. 2015. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Orientador: Marcos André Gonçalves.
4. Felipe Viegas. EXPLOITING EFFICIENT AND EFFECTIVE BAYESIAN STRATEGIES FOR TEXT CLASSIFICATION. 2015. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
5. Eder Ferreira Martins. Recomendação Associativa de Tags na Ausência de Informação Previa. 2013. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Coorientador: Marcos André Gonçalves.

Tese de doutorado

1. Daniel Hasan Dalip. A Multi-View Approach for Estimating the Quality of Collaborative Content on the Web 2.0. 2015. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
2. Anderson Ferreira. Contributions for Solving the Author Name Ambiguity Problem in Bibliographic Citations. 2012. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
3. Moises Gomes de Carvalho. Abordagens evolucionárias para problemas relacionados a integração de dados. 2009. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Coorientador: Marcos André Gonçalves.
4. Leonardo Rocha. Uso de Contextos Temporais para Classificação Automática de Documentos. 2009. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Coorientador: Marcos André Gonçalves.
5. Guilherme Tavares de Assis. Uma Abordagem Baseada em Gênero para Coleta Temática. 2008. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Coorientador: Marcos André Gonçalves.

Vale ressaltar que, cerca de 85% dos currículos considerados possuem a seção de

orientação de doutorado em branco. Para o mestrado esse número era de 69%, mas ainda assim mais da metade dos doutores não possuia orientações informadas em seus currículos Lattes.

A Figura 3.2 mostra parte dessa seção contendo trechos das listas de orientações concluídas também do Prof. Marcos André Gonçalves. Assim, a partir dessas duas listas é possível obter a relação de todos os descendentes acadêmicos diretos de um pesquisador.

Dados Gerais	Total
Doutorados Informados	261.773
Mestrados Informados	229.390
Orientações de Doutorado	233.103
Orientações de Mestrado	778.847
Coorientações de Doutorado	60.283
Coorientações de Mestrado	151.523

Tabela 3.1. Total de titulações (graus acadêmicos) e orientações presentes nos currículos.

A Tabela 3.1 apresenta dados gerais sobre o total de titulações (Doutorado/Mestrado), bem como de orientações e coorientações (Doutorado/Mestrado), registradas nos currículos coletados. Analisando esses dados, é possível perceber que os currículos coletados apresentam uma média de 1,02 doutorado por currículo, ou seja, cerca de 2% dos pesquisadores possuem mais de um doutorado informado no currículo. Por outro lado, cerca de 10% dos currículos não indicam formação de mestrado.

Além disso, cerca de 12% dos pesquisadores estão vinculados a instituições localizadas no eixo SP-RJ-MG, enquanto os demais estão distribuídos entre instituições espalhadas pelo Brasil. Entretanto, é importante ressaltar que 42.674 pesquisadores (16,6% do total) não informaram qualquer vinculação institucional. A Tabela 3.2 apresenta a relação das 20 instituições com o maior número de doutores.

Ao preencher o seu currículo Lattes, um pesquisador pode informar até três áreas do conhecimento para a sua tese ou dissertação, utilizando para isso o esquema de classificação definido pelo CNPq[1] que abrange quatro níveis: Grande Área, Área, Subárea e Especialidade. Entretanto, para fins de análise, nesta dissertação é considerada apenas a classificação atribuída pelo pesquisador à sua tese de doutorado nos dois primeiros níveis. Vale ressaltar que esses dois níveis incluem no total nove Grandes Áreas e 99 Áreas.

A Tabela 3.3 apresenta a distribuição dos currículos por grande área. A grande área mais popular é a de Ciências Humanas e a menos popular, desconsiderando a

[1]http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf

Posição	Instituição	UF	# Doutores
1	Universidade de São Paulo	SP	12.252
2	Universidade Est. Paulista Júlio de Mesquita Filho	SP	5.725
3	Universidade Federal do Rio de Janeiro	RJ	5.661
4	Universidade Estadual de Campinas	SP	4.551
5	Universidade Federal de Minas Gerais	MG	4.079
6	Universidade Federal do Rio Grande do Sul	RS	3.749
7	Universidade Federal de Santa Catarina	SC	3.282
8	Universidade de Brasília	DF	2.983
9	Universidade Federal Fluminense	RJ	2.890
10	Universidade Federal de Pernambuco	PE	2.807
11	Universidade Federal do Paraná	PR	2.661
12	Universidade Federal de São Paulo	SP	2.650
13	Universidade do Estado do Rio de Janeiro	RJ	2.416
14	Empresa Brasileira de Pesquisa Agropecuária†	–	2.284
15	Universidade Federal da Bahia	BA	2.259
16	Universidade Federal do Rio Grande do Norte	RN	2.234
17	Universidade Federal da Paraíba	PB	2.188
18	Universidade Federal do Ceará	CE	2.178
19	Fundação Oswaldo Cruz†	–	2.148
20	Universidade Federal de Goiás	GO	2.130

Tabela 3.2. Relação das 20 instituições com maior número de doutores (†Intituições com unidades localizadas em mais de um estado).

grande área Outros, é a de Linguística, Letras e Artes. Entretanto, é importante notar que 114.630 pesquisadores (69%) não indicaram nem mesmo a grande área de sua tese.

Grande Área	Quantidade
Ciências Humanas	76.796
Ciências Exatas e da Terra	62.523
Ciências da Saúde	60.824
Ciências Biológicas	58.832
Engenharias	46.697
Ciências Agrárias	44.306
Ciências Sociais	41.740
Linguística, Letras e Artes	27.233
Outros	2.647
Não Informada	114.630

Tabela 3.3. Distribuição dos currículos por grande área.

Após indicar a grande área de sua tese, o pesquisador tem a opção de indicar também a respectiva área específica. Entretanto, nos currículos coletados, mais de 55% dos pesquisadores não fizeram essa indicação. Entre as áreas mais indicadas,

Educação, Medicina e Agronomia despontam como as mais frequentes nos currículos (Tabela 3.4). Vale ainda ressaltar que oito dessas áreas mais populares (Educação, Medicina, Química, Psicologia, Física, Letras, História e Ciência da Computação) estão vinculadas às três grandes áreas mais indicadas nos currículos: Ciências Humanas, Ciências Exatas e da Terra e Ciências da Saúde (Tabela 3.3).

Área	Quantidade
Educação	14.558
Medicina	14.032
Agronomia	11.710
Química	10.767
Psicologia	8.515
Física	7.593
Letras	7.584
Bioquímica	7.108
História	6.949
Ciência da Computação	6.730

Tabela 3.4. Distribuição dos currículos por área para as 10 áreas mais indicadas.

3.2 Tratamento dos Dados

3.2.1 Extração e Limpeza dos Dados

O primeiro passo para a extração dos dados dos currículos coletados foi identificar na estrutura do documento XML, gerado a partir da coleta, todas as marcações correspondentes a cada uma das seções contendo os dados necessários para a geração das árvores genealógicas acadêmicas. A Figura 3.3 mostra o extrato de um documento XML contendo as marcações que incluem a identificação e o endereço do pesquisador. Para identificação interna dos pesquisadores ao gerar os nodos das árvores foram utilizados seus dados básicos de identificação existentes nos currículos, nome e identificador Lattes, e também o nome de sua instituição.

As marcações referentes à seção de formação acadêmica são divididas de acordo com as titulações obtidas pelo pesquisador. A Figura 3.4 mostra o extrato de um documento XML que inclui dados da formação acadêmica (mestrado e doutorado) de um pesquisador.

A partir das marcações existentes, é possível obter dados da instituição onde o pesquisador estudou, o título de sua tese ou dissertação, o nome do orientador, o ano de obtenção do título e o número de identificação do orientador, se houver. Esses dados

Figura 3.3. Extrato de um documento XML contendo os dados da seção Identificação do currículo Lattes de um pesquisador.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" NUMERO-IDENTIFICADOR="
    1162362624079364" DATA-ATUALIZACAO="27102017" HORA-ATUALIZACAO="160302"><
    DADOS-GERAIS NOME-COMPLETO="Rodrygo Luis Teodoro Santos" NOME-EM-CITACOES-
    BIBLIOGRAFICAS="SANTOS, R. L. T.;Santos, Rodrygo L. T.;Santos, Rodrygo L.T
    .;SANTOS, R" NACIONALIDADE="B" PAIS-DE-NASCIMENTO="Brasil" UF-NASCIMENTO="
    MG" CIDADE-NASCIMENTO="Divinópolis" PERMISSAO-DE-DIVULGACAO="NAO" DATA-
    FALECIMENTO="" SIGLA-PAIS-NACIONALIDADE="BRA" PAIS-DE-NACIONALIDADE="Brasil
    ">
<ENDERECO FLAG-DE-PREFERENCIA="ENDERECO_INSTITUCIONAL">
<ENDERECO-PROFISSIONAL CODIGO-INSTITUICAO-EMPRESA="033300000002" NOME-
    INSTITUICAO-EMPRESA="Universidade Federal de Minas Gerais" CODIGO-ORGAO="
    033383000005" NOME-ORGAO="Instituto de Ciências Exatas" CODIGO-UNIDADE="
    033383004000" NOME-UNIDADE="Departamento de Ciência da Computação"
    LOGRADOURO-COMPLEMENTO="Av. Antônio Carlos, 6627 - Prédio do ICEx - Sala
    3001" PAIS="Brasil" UF="MG" CEP="31270010" CIDADE="Belo Horizonte" BAIRRO="
    Pampulha" DDD="31" TELEFONE="34096578" RAMAL="" FAX="" CAIXA-POSTAL="" HOME
    -PAGE="http://www.dcc.ufmg.br/~rodrygo"/>
</ENDERECO>
</DADOS-GERAIS>
</CURRICULO-VITAE>
```

são utilizados para ligar o nodo do pesquisador aos nodos de seus orientadores, caso já existam, ou para criá-los caso ainda não existam.

Entretanto, para a correta extração dos dados dos documentos XML foram necessários alguns passos adicionais para garantir a padronização desses dados, evitando-se assim o processamento de conteúdo indesejado. Isso porque, devido à entrada de dados na Plataforma Lattes ser livre, em muitas situações titulações, comentários e datas são incluídos junto aos nomes dos orientadores ou dos orientandos. Outro problema é a falta de padronização na forma como os dados são inseridos, já que, muitas vezes, pré-nomes são invertidos em relação aos sobrenomes ou sobrenomes são simplesmente omitidos, o que, juntamente com erros de ortografia, dificultam enormemente o correto casamento de nomes próprios. Tudo isso dificulta a correta identificação dos pesquisadores, razão pela qual foram necessários alguns passos específicos para tratamento dos dados, explicados nos parágrafos seguintes, durante a extração dos dados dos currículos.

Na seção de Formação acadêmica/titulação, o campo contendo o nome do orientador foi o que exigiu o maior esforço em termos de limpeza dos dados. Isso deveu-se, em muitos casos, a diversos termos adicionados ao próprio nome do orientador. Entre esses termos, podemos citar pronomes de tratamento ou titulações como "Dr.", "Ph.D.", "Prof." e "Sr.", o papel desempenhado na orientação ("orientador" ou "coorientador") e, até mesmo, comentários como departamento do orientador, datas diversas e, em muitos casos, o nome do coorientador seguido ao do próprio orientador. A Tabela 3.5

Figura 3.4. Extrato de um documento XML contendo a seção Formação acadêmica/titulação do currículo Lattes de um pesquisador.

