CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

GENEALOGIA ACADÊMICA BRASILEIRA: UMA CARACTERIZAÇÃO DA RELAÇÃO ORIENTADOR-ORIENTADO NO BRASIL

TALES HENRIQUE JOSÉ MOREIRA

Orientador: Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Agosto de 2018

TALES HENRIQUE JOSÉ MOREIRA

GENEALOGIA ACADÊMICA BRASILEIRA: UMA CARACTERIZAÇÃO DA RELAÇÃO ORIENTADOR-ORIENTADO NO BRASIL

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Sistemas Inteligentes

Orientador: Prof. Dr. Gray Farias Moita
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Agosto de 2018

Moreira, Tales Henrique José
M838g Genealogia acadêmica brasileira: uma caracterização da relação orientador-orientando no Brasil / Tales Henrique José Moreira. – 2018.
70 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Gray Farias Moita.
Coorientador: Thiago Magela Rodrigues Dias.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Matemáticos – Genealogia – Teses. 2. Curriculum vitae – Análise – Teses. 3. Orientadores educacionais – Teses. 4. Professores e orientandos – Teses. 5. Mineração de dados (Computação) – Teses. I. Moita, Gray Farias. II. Dias, Thiago Magela Rodrigues. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 510.92

Elaboração da ficha catalográfica pela Biblioteca-Campus II / CEFET-MG

Esta folha deverá ser substituída pela cópia digitalizada da folha de aprovação fornecida pelo Programa de Pós-graduação.

Dedico esta dissertação à minha
família e amigos.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”.
(Albert Einstein)

AGRADECIMENTOS

Ao meu orientador, Professor Gray Farias Moita, pelo empenho e dedicação para que este trabalho fosse desenvolvido da melhor forma possível e por sua ajuda e compreensão em momentos difíceis.

Ao co-orientador, Professor Thiago Magela Rodrigues Dias, também pelo empenho e dedicação, além de todo o incentivo dado desde o início. Sem a colaboração e incentivo dado pelo Thiago, certamente este trabalho não existiria, já que foi o grande incentivador para que eu ingressasse no meio acadêmico, sendo o responsável inclusive por me apresentar o Professor Gray, a quem tive o privilégio de conhecer e receber suas orientações.

Ambos foram essenciais e decisivos para que este trabalho chegasse no estágio final. Além de todas as orientações que recebi durante o mestrado, fica o meu agradecimento especial pelas revisões finais que tanto me ajudaram.

Aos membros da banca (internos e externo), por aceitarem o convite e efetivamente deixarem sua contribuição em meu trabalho.

Agradeço ainda ao CEFET-MG, por toda a sua estrutura, corpo docente e servidores, que proporcionaram um ambiente propício para o bom desenvolvimento do trabalho. Um agradecimento especial ao PPGMMC (Programa de Pós-Graduação em Modelagem Matemática e Computacional) e todo o seu corpo docente, com quem muito aprendi.

Por fim, um agradecimento aos meus pais, irmã e amigos, que incentivaram em momentos de dificuldade e me ajudaram a superar todas as dificuldades encontradas durante o mestrado. Em especial a minha irmã, que já se encontra em sua jornada também na computação.

RESUMO

A relação orientador-orientado, que se caracteriza como uma forma de propagação do conhecimento, como resultado pode proporcionar o surgimento de novos orientadores. Especialmente em programas de pós-graduação, onde orientados tendem a permanecer no meio acadêmico, repassando o conhecimento adquirido de seus orientadores durante todo o seu período acadêmico a seus futuros orientados. Portanto, caracterizar como ocorre o processo de orientação resultante dessa relação é mais um importante meio de análise da colaboração cientifica nas diversas áreas do conhecimento. Neste trabalho, são utilizados os dados de orientações extraídos dos currículos cadastrados na Plataforma Lattes com o objetivo de obter-se, além de uma visão geral do processo de orientação no Brasil, uma análise das maiores árvores genealógicas encontradas e da floresta contendo as 100 maiores árvores, identificando dessa forma como o conhecimento científico Brasileiro tem se propagado.

PALAVRAS-CHAVE: Genealogia Acadêmica, Plataforma Lattes, Orientação, Mineração de dados.

ABSTRACT

The advisor-advised relationship is characterized as a form of propagation of the knowledge and, as a result, it can provide the emergence of new advisors. It is particularly true in postgraduate programs, where they tend to remain in the academic environment, passing on the knowledge gained from their advisors throughout their academic period to their future advised. Therefore, characterizing how the orientation process resulting from this relationship occurs is another important means of analyzing scientific collaboration in the various areas of knowledge. In this project, the data of supervisions extracted from the curricula in Lattes Platform are used with the objective of obtaining besides an overview of the orientation process in Brazil, an analysis of the largest genealogical trees found and the forest containing the 100 largest trees, thus, identifying how knowledge is been spread.

KEYWORDS: Academic Genealogy, Lattes Platform, Supervision, Datamining.

LISTA DE FIGURAS

FIGURA 1: REDE DE COAUTORIA EXEMPLIFICANDO OS CONCEITOS BÁSICOS APRESENTADOS ACIMA 22
FIGURA 2: CRESCIMENTO DA PLATAFORMA LATTES ENTRE OS MESES DE MAIO DE 2014 E JUNHO DE 2016 31
FIGURA 3: VISÃO GERAL DO LATTESDATAXPLORER 31
FIGURA 4: PROCESSO DE EXTRAÇÃO E SELEÇÃO DE DADOS 33
FIGURA 5: VISÃO GERAL DO LATTESDATAXPLORER EXPANDIDO 34
FIGURA 6: PROCESSO DE SUMARIZAÇÃO DOS DADOS 35
FIGURA 7: VÍNCULOS DA PLATAFORMA LATTES 36
FIGURA 8: PROCESSO DE DESAMBIGUAÇÃO DE NOMES DA PLATAFORMA LATTES 37
FIGURA 9: EXEMPLO DE DICIONÁRIO UTILIZADO NO PROCESSO DE DESAMBIGUAÇÃO 38
FIGURA 10: APLICAÇÃO DO DESAMBIGUADOR EM UM REGISTRO DE ORIENTAÇÃO 39
FIGURA 11: APLICAÇÃO DO DESAMBIGUADOR EM UM REGISTRO DE ORIENTAÇÃO COM USO DO RESOLVEDOR DE HOMÔNIMOS 39
FIGURA 12: VINCULAÇÕES ANTES (A) E DEPOIS (B) DO PROCESSO DE DESAMBIGUAÇÃO. 41
FIGURA 13: DISTRIBUIÇÃO DAS ORIENTAÇÕES POR ESTADO 44
FIGURA 14: DISTRIBUIÇÃO DAS ORIENTAÇÕES POR NATUREZA DE TODOS OS INDIVÍDUOS 45
FIGURA 15: GRÁFICO DE ORIENTAÇÕES POR ANO CONSIDERANDO TODOS OS INDIVÍDUOS ATÉ O ANO DE 2016 45
FIGURA 16: DISTRIBUIÇÃO DAS ORIENTAÇÕES POR FORMAÇÕES 47
FIGURA 17: GRÁFICO DE ORIENTAÇÕES POR GRANDES ÁREAS 48
FIGURA 18: GRÁFICO DE ORIENTAÇÕES POR GRANDES ÁREAS APÓS USO DE INFORMAÇÕES DO ORIENTADOR 49
FIGURA 19: GRÁFICO DE ORIENTAÇÕES DISTRIBUÍDOS POR NATUREZA DA ORIENTAÇÃO E GRANDES ÁREAS 49
FIGURA 20: GRÁFICO DE ORIENTAÇÕES POR ANO 50
FIGURA 21: DISTRIBUIÇÃO DAS ORIENTAÇÕES POR GRANDE ÁREA 51
FIGURA 22: ÁRVORE GENEALÓGICA DO ORIENTADOR COM A MAIOR QUANTIDADE DE ORIENTAÇÕES 53
FIGURA 23: ÁRVORE GENEALÓGICA DO ORIENTADOR COM A MAIOR QUANTIDADE DE GERAÇÕES E DESCENDENTES 54
FIGURA 24: ÁRVORE GENEALÓGICA DO ORIENTADOR COM A MAIOR QUANTIDADE DE DESCENDENTES 55
FIGURA 25: DISTRIBUIÇÃO DOS NÓS RAÍZES POR GRANDES ÁREAS DE ATUAÇÃO 56
FIGURA 26: DISTRIBUIÇÃO POR ESTADO DE ORIGEM DOS INDIVÍDUOS DA FLORESTA 57
FIGURA 27: RELAÇÃO ENTRE O TAMANHO DAS ÁRVORES E QUANTIDADE DE GERAÇÕES 59
FIGURA 28: FLORESTA COM AS 100 MAIORES ÁRVORES IDENTIFICADAS 60

LISTA DE TABELAS

QUADRO 1: COMPARAÇÃO ENTRE LATTESDATAXPLORER E SCRIPTLATTES. 32
TABELA 1: RESULTADO DA SUMARIZAÇÃO DE TODOS OS CURRÍCULOS ANALISADOS. 36
TABELA 2: RESULTADO DO DESAMBIGUADOR 40
TABELA 3: RANKING COM OS DEZ PAÍSES COM MAIOR QUANTIDADE DE INDIVÍDUOS ESTRANGEIROS ORIENTADORES. 44
TABELA 4: TOTAL DE ORIENTAÇÕES E TAXA DE CRESCIMENTO POR ANO DE TODOS OS INDIVÍDUOS 46
TABELA 5: INDIVÍDUO COM MAIOR QUANTIDADE DE ORIENTAÇÕES PARA CADA NÍVEL DE CAPACITAÇÃO..... 47
TABELA 6: INDIVÍDUO COM MAIOR QUANTIDADE DE ORIENTAÇÕES EM CADA GRANDE ÁREA. 50
TABELA 7: RANKING DOS INDIVÍDUOS COM MAIOR QUANTIDADE DE ORIENTAÇÕES DE PÓS-GRADUAÇÃO. 52
TABELA 8: INDIVÍDUOS COM MAIORES REDES DO CONJUNTO DE DADOS. 52
TABELA 9: ORIENTADORES ESTRANGEIROS COM MAIOR QUANTIDADE DE ORIENTAÇÕES NA FLORESTA..... 58
TABELA 10: MAIORES RAÍZES DA FLORESTA 58
TABELA 11: MÉTRICAS DA FLORESTA CARACTERIZADA..... 60

LISTA DE ABREVIATURAS E SIGLAS

CA	Comitê de Assessoramento
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSV	Comma Separated Values
C&T	Ciência e Tecnologia
HTML	HyperText Markup Language
PG	Pós-Graduação
PPGs	Programas de Pós-Graduação
URL	*Uniform Resource Locator*
XML	*eXtensible Markup Language*
XPath	*XML Path Language*
Web	*World Wide Web*
WOS	*Web of Science*

SUMÁRIO

1 INTRODUÇÃO 14

1.1 DESCRIÇÃO DO PROBLEMA 16
1.2 OBJETIVOS E CONTRIBUIÇÕES 17

2 FUNDAMENTAÇÃO CONCEITUAL 19

2.1 BIBLIOMETRIA 19
2.2 ANÁLISE DE REDES SOCIAIS 21
2.2.1 Conceitos Básicos 21

3 TRABALHOS RELACIONADOS 25

3.1 TRABALHOS ENVOLVENDO DADOS DA PLATAFORMA LATTES 25
3.2 TRABALHOS SOBRE GENEALOGIA ACADÊMICA 27

4 COLETA E TRATAMENTO DE DADOS 30

4.1 MOTIVAÇÕES 30
4.2 LATTESDATAXPLORER 31
4.2.1 Sumarização dos Dados 34
4.2.2 Desambiguação de Nomes 36
4.2.3 Dados Estratificados 41

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS 43

5.1 ESTATÍSTICAS GERAIS 43
5.2 ESTATÍSTICAS POR GRANDES ÁREAS 48
5.3 ÁRVORES GENEALÓGICAS 51
5.4 FLORESTAS DE ÁRVORES DE GENEALOGIA ACADÊMICA 56

6 CONSIDERAÇÕES FINAIS 62

6.1 PUBLICAÇÕES 63
6.2 COMPLEMENTAÇÃO DA PESQUISA 65

REFERÊNCIAS 66

Capítulo 1

Introdução

Há algum tempo, a *World Wide Web (Web)* vem se consolidando como uma importante fonte de informações sobre diversas áreas do conhecimento, permitindo a concepção de enormes repositórios de dados. Tais dados, por exemplo, podem ter sua origem em ferramentas, como, bibliotecas digitais, redes sociais acadêmicas e sítios para registro individual de produção acadêmica em que, além de acessar seu conteúdo, seus usuários podem ser responsáveis por registrar sua própria produção acadêmica (DIAS, 2016).

Nas últimas décadas do século XX, em razão da construção, manutenção e informatização de repositórios de dados científicos, tornou-se realidade a produção de indicadores bibliométricos de maior representatividade (MUGNAINI; JANNUZZI; QUONIAM, 2004). Neste mesmo período, diversos outros estudos têm procurado compreender a evolução da ciência e, principalmente, como ocorre a colaboração científica entre indivíduos. Assim sendo, recentemente tem aumentado o surgimento de técnicas, como, por exemplo, métricas baseadas em análise de redes sociais, análises bibliométricas e cientométricas com o objetivo de auxiliar na análise destes dados (DIAS, 2016).

Percebe-se neste contexto que a evolução das pesquisas científicas tem forte influência no processo de formação, em que pesquisadores orientadores inserem novos pesquisadores que contribuem para que novos estudos sejam realizados em diversas áreas do conhecimento. Destaca-se que grande parte dos trabalhos realizados com orientação no Brasil são decorrentes de Programas de Pós-Graduação (PPGs), impulsionados pela necessidade de capacitação e titulação de pesquisadores.

Os termos bibliometria e cientometria são importantes para a compreensão e análise das atividades de orientação e de produção acadêmica (ARAÚJO, 2006). Para Nicholas e Ritchie (1978), a bibliometria é utilizada para realizar uma avaliação objetiva da produção através de métodos quantitativos. Neste caso, a bibliometria aplica-se a livros, documentos, revistas, artigos, autores e usuários. A partir desses conceitos, é possível utilizar dados de orientações acadêmicas com o objetivo de analisar como ocorre o processo de orientação em uma instituição, em um Programa de Pós-Graduação ou mesmo em determinadas áreas

do conhecimento, além de possibilitar que a dispersão do conhecimento científico por meio das orientações seja observada. Logo, os dados sobre o processo de orientação acadêmica se caracterizam como um novo e importante objeto de estudo para compreender o processo de formação através da genealogia acadêmica, já que possibilita compreender e analisar a propagação do conhecimento.

Para Sugimoto (2014), genealogia acadêmica é um estudo quantitativo da herança intelectual através da relação orientador-orientado. Para Ferreira, Furtado e Silveira (2009), o binômio (ou díade) orientador-orientado é indubitavelmente a base dos PPGs, o que determina o crescimento e a expansão dos cursos de Pós-Graduação (PG) e a demanda de orientação. Além disso, os autores ressaltam que o aluno de PG é um pesquisador em potencial em estágio avançado de desenvolvimento, ou seja, a caminho da autonomia científica, mas ainda dependente de um professor, o que justifica as atividades de orientação como efetivamente necessárias. Adicionalmente aos dados básicos que caracterizam os vínculos sobre o processo de orientação, informações como, por exemplo, publicações e área de atuação podem ser analisadas com o intuito de se compreender o perfil de orientação de um orientador em particular ou de uma área do conhecimento.

Tendo em vista as possibilidades de visualização e entendimento do histórico de orientação e, consequentemente, a difusão do conhecimento, realizar a modelagem e a caracterização de árvores genealógicas acadêmicas surge como uma alternativa interessante para a análise de como a ciência brasileira tem se propagado utilizando-se para tanto, dados de orientações. Para isso, as árvores genealógicas acadêmicas podem também ser caracterizadas, facilitando o entendimento do processo de formação com a análise destas árvores.

De modo geral, a caracterização de uma árvore se dá por um grafo composto por um conjunto de nós (vértices) e conexões (arestas) entre os nós, podendo ser direcionadas ou não (SZWARCFITER, 1984). Esta é uma abordagem que teve origem na sociologia, na psicologia social e na antropologia (SCOTT, 2012). A análise realizada a partir de grafos, na qual as redes produzidas neste trabalho (árvores e florestas de árvores genealógicas) podem ser inseridas, interessa a diversos pesquisadores, independentemente de sua área de atuação.

A tentativa de compreender o impacto da análise de redes no meio social no qual está inserido originou diversas metodologias de análise baseadas nas relações entre os indivíduos estruturadas em forma de redes. Tais redes são sistemas baseados nas ciências sociais, representando sujeitos sociais (indivíduos, grupos, organizações, etc.) através de nós, conectados por algum tipo de relação (MARTELETO; SILVA, 2004).

Árvores genealógicas podem ser definidas como uma estrutura que representa todo ou parte do histórico dos antepassados de um indivíduo. Trata-se de uma representação gráfica que apresenta, de forma hierárquica, os antepassados, podendo ou não ter informações complementares que visam permitir um melhor entendimento do histórico de um indivíduo. Diante disto, as árvores genealógicas acadêmicas são caracterizadas como árvores que representam hierarquicamente o histórico de um orientador e todos os seus orientados. Logo, caracterizando-se uma árvore genealógica acadêmica, é possível observar como o seu conhecimento foi repassado ao longo do tempo.

Com isso, destaca-se que dados sobre orientações, bem como sobre características individuais de cada um dos indivíduos a serem analisados, são de extrema importância para a análise e compreensão da relação orientador-orientado. Neste trabalho, os dados utilizados são os currículos que compõem a Plataforma Lattes[1] sob a coordenação do CNPq[2] (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Para tanto, os currículos cadastrados na Plataforma Lattes dos indivíduos que têm atuado como orientadores são analisados com o objetivo de se extrair informações de interesse e, dessa forma, obter uma visão geral do processo de orientação deste conjunto e, ainda, produzir árvores genealógicas acadêmicas, onde é possível verificar como o conhecimento tem se propagado nas diversas áreas do conhecimento.

Diversos trabalhos vêm sendo desenvolvidos com foco em genealogia acadêmica e utilizando como fontes de dados outras plataformas *web*. O *Mathematics Genealogy Project*[3] e o projeto *Neurotree*[4] são importantes fontes de dados genealógicos das áreas de Matemática e Neurociência, respectivamente, além da comunidade científica dos Físicos[5]e do PhDTree[6], para os acadêmicos titulados com doutorado (ROSSI; MENA-CHALCO, 2015). Tais projetos, disponibilizados livremente para acesso, são extremamente importantes para a documentação histórica de seus membros, porém são especializados em determinadas áreas do conhecimento ou repositórios de dados internacionais, não representando, assim, a realidade de todas as áreas do conhecimento ou de uma comunidade, como no caso de um determinado país, por exemplo.

1.1 Descrição do Problema

Para Mugnaini e Digiampietri (2014), o levantamento da produção científica de um país permite estudar aspectos que podem ser considerados como resultados mensuráveis através do sistema de ciência, tecnologia e inovação. Tais aspectos também podem ser estudados pelo levantamento de orientações acadêmicas, podendo mensurá-las da mesma maneira.

Dados de orientações acadêmicas se tornaram uma importante fonte de estudo quanto ao processo de orientação. Porém, obter dados para a realização de tais estudos não é uma tarefa trivial devido à existência de diferentes formatos empregados e diferentes repositórios, como os projetos *Matematics Genealogy Project* e *Neurotree*, mencionados acima. Além disso, grande parte dos repositórios existentes com dados sobre orientações são especializados em determinadas áreas, sendo complexa uma possível integração entre estes, visto que cada repositório cria e mantém seus dados com padrões estruturais próprios sem nenhuma padronização entre si. Diante disto, muitas vezes se faz necessária

[1] Plataforma Lattes: http://lattes.cnpq.br

[2] CNPq: http://www.cnpq.br

[3] Mathematics Genealogy Project: http://genealogy.math.ndsu.nodak.edu

[4] Neurotree: http: //neurotree.org/neurotree

[5] Comunidade Científica dos Físicos: http://academictree.org/physics

[6] PhDTree: http://phdtree.org

a utilização de técnicas específicas para integração e conciliação de dados originados em diferentes repositórios (CHRISTEN, 2012).

Além de especializados em determinadas áreas, os principais repositórios contendo dados de orientações são, em sua maioria, internacionais, possuindo pouca ou nenhuma informação de orientações brasileiras. Logo, fica evidente a dificuldade em analisar dados de orientações de modo abrangente sobre todos os registros de orientação de um grande conjunto de indivíduos como, por exemplo, todos os orientadores de um país. Diante disso, optou-se pelo uso dos currículos cadastrados na Plataforma Lattes, que possui um grande conjunto de registros de orientações, com aproximadamente 8 milhões de orientações concluídas, independentemente da área ou natureza da orientação. Diferentemente da maioria dos repositórios internacionais citados anteriormente, os próprios usuários da plataforma são encarregados de cadastrar suas orientações.

Apesar de possuir um conjunto de dados significativo e interessante para o desenvolvimento deste trabalho, os currículos da Plataforma Lattes possuem algumas limitações, como a não vinculação implícita de todos os orientados com identificadores da plataforma. Ou seja, em muitos dos casos, como a análise aqui descrita é realizada verificando a seção de orientações concluídas de um determinado indivíduo, este indivíduo (que é o orientador) já tem seu identificador definido. No entanto, nem sempre o identificador de seus orientados está implicitamente vinculado aos nomes deles, na seção de orientações concluídas. Tal fato está relacionado a diversos desafios para caracterização de grandes árvores, como, por exemplo, a desambiguação de nomes, o que torna o processo de construção das árvores uma tarefa não trivial. Além disso, erros de digitação durante a inserção das informações na Plataforma Lattes podem reduzir as chances de sucesso na desambiguação a partir do nome ou citação.

Diante disso, este trabalho apresenta um estudo sobre os dados de orientação acadêmica registrados nos currículos cadastrados na Plataforma Lattes, propondo uma estratégia para identificar orientações que não estão implícitas nos currículos, permitindo caracterizar de forma inédita grandes árvores genealógicas, possibilitando, dessa maneira apresentar um estudo sobre o histórico de todos os indivíduos com orientações concluídas em cursos de pós-graduação. Os resultados apresentados são inéditos, tendo em vista a quantidade de registros analisados, devido a adoção da Plataforma Lattes como fonte de dados e ao processo de desambiguação utilizado.

1.2 Objetivos e Contribuições

Buscando um maior entendimento sobre como tem se desenvolvido o processo de orientação acadêmica brasileira nas diversas áreas do conhecimento, esta pesquisa tem como objetivo geral caracterizar as orientações em programas de pós-graduação, a partir de análises bibliométricas e baseadas em análises de redes sociais realizadas sobre dados curriculares disponíveis na Plataforma Lattes.

