XIX encontro nacional de pesquisa em ciência da informação
ENANCIB
// SUJEITO INFORMACIONAL E AS PERSPECTIVAS ATUAIS EM CIÊNCIA DA INFORMAÇÃO. //
22-26 OUTUBRO 2018
LONDRINA/PR
Universidade Estadual de Londrina

XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018

GT-8 – Informação e Tecnologia

UMA ESTRATÉGICA PARA A CARACTERIZAÇÃO E ANÁLISE DE ÁRVORES GENEALÓGICAS ACADÊMICAS EM GRANDES REPOSITÓRIOS DE DADOS

Tales Henrique José Moreira – Centro Federal de Educação Tecnológica de Minas Gerais

Thiago Magela Rodrigues Dias – Centro Federal de Educação Tecnológica de Minas Gerais

Patrícia Mascarenhas Dias – Centro Federal de Educação Tecnológica de Minas Gerais

Gray Farias Moita – Centro Federal de Educação Tecnológica de Minas Gerais

A STRATEGIC FOR THE CHARACTERIZATION AND ANALYSIS OF ACADEMIC GENEALOGICAL TREES IN LARGE DATA REPOSITORIES

Modalidade da Apresentação: Comunicação Oral

Resumo: A relação orientador-orientado, que pode se caracterizar como uma forma de propagação do conhecimento, como resultado pode proporcionar um aumento na produção científica dos orientadores e orientados. Especificamente em programas de pós-graduação, onde orientados submetem artigos científicos em diversos meios de publicação em coautoria com seus orientados, estes acabam por impulsionar a produção científica de seu orientador, já que em geral os orientadores surgem como coautores das publicações. Portanto, caracterizar como ocorre o processo de orientação e a produção científica 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 intuito de obter-se, além de uma visão geral do processo de orientação no Brasil, uma análise advinda desse processo de orientação, identificando dessa forma como o conhecimento científico Brasileiro tem se propagado. 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.

Palavras-Chave: Genealogia Acadêmica; Plataforma Lattes; Orientação.

Abstract: The guiding-oriented relationship, which can be characterized as a way of propagating knowledge, as a result can provide an increase in the scientific output of guiding and oriented. Specifically, in postgraduate programs, where the subjects submit scientific articles in several media of co authorship with their advised ones, these end up impelling the scientific production of its advisor, since in general the advisors appear as coauthor of the publications. Therefore, characterizing how the orientation process occurs and the scientific production resulting from this relationship is another important means of analyzing scientific collaboration in the various areas of knowledge. In this work, the data of guidelines extracted from the curricula registered in the Lattes Platform are used in order to obtain, besides an overview of the orientation process in Brazil, an analysis derived from this orientation process, thus identifying the knowledge Brazilian science has spread. Seeking a greater understanding of how the process of Brazilian academic orientation has developed in the different areas of knowledge, this research has as general objective to characterize the orientations in postgraduate programs, based on bibliometric analyzes and based on analyzes of social networks carried out curricular data available on the Lattes Platform.

Keywords: Academic Genealogy; Lattes Platform; Orientation.

1 INTRODUÇÃO

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.

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. A presente abordagem é, até então, inédita, tendo em vista a abrangência de indivíduos e de dados analisados.

2 TRABALHOS CORRELATOS

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.

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

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.

Dores e Laender (2016) optaram por utilizar em seu trabalho a base de dados *NDLTD* (*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 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*. 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.

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.

Já Tuesta *et al.* (2012) seguem em uma linha diferente, apresentando 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.

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.

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. Ao analisar todo o conjunto de currículos, propondo para isso técnicas e métodos para processamento de grande volume de dados, uma visão inédita do processo de orientação brasileiro é apresentada, detalhando como ocorre tal processo, inclusive por grandes áreas do conhecimento.

3 DESENVOLVIMENTO

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.