```
<FORMACAO-ACADEMICA-TITULACAO>
<MESTRADO SEQUENCIA-FORMACAO="1" NIVEL="3" CODIGO-INSTITUICAO="033300000002"
    NOME-INSTITUICAO="Universidade Federal de Minas Gerais" CODIGO-ORGAO=""
    NOME-ORGAO="" CODIGO-CURSO="32000049" NOME-CURSO="Ciências da Computação"
    CODIGO-AREA-CURSO="10300007" STATUS-DO-CURSO="CONCLUIDO" ANO-DE-INICIO="
    2006" ANO-DE-CONCLUSAO="2007" FLAG-BOLSA="SIM" CODIGO-AGENCIA-FINANCIADORA=
    "002200000000" NOME-AGENCIA="Conselho Nacional de Desenvolvimento Científico e Tecnológico" ANO-DE-OBTENCAO-DO-TITULO="2007" TITULO-DA-DISSERTACAO-
    TESE="WhizKEY: Um Ambiente para Instalação de Bibliotecas Digitais" NOME-
    COMPLETO-DO-ORIENTADOR="Marcos André Gonçalves" TIPO-MESTRADO="N" NUMERO-ID
    -ORIENTADOR="3457219624656691" CODIGO-CURSO-CAPES="32001010004P6" TITULO-DA
    -DISSERTACAO-TESE-INGLES="" NOME-CURSO-INGLES="Computer Science" NOME-DO-CO
    -ORIENTADOR="" CODIGO-INSTITUICAO-DOUT="" NOME-INSTITUICAO-DOUT="" CODIGO-
    INSTITUICAO-OUTRA-DOUT="" NOME-INSTITUICAO-OUTRA-DOUT="" NOME-ORIENTADOR-
    DOUT=""/>
<DOUTORADO SEQUENCIA-FORMACAO="3" NIVEL="4" CODIGO-INSTITUICAO="085700000002"
    NOME-INSTITUICAO="University of Glasgow" CODIGO-ORGAO="" NOME-ORGAO=""
    CODIGO-CURSO="90000002" NOME-CURSO="Ciência da Computação" CODIGO-AREA-
    CURSO="90000002" STATUS-DO-CURSO="CONCLUIDO" ANO-DE-INICIO="2008" ANO-DE-
    CONCLUSAO="2013" FLAG-BOLSA="SIM" CODIGO-AGENCIA-FINANCIADORA="000100000991
    " NOME-AGENCIA="Universities UK" ANO-DE-OBTENCAO-DO-TITULO="2013" TITULO-DA
    -DISSERTACAO-TESE="Explicit Web Search Result Diversification" NOME-
    COMPLETO-DO-ORIENTADOR="Iadh Ounis" TIPO-DOUTORADO="N" CODIGO-INSTITUICAO-
    DOUT="" NOME-INSTITUICAO-DOUT="" CODIGO-INSTITUICAO-OUTRA-DOUT="" NOME-
    INSTITUICAO-OUTRA-DOUT="" NOME-ORIENTADOR-DOUT="" NUMERO-ID-ORIENTADOR=""
    CODIGO-CURSO-CAPES="" TITULO-DA-DISSERTACAO-TESE-INGLES="" NOME-CURSO-
    INGLES="" NOME-DO-ORIENTADOR-CO-TUTELA="" CODIGO-INSTITUICAO-OUTRA-CO-
    TUTELA="" CODIGO-INSTITUICAO-CO-TUTELA="" NOME-DO-ORIENTADOR-SANDUICHE=""
    CODIGO-INSTITUICAO-OUTRA-SANDUICHE="" CODIGO-INSTITUICAO-SANDUICHE="" NOME-
    DO-CO-ORIENTADOR="">
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Busca na Web" PALAVRA-CHAVE-2="Diversidade em
    busca" PALAVRA-CHAVE-3="" PALAVRA-CHAVE-4="" PALAVRA-CHAVE-5="" PALAVRA-
    CHAVE-6=""/>
<AREAS-DO-CONHECIMENTO>
<AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="
    CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-CONHECIMENTO="" NOME-DA-SUB-
    AREA-DO-CONHECIMENTO="Recuperação de Informação" NOME-DA-ESPECIALIDADE=""/>
</AREAS-DO-CONHECIMENTO>
</DOUTORADO>
</FORMACAO-ACADEMICA-TITULACAO>
```

quantifica alguns dos termos mais frequentes encontrados na etapa de tratamento e limpeza de dados junto ao nome do orientador. Isso posto, a estratégia adotada foi optar por removê-los através do uso de expressões regulares para evitar a perda do mínimo possível de informação relevante.

Já a Figura 3.5 apresenta o extrato de um documento XML que inclui a lista de orientações concluídas de um pesquisador, indicando o seu papel como orientador ou coorientador. Além disso, para cada um dos orientandos presentes na lista, o documento inclui como atributo o identificador do seu currículo, se existente, o nome do orientando, o título da tese ou dissertação defendida, e o ano e a instituição onde

Tabela 3.5. Termos mais encontrados junto ao nome dos orientadores.

Termo	Total
dr	3.078
dr.	14.271
dra	1.657
dra.	4.855
prof	5.577
profa.	1.894

ocorreu a defesa. Esses dados permitem estabelecer todas as relações entre o nodo do pesquisador e os nodos de seus orientandos.

Figura 3.5. Extrato de um documento XML contendo a seção Orientações Concluídas do currículo Lattes de um pesquisador.

```
<ORIENTACOES-CONCLUIDAS>
<ORIENTACOES-CONCLUIDAS-PARA-MESTRADO SEQUENCIA-PRODUCAO="85">
<DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO NATUREZA="Dissertação de
    mestrado" TIPO="ACADEMICO" TITULO="Recomendação de etiquetas para sumariza
    ção de perfis acadêmicos" ANO="2015" PAIS="Brasil" IDIOMA="Português" HOME-
    PAGE="" FLAG-RELEVANCIA="NAO" DOI="" TITULO-INGLES=""/>
<DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO TIPO-DE-ORIENTACAO="
    CO_ORIENTADOR" NOME-DO-ORIENTADO="Isac Sandin Ribeiro" CODIGO-INSTITUICAO="
    033300000002" NOME-DA-INSTITUICAO="Universidade Federal de Minas Gerais"
    CODIGO-ORGAO="" NOME-ORGAO="" CODIGO-CURSO="32000049" NOME-DO-CURSO="Ciê
    ncias da Computação" FLAG-BOLSA="NAO" CODIGO-AGENCIA-FINANCIADORA="" NOME-
    DA-AGENCIA="" NUMERO-DE-PAGINAS="" NUMERO-ID-ORIENTADO="" NOME-DO-CURSO-
    INGLES="Computer Science"/>
</ORIENTACOES-CONCLUIDAS-PARA-MESTRADO>
<ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO SEQUENCIA-PRODUCAO="158">
<DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO NATUREZA="Tese de
    doutorado" TITULO="Similarity-enhanced collaborative filtering" ANO="2017"
    PAIS="Brasil" IDIOMA="Inglês" HOME-PAGE="" FLAG-RELEVANCIA="NAO" DOI=""
    TITULO-INGLES=""/>
<DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO TIPO-DE-ORIENTACAO="
    CO_ORIENTADOR" NOME-DO-ORIENTADO="Ramon Pereira Lopes" CODIGO-INSTITUICAO="
    033300000002" NOME-DA-INSTITUICAO="Universidade Federal de Minas Gerais"
    CODIGO-ORGAO="" NOME-ORGAO="" CODIGO-CURSO="32000049" NOME-DO-CURSO="Ciê
    ncias da Computação" FLAG-BOLSA="NAO" CODIGO-AGENCIA-FINANCIADORA="" NOME-
    DA-AGENCIA="" NUMERO-DE-PAGINAS="" NUMERO-ID-ORIENTADO="" NOME-DO-CURSO-
    INGLES="Computer Science"/>
</ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO>
</ORIENTACOES-CONCLUIDAS>
```

Outra dificuldade que levou a mais um passo adicional para tratamento dos dados foi que, muitas vezes, os nomes dos pesquisadores estão dispostos em diferentes formatos como, por exemplo, sobrenome seguido de nome separado por vírgula. Assim, para a padronização de nomes próprios foi necessário reconhecer os formatos mais comuns e padronizá-los de acordo com o formato mais usual no Brasil, ou seja, primeiro nome seguido de sobrenomes. Além disso, foram removidos acentos e

letras maiúsculas foram trocadas pelas suas equivalentes minúsculas. Também foram removidas dos nomes as chamadas "stop words", entre elas preposições como, "das", "de", "dos" e outras. A Tabela 3.6 mostra alguns exemplos de como nomes próprios foram padronizados. Alguns complementos, como, "Junior" ou "Jr", "Filho", "Neto" e outros foram também considerados como "stop words". Porém, tais complementos não foram removidos, tendo sido anexados ao sobrenome anterior, por exemplo, como no caso de um nome contendo o sobrenome "Câmara" seguido do sufixo "Neto" e que foi tratado como sendo o sobrenome único "Câmara Neto".

Tabela 3.6. Exemplos de padronização de nomes.

Nome pré-processado	**Nome pós-processado**
Professor José da Silva Aguiar	jose silva aguiar
Dra. (coorientação) Maria Mendes	maria mendes
Menezes, César de	cesar menezes
João A. Faria Jr	joao a faria junior
Engenheira Marta Silva	marta silva
César de Menezes / Marta Silva	cesar menezes

3.2.2 Algoritmo para Construção das Árvores

A principal tarefa na construção das árvores genealógicas acadêmicas foi ligar cada pesquisador cujo currículo foi coletado na Plataforma Lattes a seus respectivos orientadores, bem como, a cada um dos orientandos presentes nas respectivas listas de orientações. Para isso, foi preciso que cada pesquisador fosse reconhecido como uma entidade única. Um pesquisador aparece no repositório da Plataforma Lattes com três papéis diferentes. Primeiro ele aparece como orientando na lista de orientações do seu orientador. Depois, ele aparece como pesquisador a partir do seu próprio currículo, listando os seus orientadores e orientandos. Por fim, ele aparece como orientador, seja de mestrado ou doutorado, no currículo de cada um de seus orientandos.

Para garantir a unicidade das entidades presentes na Plataforma Lattes, o primeiro passo foi utilizar, quando presente, o próprio código identificador único presente no currículo, um número composto por 16 dígitos. Algumas vezes a própria plataforma identifica as demais entidades presentes em um currículo, ligando aquela referência a um pesquisador ao seu próprio currículo. Na Figura 3.6, essa ligação é representada pelo ícone amarelo presente no início de cada linha que identifica um orientando de um pesquisador. Como pode ser percebido, essa ligação está presente apenas em al-

guns casos, o que significa que nem sempre é possível identificar automaticamente o pesquisador referenciado.

Figura 3.6. Exemplos de entidades reconhecidas na Plataforma Lattes (as duas que contêm a logomarca do Lattes à frente).

Tese de doutorado

1. Daniel Hasan Dalip. A Multi-View Approach for Estimating the Quality of Collaborative Content on the Web 2.0. 2015. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
2. Anderson Ferreira. Contributions for Solving the Author Name Ambiguity Problem in Bibliographic Citations. 2012. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Orientador: Marcos André Gonçalves.
3. Moises Gomes de Carvalho. Abordagens evolucionárias para problemas relacionados a integração de dados. 2009. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Coorientador: Marcos André Gonçalves.
4. Leonardo Rocha. Uso de Contextos Temporais para Classificação Automática de Documentos. 2009. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, . Coorientador: Marcos André Gonçalves.
5. Guilherme Tavares de Assis. Uma Abordagem Baseada em Gênero para Coleta Temática. 2008. Tese (Doutorado em Ciências da Computação) - Universidade Federal de Minas Gerais, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Coorientador: Marcos André Gonçalves.

Para aquelas entidades ainda sem identificação, é criado um identificador provisório até que o identificador original seja encontrado, sendo que, em algumas situações, esse identificador é mantido posteriormente, como no caso das entidades que não possuam o seu currículo na Plataforma Lattes (por exemplo, pesquisadores estrangeiros que orientaram brasileiros em programas no exterior). Esse identificador é formado da seguinte maneira:

- Para orientadores de doutorado: "(Lattes(la)) + (posição(1,2,3...)) + (papel orientador(or)) + (doutorado(dr)) + identificador do pesquisador" (ex.: la1ordr8675903095837164);
- Para orientadores de mestrado: "(Lattes(la)) + (posição(1,2,3...)) + (papel orientador(or)) + (mestrado(ms)) + identificador do pesquisador" (ex.: la2orms8675903027583041);
- Para alunos de doutorado: "(Lattes(la)) + (aluno(al)) + (doutorado(dr)) + (papel orientador(or/coor)) + (posição(1,2,3...)) + (Lattes(la)) + identificador do pesquisador" (ex.: laaldror2la8675903027583041);
- Para alunos de mestrado: "(Lattes(la)) + (aluno(al)) + (mestrado(ms)) + (papel orientador(or/coor)) + (posição(1,2,3...)) + (Lattes(la)) + identificador do pesquisador" (ex.: laalcoorms7la8675903095837164).

Com isso, foi possível não só criar um identificador único para cada pesquisador que não possuia um identificador Lattes conhecido, como também rastrear a origem da-

quela identificação. Por exemplo, o identificador "laalcoorms7la8675903095837164" indica que ele pertence a um aluno de mestrado, listado na posição 7 da lista de orientandos do pesquisador cujo identificador Lattes (la) é 8675903095837164.

Após esse processo foi possível iniciar a construção das árvores seguindo o Algoritmo 1 proposto para essa finalidade [Dores et al., 2017]. Esse algoritmo recebe como entrada os dados extraídos e processados de cada currículo coletado, retornando como resultado um Grafo Dirigido Acíclico (GDA) [Bollobas, 1998] de arestas múltiplas. Um GDA nada mais é do que um grafo com arestas dirigidas e que não contém ciclos. Neste caso, suas arestas podem ser múltiplas porque um pesquisador pode orientar (ou coorientar) um mesmo aluno mais de uma vez (por exemplo, no mestrado e no doutorado). As arestas são dirigidas para representar a relação de orientação e não se espera ciclos já que existe uma hierarquia no processo de orientação acadêmica.

Algorithm 1: Processo de construção das árvores genealógicas acadêmicas

Entrada: Um conjunto C de currículos Lattes;
Saída: Um grafo G com todas as árvores construídas;