Especificamente, para alcançar tal objetivo, são abordados os seguintes desafios:

- Obtenção de todos os currículos da Plataforma Lattes;

- Sumarização dos dados de orientação obtidos através de todos os currículos da Plataforma Lattes;
- Desambiguação de nomes de indivíduos que não tenham vínculos explícitos nos registros de orientação;
- Construção de árvores de genealogia acadêmicas a partir dos dados sumarizados;
- Adoção de métricas de análise de redes sociais para extração de conhecimento sobre as árvores identificadas.

Neste contexto, este trabalho procura explorar dados dos currículos cadastrados na Plataforma Lattes de modo a analisar como ocorre a orientação acadêmica no Brasil nas diversas áreas do conhecimento. Para isso, diferentes levantamentos e análises de dados são realizados com o intuito de avaliar diversos indicadores que são considerados. Assim sendo, destacam-se como principais contribuições da pesquisa:

- A construção de um conjunto de dados sobre orientações a partir da análise de todo o conjunto de currículos cadastrados na Plataforma Lattes;
- Uma caracterização geral dos dados de orientação contidos no repositório construído;
- Uma análise do perfil dos orientadores brasileiros, considerando somente os indivíduos com orientações em cursos de pós-graduação;
- Identificação de indivíduos e grandes áreas com maior representatividade em orientações;
- Geração de árvores e floresta de árvores genealógicas de orientadores.

Nesse cenário, abordagens considerando a adoção de métricas de análises de redes sociais e de indicadores bibliométricos podem ser de grande valia para a compreensão acerca da orientação acadêmica brasileira nas diversas áreas do conhecimento. Assim, este estudo explorará todo o conjunto de currículos cadastrados na Plataforma Lattes, obtendo toda a genealogia acadêmica brasileira, .

Capítulo 2

Fundamentação Conceitual

Neste capítulo é exibido um estudo bibliográfico referente ao tema da pesquisa. Na Seção 2.1, são apresentados conceitos de bibliometria, bem como as principais métricas adotadas para análise bibliométrica. Além disso, também são apresentados conceitos relacionados à análise de redes sociais (Seção 2.2), tendo em vista que algumas análises realizadas neste trabalho fazem uso de métricas aplicadas a redes sociais. Por fim, são apresentados conceitos relacionados a desambiguação de nomes para produção das árvores genealógicas através dos currículos cadastrados na Plataforma Lattes (Seção 2.3).

2.1 Bibliometria

Com o aumento da disponibilidade de conhecimento produzido, são necessários recursos para controlar e monitorar este conhecimento. Por meio de diversas técnicas, a Bibliometria procura estudar os aspectos quantitativos da ciência e da produção científica, podendo tal estudo ser feito por meio de uma disciplina ou de uma atividade que envolve aspectos sociais, econômicos e políticos. Ambas fornecem um ferramental que auxilia no processo de mapear o campo científico e de extrair informações úteis para compreender como a produção científica tem sido realizada (HAYASHI, 2013).

O termo bibliometria é definido por Araújo (2006) como uma ferramenta quantitativa que pode ser utilizada para analisar e medir índices de produção científica. O termo também é definido por Pritchard (1969) como sendo todos os estudos que se destinam a qualificar os processos de comunicação. Diante disto, a bibliometria pode ser apontada como a área responsável por avaliar a ciência, baseando-se nos dados de pesquisadores. Sendo assim, esta pode ser utilizada para mensurar produção científica de indivíduos, formação acadêmica, instituições, e, no caso deste estudo, registros de orientações.

A bibliometria possui três leis básicas, segundo Guedes e Borschiver (2005):

- A Lei de Bradford normalmente é aplicada em conjuntos de periódicos, com o objetivo de identificar a extensão na qual artigos científicos de assunto específico estavam presentes em periódicos destinados a outros assuntos. Desta forma, era estudado a distribuição dos artigos de acordo com variáveis de proximidade ou de afastamento, onde Bradford realizou uma série de estudos que culminaram, em 1934, com a lei da dispersão. A lei de Bradford permite estimar o grau de relevância de periódicos em uma dada área do conhecimento e que os periódicos de maior qualidade ou relevância para aquela área produzem um maior número de artigos sobre um dado assunto, formando um núcleo (ARAÚJO, 2006);

- A Lei de Lotka, também conhecida como Lei do Quadrado Inverso por sua premissa, é determinada pela descoberta de que grande parte dos trabalhos científicos são produzidos por um pequeno número de pesquisadores (que publicam muito). No entanto, em termos de produção, um grande número de pesquisadores (que pouco publicam) se igualam ao pequeno número de pesquisadores mais produtivos, surgindo, por causa disto, a Lei do Quadrado Inverso $1/n^2$ (SANTOS, 2010);

- A Lei de Zipf descreve a relação entre palavras de um determinado texto e a ordem destas. Em um de seus estudos, Zipf encontrou correlação entre o número de palavras distintas e a frequência de seu uso, concluindo que existe uma regularidade na seleção e uso destas, em que um pequeno número de palavras é usado com maior frequência. Diante disto, formulou-se o princípio do menor esforço: o assunto do documento é indicado pelas palavras com maior frequência (ARAÚJO, 2006).

Goffman e Newill (1964) aplicaram os conceitos obtidos através das três leis básicas agregando a outros estudos e conceitos e estudos relacionados à bibliometria. Dentre estes estudos, se destaca a teoria epidêmica de transmissão de ideias, onde a propagação de ideias dentro de uma determinada comunidade é semelhante à transmissão de doenças infecciosas, onde as ideias de um trabalho são referentes ao material infectante e seus leitores são os indivíduos suscetíveis à doença.

Um exemplo de estudo relacionado à bibliometria com foco em orientações é o índice-h genealógico expandido. Ele é uma adaptação do índice-h, medida de desempenho que classifica pesquisadores em função do número de publicações e suas citações correspondentes. Diferentemente do original, a adaptação tem como objetivo identificar o impacto ou influência de orientadores, levando em consideração a amplitude dos relacionamentos de orientação (quantidade de orientados diretos) e expandindo a abrangência da métrica a todas as possíveis gerações identificadas (ROSSI; MENA-CHALCO, 2015).

Diante do apresentado, é possível observar que estudos quantitativos e estatísticos envolvendo diversas técnicas são possíveis através de métricas de bibliometria. Tais métricas podem ser aplicadas em distintas áreas do conhecimento, possibilitando medir a produção científica e o processo de orientações acadêmicas através de uma análise quantitativa.

2.2 Análise de redes sociais

De acordo com Digiampietri et al. (2014), a maioria das tarefas complexas não são realizadas de forma individual, mas sim por grupos de indivíduos. Diante disto, a análise de redes sociais tem sido considerada tão importante (ou mais) que a simples análise de indicadores individuais.

O tipo de uma rede social é definido pelo contexto em que esta está inserida. Geralmente, o tipo de uma rede está diretamente ligado ao tipo de relacionamento envolvido nessa rede. Assim, de acordo com os relacionamentos, pode-se obter diferentes tipos de redes. Um exemplo são nós que representam indivíduos interligados através de vínculos familiares, coautorias ou relacionamento afetivo, alterando apenas o tipo de relacionamento entre eles (MENEZES, 2012). Ao aplicar este conceito ao contexto desta pesquisa, os vínculos mais comuns são de coautoria e orientação, caracterizando assim dois tipos de redes sociais: redes de coautoria, obtidas através do relacionamento entre indivíduos em publicações, e redes de orientações, obtidas através da orientação entre orientador e orientado.

Para Lopes (2012), uma análise que considera todos os elementos da rede social é a mais importante. Todas as informações da rede, como informações de todos os nós e informações de todos os relacionamentos, são utilizadas a fim de explicar a estrutura social da rede como um todo. Desta maneira, a rede se torna mais complexa, porém, ao mesmo tempo, mais completa.

No contexto desta pesquisa, são produzidas e analisadas as árvores e a floresta de orientação, que, por sua natureza, podem ser consideradas como uma rede. A análise deste tipo de rede tem despertado o interesse da comunidade devido à sua capacidade de representação de problemas complexos, subsidiando o estudo das propriedades e comportamentos dos elementos que as compõem (DIAS, 2016).

A análise de redes científicas permite identificar diversas informações relevantes, como relacionamento entre pesquisadores e instituições, área de atuação, localização geográfica, além do tipo de relacionamento entre os nós.

2.2.1 Conceitos Básicos

Para um maior entendimento dos conceitos básicos, a Figura 1 refere-se a uma rede social fictícia com a capacidade de exemplificar os diversos conceitos baseados em Wasserman e Faust (1994) aplicados na análise de redes. Esta pode ser considerada uma árvore genealógica, onde atores orientam trabalhos de seus alunos (vínculos), mesmo às vezes pertencendo a áreas distintas (subgrupos). Logo, o conjunto de todos os orientadores e orientados formam o grupo e a coleção de vínculos representam a árvore genealógica.

Atores são entidades sociais que participam de uma rede. Estes possuem a capacidade de agir e se vincular com outros atores da rede. Podem representar um indivíduo, uma organização ou até mesmo um coletivo social, como, por exemplo, uma classe escolar. As redes são denominadas monomodal quando todos os atores pertencem a um mesmo tipo.

No entanto, também existem redes com atores de tipos distintos. No exemplo da Figura 1, os autores são indivíduos que orientam em diversos níveis de formação, formando vínculos na rede.

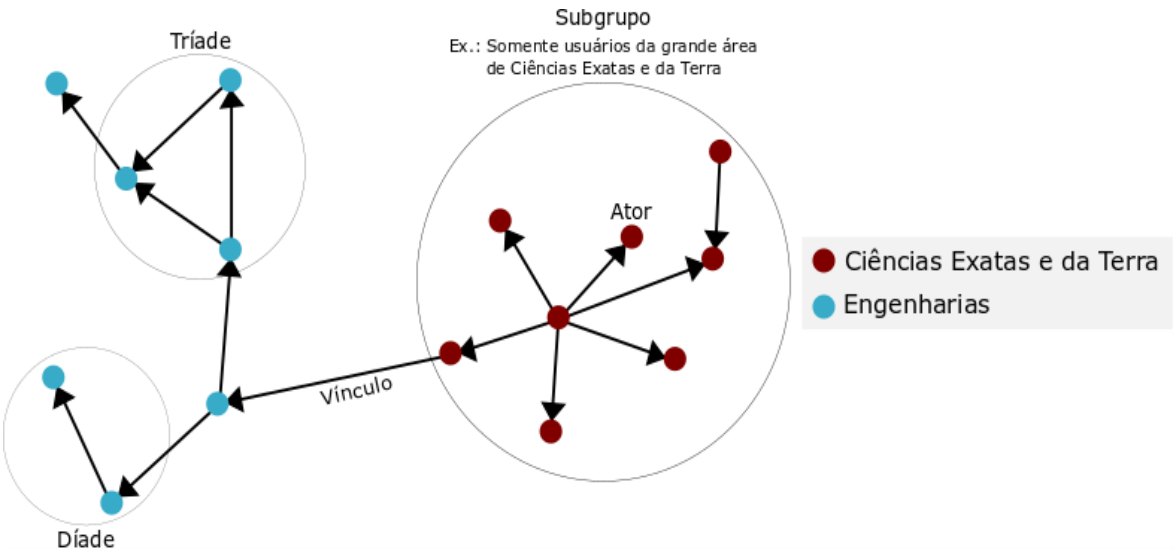


Figura 1: Rede de coautoria exemplificando os conceitos básicos apresentados acima

Vínculos referem-se à conexão entre dois atores. Estes vínculos podem ser de diversos tipos, como, por exemplo, rede de amizade, estradas em uma rede de cidades ou rotas aéreas em uma rede de aeroportos. Tais vínculos podem determinar o tipo da rede, e podem, opcionalmente, possuir valor e serem direcionados.

Outros conceitos importantes observados na Figura 1 são as Díades e Tríades. Uma Díade é um par de atores com um possível vínculo entre eles. Este vínculo pertence ao par de atores, não pertencendo a um ator individualmente. Geralmente, é a unidade de estudo em análises estatística em redes sociais. Já uma Tríade é um subconjunto formado por três atores com os possíveis vínculos entre si. Essencialmente, em uma tríade, há duas díades. Em redes sociais, isso pode ser um fator importante, pois se o Ator 1 possui vínculo com o Ator 2 e este com o Ator 3, existe um possível caminho, passando pelo Ator 2, permitindo que o Ator 1 tenha contato com o Ator 3.

Os atores e seus vínculos podem formar Grupos e Subgrupos. Um Grupo é conjunto de atores e seus vínculos, definido através de um limite. Consiste em um conjunto finito de atores que podem estabelecer vínculos. Já um Subgrupo é um subconjunto de atores e todos os vínculos entre eles. Se tratando de redes sociais complexas e extensas, pode ser de extrema importância analisá-las em forma de subgrupos específicos, delimitados dentro do grupo.

Uma Relação é um conjunto de vínculos de um determinado tipo. Enquanto um vínculo se dá somente entre dois atores, a relação é definida para todo o conjunto de vínculos. Ou seja, uma relação é uma coleção de vínculos de um determinado tipo.

Todos os conceitos citados acima são utilizados para representar uma Rede Social, definida como um conjunto de atores e seus vínculos. É toda uma estrutura social onde atores

podem se relacionar através de vínculos de diversos tipos, por exemplo, orientação e coautoria.

2.3 Desambiguação de Nomes

O uso de repositórios de dados digitais acadêmicos, como por exemplo, a Plataforma Lattes, são importantes fontes para a descoberta de conhecimento, proporcionando diversas análises e estudos. Porém, diversos problemas podem ser encontrados em repositórios desse tipo, tais como, erros de digitação, falta de padronização e ambiguidade (Lee et al. 2007).

Dentre os problemas citados, a ambiguidade afeta diretamente o estudo aqui apresentado, já que tem grande impacto nas árvores genealógicas produzidas. O problema ocorre devido à falta de um vínculo explicito entre indivíduos. Quando esse vinculo é inexistente, é possível produzi-lo através dos nomes dos indivíduos, porém, estes podem ser ambíguos, ou seja, quando um indivíduo possui variações em seu nome em diferentes artigos, como por exemplo, através do uso de citações ou abreviações, ou diferentes indivíduos possuem o mesmo nome o que faz com que sejam confundidos (McKay et al. 2010).

Para resolver este problema, se faz necessário o uso de algoritmos de desambiguação. A seguir são apresentados os algoritmos mais conhecidos (Ferreira et al. 2012):

- HHC (*Heuristic-based Hierarchical Clustering*): método não supervisionado baseado em heurísticas computacionais;

- SAND (*Self-training Author Name Disambiguator*): método de auto-treinamento para desambiguação de nomes de citações em 3 etapas.

- INDi (*Incremental Name Disambiguation*): tem como vantagens sua simplicidade e o fato de desambiguar apenas registros não identificados em seus vínculos. Ou seja, vínculos realizados através de nomes, serão desambiguados e vínculos realizados através de um identificador único são mantidos (Carvalho et al. 2011).

A Plataforma Lattes quando concebida, não implementava nenhuma forma de vínculos entre indivíduos, como por exemplo, de orientação e coautoria. Com o passar do tempo, o recurso para informar os vínculos entre indivíduos foi implementado, porém, dados anteriores permanecem sem vínculos. Além disso, a informação do vínculo não é obrigatória, visto que nem sempre o indivíduo a ser vinculado está cadastrado na Plataforma Lattes.

Diante disso, um modelo incremental se torna interessante, mantendo todos os vínculos informados nos próprios currículos e desambiguando os demais. Este trabalho apresenta um método que possui as mesmas vantagens do modelo INDi, além de desempenho e resultados satisfatórios, especialmente no caso da Plataforma Lattes. Diferentemente de outras bases que disponibilizam apenas as informações contidas na referência do trabalho, a Plataforma Lattes oferece uma grande variedade de informações, como instituição de

ensino, área de atuação, formação, entre outras, o que pode ser usado para distinguir autores e facilitar o processo de desambiguação, dispensando o uso de algoritmos mais complexos.

Capítulo 3

Trabalhos Relacionados

Este capítulo enumera uma série de trabalhos relacionados a este estudo. Inicialmente, trabalhos que possuem como principal fonte dados os currículos cadastrados na Plataforma Lattes são apresentados (Seção 3.1). Em seguida, são apresentados trabalhos com foco em Genealogia Acadêmica que fazem uso de dados provenientes tanto da Plataforma Lattes quanto de outras fontes de dados (Seção 3.2).

3.1 Trabalhos Envolvendo Dados da Plataforma Lattes

Diversos trabalhos utilizam a Plataforma Lattes como fonte de dados para diversas análises. Isso ocorre não só devido a sua grande quantidade de currículos cadastrados, mas também pela variedade de dados contidos nos mesmos, como por exemplo, formação acadêmica, idiomas, projetos, produção técnica e bibliográfica. Diante disso, esta seção apresenta trabalhos que utilizam dados da Plataforma Lattes para os mais diversos tipos de análises.

Lane (2010), em artigo publicado na revista *Nature*, descreve que medir e avaliar o desempenho acadêmico já é uma realidade. São medições que vão desde um simples ranqueamento até as que influenciam no financiamento da pesquisa nas universidades. Mesmo com toda a importância, sistemas de medições existentes são limitados, sendo descritos diversos problemas provenientes no uso das métricas atuais. A autora apresenta uma gama de esforços no sentido de construir infraestruturas confiáveis, que, apesar de úteis, são trabalhosas de manter. Porém, um bom exemplo de boas práticas citado pela autora é a experiência Brasileira com a Plataforma Lattes, descrevendo diversos esforços que foram realizados que a tornaram um dos sistemas de dados acadêmicos mais limpos que existem, fornecendo dados de qualidade.

Mugnaini, Leite e Leta (2012) ressaltam que apesar de não possuir indexação e catalogação de periódicos como ocorre em diversas bases de indexação, a Plataforma Lattes é uma fonte inesgotável de informações sobre a ciência brasileira de maneira geral. É destacado ainda pelos autores que, apesar da grande quantidade de informações presente na plataforma, é observado uma baixa frequência de estudos cientométricos. Isso é um reflexo das limitações e da dificuldade imposta pela plataforma para recuperação e

extração das informações, estabelecendo-se como obstáculos para tais estudos. Apesar disso, a grande quantidade de informações pessoais, acadêmicas e diversos tipos de produção, acessada livremente, deve ser um estímulo para uso deste repositório por pesquisadores brasileiros da área. Destaca-se ainda, o fato de a Plataforma Lattes reunir em um único repositório, informações de toda a produção científica brasileira, permitindo a análise que, por vezes, só seriam possíveis através de repositórios internacionais. Além disso, diversas informações deixariam de ser analisadas, já que muitas delas somente estão presentes na Plataforma Lattes.

Para Ferraz et al. (2014), até o momento, não existe uma base curricular tão completa quanto a Plataforma Lattes, tornando-a única. Apesar de já existirem bases semelhantes em outros países, como, por exemplo, Portugal e países latino-americanos, no Brasil ela é fortemente utilizada, já que o cadastro é obrigatório tanto para concessão de fomento quanto para oficialização de vínculo profissional em cursos de nível superior. Diferentemente da Plataforma Lattes, onde são registrados todos os dados curriculares do indivíduo, a maior parte dos demais repositórios existentes disponibilizam livremente apenas dados referenciais e informações secundárias.

Mena-Chalco, Digiampietri e Cesar-Jr (2012) apresentam dois importantes motivos que reforçam a importância da Plataforma Lattes, sendo eles: ela se tornou um padrão nacional e vêm sendo utilizada na avaliação das atividades dos pesquisadores brasileiros e que pesquisadores das mais diversas áreas estão cadastrados na Plataforma.

Diante da grande relevância conquistada pela Plataforma Lattes, diversos trabalhos exploraram seus dados curriculares a fim de obter informações relevantes para compreender o estado atual da pesquisa no Brasil. Trabalhando com grupos de indivíduos, por área de atuação, atividade profissional, produção bibliográfica e técnica ou orientações, o objetivo quase sempre é traçar o perfil cientométrico deste grupo de modo quantitativo e/ou qualitativo, como pode ser visto em Belém et al. (2018), Ramos e Razera (2017) e Pupim et al. (2017).

A Plataforma Lattes passou a ser utilizada em diversos trabalhos para analisar pequenos conjuntos de pesquisadores, como por exemplo, no trabalho de Alves, Yanasse e Soma (2011), em que é analisado o perfil dos bolsistas de produtividade em pesquisa das áreas de engenharia de produção e transportes e enfoque na subárea de pesquisa operacional. Sua análise consiste em extrair dados de um conjunto de 186 pesquisadores e os caracteriza com o intuito de apresentar o perfil destes pesquisadores. Interessante observar que os dados utilizados neste trabalho, em sua grande maioria, são encontrados apenas na Plataforma Lattes, destacando ainda mais sua relevância.

Uma grande diferença entre estes trabalhos citados em relação ao desta dissertação é justamente o fato de que a grande maioria trabalha com pequenos grupos de indivíduos, seja pela dificuldade de obtenção da base completa, ou mesmo pelo tratamento e processamento de todos os currículos cadastrados na Plataforma Lattes. Os trabalhos de Dias (2016) e Dias et al. (2013) exploram bem a questão da extração de dados, expondo as dificuldades enfrentadas durante o processo de extração e tratamento dos currículos.

Santos (2015) utiliza os dados dos currículos cadastrados na Plataforma Lattes para comparar produções dos conjuntos dos pesquisadores bolsistas e não bolsistas de produtividade através de análises estatísticas. O objetivo é contribuir na avaliação da ciência através de métodos quantitativos aplicados na produção científica, além do índice-h como indicador de qualidade. Além de utilizar a Plataforma Lattes como fonte de dados de dados, o estudo de Santos (2015) utiliza o arcabouço para coleta e tratamento dos dados desenvolvido por Dias et al. (2013) descrito no capítulo 4, também utilizado e adaptado para este trabalho.

Porém, apesar de todas as vantagens apresentadas pela Plataforma Lattes, obter todo o conjunto de dados não é uma tarefa trivial, apesar do arcabouço desenvolvido por Dias et al. (2013) simplificar o processo. A Plataforma Lattes vem sofrendo constantes modificações, e somente a pouco tempo passou a disponibilizar uma lista com todos os identificadores de currículos cadastrados. Mesmo assim, a tarefa de extração ainda é complexa, visto que para o download de cada currículo, é necessário seguir várias etapas até a aquisição.

3.2 Trabalhos Sobre Genealogia Acadêmica

Nesta seção são apresentados trabalhos com foco em genealogia acadêmica, tendo como fonte de dados diversas bases, como banco de teses e dissertações, base de genealogia ou mesmo a Plataforma Lattes. Diante disso, é apresentado resultados existentes de análises da relação orientador-orientado e as diferenças em relação ao presente trabalho.

Antes mesmo da Plataforma Lattes obter destaque como fonte de dados para análises, já se destacavam alguns trabalhos relacionados a genealogia acadêmica. Em Leite Filho e Martins (2006), são verificadas as influências da relação orientador-orientado no processo de produção de teses e dissertações dos PPGs em Contabilidade da cidade de São Paulo. Os autores citam como justificativas para o estudo a importância de se analisar aspectos que teriam ligação com a construção do conhecimento, especificamente em se tratando da área de Contabilidade, e a tentativa de sinalizar a importância da temática orientação.