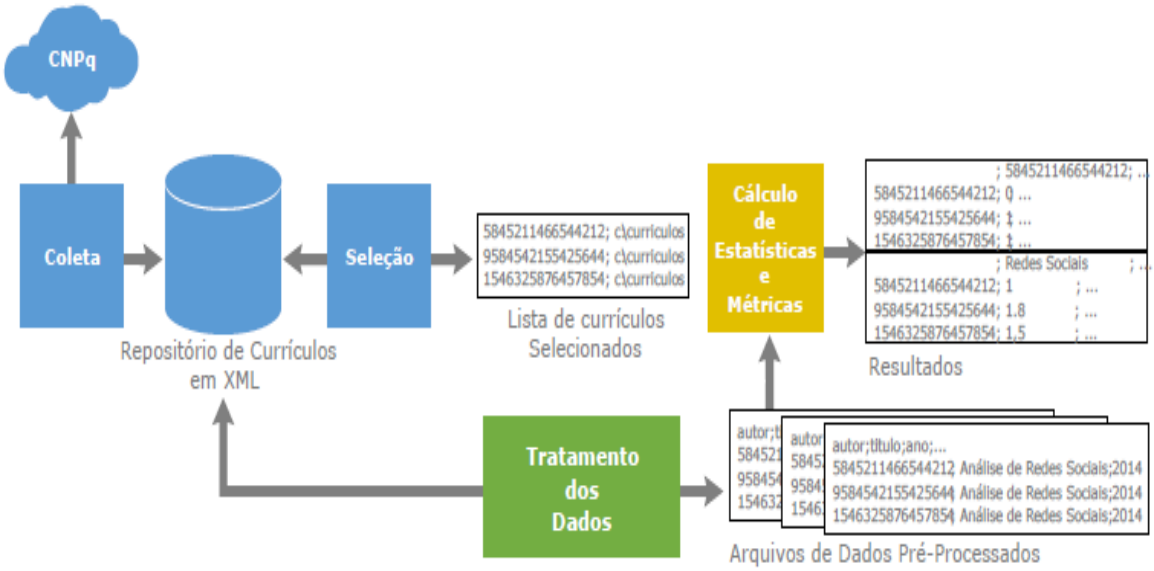
Mugnaini, Leite e Leta (2012) ressaltam que apesar de não possuir indexação e catalogação de periódicos como ocorre em diversas bases de indexação, a Plataforma Lattes é uma fonte inesgotável de informações sobre a ciência brasileira de maneira geral. É destacado ainda pelos autores que, apesar da grande 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.

Este estudo tem como principal fonte de dados os currículos cadastrados na Plataforma Lattes. Os currículos se tornaram um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do país. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia (CNPQ, 2017).

A Plataforma Lattes ganhou nos últimos anos uma projeção internacional, desde então tendo grande relevância para a produção técnico científica brasileira, o que possibilitou diversas parcerias com países da América e Europa, formando a Rede ScienTI (DA SILVA; DO NASCIMENTO, 2006).

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 1).

Figura 1: Visão geral do LattesDataXplorer.