```
1  Ordena C pelo ano de obtenção do grau de doutor;
2  Definir G vazio;
3  foreach Currículo c em C do
4  |  Busca em G pelo nodo do pesquisador n;
5  |  if Se não existe o nodo n em G then
6  |  |  Cria o nodo n;
7  |  else
8  |  |  Atualiza os atributos acadêmicos de n;
9  |  end
10 |  Busca em G pelos nodos p e m dos orientadores de Mestrado e Doutorado;
11 |  if Se p ou m não foram encontrados then
12 |  |  Cria p, m ou ambos;
13 |  else
14 |  |  Atualiza os atributos acadêmicos de p e m;
15 |  end
16 |  Liga os nodos p e m ao nodo n;
17 |  foreach aluno orientado em c do
18 |  |  Busca em G pelo nodo do aluno orientado a;
19 |  |  if Se não existe o nodo a em G then
20 |  |  |  Cria o nodo a;
21 |  |  else
22 |  |  |  Atualiza os atributos acadêmicos de a;
23 |  |  end
24 |  |  Liga o nodo a ao nodo n;
25 |  end
26 end
```

Seguindo o Algoritmo 1, a primeira etapa para a construção das árvores genealógicas acadêmicas consiste em ordenar o conjunto C de currículos de acordo com o ano em que o pesquisador obteve o seu título de doutor (linha 1). Ordenar os currículos em ordem cronológica contribui para evitar a comparação desnecessária de dados dos currículos antigos com os de currículos mais recentes, permitindo gerar árvores mais uniformemente, pois garante a criação de todos os nodos antecessores antes dos seus sucessores. O próximo passo consiste em criar um grafo vazio (linha 2), que será preenchido com nodos representando os pesquisadores e arestas representando as relações de orientação entre esses pesquisadores.

A seguir, para cada currículo presente no conjunto C (linhas 3 a 26), são executados três passos, listados a seguir:

1. Procura em G pelo nodo do pesquisador, criando um novo nodo caso ele ainda não exista ou atualizando-o caso contrário (linhas 4 a 9);

2. Procura em G pelos nodos correspondentes aos orientadores de mestrado e doutorado do pesquisador, criando-os caso ainda não existam ou atualizando-os com alguma informação relevante caso contrário. A seguir, conecta os nodos dos orientadores ao nodo do pesquisador (linhas 10 a 16);

3. Para cada aluno orientado pelo pesquisador, procura em G pelo respectivo nodo, criando-o caso não exista ou atualizando-o caso necessário. A seguir conecta o nodo do aluno ao nodo do pesquisador (linhas 17 a 25).

O primeiro passo procura determinar se o pesquisador que está sendo tratado já foi processado anteriormente, isto é, se ele foi citado em um currículo já processado. Por exemplo, normalmente, a primeira citação de um pesquisador vem da lista de alunos do seu orientador, que por ser um pesquisador mais sênior já teve o seu currículo processado antes. Assim, se esse pesquisador já foi processado anteriormente, esse nodo é recuperado e atualizado com novos dados, por exemplo, novas formações que estão presentes no currículo sendo processado.

O segundo passo tem como o objetivo criar os nodos que representam os orientadores do pesquisador cujo currículo está sendo processado, sejam eles de mestrado ou doutorado. Da mesma forma que no passo anterior, o algoritmo busca entre nodos já criados se há algum que corresponde aos orientadores presentes no currículo sendo processado. Caso verdadeiro, verifica-se se há algum dado a ser atualizado. Caso contrário, um novo nodo é criado representando o orientador do pesquisador. Ao fim desse passo é criada, para cada relação desse pesquisador com seus orientadores, uma aresta

dirigida no sentido orientador-pesquisador e a ela são inseridos dados adicionais relativos a essa orientação. Por exemplo, uma orientação de mestrado possui uma data de conclusão, o nome da instituição onde foi realizada, o título da dissertação defendida e a área do conhecimento associada.

Por fim, o terceiro passo tem como objetivo processar as listas contendo os alunos de mestrado e doutorado orientados por esse pesquisador. Mais uma vez, há a busca por nodos já criados ou novos nodos são criados e, para cada nodo recuperado ou criado, é gerada uma aresta, agora no sentido pesquisador-orientando, sendo inseridos dados sobre o mestrado ou doutorado desses orientandos. Esses três passos são repetidos para cada um dos currículos coletados até que todos tenham sido processados.

3.2.3 Processo de Desambiguação de Nomes

Um componente crítico do nosso algoritmo é a função que busca por nodos já criados utilizada nas linhas 4, 10 e 18. Apesar de a Plataforma Lattes prover um identificador único para cada pesquisador, como mencionado anteriormente, nem sempre é possível encontrar instantaneamente o nodo do pesquisador que é referenciado apenas pelo seu nome em outro currículo. Assim, para lidar com esse problema, foi implementado um processo simples, porém bastante efetivo, para realizar a desambiguação de nomes ao processar os dados dos pesquisadores.

Como descrito por Ferreira et al. [2012] e Smalheiser & Torvik [2009], os dois principais desafios para o processo de desambiguação de nomes são:

1. O fato de um mesmo pesquisador poder aparecer no repositório com diferentes nomes devido a abreviações ou alterações ocorridas em razão de casamento, motivos religiosos ou mudança de gênero;

2. A possibilidade de pesquisadores distintos terem nomes similares (polissemia).

Para lidar com esse problema de forma bastante ampla considerou-se diferentes situações. Para aquelas entidades sem um código de identificação padrão do Lattes, a função de busca inclui um passo extra. Esse passo foi inspirado no trabalho de Cota et al. [2010] que combina funções de similaridade aplicadas a atributos presentes em cada currículo com algumas heurísticas usadas para desambiguar nomes de autores em artigos científicos. Nesta dissertação foi necessário apenas adaptar essas funções e heurísticas para o contexto de orientações acadêmicas. Para isso, foram considerados outros dados que pudessem agir como atributos que auxiliassem na identificação dos pesquisadores. Esses atributos são, além do nome do pesquisador, o nome da instituição

onde obteve o seu título acadêmico, o título da tese ou dissertação defendida e o ano em que ocorreu a respectiva defesa.

Cada vez que um currículo é processado, cada nodo criado é inserido em um índice com uma chave formada pela primeira letra do primeiro nome do pesquisador juntamente com o seu último nome. Esse índice tem como objetivo realizar um primeiro filtro dos nodos candidatos a uma fusão, evitando a criação de um novo nodo caso ele já exista no repositório. Vale ressaltar que muitas vezes, em um primeiro momento, são criados vários nodos para um mesmo pesquisador e somente depois esses nodos são fundidos em um nodo único. Isso porque, ao processar os dados de um pesquisador pela primeira vez, ainda não existem atributos suficientes para unificar os seus nodos. Normalmente, nodos criados a partir de dados provenientes da seção de orientações concluídas de seus orientadores de mestrado e doutorado terão diferentes valores de atributos, como título e ano de defesa, para cada uma das formações, mesmo tratando-se do mesmo pesquisador. Porém, no processamento do currículo desse pesquisador, tanto os dados sobre o mestrado quanto os dados sobre o doutorado estão presentes na sua seção de titulações acadêmicas. Assim, ao processar o seu currículo todos os nodos referentes a esse pesquisador são identificados e fundidos em um único nodo.

O processo proposto para desambiguação de nomes pode ser dividido em duas etapas. A primeira busca, por meio da estratégia de comparação por fragmentos [Oliveira et al., 2005], verificar se dois nomes são similares. Nessa etapa, os nomes são divididos em fragmentos e cada um deles é comparado com o seu fragmento equivalente no outro nome, utilizando para essa comparação a distância de Levenshtein [1966]. Um limite para essa distância é definido e caso a distância entre os dois nomes seja menor que esse limite, o fragmento é marcado. Cabe destacar que a abreviação de um nome é comparada com a primeira letra de um fragmento. Ao final, se em pelo menos uma das cadeias de caracteres comparadas houver fragmentos marcados, então os nomes são considerados compatíveis. Caso contrário, eles são considerados incompatíveis.

Um diferencial da comparação por fragmentos é permitir a comparação de abreviações. Diferentemente de artigos científicos que geralmente incluem o chamado "nome de citação", nos currículos não há um padrão para a inserção dos nomes dos pesquisadores. Essa função busca resolver o problema de nomes que sejam na verdade "sinônimos", caso das abreviações de nomes próprios que geralmente aparecem em currículos.

A segunda parte do processo procura resolver o problema de homônimos, ou seja, dois pesquisadores distintos mas que possuem o mesmo nome ou nomes muito parecidos. O fato de haver muitas abreviações nos currículos agrava ainda mais esse problema. Para isso, os atributos correspondentes são comparados de forma a verificar

se dois nomes parecidos correspondem a um mesmo pesquisador ou a pesquisadores diferentes. No caso de outros atributos, a comparação é feita de acordo com as suas características específicas. Por exemplo, para o título da tese ou dissertação é utilizada a distância do cosseno [Baeza-Yates & Ribeiro-Neto, 1999], pois percebeu-se que muitas vezes há pequenas variações entre a grafia dos títulos existentes no currículo de um pesquisador e nos currículos de seus orientadores. Para o ano de conclusão há um peso caso ele seja o mesmo e outro para o caso em que haja uma diferença de no máximo um ano entre os anos comparados.

Muitas vezes, os currículos não incluem todos os atributos necessários para a segunda etapa de comparação, prejudicando a comparação entre os diferentes nodos. Para contornar essa dificuldade, ao processar a seção de titulações do currículo de um pesquisador, é verificado se a lista de orientandos presente na seção de orientações concluídas do currículo do seu orientador inclui o seu nome. Essa estratégia auxilia, por exemplo, quando currículo do orientador não lista o título da tese desse pesquisador.

3.3 Armazenamento dos Dados

Por fim, concluído o processo de tratamento de dados, eles foram armazenados em um banco de dados orientado a grafos implementado utilizando-se o sistema Neo4j [Webber, 2012]. Além de facilitar a especificação de consultas sofisticadas, essa tecnologia permite alterar a estrutura dos dados sem que o banco de dados seja diretamente afetado. Isso porque, diferentemente dos bancos de dados relacionais, em um banco de dados não relacional (NoSQL) os relacionamentos são parte dos dados, o que permite alterá-los sem afetar a estrutura do banco de dados como um todo.

Especificamente, o sistema Neo4j armazena os dados como vértices e arestas, ou seja, na sua terminologia, nodos e relacionamentos. Assim, entidades são representadas como nodos e associações entre elas são representadas como relacionamentos entre os nodos. A Figura 3.7 apresenta a representação dos dados referentes às árvores genealógicas acadêmicas, na qual o nodo representa pesquisadores e possui diversos atributos, como nome, id, e-mail, entre outros, e o auto-relacionamento estabelece arestas entre pesquisadores que ligam um pesquisador a outro. Cada aresta possui um atributo denominado Tipo que tem o objetivo de identificar o relacionamento em função do papel desempenhado pelo pesquisador, ou seja, orientador ou coorientador, bem como se esse relacionamento ocorreu durante o mestrado ou doutorado. Assim, a estrutura do banco de dados Neo4j que armazena as árvores genealógicas acadêmicas é derivada diretamente do Algoritmo 1, ou seja, a estrutura do grafo retornado pelo

algoritmo é diretamente armazenada no banco de dados.

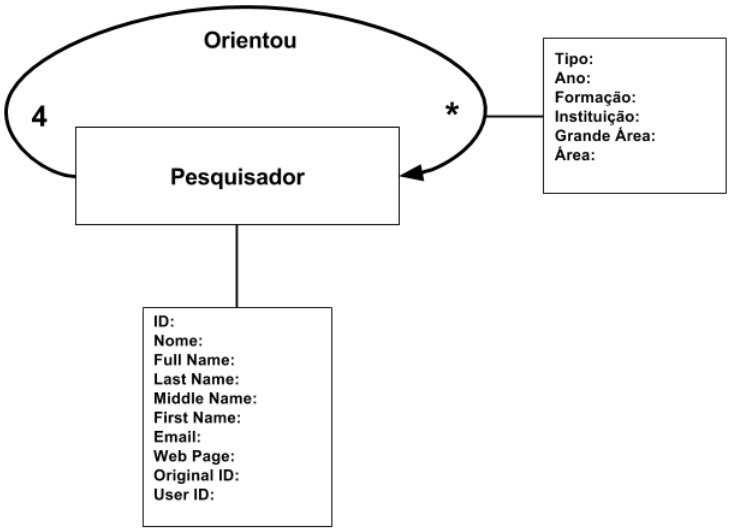

Figura 3.7. Representação gráfica da estrutura do banco de dados armazenado pelo sistema Neo4j.

A principal vantagem de um banco de dados implementado usando-se o sistema Neo4j é permitir a visualização dos dados como um grafo, possibilitando a especificação de consultas que permitem percorrê-lo de modo a recuperar todos os nodos e arestas que satisfaçam tais consultas. A Figura 3.8 apresenta o exemplo de uma consulta especificada sobre o banco de dados que armazena as árvores genealógicas acadêmicas de acordo com a estrutura representada pelo diagrama da Figura 3.7.

Figura 3.8. Exemplo de uma consulta especificada de acordo com a linguagem do sistema Neo4j.