Diversas outras bases também têm sido utilizadas, seja pela forma de acesso ao conjunto de dados, pela quantidade de informações e existente ou pela presença de vínculos entre todos os indivíduos, possibilitando produzir toda a árvore genealógica da base.

Um exemplo é o trabalho de Dores e Laender (2016), que optaram por utilizar em seu trabalho a base de dados NDLTD[7] (*Networked Digital Library of Theses and Dissertations*) para fornecer uma serie de análises com base em estruturas de árvores genealógicas geradas. Apesar de ser uma base relevante, a mesma possui algumas desvantagens em relação aos currículos da Plataforma Lattes, sendo elas: menor quantidade de registros (já que esta é uma base de teses e dissertações, descartando os demais níveis de capacitação, além de depender das instituições submeterem seus dados, diferentemente da Plataforma Lattes, em que cada indivíduo é responsável por atualizar seus dados) e problema de

[7] http://www.ndltd.org/

desambiguação (sendo que o mesmo também ocorre com os currículos cadastrados na Plataforma Lattes, porém, em menor escala, já que na Plataforma Lattes existe a possibilidade de criação de vínculos entre indivíduos).

No trabalho de Rossi e Mena-Chalco (2014), denominado “Aos ombros de gigantes...”, os autores tentam caracterizar um seleto grupo de doutores em matemática titulados no Brasil. Neste trabalho foi utilizado os registros disponíveis no *Mathematics Genealogy Project* (MGP). Os autores ressaltam a importância da orientação acadêmica para a ampliação de comunidades científicas, contribuindo diretamente no crescimento dos indivíduos e seus respectivos grupos. Ainda, segundo os autores, descrever a comunidade de matemáticos com formação em instituições brasileiras é importante para a documentação da história e análise da trajetória da formação, relevância e influência de uma seleta área acadêmica do Brasil.

Apesar de existirem alternativas que facilitam todo o processo de preparação dos dados para análise, a Plataforma Lattes segue como uma importante fonte de dados devido a sua grande quantidade de dados. Diversos outros trabalhos utilizam os currículos cadastrados na Plataforma Lattes como fonte de dados.

Tuesta et al. (2012) por exemplo, apresenta uma análise temporal da relação orientador-orientado, com um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação, extraindo os dados de análise dos currículos cadastrados na Plataforma Lattes. No trabalho são analisadas as principais características do grupo e as relações de coautoria. Diante disso, os autores identificam que a duração média do período de colaboração é superior a 3 anos após a data da primeira publicação, concluindo que a média de tempo de doutoramento tem diminuído, possivelmente devido às facilidades atuais, trazidas principalmente pela internet.

Elias, Floeter-Winter e Mena-Chalco (2016) utilizou dados da Plataforma Lattes para identificar grafos genealógicos de pesquisadores respeitados no Brasil em uma determinada área do conhecimento. Neste trabalho, os dados foram coletados de forma semi-manual. O resultado do trabalho pode possuir uma maior precisão, já que os pesquisadores em estudo têm relevância no âmbito nacional, tendendo a ter os seus currículos mais atualizados e melhor preenchidos. O processo manual/semi-manual também pode ajudar a corrigir dados imprecisos nos currículos.

Delgado et al. (2015) usou os dados da Plataforma Lattes para analisar as principais características de um conjunto de pesquisadores docentes dos cursos de pós-graduação em Ciências da Computação no Brasil durante o triênio de 2007-2009 e suas relações de coautoria. Neste caso, os currículos foram obtidos através do ScriptLattes proposto por Mena-Chalco e Junior (2009). O objetivo do trabalho é analisar os pares formados por orientador-orientado, possibilitando analisar estatisticamente a relação entre eles. Além de estatísticas gerais da relação orientador-orientado, como, por exemplo, primeira publicação de orientados e número de publicações em conjuntas, foram apresentadas as redes de orientações acadêmicas e de coautoria entre orientador e orientado.

Trabalhos com base em genealogia acadêmica podem ter como objetivo propor técnicas para mensurar quantitativamente e/ou qualitativamente as orientações realizadas e,

consequentemente, as árvores genealógicas. Rossi e Mena-Chalco (2015) propõe um índice-h genealógico expandido com base na medida índice-h conhecida na Bibliometria/Cientometria. Esta derivação tem como objetivo caracterizar grafos de orientações permitindo o estudo em função do seu desempenho no papel de formador de recursos humanos.

Um dos primeiros trabalhos encontrados que tratam a genealogia acadêmica com base nos dados da Plataforma Lattes é proposto por Miyahara (2011). Nele, são construídas árvores genealógicas de pesquisadores, considerando para isso as relações de orientações. Para o autor, as árvores genealógicas podem indicar todo o histórico de um determinado pesquisador. A ferramenta proposta utiliza os currículos cadastrados na Plataforma Lattes para a caracterização das árvores. Entre todos os trabalhos citados acima, apenas este apresenta um método para desambiguar os dados de orientações dos currículos cadastrados na Plataforma Lattes. Este é o que mais se aproxima do estudo aqui apresentado, mas com alguns pontos a serem observados. São eles:

- O trabalho já se encontra defasado, tendo sido apresentado no ano de 2011, sendo que desta data até a data atual, a Plataforma Lattes obteve um acréscimo de mais de 2.000.000 de currículos, além do acréscimo na quantidade de orientações.
- A técnica usada para desambiguação é diferente, já que o autor tenta encontrar similaridade entre os nomes com uso de um *array*, fazendo com que a complexidade do algoritmo seja maior. Diferentemente, o estudo aqui apresentado faz uso de um dicionário (tabela *hash*), possibilitando um acesso direto ao elemento, além de outras melhorias que serão detalhadas no próximo capítulo.
- O autor apresenta poucos resultados, produzindo e exibindo apenas pequenas redes.

Logo, utilizar dados curriculares da Plataforma Lattes para análise do processo de orientação, caracterizando para isso árvores e floresta de genealogia, possibilita obter uma visão ampla sobre como tem evoluído o processo de orientação no Brasil, principalmente nos níveis mais altos de capacitação.

A grande diferença entre este trabalho e todos os citados acima é a quantidade de dados analisada (toda a plataforma e em alguns casos, registros de orientações em programas de pós-graduação). Diversos trabalhos focam em uma área de atuação ou um indivíduo como raiz. Este estudo pretende analisar todo o conjunto de currículos, em que uma visão inédita do processo de orientação brasileiro é apresentada, detalhando como ocorre tal processo, inclusive por grandes áreas do conhecimento.

Capítulo 4

Coleta e tratamento de dados

Neste capítulo, é descrito o processo de coleta e tratamento dos dados, englobando todo o processamento necessário para a análise da genealogia acadêmica brasileira. São apresentadas algumas motivações que tornam a Plataforma Lattes um importante repositório para a análise de dados científicos, inclusive sobre a genealogia acadêmica, e que serviram de motivação para este trabalho (Seção 4.1). Após, é apresentada a arquitetura geral do LattesDataXplorer (DIAS, 2016), bem como seu processo de coleta dos dados (Seção 4.2). São ainda detalhadas suas principais funcionalidades, como a sumarização dos dados (Subseção 4.2.1), a desambiguação de nomes dos indivíduos (Subseção 4.2.2) e, por fim, as árvores caracterizadas, bem como, estatísticas e métricas calculadas (Subseção 4.2.3).

4.1 Motivações

A escolha da Plataforma Lattes para a extração como fontes de dados, está relacionada ao fato de que a Plataforma Lattes é a única quando se trata da integração de dados científicos de currículos e de instituições da área de C&T, registrando os dados acadêmicos e as produções científicas dos pesquisadores e instituições, permitindo que a atualização dos dados seja realizada pelos próprios pesquisadores.

Os currículos agregam dados de pesquisadores e de suas atividades acadêmicas, tais como publicações, orientações e atividades profissionais, entre outras. Tais dados se tornaram um padrão nacional de avaliação individual dos pesquisadores, utilizados para investigar e compreender as atividades de diversos grupos de pesquisa, visto que esta é uma plataforma com informações valiosas a respeito da atividade acadêmica brasileira. (DIGIAMPIETRI et al., 2012).

Atualmente a Plataforma Lattes conta com aproximadamente 5.200.000 currículos de docentes, estudantes e pesquisadores, entre outros, com crescimento, conforme Figura 2 com dados coletados por Dias (2016) entre maio de 2014 e junho de 2016.

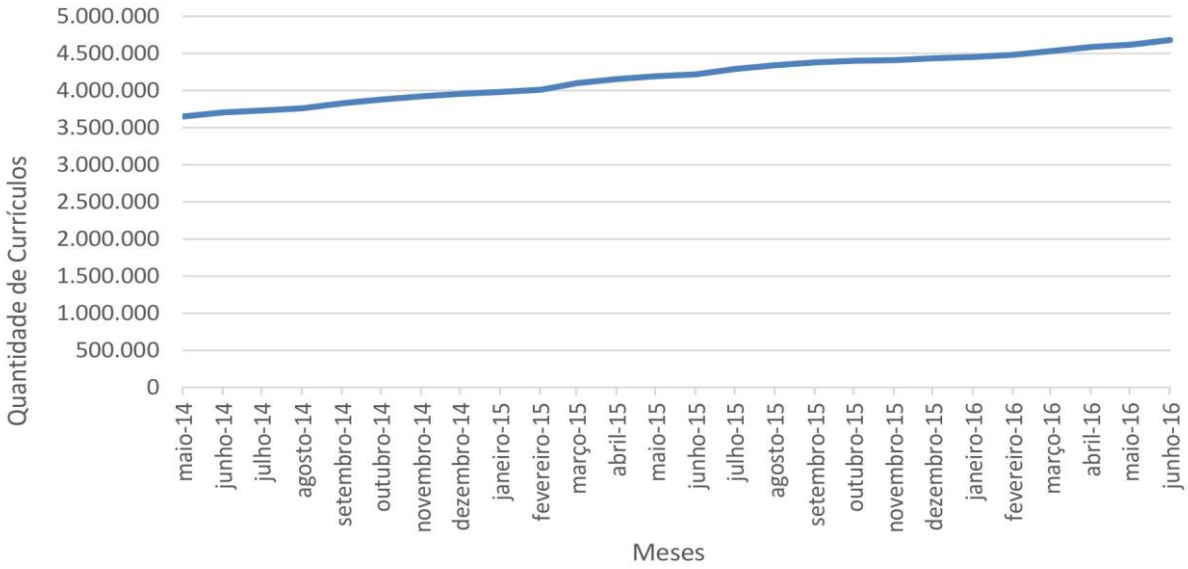


Figura 2: Crescimento da Plataforma Lattes entre os meses de maio de 2014 e junho de 2016

Fonte: Dias (2016)

Tais currículos podem ser acessados no formato HTML (*HyperText Markup Language*) através da página de pesquisa da Plataforma ou diretamente através de um link com o identificador único de cada indivíduo. Além disso é possível obter uma versão do currículo em formato XML (*eXtensible Markup Language*).

4.2 LattesDataXplorer

De posse dos currículos, é possível aplicar a eles diversas transformações e processá-los, a fim de se obter resultados importantes. Para a coleta dos currículos que compõem a Plataforma Lattes utilizados neste trabalho, um arcabouço denominado LattesDataXplorer (DIAS, 2016) foi utilizado (Figura 3).

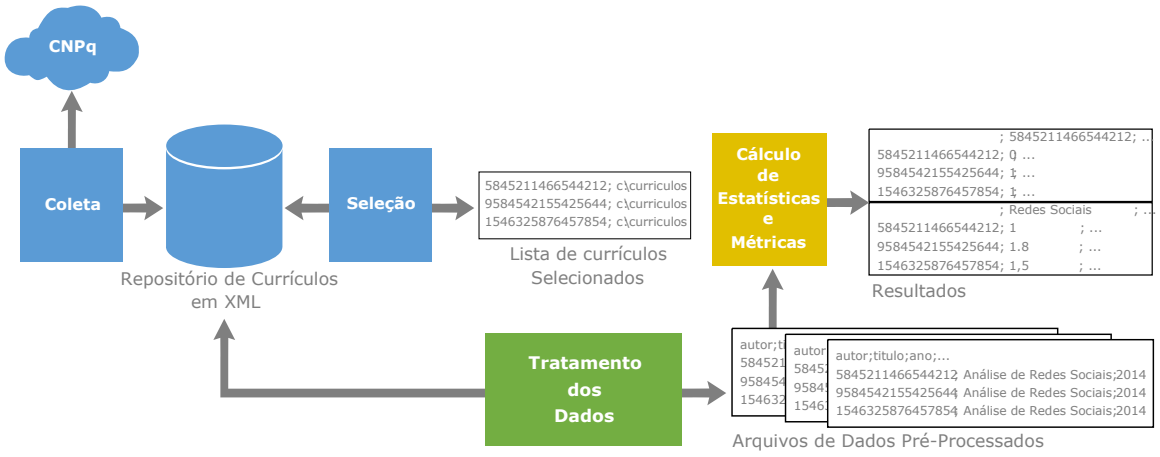


Figura 3: Visão geral do LattesDataXplorer

Fonte: (DIAS, 2016)

O LattesDataXplorer é responsável por englobar todo o conjunto de técnicas e métodos para a coleta, tratamento e análise dos dados utilizados nesta dissertação. Ele é composto por diversos módulos, responsáveis por todo o processo de coleta e tratamento dos dados (DIAS, 2016). Além de obter os currículos, o arcabouço produz diversos arquivos com resultados de métricas, estatísticas e seleção de currículos.

O arcabouço LattesDataXplorer tem como principal alternativa o ScriptLattes, desenvolvido por Mena-Chalco e Junior (2009), responsável também por obter e processar os dados da Plataforma Lattes. A Quadro 1 apresenta uma comparação entre as duas ferramentas endossando a escolha do LattesDataXplorer para o processo de obtenção dos currículos.

Quadro 1: Comparação entre LattesDataXplorer e ScriptLattes.

Descrição	LattesDataXplorer	ScriptLattes
Obtenção dos Identificadores de todos os currículos para extração	Sim	Não
Resolução de *Captcha* para acesso aos currículos	Sim	Não. É necessário baixar todos os currículos em formato HTML antes de usa-lo
Parser dos currículos em formato HTML	Sim	Sim
Download dos currículos em formato XML	Sim	Não
Identificação de Colaboração Científica	Sim (Método Linear)	Sim (Método Quadrático utilizando a distância de Levenshtein)
Produção de relatórios	Não	Sim

Fontes: (Dias et al. (2013) e ScriptLattes, 2018)

Na Figura 4, é apresentado o módulo de coleta e seleção dos dados. O processo de extração de todos os dados curriculares da Plataforma Lattes é dividido em três componentes, necessários para a obtenção dos currículos:

- Download dos códigos de acesso (um identificador secundário utilizado para acessar o currículo), obtidos através da Plataforma Lattes, possibilitando o acesso a todos os currículos através da Plataforma *Web* do CNPq;
- Extração de um identificador de acesso necessário para download do currículo em XML, e;
- Por último, a extração do currículo em formato XML, sendo este componente responsável pelo armazenamento local do currículo.

Todo este processo é baseado em *crawlers* que realizam uma varredura em páginas a procura dos códigos de acesso. Por fim, fazendo uso de multiprocessamento, o download dos currículos é realizado.

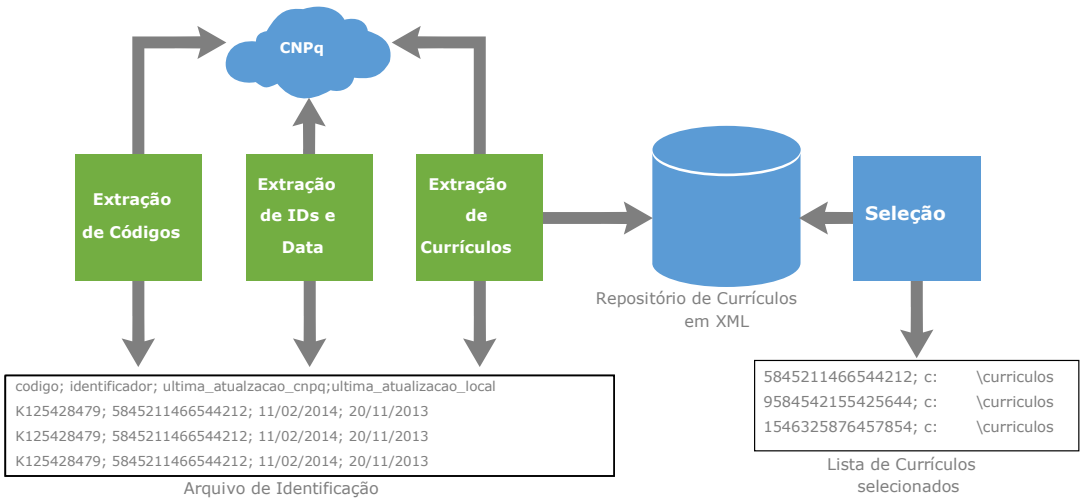


Figura 4: Processo de extração e seleção de dados

Fonte: (DIAS, 2016)

A primeira etapa, consiste na extração dos códigos de acesso através da listagem de todos os currículos da Plataforma Lattes (Módulo Extração de Códigos da Figura 4). De posse dos códigos, os mesmos são armazenados em um arquivo de identificação em formato CSV (*Comma Separated Values*).

De posse dos códigos obtidos, os mesmos são utilizados para extrair os identificadores e a data da última atualização dos currículos (Módulo Extração de Ids e Data da Figura 4). Este identificador é necessário para realizar o *download* do currículo em formato XML, disponibilizado pelo CNPq.

Por fim, utilizando o arquivo de identificação, é realizado o *download* de todos os currículos e os mesmos são agora armazenados em forma de arquivo XML em um repositório de currículos (Módulo Extração de Currículos da Figura 4). Durante o processo de atualização do repositório de currículos, todo o currículo é substituído, visto que o processo de atualização de seções específicas do currículo é uma tarefa complexa. Sendo assim, todo o currículo é armazenado como um arquivo XML em disco e, no caso de atualização, o mesmo é sobrescrito.

Considerando a necessidade de análise de grupos específicos, como pesquisadores de uma determinada instituição ou área de atuação, por exemplo, um módulo de seleção foi desenvolvido com o objetivo de produzir uma lista de currículos correspondentes a determinados critérios de consulta (DIAS, 2016). Este módulo facilita a criação de grupos com base em quaisquer critérios, desde que estes sejam encontrados nos currículos, possibilitando, inclusive, o uso de operadores lógicos (*AND* e *OR*) para uma busca mais específica.

De posse do arcabouço desenvolvido por Dias (2016), foi realizada uma expansão do mesmo para atender as necessidades do presente estudo (Figura 5).

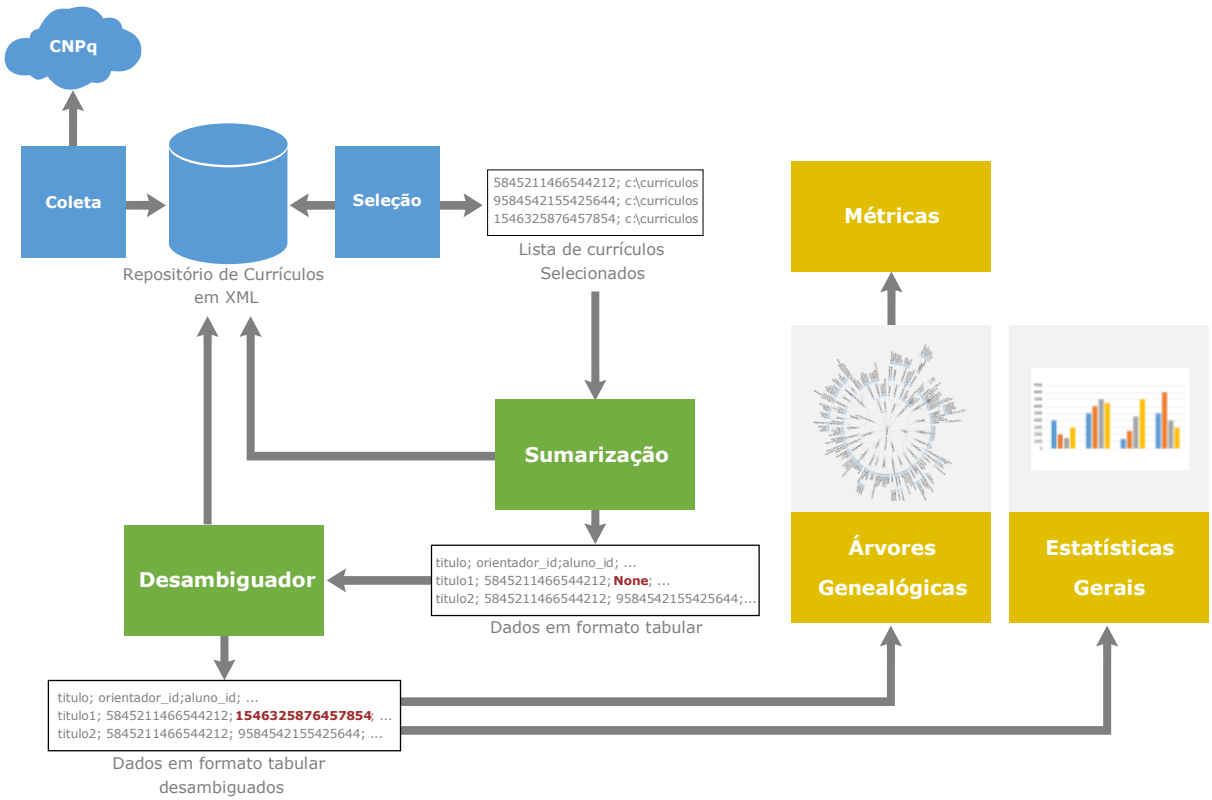


Figura 5: Visão geral do LattesDataXplorer expandido

Fonte: Próprio Autor

O arcabouço inicialmente proposto, que engloba toda a parte de extração e seleção dos dados, foi expandido de tal maneira que passou a incorporar módulos para sumarização dos dados, desambiguação dos registros de orientação e produção de árvores genealógicas, além do cálculo de estatísticas gerais, que são fundamentais para a obtenção dos resultados aqui apresentados. Tais módulos são fundamentais para desambiguar os dados de orientação acadêmica e, consequentemente, produzir árvores genealógicas acadêmicas que incluem indivíduos antes desconhecidos. Estes mesmos dados desambiguados também são utilizados na obtenção de análises estatísticas gerais. Os módulos adicionados na expansão do arcabouço são detalhados nas próximas subseções.