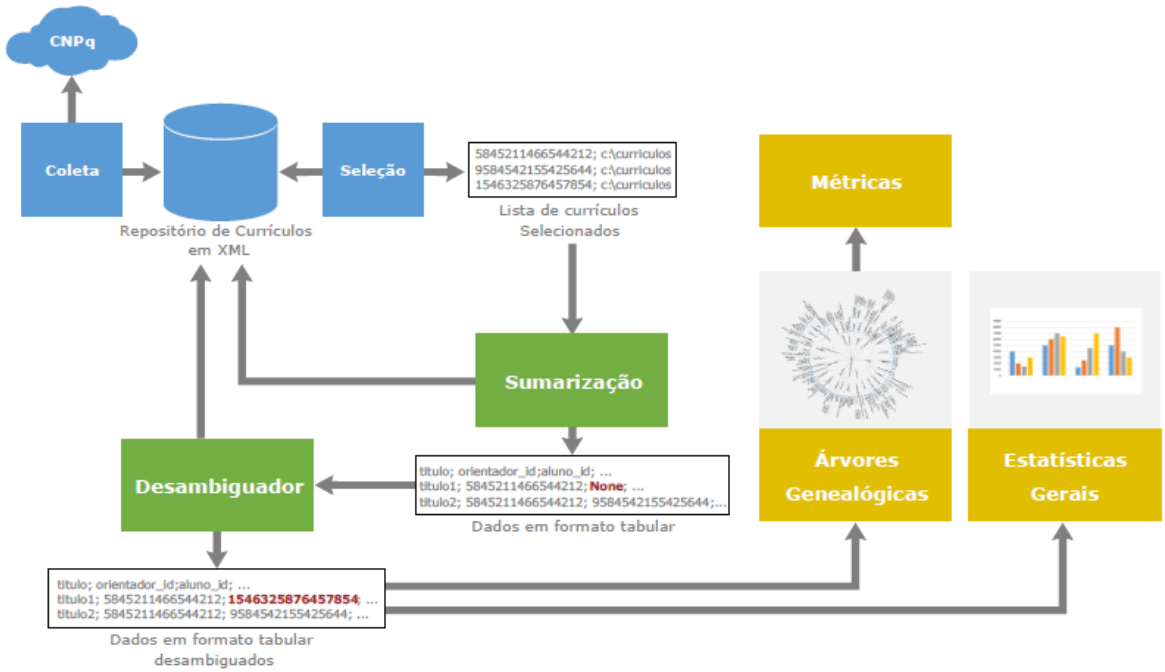
Fonte: DIAS – 2016.

O LattesDataXplorer é responsável por englobar todo o conjunto de técnicas e métodos para a coleta, tratamento e análise dos dados utilizados neste trabalho. Ele é composto por

diversos módulos, responsáveis por todo o processo de coleta e tratamento dos dados (DIAS, 2016). O processo de extração de todos os dados curriculares da Plataforma Lattes é dividido em etapas, necessárias para a obtenção dos currículos.

De posse do arcabouço desenvolvido por Dias (2016), foi realizada uma expansão do mesmo para atender as necessidades do presente estudo (Figura 2).

Figura 2: Visão geral do LattesDataXplorer expandido.



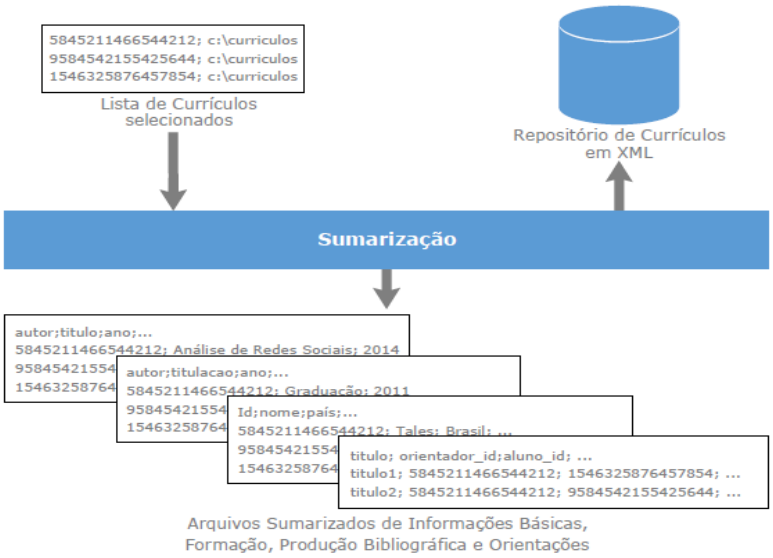
Fonte: Próprio Autor.

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.

De posse dos currículos, os dados podem 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 3. 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 3: Processo de sumarização dos dados.



Fonte: Criado pelo autor.