```
MATCH (virgilio:Person { name:"virgilio augusto fernandes almeida" }),
(alberto:Person { name:"alberto henrique frade laender" }),
(virgilio)-[r:SUPERVISED]->(pesq),
p = shortestPath((alberto)-[:SUPERVISED*..4]-(pesq))
WHERE toLower(r.university_name) = 'universidade federal de minas gerais'
AND length(p) < 5 AND NOT ANY(x IN NODES(p)
WHERE x.name = "virgilio augusto fernandes almeida")
RETURN DISTINCT pesq.name as Nome, length(p) as Distância
ORDER BY length(p)
LIMIT 10
```

A consulta especificada retorna o respectivo nome e a menor distância entre os nodos dos pesquisadores orientados pelo professor Virgílio Augusto Fernandes Almeida

Tabela 3.7. Resultado da consulta exemplo.

Nome	Distância
Luiz Henrique Gomes	1
Fabricio Benevenuto de Souza	2
Lucila Ishitani	2
Jussara Marques de Almeida	2
Marisa Affonso Vasconcelos	3
Humberto Torres Marques Neto	3
Fatima de Lima Procópio Duarte Figueiredo	3
Cristina Duarte Murta	3
Luciano Pereira Gomes	4
Leandro Faria Freitas	4

e o nodo do professor Alberto Henrique Frade Laender, ambos da Universidade Federal de Minas Gerais. O seu resultado, limitado a 10 ocorrências, pode ser visto na Tabela 3.7, onde a coluna Nome corresponde ao nome do pesquisador orientado pelo professor Virgílio Almeida e a coluna Distância indica o tamanho (número de passos) do menor caminho até o nodo do professor Alberto Laender.

Capítulo 4

Caracterização e Análise das Árvores

Este capítulo apresenta uma série de análises sobre as árvores genealógicas acadêmicas construídas como resultado desta dissertação. Inicialmente, são apresentadas algumas estatísticas gerais dessas árvores. Em seguida, é realizada uma análise de suas propriedades estruturais levando em consideração as grandes áreas do conhecimento definidas pelo CNPq. Finalmente, é apresentada uma visão geral do portal Science Tree que foi desenvolvido para possibilitar a consulta, visualização e exploração das árvores construídas.

4.1 Estatísticas Gerais

A fim de entender a estrutura das árvores genealógicas acadêmicas dos pesquisadores envolvidos neste estudo, procurou-se medir e caracterizar diversos aspectos relacionados a essas árvores, entre eles o número de descendentes presentes em cada uma delas, a fecundidade acadêmica desses pesquisadores e o tamanho de suas respectivas linhagens. Com isso, obteve-se uma ampla visão do estado em que se encontra a rede que envolve as relações de orientações acadêmicas no Brasil.

A Tabela 4.1 apresenta algumas estatísticas gerais relativas às árvores construídas. Inicialmente, nota-se que os dados extraídos dos mais de 250 mil currículos coletados da Plataforma Lattes permitiram a criação dos mais de um milhão de nodos que compõem essas árvores. Isso mostra o volume de informação contido em cada currículo, já que, normalmente, cada um deles contém dados sobre o pesquisador, seus orientadores e seus orientados.

O conjunto de árvores construídas possui um número de componentes muito menor do que o número total de árvores. Isso indica que muitas dessas árvores estão unidas por uma ou mais arestas. Isto é, em algum momento essas árvores uniram-se em razão da orientação de um pesquisador que não possuía qualquer relação com os demais. Por exemplo, quando um pesquisador realiza o mestrado e o doutorado em grupos distintos, há uma grande chance de duas árvores diferentes unirem-se por meio dos nodos de seus orientadores. Investigando um pouco mais os componentes formados, temos que o maior deles possui 981.566 nodos, isto é, cerca de 94% de todos os nodos existentes fazem parte desse componente. Esse é um número bastante inesperado, pois indica que, de alguma forma, os pesquisadores presentes no componente gigante estão conectados indiretamente. Uma hipótese para esse fato é a interdisciplinaridade gerada por algumas áreas. Por exemplo, é muito comum encontrar orientações envolvendo pesquisadores de áreas das Ciências Sociais Aplicadas e das Ciências Humanas, bem como de áreas das Ciências Exatas e da Terra e das Ciências Sociais Aplicadas. Isso torna a grande área de Ciências Sociais Aplicadas uma ponte entre as Ciências Humanas e as Ciências Exatas e da Terra.

Camadas	Quantidade
Nodos	1.041.339
Arestas	1.330.321
Componentes	18.780
Árvores	72.174

Tabela 4.1. Estatísticas gerais sobre as árvores genealógicas acadêmicas.

Separando os componentes em árvores individuais, nas quais há uma relação direta entre seus nodos, é possível analisar o tamanho dessas árvores como se fosse a sua população. Como mostrado na Figura 4.1, cerca de 60% dos orientadores possuem apenas um descendente direto. Novamente, como o número de componentes é menor que o número de árvores, essas árvores de tamanho dois estão inseridas em sua grande maioria dentro desses componentes.

Essas árvores de tamanho dois são normalmente derivadas de situações onde, após processar o currículo de um pesquisador, não foi possível encontrar o currículo de nenhum de seus antecessores ou descendentes. Incluem-se nesses casos pesquisadores que não possuem seus currículos cadastrados na Plataforma Lattes como, por exemplo, aqueles vinculados a instituições fora do país. Outra situação que leva ao surgimento dessas árvores são erros contidos nos currículos. Por exemplo, erros de grafia que impedem o casamento entre o nome existente no currículo do orientador e o nome

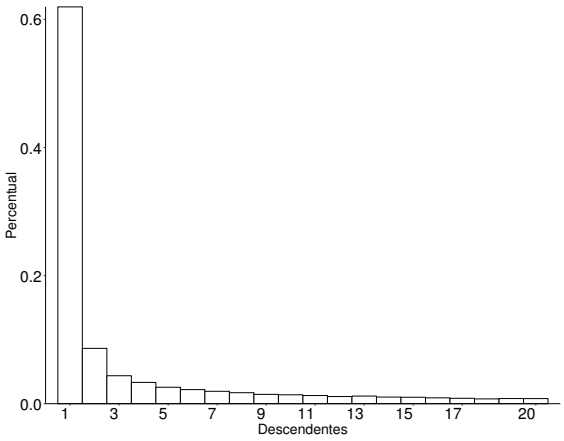

Figura 4.1. Distribuição do número de descendentes (tamanho) das árvores em relação ao nodo raiz, até 20 descendentes.

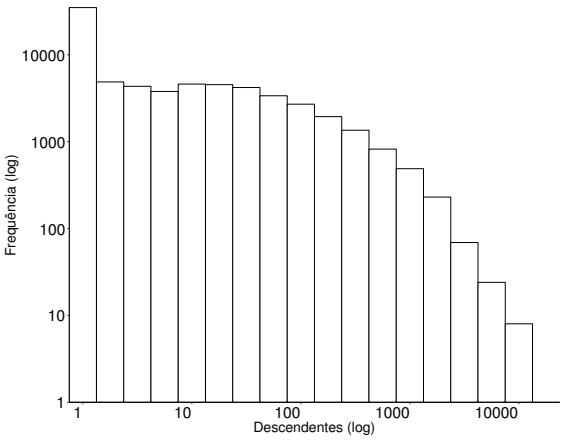

Figura 4.2. Distribuição log-log do número de descendentes (tamanho) das árvores em relação ao nodo raiz.

do orientador informado no currículo do pesquisador. O mesmo pode ocorrer para os orientandos de um pesquisador.

Pesquisador	Instituição	Tamanho
Joel Martins	PUC-SP	15.091
Jorge Pereira Lima	UFRGS	12.700
Eduardo Oliveira França	USP	10.371
Andre Dreyfus	USP	9.958
Annita Castilho	USP	9050
Florestan Fernandes	USP	9.452
Tamara Dembo	NSSR	8.987
F. G. Brieger	USP	8.317
Joaquim Campos	UFV	7.911
José Theóphylo do Amaral Gurgel	USP	7.233

Tabela 4.2. Relação das 10 árvores mais populosas.

Por outro lado, nota-se que o tamanho das árvores segue um padrão semelhante ao de distribuições de cauda pesada [Anderson, 2008], onde um pequeno número de árvores possui um número muito maior de descendentes do que as demais. Calculando o coeficiente de curtose [Westfall, 2014] para essa distribuição tem-se um valor de 754,65. Distribuições com valores de curtose maior que três já podem ser consideradas como sendo de cauda longa, quando comparadas à distribuição normal. A Figura 4.1 mostra essa distribuição apenas até as árvores de tamanho 20 que correspondem à maior parte das árvores construídas. As maiores árvores incluem milhares de nodos, como pode ser observado pela Figura 4.2, sendo em sua maioria de pesquisadores vinculados a instituições brasileiras, como mostra a Tabela 4.2. Além disso, boa parte dessas árvores

começaram a se formar na década de 1960 e são bem distintas umas das outras. Cabe ainda ressaltar que as duas maiores árvores se fundiram, embora compartilhem entre si apenas 359 nodos.

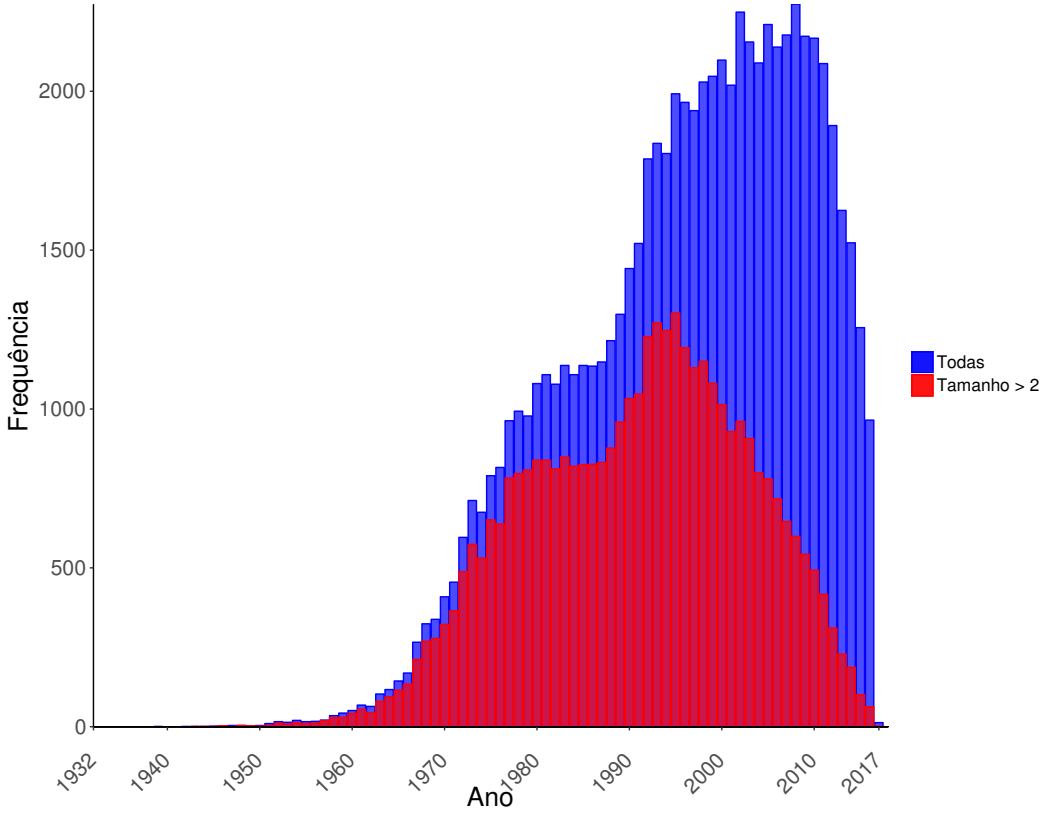

Figura 4.3. Distribuição das árvores pelo ano da orientação mais antiga.

O gráfico da Figura 4.3 apresenta a distribuição das árvores de acordo com o ano da orientação mais antiga. Como pode ser observado, grande parte das árvores de tamanho dois, destacadas em azul, foram formadas após o ano 2000. Isso mostra que essas árvores são relativamente recentes e podem estar ainda em processo de evolução. Logo, a partir do grupo de pesquisadores que fazem parte dessas árvores, que engloba 60% dos doutores, estão sendo formadas as novas árvores que se firmam na genealogia acadêmica brasileira. Analisando um pouco mais esse gráfico, é possível perceber um crescimento acentuado do número de árvores entre as décadas de 1960 e 1970 devido a diversos programas de incentivo ao desenvolvimento da pós-graduação. Na década de 1980 a 1990 houve uma estagnação no crescimento do número de árvores, possivelmente devido à recessão econômica ocorrida no Brasil naquele período, havendo novamente um pico de crescimento entre os anos de 1990 e 2000. Detalhes de tais eventos que influenciaram diretamente a pós-graduação no país foram abordados por dos Santos & de Azevedo [2009]. É possível observar ainda uma queda no número de novas árvores a