4.2.1 Sumarização dos Dados

De posse dos currículos, o módulo de seleção do arcabouço pode ser aplicado para obter conjuntos específicos de acordo com a necessidade, como por exemplo, o conjunto de currículos de indivíduos que orientam em cursos de Pós-graduação, e por fim, sumarizados, criando arquivos resumidos contendo as informações de interesse extraídas através de todo o currículo XML, o que possibilita a produção de quatro extratos de dados distintos, dependendo da necessidade, conforme a Figura 6. Todos estes dados são armazenados em formato tabular, facilitando a seu posterior processamento. Novos conjuntos podem ser definidos de acordo com as necessidades apresentadas.

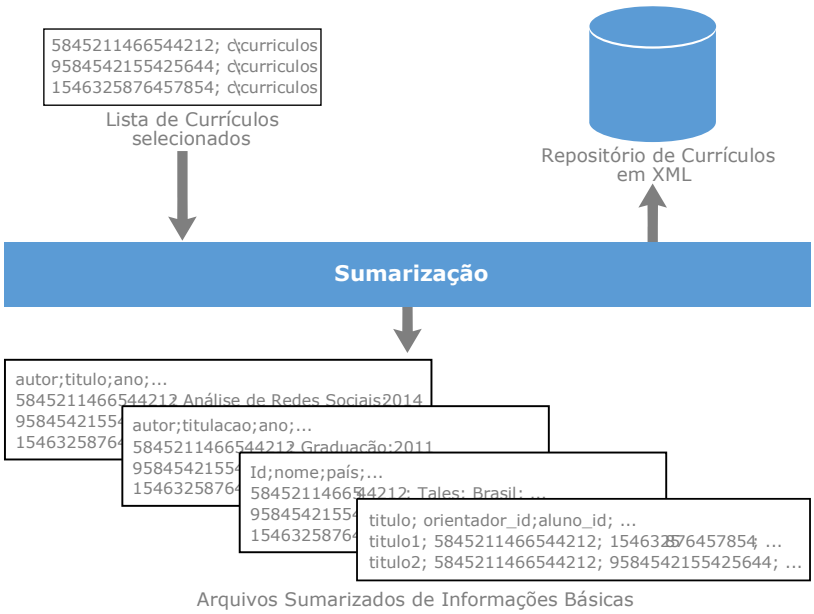


Figura 6: Processo de sumarização dos dados

Fonte: Criado pelo autor

Os dados de informações básicas são compostos por: identificador de 16 algarismos; nome; nome em citações; dados de localização como país, estado e cidade; dados de suas áreas de atuação (6 registros de grande área, área, subárea, especialização). Estes dados são relevantes para aplicar a identificação dos relacionamentos, resolvendo problemas de desambiguação, e identificando novos relacionamentos não vinculados implicitamente.

Dados de orientações são imprescindíveis para o desenvolvimento desta dissertação e são a origem de diversas análises. Dentre os dados de orientações sumarizados, estão os seguintes:

- Título do trabalho de conclusão de curso, monografia, dissertação ou tese;
- Nome do aluno orientado;
- Identificador do aluno, sendo que nem todos possuem um valor para esta coluna, sendo necessária a identificação a partir do nome;
- Nome do orientador, sendo este o proprietário do currículo em questão;
- Identificador do orientador (idem acima);
- Nome da instituição;
- Identificador da instituição;
- Nome do curso;
- Identificador do curso;
- Ano em que ocorreu a orientação;
- Natureza da orientação, podendo ser: Iniciação Científica, Graduação, Especialização, Mestrado, Doutorado, Pós-doutorado ou Outra Natureza;
- Área do conhecimento (3 registros), contendo:
    - Grande Área
    - Área
    - Sub Área
    - Especialização

Os arquivos resultantes do processo são armazenados em formato CSV, simplificando o processamento dos dados. Desta forma, o processo de ler cada um dos currículos e aplicar uma consulta em XPath (*XML Path Language*) para obtenção dos dados será realizado apenas uma vez.

Os dados representados na Tabela 1 correspondem a todo o conjunto de currículos da Plataforma Lattes no momento da extração utilizada, se revelando uma grande fonte de dados que pode ser utilizada na área da descoberta de conhecimento. Um destes é a genealogia acadêmica. Esta pode ser identificada através dos dados de orientações, possibilitando a análise da vida pregressa do indivíduo. Além disso, tais dados podem ser associados aos dados de produção bibliográfica, por exemplo.

Tabela 1: Resultado da sumarização de todos os currículos analisados.

Descrição	Quantidade Total
Dados de informações pessoais	4.591.941
Dados de formação acadêmica	10.488.174
Dados de produção bibliográfica (periódicos e anais de congresso)	19.335.882
Dados de orientações concluídas	8.138.267

4.2.2 Desambiguação de Nomes

A identificação de relacionamentos em orientações não é uma tarefa trivial. Atualmente, os registros de orientações dos currículos possuem uma opção de se realizar a vinculação manual do nome do orientado ou dos coautores a seus identificadores únicos na Plataforma Lattes. No entanto, tal vínculo não é automático e, em geral, relacionamentos antigos permaneceram sem seus vínculos com os identificadores, exibindo apenas o nome no registro de orientação ou coautoria (Figura 7). Diante disso, uma estratégia de identificação de relacionamento se faz necessária para que se possa caracterizar redes com a maior quantidade possível de indivíduos.

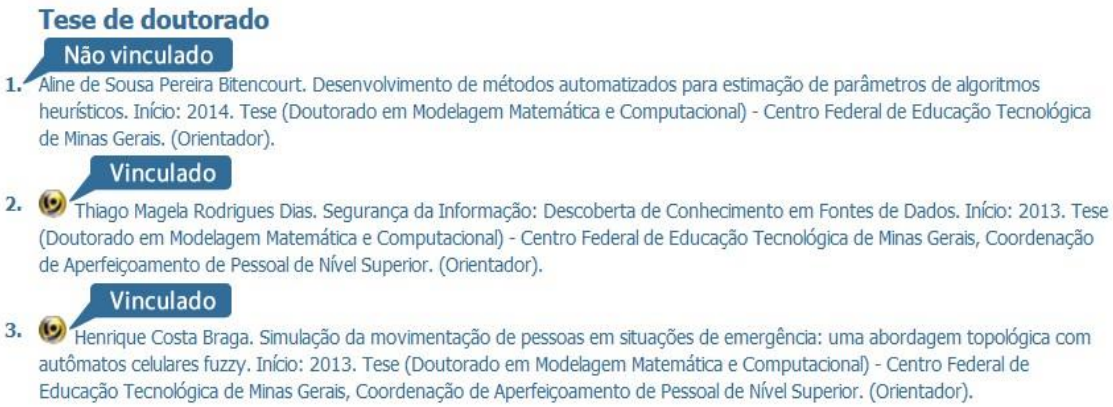


Figura 7: Vínculos da Plataforma Lattes

Fonte: (DIAS, 2016)

Para tanto, uma das principais contribuições deste trabalho é a proposta de um método eficiente para caracterização de vínculos de orientações acadêmicas em grandes repositórios de dados ambíguos, em que a primeira etapa da estratégia de identificação aqui colocada é a obtenção dos dados pessoais de cada indivíduo com currículo cadastrado na Plataforma Lattes. Com base nestes dados, é possível obter informações como nome completo e nome em citações bibliográficas, utilizadas para produção do dicionário, conforme a Figura 8. Além de nomes de citações presentes nos currículos, também são produzidos, automaticamente, possíveis nomes de citação, a partir do nome completo de cada indivíduo.

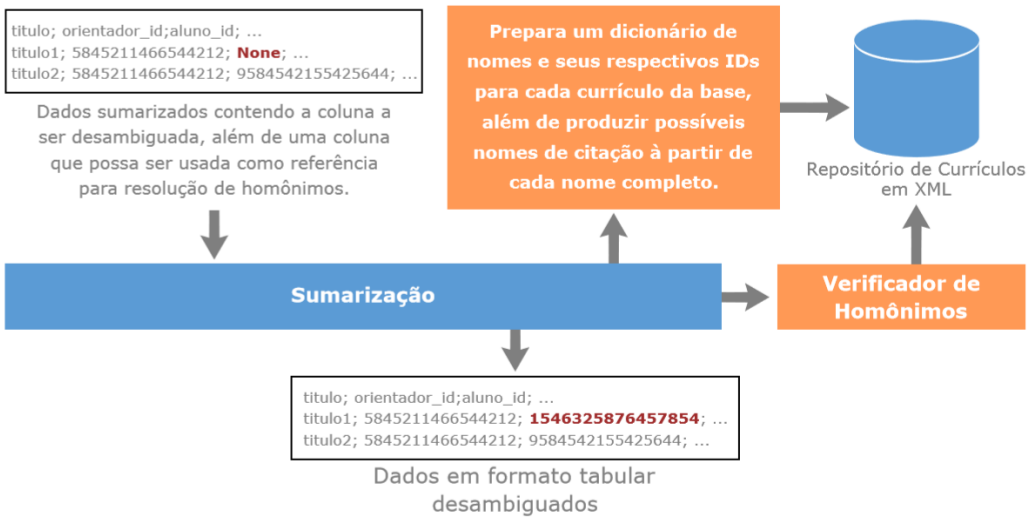


Figura 8: Processo de desambiguação de nomes da Plataforma Lattes

Fonte: Próprio Autor

De posse dos dados citados, é criado um dicionário de registros, como exemplifica a Figura 9, contendo o nome de citação como chave do dicionário e uma lista vinculada a esta chave, contendo uma lista com os identificadores únicos de cada indivíduo.

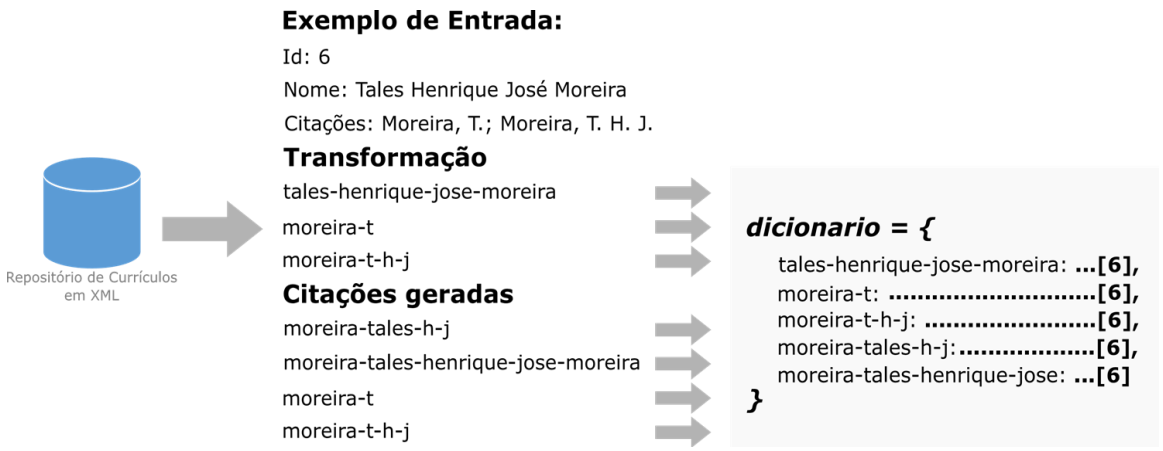


Figura 9: Exemplo de dicionário utilizado no processo de desambiguação

Fonte: Próprio Autor

Observando a Figura 10, percebe-se que os nomes marcados com o número 1 (um) referem-se ao nome completo de cada indivíduo da Plataforma. O nome de citação marcado com o número 2 (dois) é um exemplo de nome obtido através do currículo. Ou seja, o usuário proprietário deste nome o cadastrou como nome de citação. Todos os demais são nomes gerados automaticamente a partir do nome completo, fazendo as possíveis combinações.

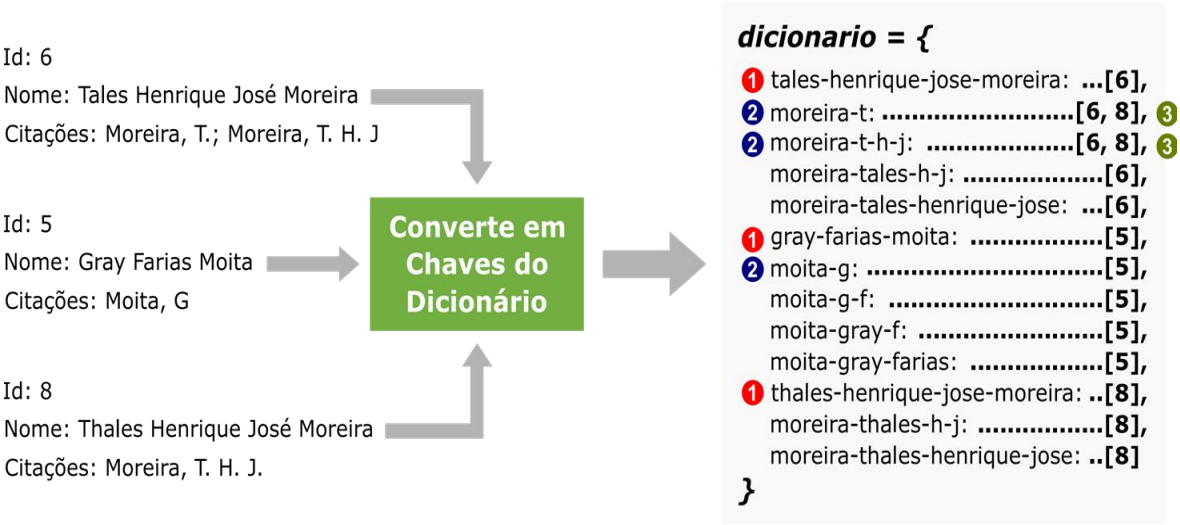


Figura 10: Inserindo indivíduos para construção do dicionário

Fonte: Próprio Autor

Todos estes nomes são chaves deste dicionário. Observando seus valores, percebe-se que a lista marcada pelo número 3 (três) possui mais de um valor (neste exemplo, possui dois valores, sendo eles 6 e 8). Isso quer dizer que o nome referente a estes valores (chave do dicionário) é um homônimo e os valores correspondem ao identificador de cada indivíduo. Neste caso, no momento da resolução do nome, o desambiguador fará uma consulta em cada um destes dois currículos procurando por uma referência. No caso de orientações, a referência corresponde ao nome e identificador do orientador. Porém, além disso, pode-se utilizar outras informações, como, por exemplo, área de atuação e instituição onde ocorreu a orientação.

As referências citadas são uma verificação cruzada para saber se os orientados possuem alguma ligação com o orientador em questão. Após aplicar a referência no desambiguador, é selecionado o de maior relevância. Ou seja, aquele que teve a maior quantidade de referências encontradas em seu currículo, possuindo a maior chance de ser o indivíduo correto. No momento da desambiguação, de posse do dicionário contendo dados de todos os indivíduos da Plataforma Lattes, o nome a ser desambiguado é, acompanhado de uma ou mais referências a serem utilizadas em caso de homônimo, passado como parâmetro para o desambiguador, conforme a Figura 11.

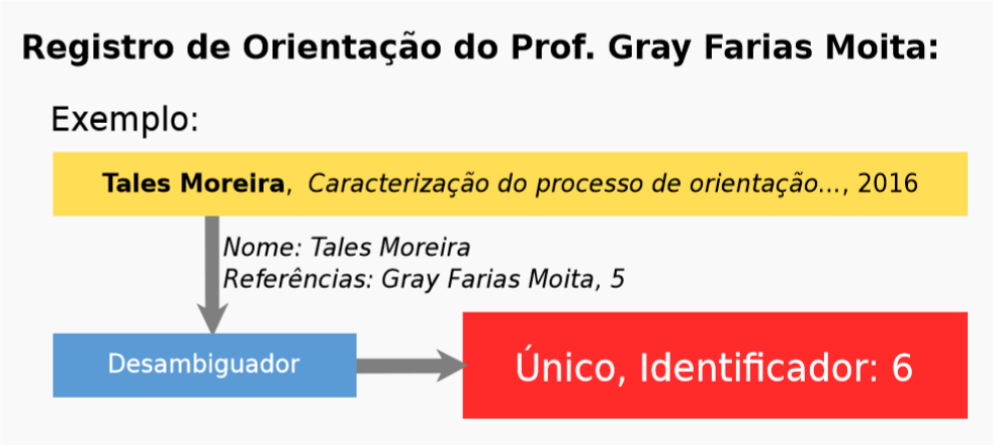


Figura 11: Aplicação do desambiguador em um registro de orientação

Fonte: Próprio Autor

Neste caso, utilizando o dicionário de registros da Figura 9, o desambiguador encontrou o identificador correspondente facilmente, já que a lista de possíveis identificadores para este nome possui apenas um único id. Deste modo, este é retornado diretamente, sendo considerado como o identificador correspondente para este nome. Vale ressaltar que, neste caso, a referência passada sequer foi utilizada. Porém, caso o nome do aluno orientado fosse "Moreira, T.", conforme Figura 12, o processo seria diferente. Neste segundo caso, o dicionário de registros possui dois possíveis identificadores para este nome (6 e 8). Diante disso, o desambiguador terá que resolver qual identificador retornar. Assim sendo, os dois currículos são visitados e ranqueados de acordo com a quantidade de vezes que a referência foi encontrada no mesmo. Ou seja, neste exemplo será pesquisado o nome "Gray Farias Moita" e o id 5 nos dois currículos, já que, se este orientou um aluno, seu nome deve constar no currículo deste, seja na seção de formação ou mesmo na seção de produção bibliográfica, retornando o de maior ocorrência.

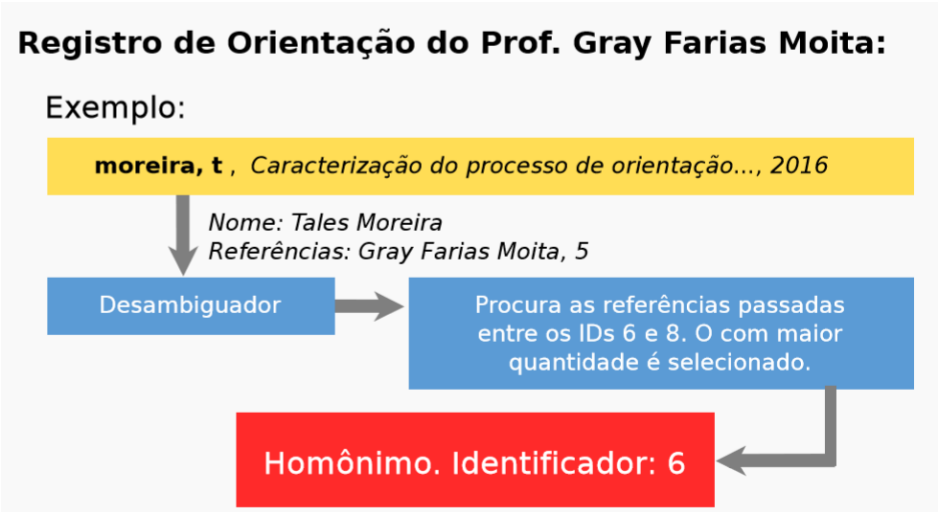


Figura 11: Aplicação do desambiguador em um registro de orientação com uso do resolvedor de homônimos

Diante disto, o desambiguador é inicializado, gerando um dicionário com informações de todos os indivíduos (cerca de 20.000.000 de nomes de citações distintos) e, posteriormente, aplicado em todo o conjunto de orientações a serem desambiguadas,

resolvendo o nome de orientados não vinculados implicitamente em cada orientação concluída informada nos currículos dos orientadores. Em caso de um único identificador para este orientado no dicionário, o identificador é atribuído a ele. Em situações de homônimos, outras informações como nome e identificador do orientador são utilizadas para tentar localizar o correto identificador do orientado. Isso possibilita a geração de árvores com um número maior de elementos, obtendo, assim, um melhor resultado.

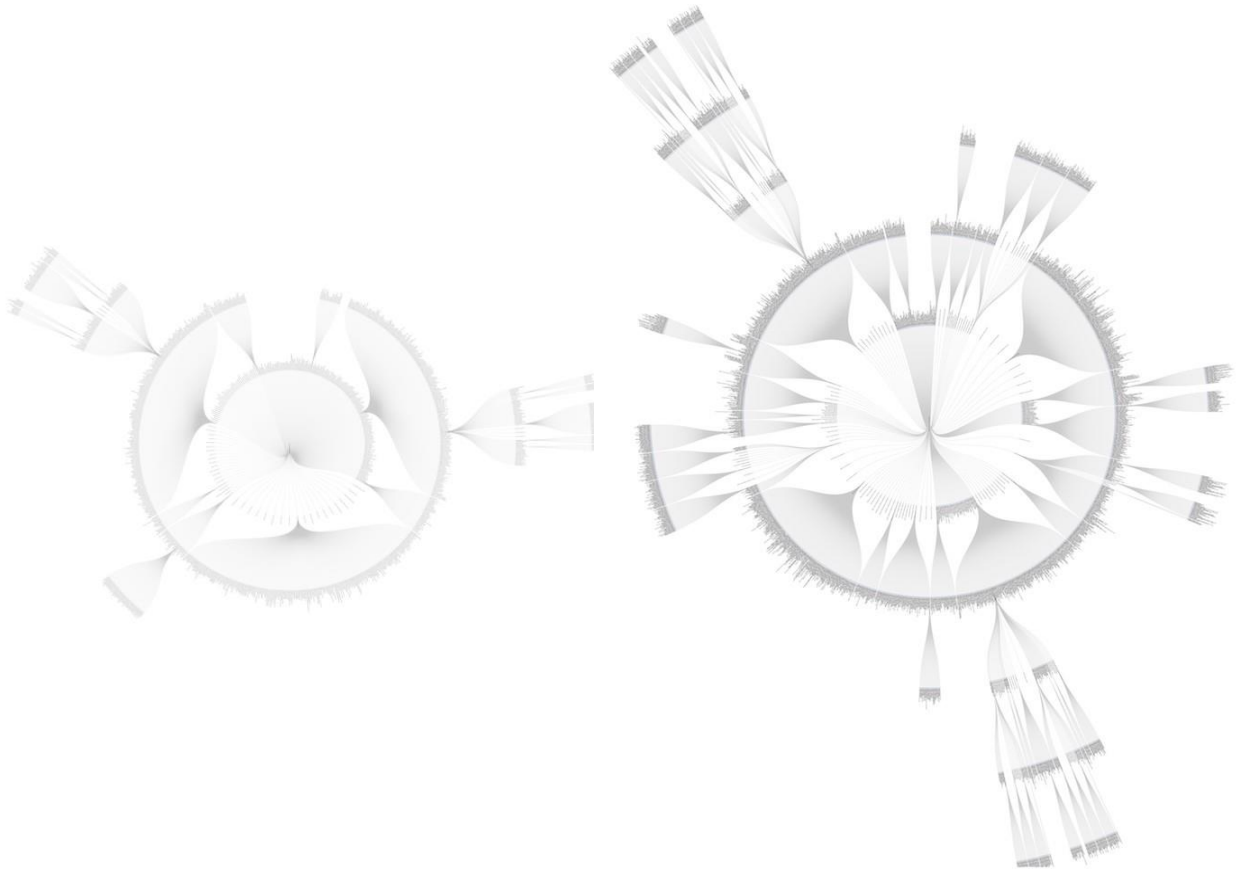
Consequentemente, diante do exposto, nem todos os relacionamentos podem ser identificados, seja por erros de digitação na definição do nome dos alunos orientados ou mesmo pela inexistência do currículo na Plataforma Lattes. Neste caso, estes serão transformados em nós folhas e seus descendentes, mesmo que existam, não são incorporados, já que não é possível analisar seus currículos pela falta do identificador. Cabe ressaltar que o mesmo desambiguador, além de orientações, pode ser aplicado em qualquer sessão do currículo, como formações acadêmicas, produções bibliográficas, entre outras.