Os arquivos resultantes de todo o processo são armazenados em formato CSV, simplificando o processamento dos dados. Logo, o processo de ler cada um dos currículos e aplicar uma consulta em XPath (*XML Path Language*) 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.

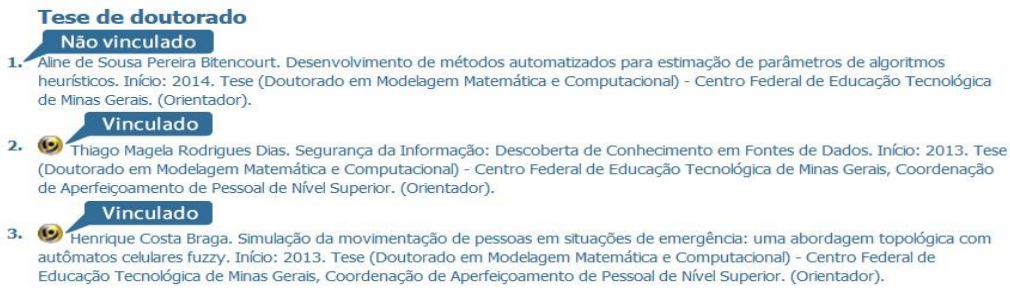
Tabela 1: Resultado da sumarização de todos os currículos analisados.

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

Fonte: Criado pelo autor.

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 4). Diante disso, uma estratégia de identificação de relacionamento se faz necessária para que se possa caracterizar redes com a maior quantidade possível de indivíduos.

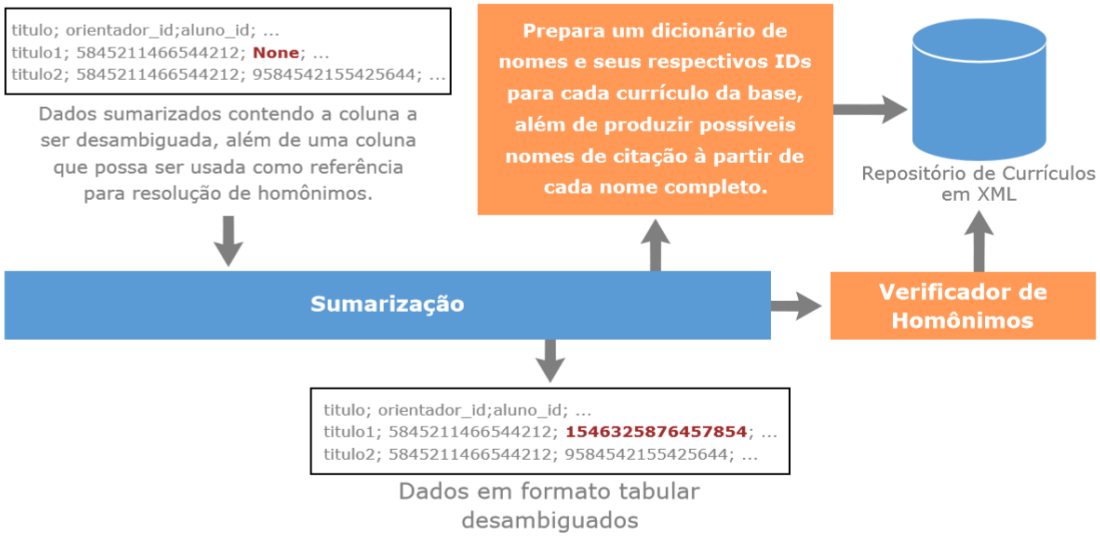
Figura 4: Vínculos da Plataforma Lattes.



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 base nestes dados, é possível obter informações como nome completo e nome em citações bibliográficas, utilizadas para produção do dicionário, conforme a Figura 5. 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 5: 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 6, contendo o nome de citação como chave do dicionário e uma lista vinculada a esta chave, contendo os identificadores únicos de cada indivíduo.