partir dos anos 2000, apesar de o número de doutores estar em pleno crescimento. Isso mostra que grande parte dos doutores que defenderam as suas teses após esse período passaram a fazer parte de árvores já existentes, tornando-as cada vez mais densas.

Outra característica importante a ser analisada é a linhagem de uma árvore, ou seja, o número de gerações acumuladas. Observando a Figura 4.4, pode-se notar que há árvores que já ultrapassam oito gerações. Em termos de anos, considerando que em média um doutor leva de três a quatro anos para se formar, são necessários cerca de 32 anos para uma árvore atingir esse número de gerações.

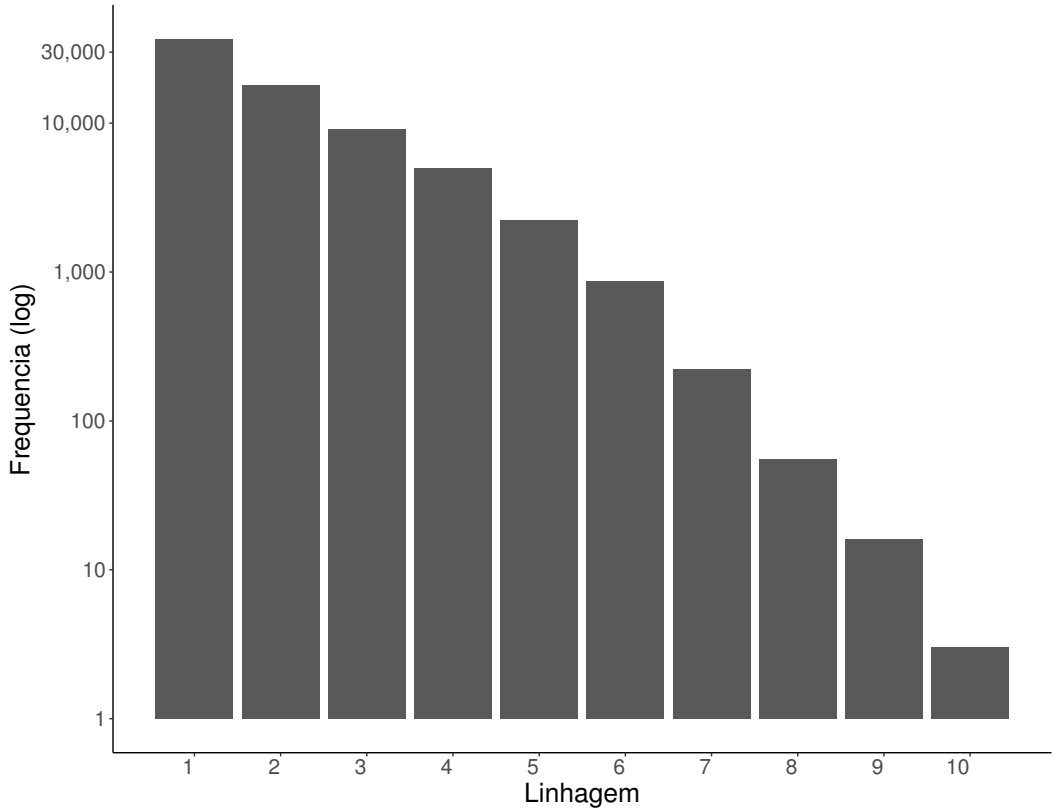

Figura 4.4. Distribuição da linguagem (profundidade) das árvores.

A Tabela 4.3 apresenta a relação dos 15 pesquisadores cujas árvores possuem as maiores linhagens. É importante ressaltar que essa relação inclui, em sua maioria, pesquisadores de instituições nacionais, predominantemente do estado de São Paulo. Vale ainda mencionar que o pesquisador André Dreyfus[1], cuja árvore é uma das três que possuem a maior linhagem, é considerado o pai da Genética brasileira, sendo também um dos fundadores da Universidade de São Paulo.

[1]http://dx.doi.org/10.1590/S0103-40141994000300017

Pesquisador	Instituição	Linhagem
André Dreyfus	USP	10
Celso D. Albuquerque Mello	UFRJ	10
Elisaldo Luiz Araújo Carlini	UNIFESP	10
Arlie C. Todd	University of Wisconsin-Madison	9
Charles Whitehair	Michigan State University	9
Gerhard Salinger	USP	9
James Sommerville	Fordham University	9
José Mario Braga	UFV	9
Lenom J. Cajuste	Colegio de Postgraduados	9
Luiz Carlos Junqueira	USP	9
Raymond W. Fahien	UFRJ	9
Samuel P. Huntington	Harvard University	9
Virginio Pessoa Delgado Filho	USP	9
Werner Marx	Universität Freiburg	9
Wilhelm Otto Daniel Martin Neitz	UFRRJ	9

Tabela 4.3. Relação dos 15 pesquisadores com as maiores linhagens.

É interessante notar que considerando tanto a métrica descendência quanto a linhagem, a maior parte das árvores analisadas nasceu em instituições brasileiras. Isso é uma evidência que a ciência brasileira é bastante estruturada e fundamenta-se nos programas de pós-graduação nacionais. Entretanto, vale ressaltar que seis das 15 árvores com maiores linhagens provêm de instituições estrangeiras (ver Tabela 4.3), algumas de grande renome mundial. Também vale notar que as três árvores de maior linhagem foram iniciadas por pesquisadores que não possuiam nenhuma titulação formal, já que na época os programas de doutorado ainda não eram difundidos no país.

Pesquisador	# Descendentes	Linhagem	Fecundidade
Marcel Kadima Kamuleta	182	2	91,0
Urbano Kurylo	301	4	79,5
Ismail A. Ghazalah	301	4	79,5
André Coupez	202	2	67,3
Gilles Olive	117	2	58,5
Patrick Depecker	117	2	58,5
Paul Brejon	148	2	49,3
Rogério Bastos Vale	133	3	44,3
T. Dracos	86	2	43,0
Antonio Gouveia Sousa	117	2	42,3

Tabela 4.4. Relação dos 10 pesquisadores com árvores mais fecundas.

A Tabela 4.4 apresenta as 10 árvores mais fecundas, de acordo com a métrica definida no Capítulo 2 (Equação 2.1). Com essa métrica é possível medir, de certa

forma, a qualidade da árvore de um pesquisador, já que espera-se que um pesquisador forme não só novos pesquisadores, mas que esses pesquisadores também se tornem orientadores. Como é possível perceber, as árvores mostradas não são nem as maiores e nem as mais profundas, pois a métrica de fecundidade procura identificar as árvores que possuem uma maior pré-disposição para se desenvolver.

4.2 Análise das Árvores Agrupadas pelas Grandes Áreas do Conhecimento

Esta seção tem como objetivo analisar, utilizando as métricas anteriores, as diferenças entre as árvores pertencentes a cada uma das grandes áreas do conhecimento, segundo o esquema de classificação definido pelo CNPq e introduzido no Capítulo 3. Para isso, foi necessário classificar as árvores construídas de acordo com as grandes áreas indicadas pelos pesquisadores em seus currículos para as suas respectivas teses e dissertações.

Grande Área	Total de Árvores
Ciências Humanas	11.738
Ciências Exatas e da Terra	8.132
Ciências da Saúde	7.055
Engenharias	6.831
Ciências Sociais Aplicadas	6.270
Ciências Agrárias	4.773
Ciências Biológicas	4.772
Linguística, Letras e Artes	3.537
Outros	580
Não Identificada	18.486
Total	72.174

Tabela 4.5. Total de árvores em cada grande área do conhecimento.

Assim, após identificar a grande área de cada uma das orientações presentes nas árvores, elas foram agrupadas de acordo com a grande área definida na maioria das teses e dissertações resultantes dessas orientações. A opção pela maioria deveu-se ao fato de que uma árvore pode incluir pesquisadores cujas teses e dissertações foram por eles classificadas em diferentes grandes áreas. Assim, ao final, as árvores foram separadas em diferentes grupos de acordo com a grande área do conhecimento predominante.

Como resultado, as árvores foram distribuídas em 10 grupos, como pode ser visto na Tabela 4.5. É importante observar que, como a grande área de Ciências Humanas é a que possui o maior número de teses e dissertações defendidas (ver Tabela 3.3), ela é também a grande área com o maior número de árvores. Além disso, para 18.486

árvores não foi possível identificar a sua grande área específica, de modo que elas foram desconsideradas para fins de análise. Vale ressaltar ainda que, em sua grande maioria, essas árvores incluíam uma única relação de orientação sem qualquer grande área identificada. Já a grande área Outros, além de pouco representativa em termos de abrangência, inclui apenas 580 árvores, número quase seis vezes menor do que o da grande área de Linguística, Letras e Artes que possui o menor número de árvores.

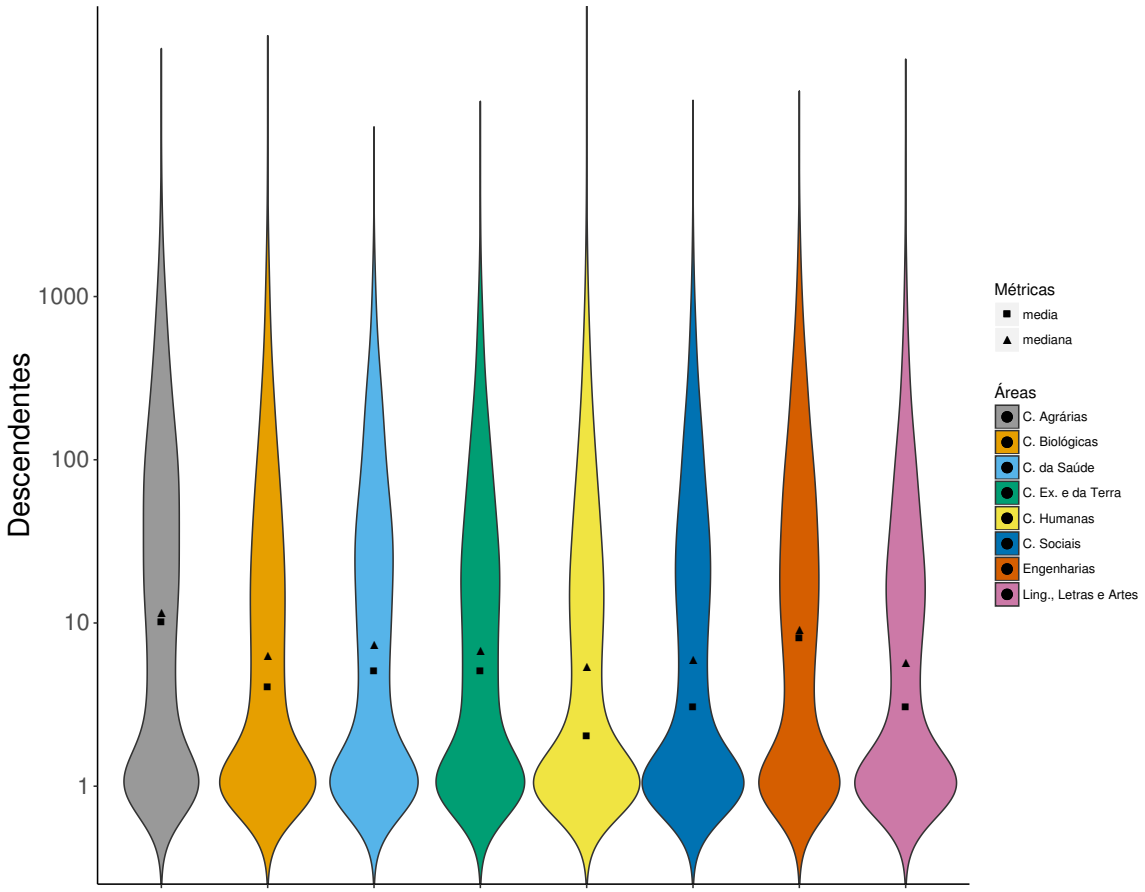

Figura 4.5. Distribuição do número de descendentes das árvores agrupadas pelas grandes áreas.

Analisando mais detalhadamente as árvores de cada grande área, a primeira métrica considerada foi a descendência. Essa foi uma das métricas com menor diferença entre as grandes áreas. Na Figura 4.5, os gráficos de violino mostram a distribuição do número de descendentes e a densidade dessa distribuição nas árvores para cada uma das grandes áreas. Como pode ser visto, a diferença entre as grandes áreas é bem sutil e em todas elas há uma grande concentração de árvores com apenas um descendente, sendo a Ciências Humanas a grande área em que essa concentração é maior. Uma hipótese para explicar esse fenômeno é que, independentemente da grande área,

as árvores em geral possuem poucos descendentes. Isso deve-se não só à juventude dessas árvores, como também ao fenômeno denominado "rico fica mais rico", também conhecido por ligação preferencial (*preferential attachment*) [Newman, 2001c], que faz com que aqueles pesquisadores que já tenham concluído mais orientações atraiam mais novos orientandos.

Nesse mesmo gráfico pode-se ver ainda a média e a mediana para cada grande área, o que auxilia na comparação entre elas. As medianas possuem uma diferença bem menor quando comparadas com as respectivas médias em todas as grandes áreas. Isso porque a mediana mostra que 50% das árvores possuem um tamanho menor do que o seu valor. Já a média é influenciada pela quantidade de árvores, de modo que grandes áreas com uma maior quantidade de árvores tendem a ter uma média menor do que a sua mediana. É interessante notar que, nas grandes áreas de Ciências Agrárias e Engenharias, a média e a mediana são bastante próximas, mostrando que nessas grandes áreas existe um maior equilíbrio entre a quantidade de árvores e o número de descendentes presentes em cada uma delas.