A Plataforma Lattes conta hoje com mais de 5.200.000 currículos, totalizando cerca de 8.449.497 orientações, independentemente de sua natureza (Iniciação Científica, Graduação, Especialização, Mestrado, Doutorado, Pós-doutorado e Outra Natureza). Antes da execução do método de identificação proposto neste trabalho, apenas 621.384 possuíam relacionamentos vinculados implicitamente, conforme a Figura 7, que apresenta um trecho de um currículo. Ou seja, apenas estas orientações foram devidamente vinculadas com os orientados pelo professor orientador em seus currículos.

Tabela 2: Resultado do desambiguador

Descrição	Quantidade	%
Não identificados	4.397.692	52,04
Não identificados entre homônimos	247.742	2,93
Subtotal:	4.645.434	54,97
Previamente vinculados	621.384	7,35
Únicos identificados	2.889.545	34,19
Identificados com Homônimos	293.134	3,46
Subtotal:	3.804.063	45,03
Total:	8.449.497	100

Após a execução do método proposto, 3.804.063 relacionamentos novos foram identificados, um total de 45,03%, valor muito superior aos 7,35% encontrados antes do processo de desambiguação. A Figura 13 apresenta as redes de um orientador que se destaca pela grande quantidade de novas orientações identificadas após o processo, exibindo a rede antes e depois da execução do método proposto, demonstrando assim a grande vantagem da aplicação do método de desambiguação.



(a) Antes (b) Depois

Figura 13: Vinculações antes (a) e depois (b) do processo de desambiguação.

Fonte: Criado pelo autor

Obviamente, nem todos os indivíduos podem ser identificados, seja por erro de digitação, ou pela não existência do currículo na Plataforma Lattes. Isso explica a alta porcentagem de currículos não identificados (54,97%). Um exemplo, é um indivíduo que orientou um outro estrangeiro, que não realizou seu cadastro na plataforma. Sendo assim, não é possível encontra-lo, e sempre que isso ocorre, seu nome é usado na construção da relação orientador-orientado ([id-orientador, nome-orientado]). Normalmente, como no caso do orientador, utiliza-se para construção do vínculo o seu identificador único. No caso da produção de uma árvore, os nós que não foram identificados são considerados nos folha.

4.2.3 Dados Estratificados

Como resultados das etapas de tratamentos dos dados, são produzidos arquivos de dados pré-processados em formato tabular, podendo ou não estar desambiguados, contendo todas as informações necessárias para o cálculo de diversas métricas bibliométricas e baseadas em análise de redes sociais, como pode ser observado na arquitetura proposta (Figura 5). Os cálculos realizados a partir destes arquivos são facilitados, já que nestes constam todos os dados sem a necessidade de acesso e busca dos dados em cada um dos currículos. Essa estrutura se torna a base para cálculos estatísticos, geração de árvores e redes, e cálculos de métricas.

Exemplos de estatísticas e métricas que estão implementadas são: média de orientações por indivíduo, frequência de orientações de uma determinada área e orientadores mais relevantes considerando seus descendentes, entre outras. Todas essas métricas possibilitam explorar os dados extraídos dos currículos cadastrados na Plataforma Lattes e proporcionam uma ampla visão da genealogia acadêmica brasileira, tendo como fonte de dados principal os registros de orientações dos currículos.

No próximo capítulo, é apresentada uma caracterização e análise geral de todos os registros de orientações obtidos.

Capítulo 5

Análise e Discussão dos Resultados

Neste capítulo é apresentada uma caracterização geral dos dados de orientação contidos nos currículos cadastrados na Plataforma Lattes, coletados em setembro de 2017. São apresentadas estatísticas gerais dos dados extraídos (Seção 5.1) e estatísticas por grandes áreas do conhecimento (Seção 5.2). Além disso, são apresentadas as árvores genealógicas acadêmicas (Seção 5.3), geradas através dos registros extraídos, e, por fim, uma análise da floresta contendo as 100 maiores árvores identificadas (Seção 5.4).

5.1 Estatísticas Gerais

Os dados aqui utilizados foram coletados em setembro de 2017 e correspondem a 5.152.148 currículos cadastrados na Plataforma Lattes, distribuídos em diversas áreas do conhecimento e nos mais diversos níveis de capacitação. Todo o processo de coleta e tratamento dos dados foi descrito no Capítulo 4, possibilitando, através desta coleta, realizar uma caracterização de todos os registros de orientações contidos nos currículos em todos os níveis de orientação. Em todo o conjunto em análise, 481.624 indivíduos orientaram em algum nível de formação, ou seja, apenas 9,348% do total de todos os indivíduos com currículos cadastrados na Plataforma Lattes. As orientações consideradas para esta caracterização cobrem o período de 1900 (orientação mais antiga cadastrada) até dezembro de 2016, período este considerado, tendo em vista a coleta realizada em 2017. Logo, apesar de a coleta ter sido realizada em setembro de 2017, somente orientações até o fim do ano de 2016 foram consideradas. Desta forma, espera-se que todos os registros de orientações realizadas até o ano de 2016 estejam inseridos nos currículos analisados.

Tratando-se de orientações em um país cientificamente emergente, pressupõe-se que exista orientadores estrangeiros orientando brasileiros. Analisando os dados obtidos, constatou-se que o Brasil possui 468.046 orientadores (97% do total). A Tabela 3 apresenta um ranking com orientadores de outros países cadastrados.

Tabela 3: Ranking com os dez países com maior quantidade de indivíduos estrangeiros orientadores.

País	Quantidade	% de Orientações Estrangeiras
Argentina	1.354	9,972%
Portugal	1.243	9,154%
Peru	1.037	7,637%
Colômbia	1.028	7,571%
Cuba	869	6,400%
Estados Unidos	640	4,713%
França	637	4,691%
Espanha	633	4,661%
Itália	584	4,301%
Alemanha	573	4,220%

No caso dos orientadores brasileiros, estes foram distribuídos pelos estados da federação (Figura 14). Como já era esperado, a maior quantidade dos orientadores está situada no estado de São Paulo (102.425 orientadores). Do total de 468.046 orientadores brasileiros identificados, 16.601 não possuem estado definido no currículo (não informado).

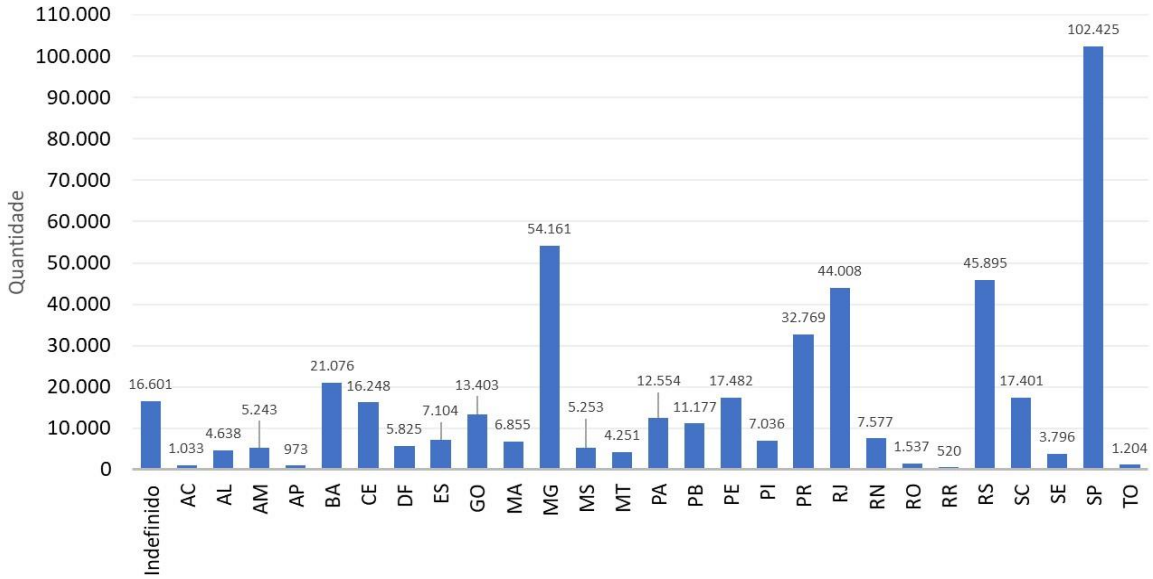


Figura 14: Distribuição das orientações por estado

Fonte: Próprio Autor

As orientações também podem ser distribuídas de acordo com o seu nível, conforme apresentado na Figura 15, sendo que grande parte destas são orientações de Graduação (44%), seguidas pelas orientações de Especialização (14%), as de Iniciação Científica (14%) e as de Outra Natureza (13%). Consequentemente, orientações mais relevantes, como, por exemplo, Mestrado, Doutorado e Pós-doutorado, correspondem a apenas a 15% do total identificado. No entanto, em diversas ocasiões ao longo deste trabalho serão utilizadas apenas orientações de pós-graduação (Mestrado, Doutorado e Pós-doutorado), devido à sua relevância e formalidade de registro destas orientações por parte dos orientadores.

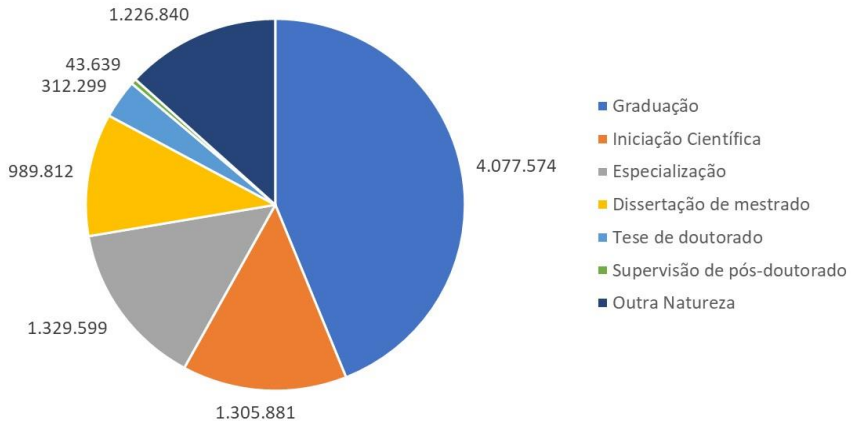


Figura 15: Distribuição das orientações por natureza de todos os indivíduos

Fonte: Próprio Autor

Conforme descrito, apesar da baixa quantidade de orientações selecionadas para as análises aqui apresentadas (1.345.750 registros de orientações em programas de Pós-graduação) se comparado ao total geral de orientações, este conjunto possui informações sobre orientações com registros mais confiáveis, o que o torna um interessante objeto de estudo.

Ao se analisar todos os registros de orientações concluídas em programas de pós-graduação identificados nos currículos da Plataforma Lattes, observou-se que a quantidade de orientações se encontra em queda mais acentuada a partir de 2014, onde foram registradas 735.869 orientações (Figura 16).

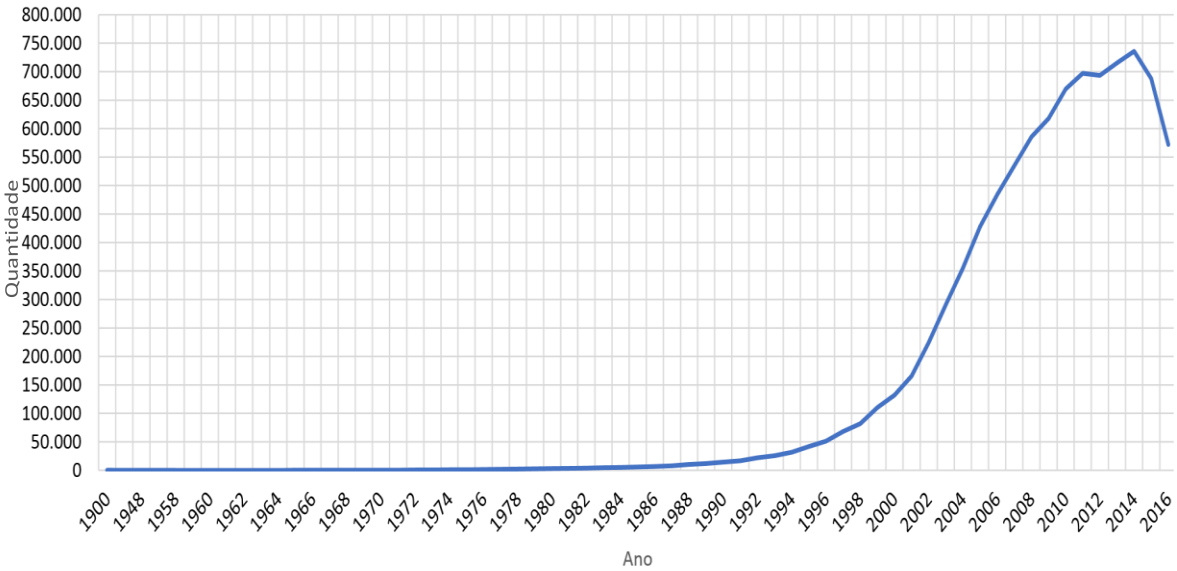


Figura 16: Gráfico de orientações por ano considerando todos os indivíduos até o ano de 2016

Fonte: Próprio Autor

Com relação à queda ocorrida em 2015, é observado na Tabela 4 que a mesma aconteceu de forma expressiva em praticamente todas as áreas, com exceção das orientações de Doutorado, onde houve quedas em 2014 (de 2,23%) e 2016 (de 1,60%), que são

relativamente pequenas em comparação com os demais níveis de capacitação. O nível com a maior queda nos últimos 5 anos foi o de Iniciação Científica, com 41,53%, seguido pela Especialização (35,95%), de Outra Natureza (31,06%) e pela Graduação (24,81%). Coincidentemente, estes são os níveis que não são utilizados em diversas partes deste estudo. Do restante, orientações de Mestrado e supervisões de Pós-doutorado tiveram queda de 17,68% e 22,03%, respectivamente.

Tabela 4: Total de orientações e taxa de crescimento por ano de todos os indivíduos

Natureza	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Iniciação Científica	69.531	73.188	77.951	89.739	97.363	104.560	108.141	107.197	99.059	66.296
	14,25%	5,26%	6,51%	15,12%	8,50%	7,39%	3,42%	-0,87%	-7,59%	-33,07%
Graduação	260.151	282.729	292.237	305.493	303.764	286.116	307.861	317.446	293.455	261.549
	8,54%	8,68%	3,36%	4,54%	-0,57%	-5,81%	7,60%	3,11%	-7,56%	-10,87%
Especialização	76.986	86.179	93.057	103.795	111.820	109.664	95.559	100.637	98.614	79.732
	13,45%	11,94%	7,98%	11,54%	7,73%	-1,93%	-12,86%	5,31%	-2,01%	-19,15%
Dissertação de mestrado	44.676	48.891	52.596	55.021	59.498	64.719	68.863	71.005	68.846	58.769
	4,68%	9,43%	7,58%	4,61%	8,14%	8,78%	6,40%	3,11%	-3,04%	-14,64%
Tese de doutorado	14.730	15.790	17.073	17.716	19.047	21.486	22.044	21.553	21.841	21.491
	7,83%	7,20%	8,13%	3,77%	7,51%	12,81%	2,60%	-2,23%	1,34%	-1,60%
Supervisão de pós-doutorado	1.843	2.392	2.794	3.332	3.570	4.040	4.636	5.069	4.943	3.977
	13,70%	29,79%	16,81%	19,26%	7,14%	13,17%	14,75%	9,34%	-2,49%	-19,54%
Outra Natureza	67.072	76.428	82.314	94.485	101.896	102.831	108.380	112.962	101.266	80.293
	16,39%	13,95%	7,70%	14,79%	7,84%	0,92%	5,40%	4,23%	-10,35%	-20,71%

Uma provável explicação para a redução das orientações nos últimos anos é possivelmente a falta de atualização. O que corrobora para isso é o fato de que orientações de Pós-graduação, onde os orientadores são mais criteriosos no preenchimento de seus currículos, possuem as menores taxas de redução na quantidade de orientações.

Verificando os dados atuais e de uma extração anterior, pôde-se observar na Tabela 5 que os valores sofreram alterações com o passar do tempo, indicando que dados de anos anteriores foram atualizados. Porém, consequências da crise econômica do país, como por exemplo, o corte de verbas para pesquisa e educação, também podem contribuir para a queda na quantidade de orientações.

Tabela 5: Comparação da quantidade de registros por ano entre dados obtidos em 2014 (anterior) e 2017 (atual)

Base	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Atual	534.989	585.597	618.022	669.581	696.958	693.416	715.484	735.869	688.024	572.107
Anterior	532.869	582.572	613.722	662.607	687.216	676.542	685.020	674.049	538.661	-
Diferença	2.120	3.025	4.300	6.974	9.742	16.874	30.464	61.820	149.363	572.107

A partir dos resultados obtidos, também pode-se observar que, em todos os níveis de capacitação, existem orientadores com quantidades acima da média geral, que acabam por distorcer a média de orientações, conforme apresentado na Tabela 6. Nela, por exemplo, na categoria de orientações de Outra Natureza existem orientadores com mais de 3.000

orientações. Diante disso, foi gerado um *boxplot*[8] que auxilia a análise dos dados de maneira mais complexa (Figura 17).

Tabela 6: Indivíduo com maior quantidade de orientações para cada nível de capacitação.

Descrição	Quantidade
Iniciação Científica	353
Graduação	814
Especialização	2.978
Dissertação de mestrado	399
Doutorado	138
Supervisão de Pós-doutorado	45
Outra Natureza	3.088

Portanto, observa-se que a mediana em todos os níveis de formação varia entre apenas 2 (duas) e 6 (seis) orientações. Entretanto, formações como Graduação e Mestrado se destacam pela quantidade de orientações por indivíduo, obtendo uma maior variação entre elas. Formações como Outra Natureza e Especialização se equiparam e, apesar de possuírem orientadores com grande quantidade de orientações, a Figura 17 demonstra que estes são valores atípicos.

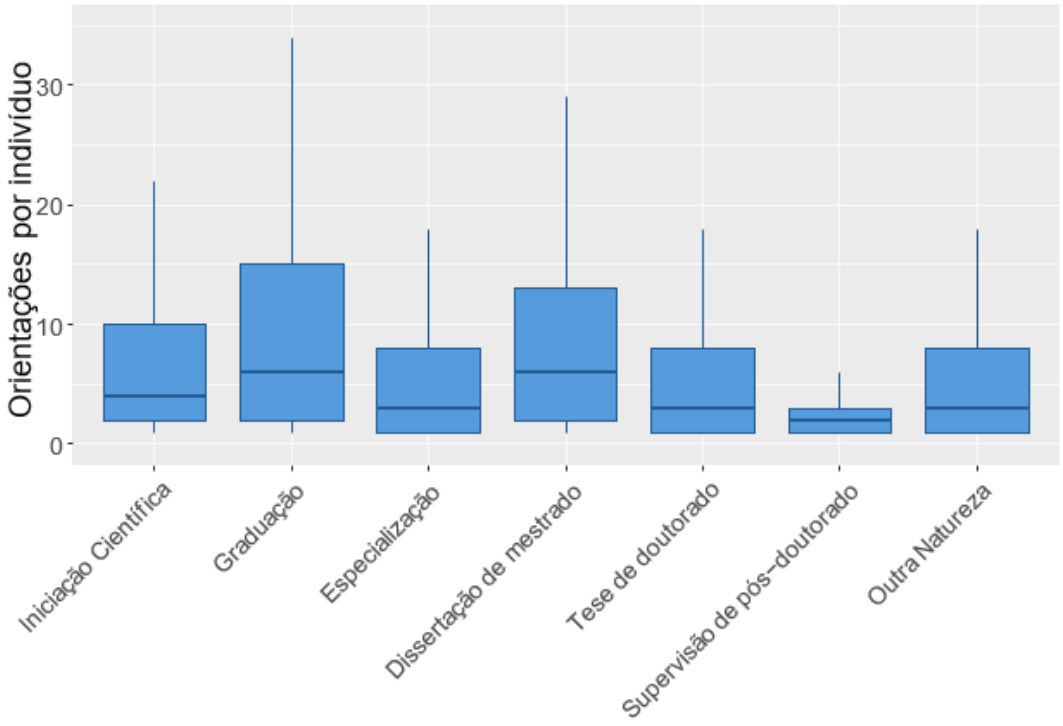


Figura 17: Distribuição das orientações por formações

Fonte: Próprio Autor

[8] Devido à grande variação entre as taxas de cada amostragem, os *outliers* de alguns diagramas de caixa foram ocultados para melhor visualização.

Os resultados apresentados acima demonstram que, apesar da importância da Plataforma Lattes para a análise do processo de orientação, a mesma possui problemas, como, por exemplo, a falta de atualização dos currículos. Isso é um dos grandes problemas da plataforma, juntamente com a falta de vínculos para diversos currículos, problema este reduzido pelo desenvolvimento do desambiguador (já descrito anteriormente). Porém, mesmo com estes problemas, pôde-se obter resultados que fornecem informações relevantes sobre as orientações cadastradas na Plataforma de maneira geral.

Na Seção 5.2, estas informações são distribuídas e apresentados em relação as suas respectivas Grandes Áreas.

5.2 Estatísticas por Grandes Áreas

Os dados analisados também foram distribuídos por grandes áreas do conhecimento, conforme apresenta a Figura 18. Do total de orientações, mais de 50% não possuem grande área informada. Entre as demais grandes áreas, a grande área com maior quantidade de orientações é a de Ciências Sociais Aplicadas.