Figura 6: Exemplo de dicionário utilizado no processo de desambiguação.

```
dicionario = {
    1 gray farias moita   ..: [5],
    2 moita, g. f.   .....: [5],
      moita, gray f.   .....: [5],
      moita, gray farias   .....: [5],

    1 tales moreira   .....: [6],
      moreira, t.   .....: [6, 8], 3
      moreira, tales   .....: [6],

    1 thiago moreira   .....: [8],
      moreira, thiago   .....: [8]
}
```

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.

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

A Plataforma Lattes conta hoje com mais de 5.200.000 currículos, totalizando cerca de 8.449.497 orientações, independentemente de sua natureza. Antes da execução do método de identificação proposto neste trabalho, apenas 621.384 possuíam relacionamentos vinculados implicitamente, conforme a Tabela 2. Ou seja, apenas estas orientações foram devidamente vinculadas com os orientados pelo orientador.

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

Fonte: Próprio Autor.

Após a execução do método proposto, 3.804.063 relacionamentos novos foram identificados, um total de 45,03%, valor bem superior aos 7,35% encontrados antes do processo de desambiguação.

4 RESULTADOS

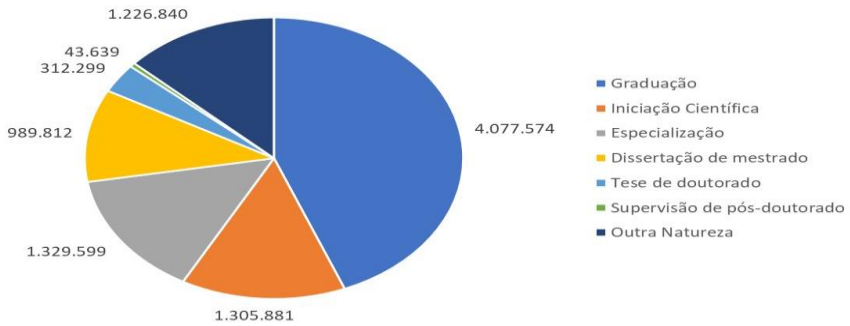
Como resultados das etapas de tratamentos dos dados, são produzidos arquivos de dados pré-processados em formato tabular, podendo ou não estar desambiguados, contendo todas as informações necessárias para o cálculo de diversas métricas bibliométricas e baseadas em análise de redes sociais, como pode ser observado na arquitetura proposta. 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.

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. 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.

As orientações também podem ser distribuídas de acordo com o seu nível, conforme apresentado na Figura 7, 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%).

Figura 7: Distribuição das orientações por natureza de todos os indivíduos.

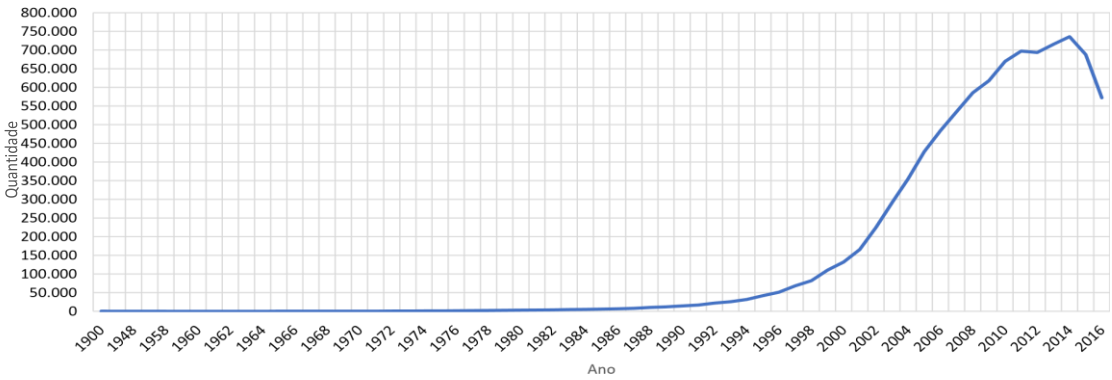


Fonte: Próprio Autor.