Outra análise do ponto de vista de cada grande área é o ano em que ocorreu a primeira orientação de cada árvore, ou seja, quando nasceram as árvores pertencentes àquela grande área. Ao observar a Figura 4.6, pode-se perceber que em parte das grandes áreas houve um decréscimo no surgimento de novas árvores após a década de 2000. Esse decréscimo, entretanto, não significa necessariamente um encolhimento dessas grandes áreas, mas pode significar a consolidação dos seus programas de pós-graduação, ou seja, cada vez mais os novos doutores dessas grandes áreas são oriundos de grupos já consolidados.

Outro fato relevante é o surgimento de alguns picos mais acentuados nas grandes áreas de Engenharias e Ciências Agrárias no fim da década de 1970 e início da década de 1980, muito provavelmente resultado da industrialização ocorrida no Brasil em meados dessas décadas [Vargas, 1997]. Durante esse período, houve um esforço do governo para a criação de novos programas de pós-graduação, inicialmente devido à necessidade de mais engenheiros no mercado, mas que depois se estendeu para as demais grandes áreas. As grandes áreas de Ciências Biológicas e Ciências Exatas e da Terra foram as únicas em que o aumento do número de novas árvores se manteve praticamente estável até 2010, lembrando que o número de doutores formados no país vem crescendo em um ritmo considerável, como atesta a recente avaliação quadrienal dos programas de pós-graduação realizada pela CAPES.

Ainda que a distribuição das árvores genealógicas acadêmicas foco desta dissertação seja dependente do cadastramento dos pesquisadores na Plataforma Lattes, que

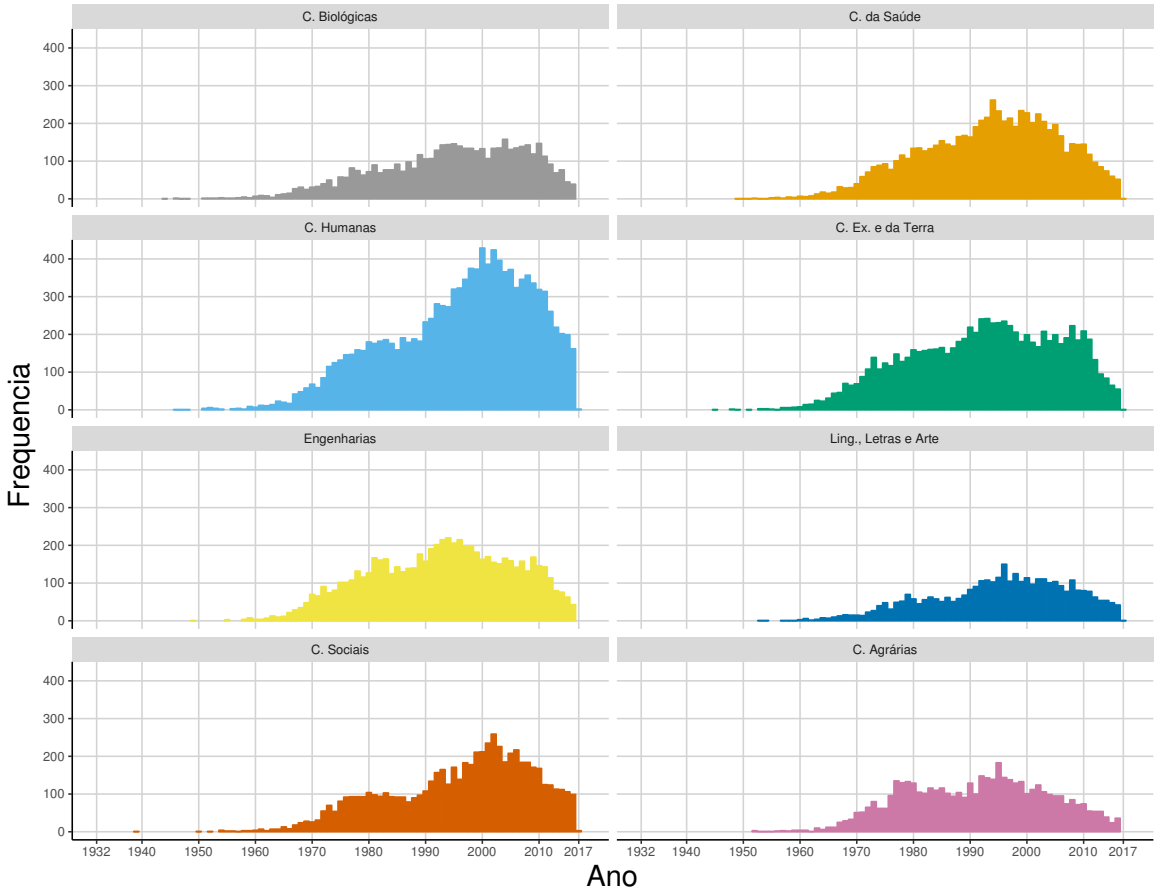

Figura 4.6. Distribuição do ano da orientação mais antiga das árvores em cada grande área.

só foi criada em meados dos anos de 1980[2], ainda assim é possível ver como se deu o desenvolvimento das diferentes grandes áreas ao longo dos anos desde a década de 1960 (ver Figura 4.3). Assim, pode-se considerar o crescimento do número de árvores em cada grande área como resultado do desenvolvimento dos respectivos programas de pós-graduação e a queda no surgimento de novas árvores como a consolidação desses programas.

Analisando as grandes áreas em termos de suas linhagens, mais uma vez verifica-se bastante similaridade entre elas. Como pode ser observado nos gráficos da Figura 4.7, não há uma grande diferença em relação ao tamanho das linhagens em cada uma das grandes áreas, ou seja, independentemente da grande área são poucas as árvores que alcançaram mais de seis linhagens. Ainda assim, observando melhor esses gráficos, é possível ver que as grandes áreas possuem diferentes proporções para a quantidade de árvores em cada tamanho de linhagem. Vale notar que a grande área de Ciências

[2]http://lattes.cnpq.br/web/plataforma-lattes/historico

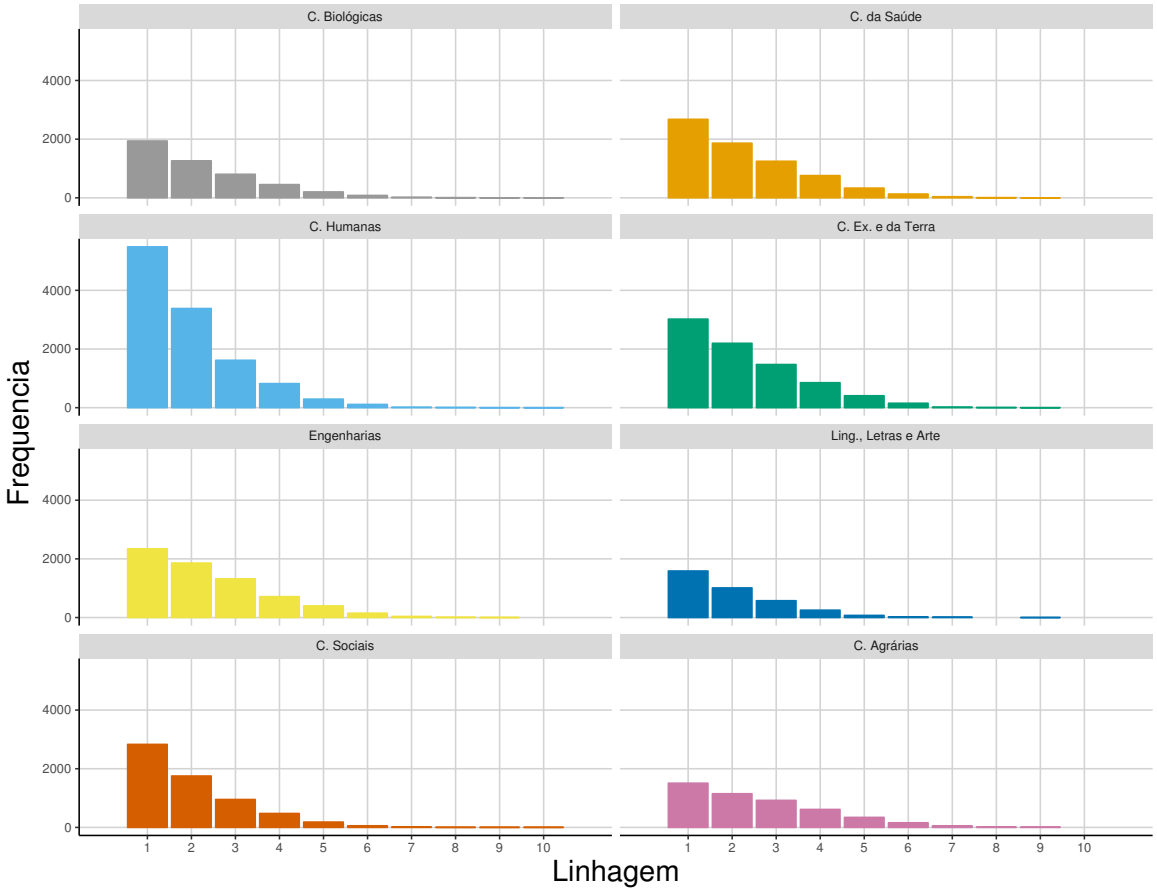

Figura 4.7. Distribuição da linhagem das árvores em cada uma das grandes áreas do conhecimento.

Agrárias é a que apresenta a melhor proporção neste caso. No caso das árvores com as maiores linhagens, os gráficos da Figura 4.7 mostram que elas pertencem às áreas de Ciências Biológicas, Ciências Humanas e Ciências Sociais Aplicadas que incluem algumas das árvores mais antigas (ver Figura 4.6).

Em relação à fecundidade das árvores em cada grande área, todas elas possuem uma grande concentração de árvores com fecundidade 1. As grandes áreas de Ciências Biológicas, Ciências Humanas e Ciências Exatas e da Terra são as que apresentam a maior concentração de árvores com fecundidade baixa. Considerando que as três são as únicas áreas que ultrapassam os 30% de árvores com fecundidade 1. Essa concentração ocorre, de modo geral, devido ao grande número de árvores com apenas um único descendente. Já a grande área de Ciências Agrárias é a única que possui uma maior concentração de árvores com valores de fecundidade mais altos, por volta de 10, o que significa que, em média cada pesquisador orientou outros 10 pesquisadores presentes na árvore. Finalmente, as grandes áreas de Ciências Humanas e Ciências Sociais Aplicadas

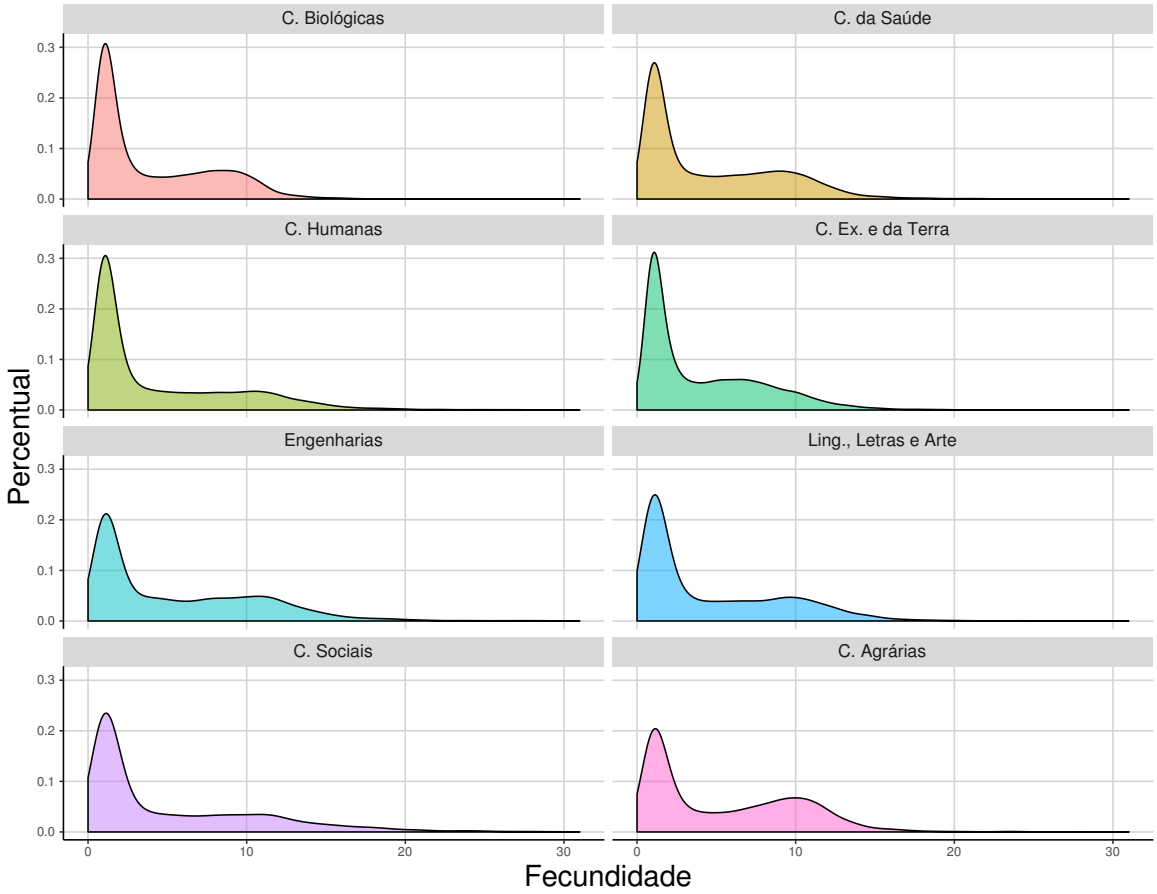

Figura 4.8. Distribuição da fecundidade das árvores em cada uma das grandes áreas do conhecimento.

são as que possuem as árvores com a menor concentração de taxas de fecundidade mais altas.

Utilizando a métrica densidade, procurou-se observar quais grandes áreas teriam as árvores mais densas em termos de orientações vindas de pesquisadores da mesma árvore, ou seja, árvores nas quais as orientações decorrem de pesquisadores que possuem um ancestral comum. A Figura 4.9 mostra a frequência das árvores para diferentes valores de densidade. As grandes áreas de Ciências Humanas e Ciências Sociais Aplicadas são as que concentram o maior número de árvores com baixa densidade, isto é, nessas grandes áreas as orientações envolvendo pesquisadores de uma mesma árvore tendem a ser mais raras, tornando essas árvores menos densas. Todas as demais grandes áreas possuem valores de densidade mais equilibrados.

Concluindo, uma hipótese para as diferenças entre algumas grandes áreas do conhecimento serem sutis para determinadas métricas é o fato de muitas árvores compartilharem diversos nodos. Assim, essas árvores possuem suas gerações divididas entre

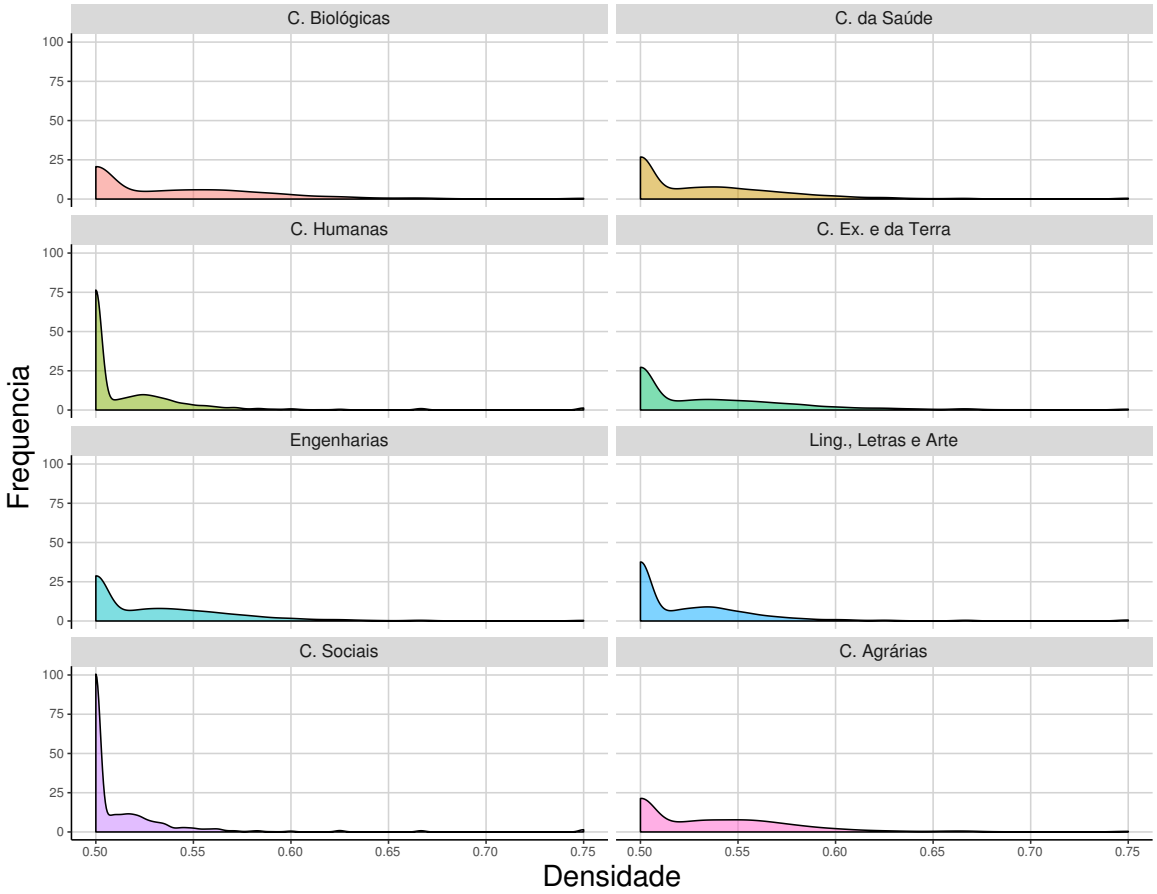

Figura 4.9. Distribuição da densidade de orientações nas árvores de cada grande área.

mais de uma grande área. O grafo da Figura 4.10 mostra como as grandes áreas do conhecimento relacionam-se entre si. Nesse grafo, um relacionamento entre duas grandes áreas acontece sempre que a grande área informada na tese do orientador é diferente da grande área informada na tese ou dissertação do orientando. Por exemplo, no caso de um orientador cuja tese foi classificada como sendo da grande área de Ciências Exatas e da Terra ter orientado um aluno cuja tese foi classificada como sendo da área de Ciências Sociais Aplicadas, essa relação é representada por uma aresta saindo da grande área de Ciências Exatas e da Terra e chegando à grande área de Ciências Sociais Aplicadas. Assim, nesse grafo, as grandes áreas que mais se relacionam entre si são Ciências da Saúde e Ciências Humanas com 5.015 relações, Ciências Biológicas e Ciências da Saúde com 4.643 relações, e Ciências Humanas e Ciências Sociais Aplicadas com 4.157 relações, mostrando que em geral esses relacionamentos se dão em razão da proximidade temática dos tópicos abordados.

De modo geral, pode-se observar que as grandes áreas do conhecimento possuem

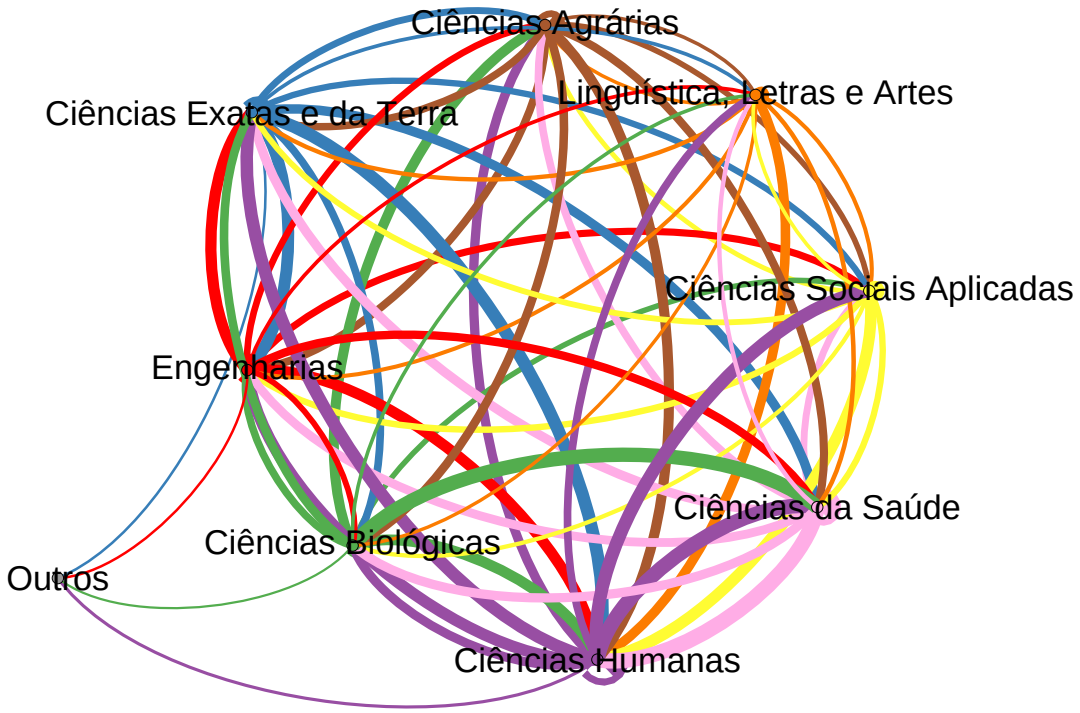

Figura 4.10. Relações interdisciplinares entre as grandes áreas do conhecimento (direção das arestas no sentido horário).

características evolucionárias semelhantes e que essas características são dinâmicas, mudando ao longo do tempo. Por exemplo, o surgimento de novos programas de pós-graduação bem como ações específicas de fomento refletem diretamente na formação das árvores de cada uma das grandes áreas. Mesmo assim, apesar de bastante diferentes entre si, as grandes áreas possuem um comportamento bastante similar em relação à formação de novos mestres e doutores. Isso porque, em termos de orientação, de modo geral todas seguem regras acadêmicas semelhantes, muitas delas definidas pela CAPES.

4.3 O Portal Science Tree

Um dos objetivos principais do projeto em que insere-se esta dissertação é possibilitar à comunidade científica brasileira acesso às árvores genealógicas acadêmicas construídas. Para isso, foi desenvolvido juntamente com esta dissertação o protótipo de um sistema que visa permitir a visualização dessas árvores e a navegação através delas. Esse

sistema está disponível por meio de um portal na WWW denominado Science Tree[3]. A Figura 4.11 mostra a página inicial do portal, a partir da qual, clicando-se na função Search e inserindo-se o nome de um pesquisador em uma caixa de busca, é possível encontrar a página que mostra o primeiro nível de sua árvore, a partir do qual pode-se navegar pelos demais níveis, caso existam.

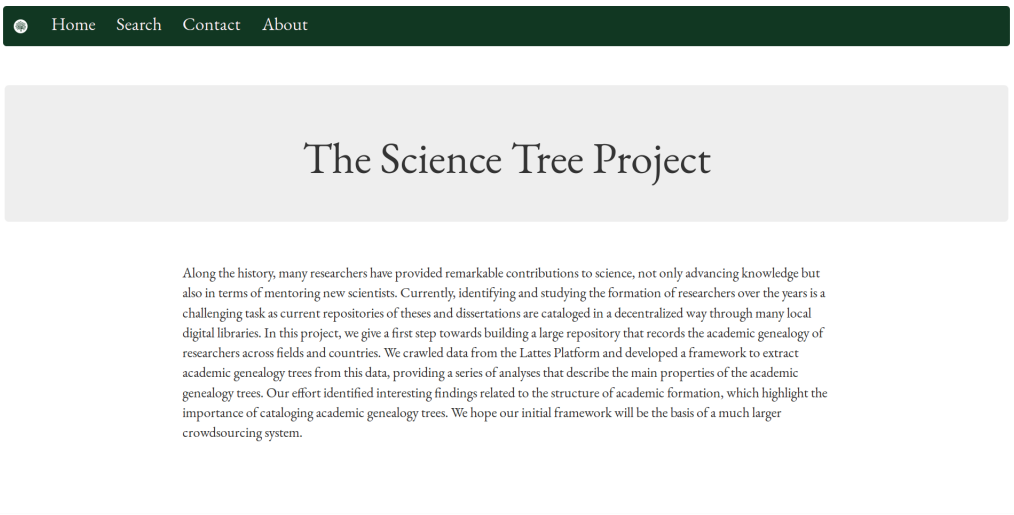

Figura 4.11. Página inicial do portal Science Tree.