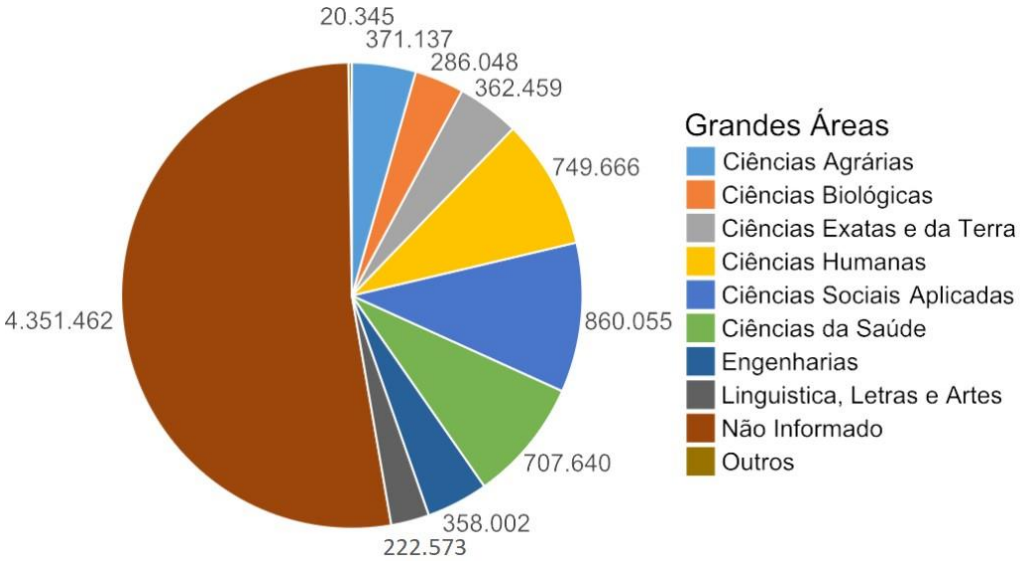


Figura 18: Gráfico de orientações por grandes áreas

Fonte: Próprio Autor

Diante da quantidade de orientações sem grande área informada, foram utilizadas as grandes áreas informadas como grande área de atuação dos orientadores, para, dessa maneira, caracterizar todas as orientações por esta característica, o que produz resultados relevantes (Figura 19).

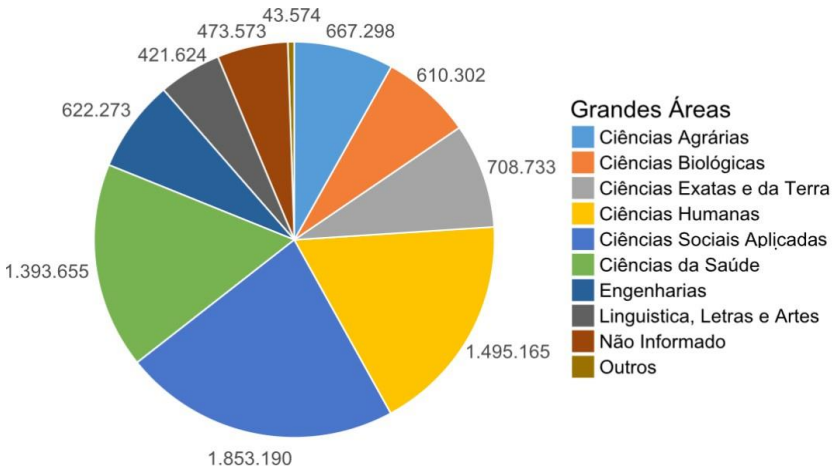


Figura 19: Gráfico de orientações por grandes áreas após uso de informações do orientador

Fonte: Próprio Autor

Pode-se observar que, apesar do uso de informações dos orientadores referente à grande área de atuação para complementação dos dados de orientações, isso não ocasionou grandes distorções, mantendo uma certa proporção no aumento de orientações em cada grande área.

Ao realizar a análise levando em consideração o nível das orientações, pode-se perceber que, na grande área de Ciências Sociais Aplicadas, em sua grande maioria, a orientações de trabalho de conclusão de curso de graduação, conforme apresentado na Figura 20. Já a grande área de Ciências Humanas, a segunda com maior quantidade de orientações, as orientações se destacam em cursos de pós-graduação. Além disso, ressalta-se também a grande área de Ciências Exatas e da Terra, com a maior quantidade de orientações de Doutorado (próximo de 100.000 orientações).

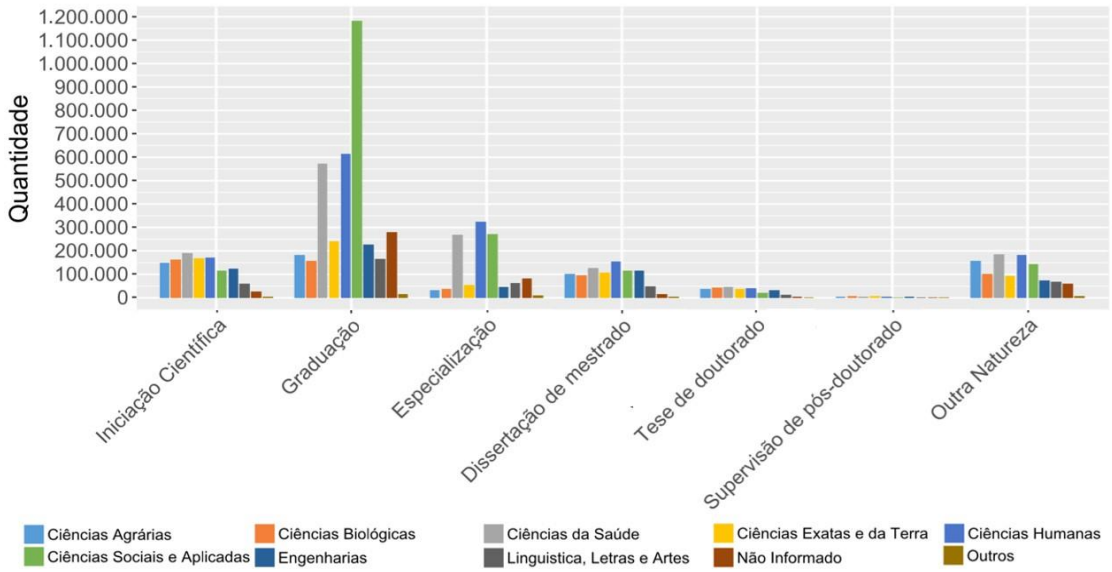


Figura 20: Gráfico de orientações distribuídos por natureza da orientação e grandes áreas

Fonte: Próprio Autor

Além do apontado, as orientações das grandes áreas foram distribuídas por grandes áreas e por ano de conclusão, mostrando a queda na quantidade de orientações já citada em cada uma das grandes áreas. A grande área “Outros” foi a que teve menor variação, mas em contrapartida, é a menor grande área com relação à quantidade de orientações. A queda mais considerável na quantidade de orientações foi a da grande área de Ciências Sociais Aplicadas, com valor superior a 30.000 orientações entre 2013 e 2015 (22%). Apesar da grande queda, a grande área de Ciências Sociais Aplicadas continua com a maior quantidade de orientações (Figura 21).

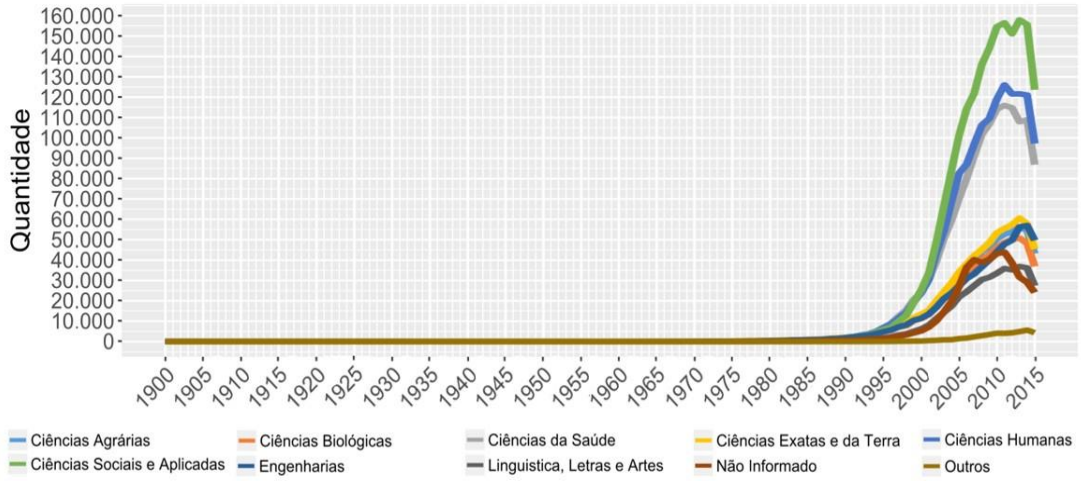


Figura 21: Gráfico de orientações por ano

Fonte: Próprio Autor

Com base nos resultados obtidos, também pode-se observar que em todas grandes áreas existem orientadores com quantidade de orientações acima da média, assim como ocorre na distribuição das orientações por nível de capacitação (Tabela 7).

Tabela 7: Indivíduo com maior quantidade de orientações em cada Grande Área.

Grande Área	Quantidade de Orientações
Ciências Agrárias	1.672
Ciências Biológicas	884
Ciências Exatas e da Terra	809
Ciências Humanas	3.130
Ciências Sociais e Aplicadas	2.968
Ciências da Saúde	1.015
Engenharias	677
Linguística, Letras e Artes	593
Não informado	521
Outros	297

Um exemplo é a grande área de Ciências Humanas, que conta com um indivíduo com mais de 3.000 orientações (o mesmo citado em relação à nível da orientação). Diante disso, foi gerado um *boxplot* que auxilia a análise dos dados de maneira mais completa (Figura 22).

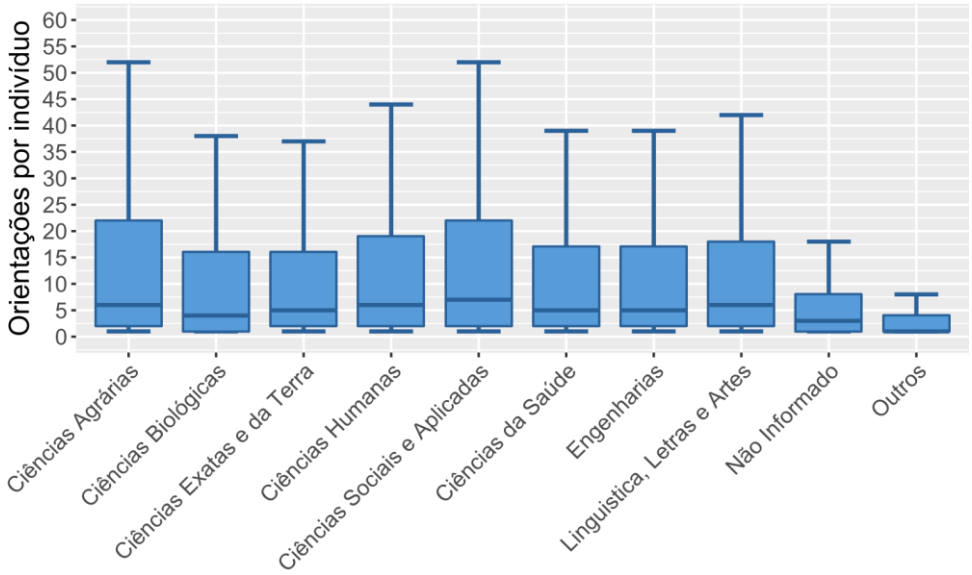


Figura 22: Distribuição das orientações por grande área

Fonte: Próprio Autor

Deste modo, é observado que a mediana em todas as naturezas varia entre apenas 1 (um) a 7 (sete) orientações. Este demonstra que a menor área (Outros), dado a quantidade de orientações, apresenta a menor mediana, além de uma pequena variação nas quantidades de orientações. Ou seja, os orientadores da área "Outros" pouco orientam. Já Ciências Sociais Aplicadas, além de possuir a maior mediana (7), possui uma maior variação. Vale também destacar a área de Ciências Agrárias, que tem quantidade de orientações similar ao da Ciências Sociais Aplicadas, apesar de sua mediana ser menor.

Pode-se observar que, independentemente da grande área, no que diz respeito às orientações todas estão em queda nos últimos anos. Foi possível identificar também que, as grandes áreas de Ciências Sociais Aplicadas e Ciências Agrárias se destacam, uma vez que seus orientadores são bastante atuantes.

5.3 Árvores Genealógicas

Nesta seção, são produzidas árvores genealógicas dos orientadores que mais orientam de acordo com os registros dos currículos cadastrados na Plataforma Lattes. Devido à quantidade de dados, são utilizados apenas dados de orientações dos cursos de pós-graduação, facilitando não só a análise, como também a visualização das árvores. Assim, são utilizados 1.196.916 registros de orientações selecionados a partir de todo o conjunto identificado.

Se aplicado o Módulo de Seleção descrito na Seção 4.2 para a obtenção apenas dos registros de pós-graduação, é possível ranquear os indivíduos do conjunto de acordo com a quantidade de orientações concluídas (Tabela 8). Assim, é apresentado nessa tabela um ranking dos indivíduos com as maiores quantidades de orientações encontrados nos currículos cadastrados na Plataforma Lattes, além de outras informações relacionadas às árvores genealógicas, como o tamanho da rede (quantidade de nós da árvore produzida) e a quantidade de gerações.

Tabela 8: Ranking dos indivíduos com maior quantidade de orientações de pós-graduação.

N	Instituição	Quantidade de Orientações Diretas	Gerações	Tamanho da Rede
1	UFSC	399	5	1.396
2	PUC-SP	351	5	1.219
3	UFV	323	3	594
3	UFV	263	7	2.592
5	COPPE - UFRJ	259	5	1.542
6	UFRJ	246	5	1.129
7	PUC-CAMPINAS	236	8	2.132
8	PUC-SP	229	4	1.541
9	UFV	224	5	1.275
10	UFV	223	3	494

As maiores quantidades de orientações nem sempre estão diretamente relacionadas à propagação de seu conhecimento para outras gerações. Em casos como o do Indivíduo Nº 1 da Tabela 8, observa-se que há uma grande quantidade de orientações. Inicialmente, pode-se presumir que seja uma das maiores redes encontradas. Porém, sua árvore genealógica possui 5 gerações de descendentes, bem inferior à maior encontrada, com 11 gerações.

Tal comportamento está relacionado a área do conhecimento, tendo algumas áreas comportamentos distintos quando avaliada a qualidade de orientações. Alunos de determinadas áreas tendem a permanecer no meio acadêmico, diferentemente de outras, em que suas carreiras tendem a ser desenvolvidas em ambiente onde as orientações não fazem parte do contexto. Geralmente, alunos de cursos de pós-graduação, notadamente no nível de doutorado (no caso do Brasil), tendem a permanecer no meio acadêmico, resultando em novos orientadores repassando seu conhecimento para outros alunos e, consequentemente, incrementando sua rede de orientações em novas gerações.

A Tabela 9 apresenta as maiores redes identificadas, ranqueadas pela maior rede encontrada e pelo maior número de gerações de descendentes, ambas com uma quantidade relativamente baixa de orientações diretas, mas com uma grande quantidade de orientações indiretas.

Tabela 9: Indivíduos com maiores redes do conjunto de dados.

Tipo	Grande Área	Gerações	Tamanho da Rede
Maior Quantidade de Gerações	Ciências Biológicas	15	5716
	Ciências Biológicas	15	5442
	Ciências Biológicas	15	5426
Tamanho máximo de Rede	Ciências Humanas	6	8318

A Figura 23 representa a árvore genealógica do Indivíduo Nº 1 da Tabela 8, com a maior quantidade de orientações diretas. Ele, que é o nó central, possui seus descendentes distribuídos nos diferentes níveis da árvore. Pode-se observar que apesar de seu nó raiz ter orientado a maior quantidade de orientações diretas (399) encontrada, seus orientados,

em geral, tiveram poucas orientações, fazendo com que sua árvore não atingisse um grande número de gerações.

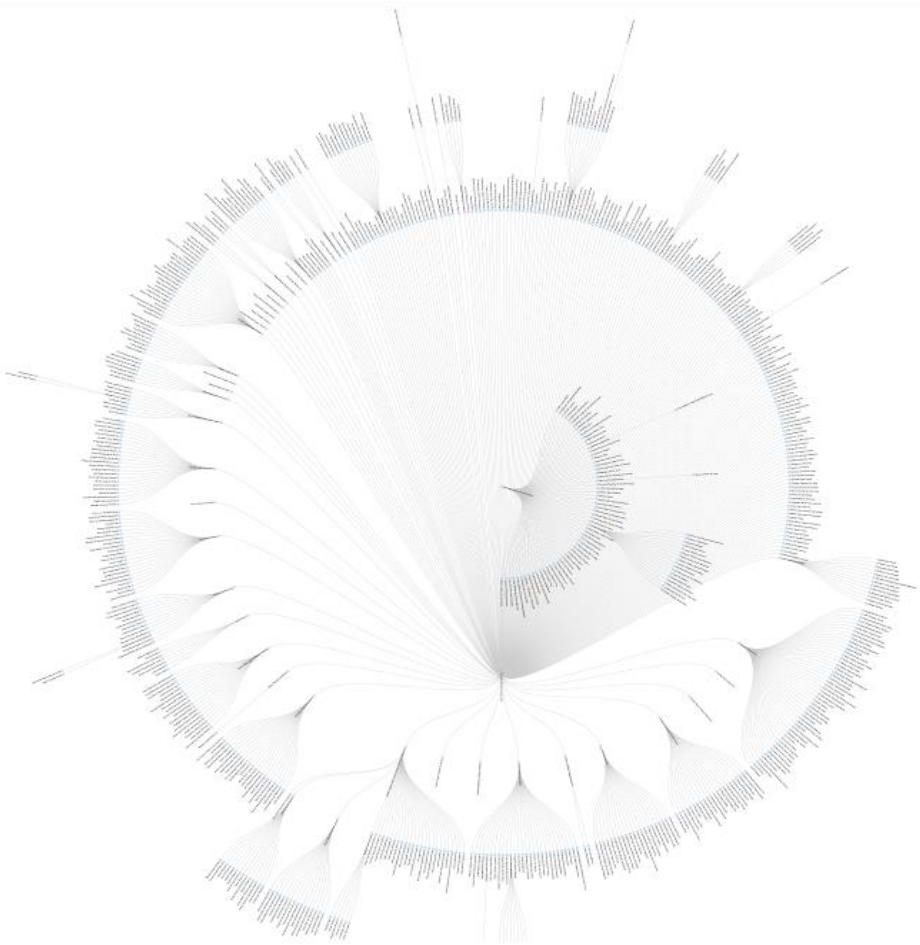


Figura 23: Árvore genealógica do orientador com a maior quantidade de orientações

Fonte: Próprio Autor

O indivíduo raiz desta árvore possui como principal grande área de atuação a grande área de Ciências Humanas e com endereço profissional na Universidade Federal de Santa Catarina, possuindo como formação máxima Doutorado em Engenharia de Produção. Além desta área, atua também nas áreas de Engenharias e Ciências Exatas e da Terra. Suas orientações iniciaram em 1994 e, além de sua grande maioria de orientações serem em programas de pós-graduação, onde 67% são de Mestrado, 28% de Doutorado e ainda possui 20 orientações de Graduação (que não estão sendo consideradas nesta seção).

Diferentemente, as árvores das figuras 24 e 25, apesar de apresentarem poucas orientações diretas, tiveram seu conhecimento propagado de forma mais intensa ao longo do tempo, já que seus orientados se tornaram orientadores, consequentemente orientando novos alunos e incrementando suas redes, aumentando o número de gerações.

A árvore da Figura 24 possui a maior quantidade de gerações encontradas nos currículos cadastrados na Plataforma Lattes (15 gerações). Além disso, possui uma quantidade de descendentes relativamente alta (5.716).

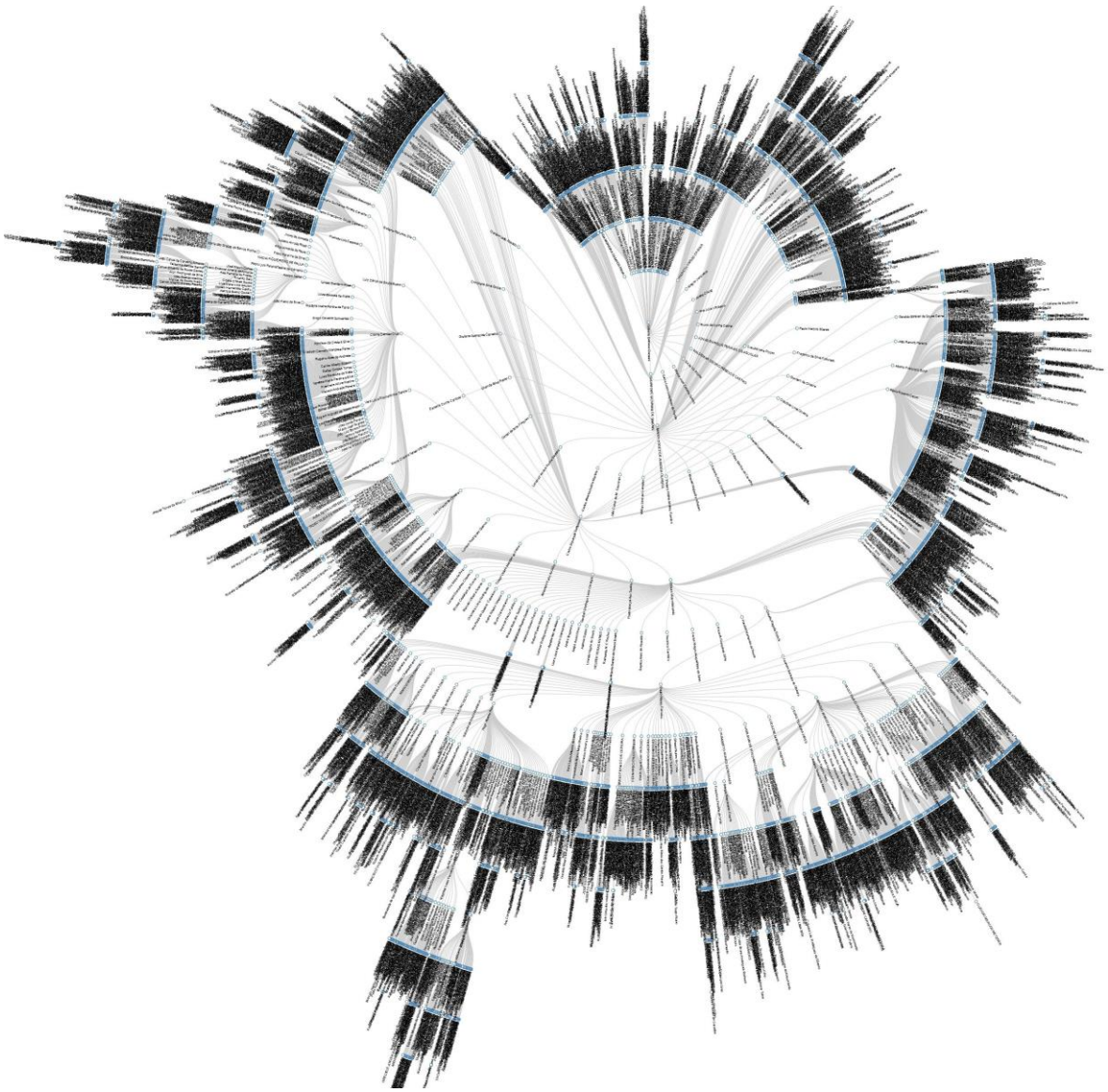


Figura 24: Árvore genealógica do orientador com a maior quantidade de gerações e descendentes

Fonte: Próprio Autor

O nó raiz (orientador principal) é bolsista de pesquisa do CNPq 1D e atua na Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). Possui como formação máxima doutorado em Bioquímica e possui como sua principal grande área de atuação Ciências Biológicas. Suas orientações iniciaram em 1970 e são distribuídas em 56% em Mestrado, 33% em Doutorado, tendo também orientado alunos de Iniciação Científica, que, a exemplo da graduação, também não está sendo considerada. Um fato interessante é que este pesquisador possui apenas 13% da quantidade de orientações diretas do indivíduo apresentado anteriormente, o que pode ser notado ao se verificar as duas figuras. Apesar disso, sua árvore é cerca de 4 vezes maior que a anterior e possui 3 vezes mais gerações. Ou seja, alguns de seus descendentes se destacaram e transformaram esta árvore em uma das mais representativas identificadas. Isso em parte também se deve ao fato de que, entre as 3 árvores analisadas, esta é a com orientações mais antigas.