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 8).

Figura 8: Gráfico de orientações por ano considerando todos os indivíduos até o ano de 2016.



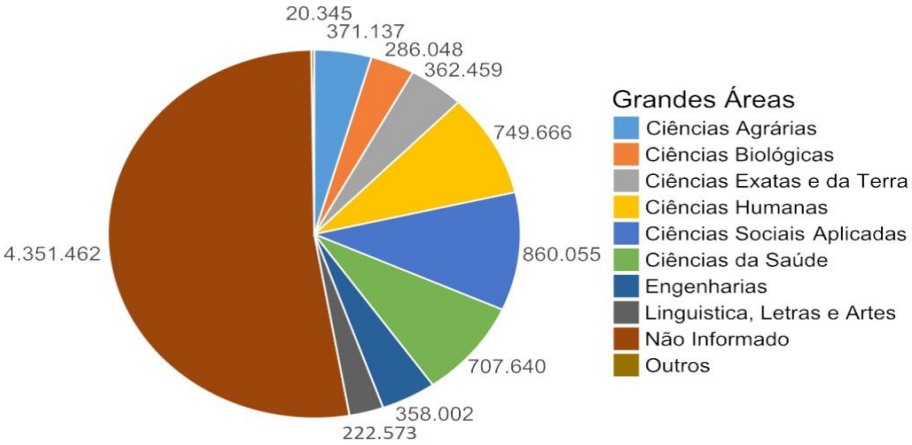
Fonte: Próprio Autor.

Uma hipótese para a redução das orientações nos últimos anos é possivelmente a falta de atualização dos currículos. Logo, 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.

Os dados analisados também foram distribuídos por grandes áreas do conhecimento, conforme apresenta a Figura 17. 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.

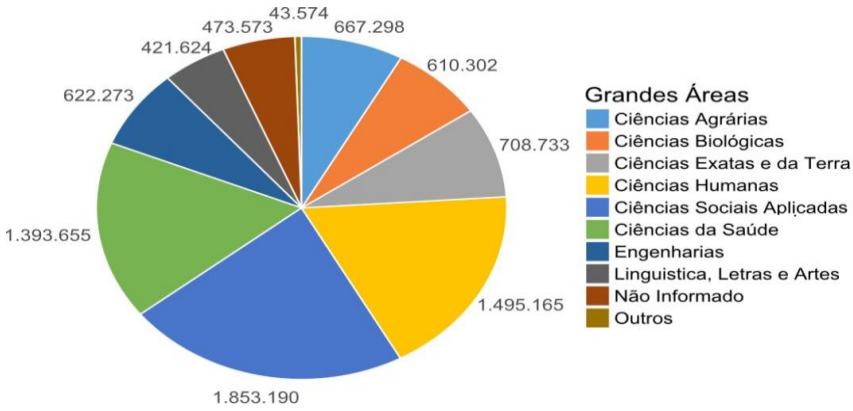
Figura 9: Gráfico de orientações por grandes áreas.



Fonte: Próprio Autor.

Diante da quantidade de orientações sem grande área informada, foram utilizadas as grandes áreas informadas pelos orientadores, para, dessa maneira, caracterizar todas as orientações por esta característica, o que produz resultados relevantes (Figura 10).

Figura 10: 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 distorções, mantendo a proporção no aumento de orientações em cada grande área.

Também foram produzidas árvores genealógicas dos orientadores que mais orientam de acordo com os registros dos currículos. 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.

A Tabela 3 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.