Esse portal foi muito útil durante a construção das árvores, pois facilitou enormemente a inspeção delas. Com isso, foi possível realizar a depuração de erros introduzidos durante o processo de extração de dados dos currículos Lattes e que impediam a correta construção das árvores (por exemplo, erros nos nomes dos pesquisadores ou nas ligações estabelecidas entre eles). Sem o portal esse trabalho de depuração das árvores teria sido bastante complexo e oneroso. Outra função em que utilizou-se o portal foi para a validação das árvores construídas. Devido ao seu fácil acesso através da World Wide Web, vários pesquisadores puderam verificar as suas próprias árvores e indicar a presença de erros, não apenas na composição das árvores, como também nas informações adicionais disponíveis.

A Figura 4.12 mostra, como exemplo, o resultado de uma busca realizada no portal Science Tree usando-se o nome do professor Cesare Lattes, o mais renomado físico brasileiro cujo sobrenome hoje identifica a plataforma do CNPq, enquanto que a Figura 4.13 mostra a página com informações acadêmicas do professor José Palazzo Moreira de Oliveira da UFRGS. Vale ressaltar que a árvore retornada a partir da busca realizada com o nome do professor Cesare Lattes, que possui 386 descendentes, não foi construída diretamente a partir dos dados do seu currículo disponível na Plataforma

[3] `http://sciencetree.net`

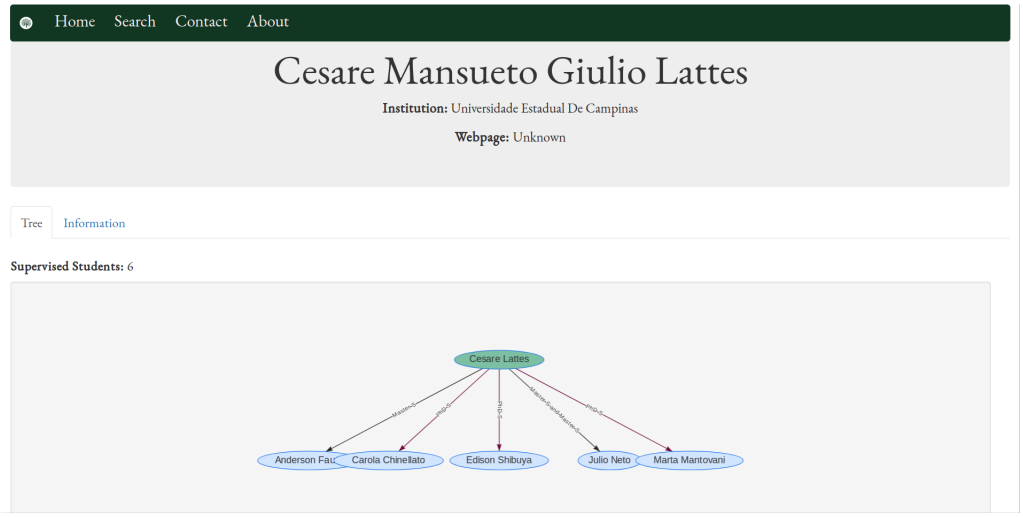

Figura 4.12. Página contendo o primeiro nível da árvore de um pesquisador.

Figura 4.13. Página com informações sobre a formação acadêmica de um pesquisador.

Lattes, pois o mesmo não fazia parte do conjunto de currículos inicialmente coletados, mas a partir dos dados dos currículos de seus orientandos disponíveis na plataforma.

Quando comparado a outros sítios que também possibilitam a visualização de árvores genealógicas acadêmicas, o portal Science Tree se mostra como um dos maiores e mais completos repositórios sobre a genealogia acadêmica brasileira com dados de mais de um milhão de pesquisadores. O portal Mathematical Genealogy Tree[4], por exemplo, possui dados de cerca de 221.000 pesquisadores da área de Matemática, enquanto que o portal Academic Tree[5] possui dados de pouco mais de 685.000 mil pesquisadores

[4]http://www.genealogy.ams.org/searh.php
[5]https://academictree.org

de diversas áreas do conhecimento. Assim, o portal Science Tree não só é uma fonte importante de consulta sobre a genealogia acadêmica brasileira, mas também serve de registro sobre como se desenvolveram as grandes áreas do conhecimento no Brasil, tornando-se assim uma importante ferramenta para análise da formação acadêmica dos pesquisadores brasileiros.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Existem diversos projetos que têm como objetivo construir as árvores genealógicas acadêmicas dos pesquisadores das mais diversas áreas do conhecimento. Em sua maioria, esses sistemas obtêm os seus dados por meio de uma estratégia de "crowdsourcing". Entretanto, muitos deles coletam esses dados de pesquisadores de uma única área do conhecimento ou têm o seu crescimento limitado pela estratégia de "crowdsourcing" adotada. Apesar dessas limitações, ultimamente têm surgido vários trabalhos que procuram analisar não só a produção científica de um grupo de pesquisadores, mas também as suas carreiras de modo geral, utilizando-se para isso de dados dos vários projetos que hoje procuram manter as árvores genealógicas acadêmicas de diversas áreas do conhecimento.

Assim, nesta dissertação foram utilizados dados extraídos dos currículos de todos os pesquisadores com título de doutor cadastrados na Plataforma Lattes para construir as suas árvores genealógicas acadêmicas. Além do trabalho realizado para tratamento dos dados extraídos dos currículos coletados, foi desenvolvido um algoritmo para processar, desambiguar e identificar as relações entre os pesquisadores informados nesses currículos, de modo a construir suas árvores genealógicas acadêmicas. Para compreender como se deu a formação dessas árvores, utilizou-se diversas métricas para caracterizar e analisar as suas estruturas individualmente, como também agrupadas pelas grandes áreas do conhecimento conforme definidas pelo CNPq.

Os resultados obtidos mostram que as árvores genealógicas acadêmicas dos pesquisadores brasileiros se desenvolveram bastante, principalmente a partir da década de 1960, quando foram criados no Brasil os primeiros programas de pós-graduação. Assim, desde essa década os programas brasileiros vêm se firmando cada vez mais como

formadores de recursos humanos qualificados, fundamentais para o desenvolvimento científico do país. Vale ressaltar, ainda, que mesmo os pesquisadores que constituem as raízes dessas árvores provêm, em sua maioria, de instituições nacionais, tendo muitos deles sido os responsáveis pela criação de importantes áreas de pesquisa no país.

A partir da análise das árvores agrupadas pelas grandes áreas do conhecimento, foi possível entender melhor as diferenças, ainda que sutis, entre as árvores de cada uma delas. Em particular, a grande área de Ciências Agrárias foi a que mais se destacou, apresentando números superiores às demais em praticamente todas as métricas consideradas.

Por meio das árvores construídas foi possível destacar o papel de grandes pesquisadores na formação acadêmica brasileira. Grande parte dos pesquisadores que constituem as raízes das árvores mais antigas são os pioneiros de suas respectivas áreas. Entre tantos, podemos destacar nomes como o de André Dreyfus, que além de ser um dos fundadores da USP, é considerado o pai da genética no Brasil, Annita Castilho, fundadora do curso de Psicologia da USP e um dos grandes nomes da Psicologia do país, e Celso D. Albuquerque Mello, pesquisador de renome na área de Direito, reconhecido pela autoria de diversos livros importantes da área.

Finalmente, foi apresentado o protótipo de um sistema disponível por meio de um portal na WWW que permite a visualização das árvores e a navegação através delas. Esse sistema permite não só uma maior interação com os resultados gerados por esta dissertação, mas também consultar e visualizar as árvores genealógicas acadêmicas construídas.

5.2 Trabalhos Futuros

Apesar das várias análises realizadas sobre as árvores construídas, tanto em termos gerais quanto agrupadas pelas grandes áreas do conhecimento, ainda existem inúmeros outros trabalhos que podem ser realizados a partir dos resultados desta dissertação. Dentre eles, pode-se mencionar um estudo sobre a evolução temporal das árvores e dos perfis dos pesquisadores mais influentes. Também seria importante analisar as árvores com base em outros aspectos como as instituições dos pesquisadores e as áreas de atuação indicadas em seus currículos, ou mesmo correlacioná-las com outras redes de colaboração acadêmica. Bem como a análise da profundidade e fecundidade das árvores geradas a partir de orientações

Um outro ponto que pode ser abordado é a expansão das árvores construídas a partir de dados coletados de outros repositórios de teses e dissertações, como a

Networked Digital Library of Theses and Dissertations (NDLTD). Essa expansão seria particularmente interessante nos casos daquelas árvores cujo nodo raiz corresponde a um pesquisador de uma instituição estrangeira. Também seria importante melhorar a visualização das árvores no portal Science Tree, tornando mais simples e direta a navegação entre as diferentes gerações de pesquisadores.

Finalmente, é importante ressaltar que os resultados desta dissertação constituem um primeiro passo para um objetivo maior que é entender como se deu o surgimento das comunidades científicas e, até mesmo, a criação das principais áreas do conhecimento. Isso permitiria uma melhor compreensão sobre a origem dessas áreas, e também sobre como se dá o processo de nascimento e morte das comunidades científicas. Além disso, a expansão das árvores com a utilização de dados de outros repositórios e iniciativas similares, como a Academic Tree, seria importante para uma melhor compreensão de como se deu a evolução da ciência e, consequentemente, da nossa própria sociedade.