Já a árvore da Figura 25 possui a maior quantidade de descendentes, apesar de possuir apenas 6 gerações (muito abaixo das 15 encontradas na árvore da Figura 24) e cerca de 30% da quantidade de orientados diretos em relação ao indivíduo da primeira árvore apresentada. O nó raiz (orientador principal) possui como titulação máxima doutorado em Filosofia da Educação e atua na Universidade Estadual de Campinas (UNICAMP), tendo como grande área de atuação Ciências Humanas. Suas orientações começaram em 1974 e são distribuídas em Doutorado (47%), Mestrado (28%) e Pós-doutorado (18%), além de algumas poucas orientações em outros níveis.

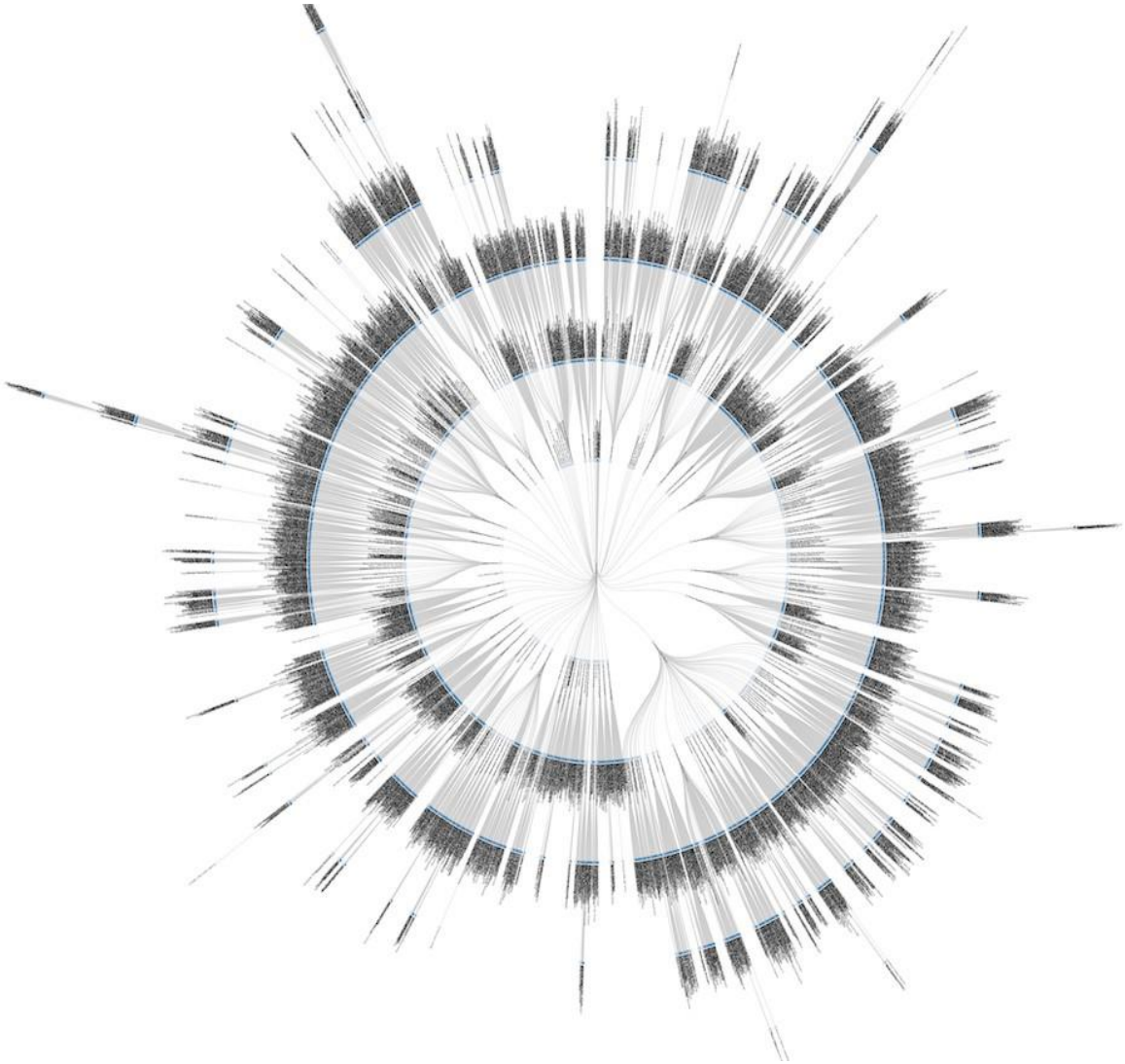


Figura 25: Árvore genealógica do orientador com a maior quantidade de descendentes

Fonte: Próprio Autor

Diferentemente das árvores anteriores, esta última é a única com supervisões de pós-doutorado, o que pode influenciar no tamanho da rede, visto que orientados de pós-doutorado já possuem doutorado concluído, e, consequentemente, já podem estar atuando em programas de pós-graduação a mais tempo que os demais.

Até o momento foram exibidas árvores genealógicas dos orientadores com relação à quantidade de orientados diretos, tamanhos das árvores (com base nos descendentes) e quantidade de gerações. Além disso, diversas outras métricas baseadas em redes podem

ser aplicadas para melhor caracterização das mesmas, conforme é mostrado na próxima seção.

5.4 Florestas de Árvores de Genealogia Acadêmica

Nessa subseção é apresentada a caracterização da floresta com as 100 maiores árvores identificadas nos currículos da Plataforma Lattes, conforme processo descrito anteriormente.

Logo, foram consideradas as maiores árvores tendo em vista a quantidade de indivíduos. Diante disso, considerando os nós raízes dessas árvores, é possível visualizar a sua distribuição pelas grandes áreas de atuação dos orientadores (Figura 26).

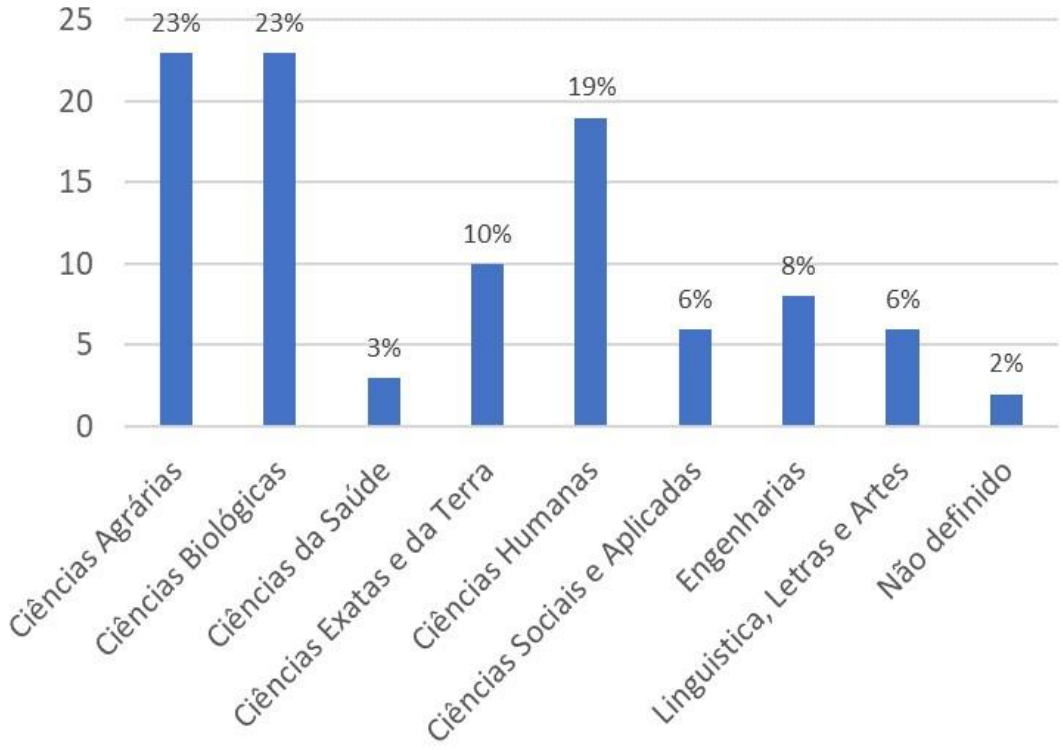


Figura 26: Distribuição dos nós raízes por grandes áreas de atuação

Fonte: Próprio Autor

Conforme pode ser observado, a grande área de Ciências Agrárias e Ciências Biológicas se destacam positivamente, sendo detentoras de 46% das principais raízes das 100 maiores árvores identificadas. Com destaque ainda para a grande área de Ciências Humanas, com considerável representatividade. Já a grande área de Ciências da Saúde possui pouca representatividade (3%), a exemplo das principais raízes que não especificaram grande área de atuação (2%).

A distribuição mostrada na Figura 26 pode estar relacionada ao fato de que algumas áreas técnicas, como, por exemplo, Engenharias, possuírem profissionais em sua maioria de perfis técnicos, diferentemente de outras áreas, onde a atuação é basicamente acadêmica

e as orientações, principalmente em níveis mais altos de capacitação, são, em geral, atividades frequentes.

Além da distribuição dos nós raízes das principais árvores identificadas pelas grandes áreas do conhecimento, também foi possível identificar a sua distribuição por estados da federação (Figura 27).

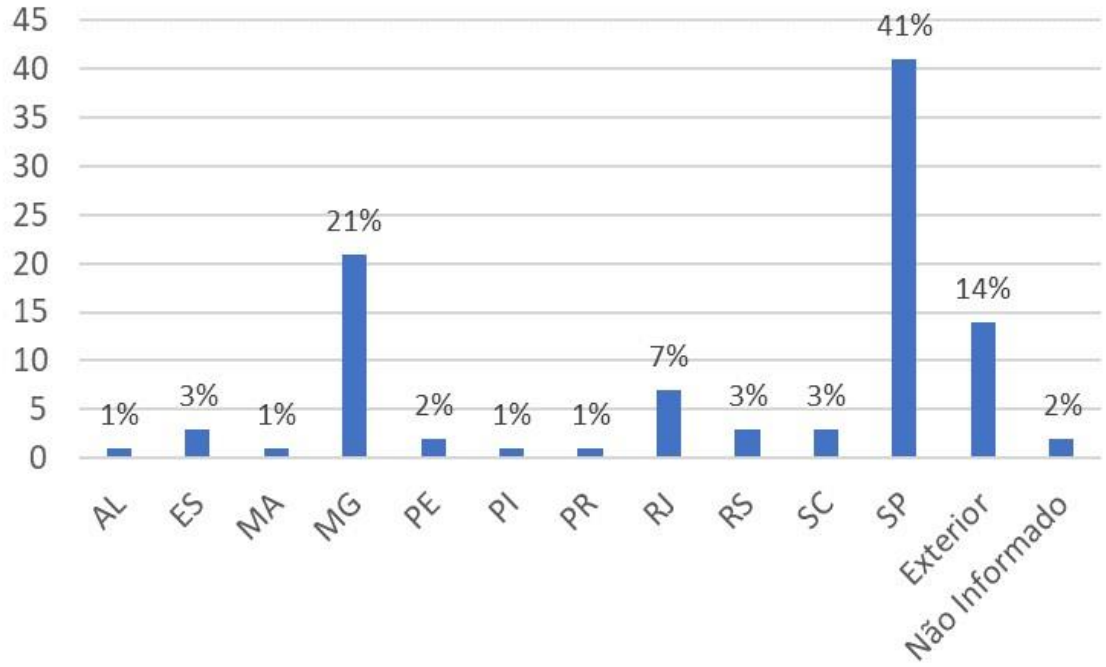


Figura 27: Distribuição por estado de origem dos indivíduos da floresta

Fonte: Próprio Autor

Observa-se que os estados de SP, MG e RJ são os estados com a maior representatividade dentre as principais raízes das árvores identificadas. Tal fato pode estar relacionado à concentração das principais instituições de ensino públicas do país, onde se agrupam os principais cursos de pós-graduação. Tais cursos, por serem tradicionais em suas respectivas instituições, possuem longo histórico de formação, resultando em representativa propagação de conhecimento dos seus percursores.

Percebe-se ainda que estados que compõem as regiões Norte e Centro-Oeste não possuem raízes nas 100 maiores árvores que compõem a floresta. Uma explicação para esta baixa representatividade está relacionada a pouca quantidade de cursos de Pós-graduação nas instituições destes estados.

Além dessa distribuição por estados da federação de origem dos nós raízes das principais árvores identificadas, também foi possível verificar que existe uma quantidade considerável de pesquisadores que são de origem estrangeira (Tabela 10).

Tabela 10: Orientadores estrangeiros (tendo como base o país de nascimento) com maior quantidade de orientações na floresta

Posição entre as 100 maiores árvores	Cidade	País	Grande Área	Área
4	New York	Estados Unidos	Ciências Biológicas	Bioquímica
11	Quito	Equador	Ciências Agrárias	Agronomia
17	Basel	Suíça	Ciências Biológicas	Bioquímica
24	Budapeste	Hungria	Ciências da Saúde	Medicina
26	Napoles	Itália	Ciências Biológicas	Bioquímica
28	Bahia Blanca	Argentina	Ciências Agrárias	Zootecnia
33	Callalli	Peru	Ciências Agrárias	Agronomia
41	Leiha	Índia	Ciências Agrárias	Agronomia
58	Não Informado	Áustria	Ciências Sociais e Aplicadas	Economia
66	Buenos Aires	Argentina	Ciências Biológicas	Farmacologia
67	Santiago	Chile	Ciências Exatas e da Terra	Química
69	Wazirabad	Paquistão	Engenharias	Engenharia Civil
73	Nova York	Estados Unidos	Linguística, Letras e Artes	Linguística
93	Marietta	Estados Unidos	Ciências Biológicas	Ecologia

Dentre as 100 maiores árvores identificadas, 14 raízes dessas árvores são de orientadores estrangeiros. Destes, destaca-se os Estados Unidos da América, seguido em menor quantidade pela Argentina. Analisando estes orientadores estrangeiros, percebe-se que tal fato está relacionado a capacitação de pesquisadores de excelência no Brasil, que foram orientados por pesquisadores destes países. Ressalta-se, ainda, que deste conjunto um grande percentual dos orientadores que são raízes das principais árvores, tem atuação nas grandes áreas de Ciências Biológicas e Ciências Agrárias.

Verificando-se os nós raízes das 10 maiores árvores, é possível observar as cidades, estados e países de origem dos maiores orientadores, bem como suas grandes áreas e áreas de atuação (Tabela 11).

Tabela 5: Maiores raízes da floresta

Descendentes	Gerações	Cidade	Estado	País	Grande Área	Área
8.318	6	Santo Antônio de Posse	SP	Brasil	Ciências Humanas	Educação
7.162	10	São Paulo	SP	Brasil	Ciências Biológicas	Genética
7.114	14	Itápolis	SP	Brasil	Ciências Agrárias	Agronomia
6.341	14	New York	-	Estados Unidos	Ciências Biológicas	Bioquímica
5.716	15	Rio de Janeiro	RJ	Brasil	Ciências Biológicas	Bioquímica
5.442	15	Rio de Janeiro	RJ	Brasil	Ciências Biológicas	Imunologia
5.426	15	Santa Bárbara	MG	Brasil	Ciências Biológicas	Bioquímica
4.860	13	Oliveira	MG	Brasil	Ciências Agrárias	Agronomia
4.424	8	Morada Nova de Minas	MG	Brasil	Ciências Exatas e da Terra	Química
4.379	10	Araguari	MG	Brasil	Ciências Biológicas	Genética

A grande área de Ciências Biológicas possui grande representatividade com 6 orientadores (raízes) dentre as 10 maiores árvores identificadas. Dentre estas, destacam-se as áreas de Bioquímica, Genética e Imunologia. Ressalta-se, também, a área de Agronomia e Química,

além da área de Educação, que contém o nó raiz da maior árvore caracterizada, contendo 8.318 descendentes.

Verificando a quantidade de gerações das 10 maiores árvores, percebe-se que a grande área de Ciências Biológicas possui os orientadores mais representativos, que apresentam a mesma quantidade de gerações identificadas em todo o conjunto de currículos cadastrados na Plataforma Lattes (15 gerações).

Com isso, um estudo importante aqui apresentado é verificar a relação entre a quantidade de gerações e a quantidade de descendentes das maiores árvores (Figura 28).

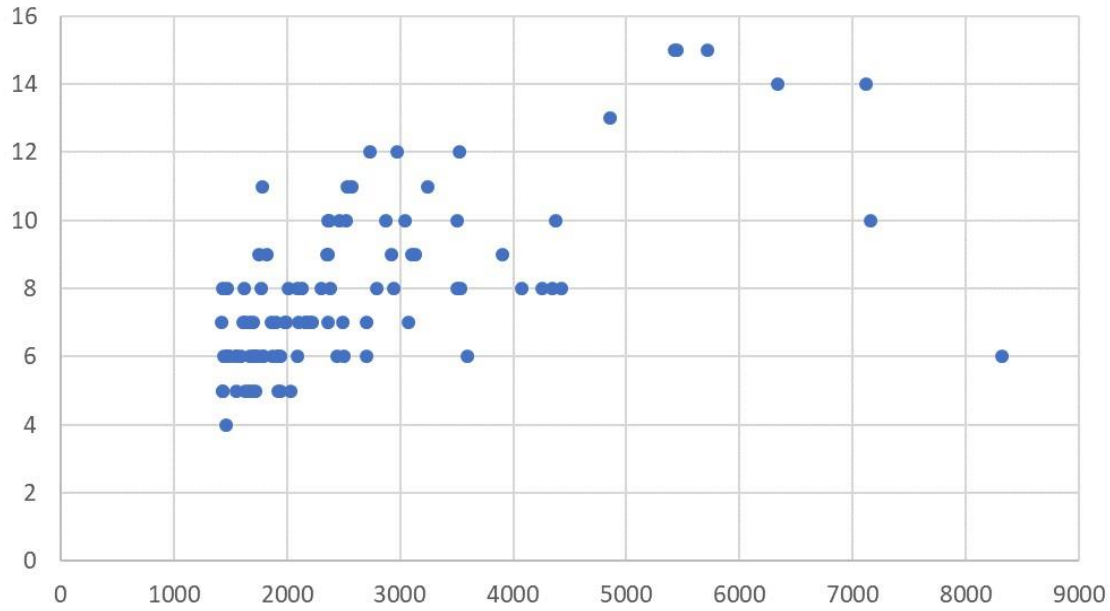


Figura 28: Relação entre o tamanho das árvores e quantidade de gerações

Fonte: Próprio Autor

Percebe-se que, em geral, existe uma correlação direta entre a quantidade de gerações e a quantidade de indivíduos nas maiores árvores identificadas (Coeficiente de Pearson de 0,62), em que as árvores com as maiores quantidades de gerações possuem grande quantidade de indivíduos. Consequentemente, dentre o conjunto analisado, as árvores com as menores quantidades de gerações possuem pequena quantidade de descendentes.

Ao se analisar as 100 maiores árvores identificadas pelo método proposto neste trabalho, foi possível caracterizar a floresta que engloba os maiores orientadores do conjunto analisado (Figura 29). Observa-se como estão conectados todos os indivíduos que integram essas 100 maiores árvores. Destacam-se, pelas cores, as grandes áreas de atuação de cada indivíduo, bem como suas interconexões.

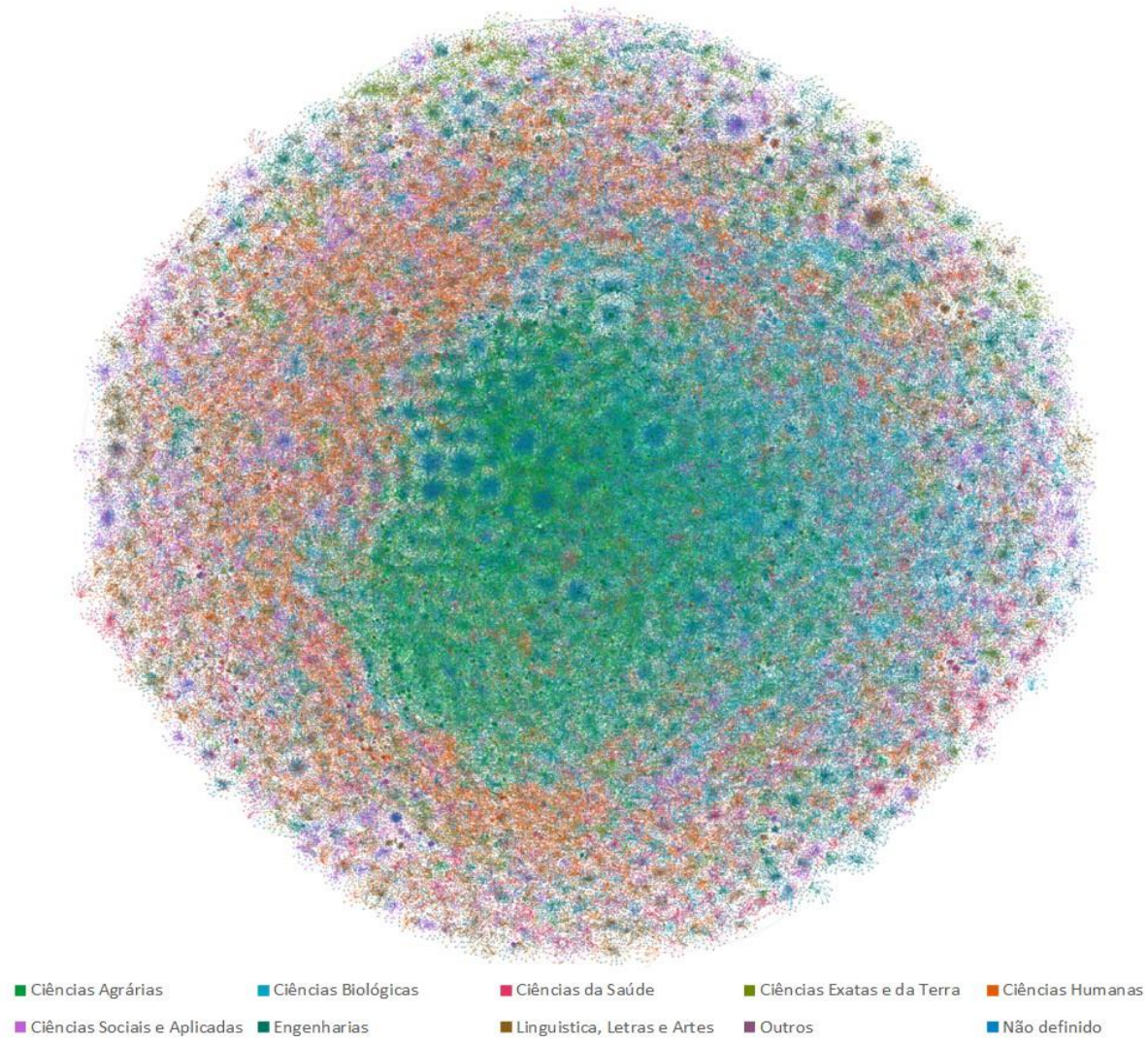


Figura 29: Floresta com as 100 maiores árvores identificadas

Fonte: Próprio Autor

Diante de tal caracterização, foi possível aplicar métricas de análise de redes para compreender como está estruturada a floresta que contempla os principais orientadores no Brasil (Tabela 12).