Referências Bibliográficas

Ali, P. A. & Panther, W. (2008). Professional development and the role of mentorship. *Nursing Standard*, 22(42):35–39.

Alves, B. L.; Benevenuto, F. & Laender, A. H. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22nd International Conference on World Wide Web*, Companion Volume, pp. 649–656, New York, NY, USA. ACM.

Anderson, C. (2008). *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion Books, New York, NY, USA.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press, New York, NY, USA.

Bang-Jensen, J. & Gutin, G. Z. (2008). *Digraphs: Theory, Algorithms and Applications*. Springer-Verlag, London, UK.

Barabási, A. L.; Jeong, H.; Néda, Z.; Ravasz, E.; Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614.

Bollobas, B. (1998). *Modern Graph Theory*. Springer-Verlag, New York, NY, USA.

Bondy, J. & Murty, U. (1976). *Graph Theory with Applications*. North Holland, Amsterdam, Netherlands.

Canto, I. & Hannah, J. (2001). A Partnership of Equals? Academic Collaboration between the United Kingdom and Brazil. *Journal of Studies in International Education*, 5(1):26–41.

Chang, S. (2003). Academic genealogy of American physicists. *AAPPS Bulletin*, 13(6):6–41.

Chen, G.; Wang, X. & Li, X. (2015). *Fundamentals of Complex Networks: Models, Structures and Dynamics*. Wiley, Hoboken, New Jersey, USA.

Coleman, T. F. & Moré, J. J. (1983). Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. *SIAM Journal on Numerical Analysis*, 20(1):187–209.

Cota, R. G.; Ferreira, A. A.; Nascimento, C.; Gonçalves, M. A. & Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, 61(9):1853–1870.

Cunningham, S. J. (2001). The birth of a field: An analysis of the 1994-2000 ACM Digital Libraries Conferences. In *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, pp. 139–146, Sydney, Australia.

Damaceno, R.; Rossi, L. & Mena-Chalco, J. (2017). Identificação do Grafo de Genealogia Acadêmica de Pesquisadores: Uma Abordagem Baseada na Plataforma Lattes. In *Anais do 32o Simpósio Brasileiro de Bancos de Dados*, pp. 76–87, Uberlândia, MG, Brasil.

David, S. V. & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE*, 7(10):e46608.

Dawson, S.; Gašević, D.; Siemens, G. & Joksimovic, S. (2014). Current State and Future Trends: A Citation Network Analysis of the Learning Analytics Field. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pp. 231–240.

Delgado-Garcia, J. F.; Laender, A. H. F. & Meira, W. (2014). Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups. In *Proceedings of the 9th Latin American Web Congress*, pp. 77–81, Ouro Preto, MG, Brasil.

Demirkan, I. & Demirkan, S. (2012). Network characteristics and patenting in biotechnology, 1990-2006. *Journal of Management*, 38(6):1892–1927.

Dias, T. M. R. (2016). *Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes.* Tese de doutorado, Programa de Pós-Graduação em Modelagem Matemática e Computacional, CEFET-MG., Belo Horizonte, Brasil.

Dores, W.; Benevenuto, F. & Laender, A. H. (2016). Extracting Academic Genealogy Trees from the Networked Digital Library of Theses and Dissertations. In *Proceedings*

of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 163–166, Newark, New Jersey, USA.

Dores, W.; Soares, E.; Benevenuto, F. & Laender, A. H. (2017). Building the Brazilian Academic Genealogy Tree. In *Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries*, pp. 537–543, Thesssaloniki, Greece.

dos Santos, A. & de Azevedo, J. (2009). A pós-graduação no Brasil, a pesquisa em educação e os estudos sobre a política educacional: os contornos da constituição de um campo acadêmico. *Revista Brasileira de Educação*, 14(42):535.

Easley, D. & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Ferreira, A. A.; Gonçalves, M. A. & Laender, A. H. (2012). A Brief Survey of Automatic Methods for Author Name Disambiguation. *ACM SIGMOD Record*, 41(2):15–26.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Havel, V. (1955). A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 80:477–480.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6):1–4.

Jackson, A. (2007). A Labor of Love: The Mathematics Genealogy Project. *Notices of the AMS*, 54(8):1002–1003.

Johnson, D. S. (1984). The Genealogy of Theoretical Computer Science: A Preliminary Report. *SIGACT News*, 16(2):36–49.

Kumar, S. & Jan, J. M. (2013). On giant components in research collaboration networks: Case of engineering disciplines in Malaysia. *Malaysian Journal of Library & Information Science*, 18(2):65–78.

Laender, A. H. F.; de Lucena, C. J. P.; Maldonado, J. C.; de Souza e Silva, E. & Ziviani, N. (2008). Assessing the Research and Education Quality of the top Brazilian Computer Science Graduate Programs. *SIGCSE Bulletin*, 40(2):135–145.

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 54–62, Roanoke, Virginia, USA.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. 10(8):707--710.

Liu, X.; Bollen, J.; Nelson, M. L. & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480.

Malmgren, R. D.; Ottino, J. M. & Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, 465(7298):622--626.

Mehlhorn, K. & Sanders, P. (2008). *Algorithms and Data Structures: The Basic Toolbox*. Springer-Verlag, Berlin, Germany.

Mena-Chalco, J. P.; Digiampietri, L. A.; Lopes, F. M. & Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, 65(7):1424–1445.

Menezes, G. V.; Ziviani, N.; Laender, A. H. & Almeida, V. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 1041–1050, Madrid, Spain.

Newman, M. E. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical review E*, 64(1):016131.

Newman, M. E. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205.

Newman, M. E. J. (2001c). Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102.

Oliveira, J. W. A.; Laender, A. H. F. & Gonçalves, M. A. (2005). Remoção de Ambiguidades na Identificação de Autoria de Objetos Bibliográficos. In *Anais do 20º Simpósio Brasileiro de Bancos de Dados*, pp. 205–219, Uberlândia, MG, Brasil.

Park, C. (2005). New variant PhD: The changing nature of the doctorate in the UK. *Journal of Higher Education Policy and Management*, 27(2):189–207.

Perc, M. (2010). Growth and structure of slovenia's scientific collaboration network. *Journal of Informetrics*, 4(4):475–482.

Rossi, L. & Mena-Chalco, J. P. (2014). Caracterização de Árvores de genealogia acadêmica por meio de métricas em grafos. In *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*, pp. 1–12, Brasília, DF, Brazil.

Sarigöl, E.; Pfitzner, R.; Scholtes, I.; Garas, A. & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3(1):9.

Schwartzman, S. (2006). A universidade primeira do Brasil: entre intelligentsia, padrão internacional e inclusão social. *Estudos Avançados*, 20:161–189.

Scott, J. (2017). *Social Network Analysis*. SAGE Publications, Thousand Oaks, CA, USA.

Silva, T. H. P.; Laender, A. H. F.; Davis, C. A.; da Silva, A. P. C. & Moro, M. M. (2017). A profile analysis of the top Brazilian Computer Science graduate programs. *Scientometrics*, 113(1):237--255.

Smalheiser, N. R. & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, 43(1):1--43.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268--276.

Tuesta, E. F.; Delgado, K. V.; Mugnaini, R.; Digiampietri, L. A.; Mena-Chalco, J. P. & Pérez-Alcázar, J. J. (2015). Analysis of an Advisor–Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil. *PLOS ONE*, 10(5):e0129065.

Uddin, S.; Hossain, L.; Abbasi, A. & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. *Scientometrics*, 90(2):687–699.

Vargas, J. I. (1997). Alguns aspectos da política nacional de ciência e tecnologia. *Química Nova*, 20:7–14.

Webber, J. (2012). A Programmatic Introduction to Neo4J. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, pp. 217–218, Tucson, Arizona, USA.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, 68(3):191–195.

Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: A co-authorship network analysis. *J. Am. Soc. Inf. Sci. Technol.*, 60(10):2107–2118.