Tabela 6: Métricas da floresta caracterizada

Descrição	Valor
Nós	194.188
Arestas	206.573
Grau Médio	1,064
Diâmetro da Rede	15
Densidade do Grafo	0,00000548
Comprimento do caminho Médio	3,752
Quantidade de nós na componente gigante	100%
Quantidade de arestas na componente gigante	100%

A floresta aqui analisada é responsável por englobar aproximadamente 3,8% de todos os indivíduos com currículos cadastrados na Plataforma Lattes. Este conjunto possui um total de 206.573 conexões, que neste estudo significam orientações, apresentando, como era de se esperar, uma pequena densidade da floresta (0,00000548) e grau médio próximo de 1 (1,064). O diâmetro da rede é igual a 15, mas com um caminho mínimo médio igual a 3,75, o que significa que dentre o conjunto analisado, um determinado indivíduo está em média muito próximo aos outros.

Destaca-se, ainda, que a componente gigante da floresta é responsável por englobar todos os indivíduos selecionados. Ou seja, ao considerar as 100 maiores árvores, nenhum indivíduo que as compõem está desconectado.

Diante do exposto, percebe-se que, de certa maneira, ao verificar as maiores árvores identificadas, elas estão interconectadas, resultando em uma floresta que contém uma única componente conexa. Tal fato está relacionado ao fato de que os indivíduos nos últimos níveis das árvores possuíram algum tipo de orientação com indivíduos de outras árvores.

Os resultados aqui apresentados ressaltam a importância da genealogia acadêmica para uma visão do processo de orientação. Inicialmente foram apresentadas análises estatísticas gerais, em que os registros de orientações obtidos através dos currículos cadastrados na Plataforma Lattes foram caracterizados, com destaque para a queda na quantidade de orientações cadastradas nos últimos anos, tendo como um possível motivo para tal a falta de atualização dos currículos.

Posteriormente as orientações foram distribuídas por grandes áreas, demonstrando características importantes de cada uma, com destaque para Ciências Sociais Aplicadas e Ciências Humanas, com grandes quantidades de orientações.

Por fim, foram produzidas árvores genealógicas e a floresta das 100 maiores árvores genealógicas, onde foram exibidas as árvores que se destacam pela quantidade de indivíduos que a compõem e pela quantidade de gerações. Para tais resultados, o desambiguador proposto foi imprescindível, já que possibilitou que outras árvores, antes isoladas, passassem a compor as maiores raízes encontradas, apresentando resultados inéditos sobre a genealogia acadêmica da pós-graduação brasileira.

Capítulo 6

Considerações Finais

Buscando um maior entendimento sobre como tem se desenvolvido o processo de orientação acadêmica brasileira, este trabalho busca caracterizar as orientações acadêmicas, a partir de análises estatísticas e baseadas em análises de redes sociais realizadas sobre dados de registros de orientações acadêmicas dos currículos cadastrados na Plataforma Lattes.

Ao analisar os dados de maneira geral, constata-se que a grande maioria das orientações estão localizadas na região Sudeste, principalmente nos estados de São Paulo, Minas Gerais e Rio de Janeiro, nessa ordem, seguidos pelos estados da região sul, o que já era esperado, já que a grande maioria dos cursos de pós-graduação estão localizados nestas regiões.

A grande maioria das orientações cadastradas na Plataforma Lattes são de cursos de Graduação, seguida por Especialização e Iniciação Científica, o que é compreensível, já que os cursos de Graduação, Iniciação Científica e Especialização são maioria no país. Porém, por outro lado, nota-se que a quantidade de orientações concluídas está em queda em todos os níveis de capacitação. Uma das possíveis explicações encontradas para isso é a falta de atualização dos currículos pelos seus responsáveis, não descartando o possível impacto da atual situação econômica do país. O que justifica essa afirmação é que, orientações de Pós-graduação (principalmente de doutorados, onde a queda foi a menor) tiveram menores quedas, onde, por exigência dos programas de pós-graduação, orientadores tendem a manter seus currículos atualizados.

Os resultados obtidos foram distribuídos por grandes áreas do conhecimento, para assim caracterizar cada uma delas. Neste contexto, destaca-se a área de Ciências Sociais Aplicadas, seguida por Ciências Humanas e Ciências da Saúde. Porém, apesar da maior quantidade de orientações da grande área Ciências Sociais Aplicadas, a grande maioria pertence a cursos de Graduação, tendo outras áreas maior destaque em Pós-graduação, como, por exemplo, Ciências Humanas e Ciências da Saúde.

Durante a análise, foram identificados orientadores com quantidades de orientação acima do normal. Diante disso, ao ignorar estes outliers, devido à distorção que estavam causando, observou-se que a grande área mais representativa continua sendo Ciências Sociais Aplicadas, mas, desta vez, seguida muito de perto por Ciências Agrárias.

Tendo em vista a necessidade de caracterizar as árvores genealógicas, as orientações cadastradas na Plataforma Lattes foram ranqueadas por quantidade de orientações diretas, gerações e tamanho da rede. Neste caso, foram utilizados apenas registros orientações de pós-graduação identificados nos currículos, o que reduz o tamanho das árvores e facilita a caracterização.

Além da árvore genealógicas do orientador com a maior quantidade de orientados diretos, da árvore com a maior quantidade de gerações e da maior árvore encontrada, foi produzida uma floresta com as 100 maiores árvores encontradas na Plataforma. A partir disso, foi possível caracterizar e aplicar métricas de análise baseadas em rede para obter informações relevantes sobre o conjunto.

Inicialmente, as raízes da rede foram distribuídas por estados. Como era de se esperar, poucos estados possuem raízes entre as 100 raízes mais bem ranqueadas, ficando restrito aos estados do Sudeste, do Sul e do Nordeste, nesta ordem. Além destes estados, foram encontradas 14 raízes de orientadores estrangeiros orientando alunos brasileiros, com destaque para os Estados Unidos e a Argentina, sendo que um dos orientadores de origem americana está entre as 4 maiores árvores.

Analisando o número de gerações e tamanho da rede, foi possível perceber que existe correção direta entre os dois valores, demonstrando que a quantidade de gerações possui impacto no incremento da quantidade de indivíduos da árvore (tamanho da rede), mesmo em raízes com poucas orientações diretas.

Por fim, foi produzida a rede com as 100 maiores raízes identificadas, onde foi aplicada diversas métricas baseadas em análise de redes. Uma métrica com resultado interessante foi a quantidade de arestas na componente gigante, onde foi constatado que todos os nós da floresta estão em uma única componente. Ou seja, a partir de qualquer indivíduo na rede é possível chegar em qualquer outro.
Diante dos resultados alcançados, pode-se concluir que caracterizar árvores genealógicas acadêmicas é importante se obter uma visão geral da relação orientador-orientado, além de toda a herança intelectual, passada de orientador para orientado. Além disso, a possibilidade de mensurar os dados de genealogia trazem resultados importantes para visualizar a influência do orientador na formação acadêmica brasileira. Além disso, este trabalho utilizou dados dos currículos cadastrados na Plataforma Lattes, demonstrando a aplicabilidade da base para os mais variados tipos de análises, podendo ser ainda mais explorada a fim de obter um aprofundamento da pesquisa da relação orientador-orientado.

6.1 Publicações

A seguir, são apresentadas as produções científicas elaboradas a partir dos estudos realizados neste trabalho.

Como resultado das pesquisas realizadas a partir dos registros de orientações obtidos através dos currículos da Plataforma Lattes, foram produzidos e publicados os seguintes artigos em periódicos:

- MOREIRA, T. H. J. et al. Uma estratégia baseada em Text Mining para compreensão do processo de orientação em repositórios de dados científicos. **JOURNAL ON ADVANCES IN THEORETICAL AND APPLIED INFORMATICS**, v. 3, p. 84–88, 2017.

- MOREIRA, T. H. J. et al. Academic genealogy: a hierarchical analysis of advisors researchers. **INTERNATIONAL JOURNAL OF DEVELOPMENT RESEARCH**, v. 7, p. 12371–12375, 2017.

- MOREIRA, T. H. J.; DIAS, T. M. R.; MOITA, G. F. UMA VISÃO GERAL DO PROCESSO DE ORIENTAÇÃO E DA PRODUÇÃO CIENTÍFICA DA RELAÇÃO ORIENTADOR-ORIENTADO NAS ÁREAS DE ENGENHARIAS. **Revista Interdisciplinar de Pesquisa em Engenharia RIPE**, v. 2, n. 10, p. 61–70, 2017.

Além disso, os seguintes artigos foram publicados em anais de congressos:

- MOREIRA, T. H. J. et al. Genealogia Científica: Uma Análise Hierárquica De Pesquisadores Orientadores. **Proceedings of the XI SIMMEC (Simpósio de Mecânica Computacional)**, Juiz de Fora, MG, Brazil, p. 9, 2014.

- MOREIRA, T. H. J.; DIAS, T. M. R.; MOITA, G. F. Uma estratégia baseada em árvores genealógicas científicas para visualização da relação orientador-orientado. In: USP (Ed.). **Encontro Brasileiro de Bibliometria e Cientometria**. São Paulo: [s.n.], 2016. p. A91—A91.

- MOREIRA, T. H. J.; DIAS, T. M. R.; MOITA, G. F. UMA VISÃO GERAL DO PROCESSO DE ORIENTAÇÃO E DA PRODUÇÃO CIENTÍFICA DA RELAÇÃO ORIENTADOR-ORIENTADO NAS ÁREAS DE ENGENHARIAS. In: **XXXVII Iberian Latin American Congress on Computational Methods in Engineering**. Brasília, DF: [s.n.], 2016.

- MOREIRA, T. H. J.; DIAS, T. M. R.; MOITA, G. F. VISÃO GERAL DO PROCESSO DE GENEALOGIA ACADÊMICA DE ALUNOS ORIENTADOS EM CURSOS DE MODELAGEM COMPUTACIONAL. In: UFPB (Ed.). **XIX Encontro Nacional de Modelagem Computacional**. João Pessoa, PB: [s.n.], 2016.

- MOREIRA, Tales HJ; DIAS, Thiago MR; MOITA, Gray F. Genealogia Acadêmica da Relação Orientador-Orientado na Área de Ciência da Computação. **Anais do Computer on the Beach**, p. 060-069, 2017.

- MOREIRA, Tales HJ et al. A text mining based strategy for understanding the process of orientation in scientific data repositories. In: **Information Systems and Technologies (CISTI)**, 2017 12th Iberian Conference on. IEEE, 2017. p. 1-5.

- MOREIRA, Tales HJ; DIAS, Thiago MR; MOITA, Gray F. An Overview of the Advice Process and the Scientific Production of the Adviser-Advised Relationship in the Areas of Engineering. **World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering**, v. 11, n. 10, p. 2265-2269, 2017.

6.2 Complementação da Pesquisa

Como apresentado anteriormente, inicialmente foi apresentada uma visão geral do conjunto de indivíduos considerando suas orientações concluídas. Posteriormente, foram realizadas análises com base árvores genealógicas caracterizadas. No entanto, tais caracterizações são apenas uma visão geral de todo o conjunto de indivíduos, sendo necessário um recorte deste conjunto para análises mais específicas.

A partir dos dados obtidos, é possível obter diversos tipos de informações que possibilitam diferentes análises e resultados. Diante disso, além da caracterização realizada até o momento, pode-se utilizar os dados de produção bibliográfica presentes nos currículos cadastrados na Plataforma Lattes para obter uma visão geral da produção científica da relação entre orientador e orientado, além de mensurar a produtividade das relações de orientação.

Diversos outros resultados podem ser obtidos, devido a quantidade e variedade de informações que podem ser extraídas dos currículos cadastrados na Plataforma Lattes para analisar e mensurar de forma quantitativa e qualitativa a ciência brasileira.

Além disso, uma comparação direta de entre os desambiguadores existentes e o método proposto para desambiguação de toda a Plataforma Lattes é de extrema importância, visando obter resultados com a maior precisão possível, além de um possível aprimoramento do método.

Por fim, a integração de dados obtidos através de agências de fomento com dados de orientação pode ser um interessante objeto de estudo, aprofundando os resultados aqui apresentados.

Referências

ALVES, A.; YANASSE, H.; SOMA, N. LattesMiner: a Multilingual DSL for Information Extraction from Lattes Platform. In: **Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE!'11, AOOPES'11, NEAT'11, & VMIL'11**. [s.n.], 2011. ISBN 9781450311830. Disponível em: <http://www.dsmforum.org/events/dsm11/Papers/alves.pdf>.

ARAÚJO, C. A. A. Bibliometria: evolução histórica e questões atuais. **Em questão**, v. 12, n. 1, 2006.

BELÉM, D. et al. Quem são e onde atuam os turismólogos brasileiros com título de doutor? **Revista Turismo & Desenvolvimento**, v. 1, n. 27/28, p. 1865–1875, 2018.

Carvalho, A. P., Ferreira, A. A., Laender, A. H. F., and Gonçalves, M. A. Incremental unsupervised name disambiguation in cleaned digital libraries. **Journal of Information and Data Management**, 2(3):289–304, 2011.

CHRISTEN, P. **Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection**. [S.l.]: Springer Data-Centric Systems and Applications, 2012. 1–279 p. ISSN 2197-9723. ISBN 978-3-642-31163-5.

CNPQ. **Sobre a plataforma Lattes**. 2017. Disponível em: <http://lattes.cnpq.br/>.

DA SILVA, L. C. R.; DO NASCIMENTO, H. A. D. **VISUALIZANDO BASES CURRICULARES**. 2006.

DELGADO, K. V. et al. Análise Comparativa da Produtividade dos Pares OrientadorOrientado em Ciência da Computação. **Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI**, v. 14, n. 1, 2015.

DIAS, T. M. et al. Modelagem e caracterização de redes científicas: um estudo sobre a plataforma lattes. In: **BRASNAM-II Brazilian Workshop on Social Network Analysis and Mining**. [S.l.: s.n.], 2013. p. 10–20.

DIAS, T. M. R. **Um Estudo Sobre a Produção Científica Brasileira a Partir de Dados da Plataforma Lattes**. 181 p. Tese (Doutorado) — Centro Federal de Educação Tecnológica de Minas Gerais, 2016.

DIAS, T. M. R. et al. Modelagem e Caracterização de Redes Científicas: Um Estudo Sobre a Plataforma Lattes. In: **BRASNAM - II Brazilian Workshop on Social Network Analysis and Mining**. [S.l.: s.n.], 2013. p. 10–20.

DIAS, T. M. R. et al. Identificação e Caracterização de Redes Científicas de Dados Curriculares. **iSys-Revista Brasileira de Sistemas de Informação**, v. 7, n. 3, p. 5–18, 2014.

DIGIAMPIETRI, L. A. et al. Análise da Rede dos Doutores que Atuam em Computação no Brasil. 2014.

DIGIAMPIETRI, L. A. et al. Minerando e Caracterizando Dados de Currículos Lattes. In: **I Brazilian Workshop on Social Network Analysis and Mining (BrasNAM)**. [S.l.: s.n.], 2012. p. 12.

DORES, W.; LAENDER, A. H. F. 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 - JCDL '16**. New York, New York, USA: ACM Press, 2016. p. 163–166. ISBN 9781450342292. Disponível em: <http://dl.acm.org/citation.cfm?doid=2910896.2910916>.

ELIAS, M.; FLOETER-WINTER, L. M.; MENA-CHALCO, J. P. The dynamics of brazilian protozoology over the past century. **Memórias do Instituto Oswaldo Cruz**, SciELO Brasil, v. 111, n. 1, p. 67–74, 2016.

FERRAZ, R. R. N. et al. Análise e gestão de análise de redes de colaboração entre pesquisadores de programas de pós-graduação stricto sensu com a utilização da ferramenta computacional scriptlattes. **Perspectivas em Gestão & Conhecimento**, v. 4, p. 133–147, 2014.

FERREIRA, L. M.; FURTADO, F.; SILVEIRA, T. S. Relação Orientador-Orientando. O Conhecimento Multiplicador. **Acta Cirúrgica Brasileira**, v. 24, n. 3, p. 170–172, 2009.

GOFFMAN, W.; NEWILL, V. A. Generalization of epidemic theory. **Nature**, v. 204, n. 4955, p. 225–228, 1964.

GUEDES, V. L. S.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. **Encontro Nacional de Ciência da Informação**, v. 6, p. 1–18, 2005.

HAYASHI, C. P. I. Sociologia da Ciência, Bibliometria e Cientometria: Contribuições para a Análise da Produção Científica. In: **SEMINÁRIO DE EPISTEMOLOGIA E TEORIAS DA EDUCAÇÃO**. [S.l.: s.n.], 2013.

LANE, J. Let's make science metrics more scientific. **Nature**, Nature Publishing Group, v. 464, n. 7288, p. 488–489, 2010.

Lee, D., Kang, J., Mitra, P., Giles, C. L., and On, B. W. Are your citations clean? **Communications of the ACM**, 50(21): 3338, 2007.

Leite Filho, G. A.; MARTINS, G. D. A. Relação orientador-orientando e suas influências na elaboração de teses e dissertações. **Revista de Administração de Empresas**, SciELO Brasil, v. 46, n. spe, p. 99–109, 2006.

LOPES, G. R. Avaliação e Recomendação de Colaborações em Redes Sociais Academicas. **Programa de Pós-Graduação em Computação, UFRS, Rio Grande do Sul**, 2012.

MARTELETO, R. M.; SILVA, A. B. d. O. Redes e capital social: o enfoque da informação para o desenvolvimento local. **Ciência da informação**, SciELO Brasil, v. 33, n. 3, p. 41–49, 2004.

McKay, D., Sanchez, S. and Parker, R. What's my name again?: sociotechnical considerations for author name management in research databases. In: **Proceedings of the 22nd Conference of the ComputerHuman Interaction Special Interest Group of Australia on ComputerHuman Interaction**. Brisbane, Australia, 2010.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JR, R. M. Caracterizando as redes de coautoria de currículos Lattes. In: **BraSNAM - Brazilian Workshop on Social Network Analysis and Minig**. [S.l.]: Brazilian Computing Society, 2012. p. 12.

MENA-CHALCO, J. P.; JUNIOR, R. M. C. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009.

MENEZES, V. S. d. A. Análise de Redes Sociais Científicas. **COPPE/UFRJ**, 2012.

MIYAHARA, E. K. **Genealogia Acadêmica Lattes**. Tese (Doutorado) — Universidade de São Paulo, 2011.

MUGNAINI, R.; DIGIAMPIETRI, L. A. Comunicação científica no Brasil (1998-2012): indexação , crescimento , fluxo e dispersão. **Transinformação-ISSN 2318-0889**, v. 26, n. January, 2014. ISSN 01033786.

MUGNAINI, R.; JANNUZZI, P. d. M.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. **Ciência da Informação**, SciELO Brasil, v. 33, n. 2, p. 123–131, 2004.

MUGNAINI, R.; LEITE, P.; LETA, J. Fontes de informação para análise de internacionalização da produção científica brasileira. **PontodeAcesso**, v. 5, n. 3, p. 87–102, 2012.

NICHOLAS, D.; RITCHIE, M. **Literature and bibliometrics**. [S.l.]: C. Bingley, 1978. ISBN 0208015418.

PRITCHARD, A. Statistical bibliography or bibliometrics. **Journal of documentation**, v. 25, p. 348, 1969.

PUPIM, L. et al. Pppg o1o-atividades profissionais e acadêmicas dos egressos da pósgraduação em odontologia na área de saúde coletiva. **ARCHIVES OF HEALTH INVESTIGATION**, v. 6, 2017.

RAMOS, L. S.; RAZERA, J. C. C. Um perfil do conjunto de pesquisadores sobre formação de professores na área brasileira de educação em ciências. **Enseñanza de las ciencias**, n. Extra, p. 2951–2958, 2017.

ROSSI, L.; MENA-CHALCO, J. P. Aos ombros de gigantes: um estudo de genealogia acadêmica dos matemáticos no Brasil. In: **Simpósio de Pesquisa do Grande ABC (SPGABC)**. [S.l.: s.n.], 2014. p. 1–2.

ROSSI, L.; MENA-CHALCO, J. P. Índice-h genealógico expandido: Uma medida de impacto em grafos de orientação acadêmica. In: **IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015)**. [S.l.: s.n.], 2015. p. 12.

SANTOS, L. R. F. **Utilização de Dados da Plataforma Lattes para a Avaliação da Distribuição da Bolsa de Produtividade em Pesquisa do CNPq**. 82 p. Dissertação (Mestrado) — Centro Federal de Educação Tecnológica de Minas Gerais, 2015.

SANTOS, L. R. F. et al. Identification of the profile of productivity researchers in CNPq research using lattes platform. **INTERNATIONAL JOURNAL OF DEVELOPMENT RESEARCH**, v. 7, p. 12366–12370, 2017.

SANTOS, S. M. dos. **Perfil dos periódicos científicos de Ciências Sociais e de Humanidades: mapeamento das caracteríticas extrínsecas**. [S.l.]: Universidade de São Paulo, 2010.

SCOTT, J. **Social network analysis**. [S.l.]: Sage, 2012. ISBN 1446259455.

ScriptLattes. **scriptLattes: Uma ferramenta para extração e visualização de conhecimento a partir de Currículos Lattes**. 2018. Disponível em: <http://scriptlattes.sourceforge.net/>.

SUGIMOTO, C. R. Academic genealogy. In: **Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact**. MIT Press, 2014. p. 365–382.

SZWARCFITER, J. L. Grafos e Algoritmos Computacionais.[SI]: Editora Campus Ltda. **Rio de Janeiro-1984**, 1984.

TUESTA, E. F. et al. Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação. In:

Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). [S.l.: s.n.], 2012. p. 11.

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**. [S.l.]: Cambridge university press, 1994. v. 8. ISBN 0521387078.