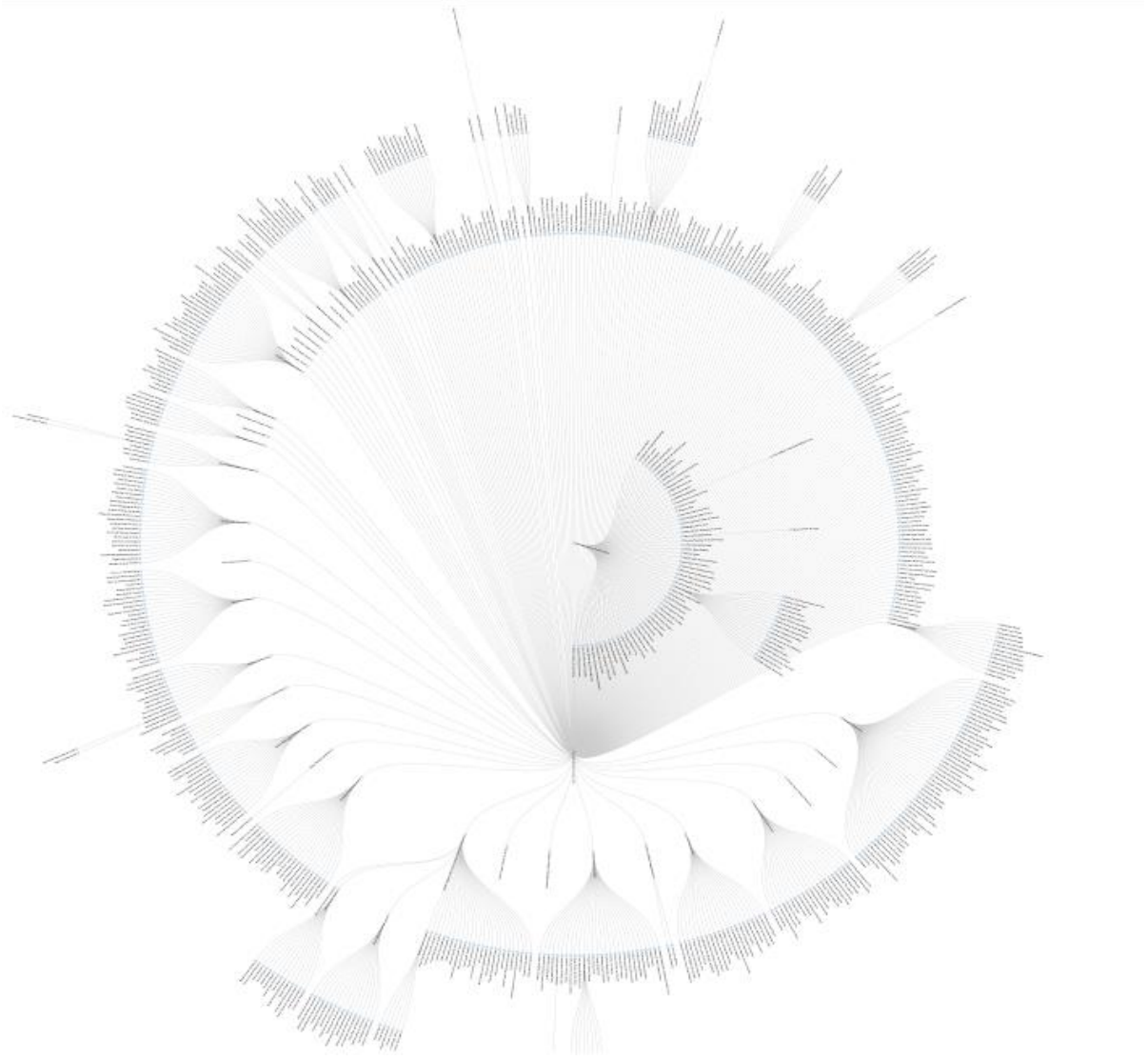
Tabela 3: Indivíduos com maiores redes do conjunto de dados.

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

Fonte: Próprio Autor.

A Figura 11 representa a árvore genealógica do Indivíduo 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 encontrada (399), seus orientados, em geral, tiveram poucas orientações, tornando a árvores com poucas gerações.

Figura 11: Árvore genealógica do orientador com a maior quantidade de orientações diretas



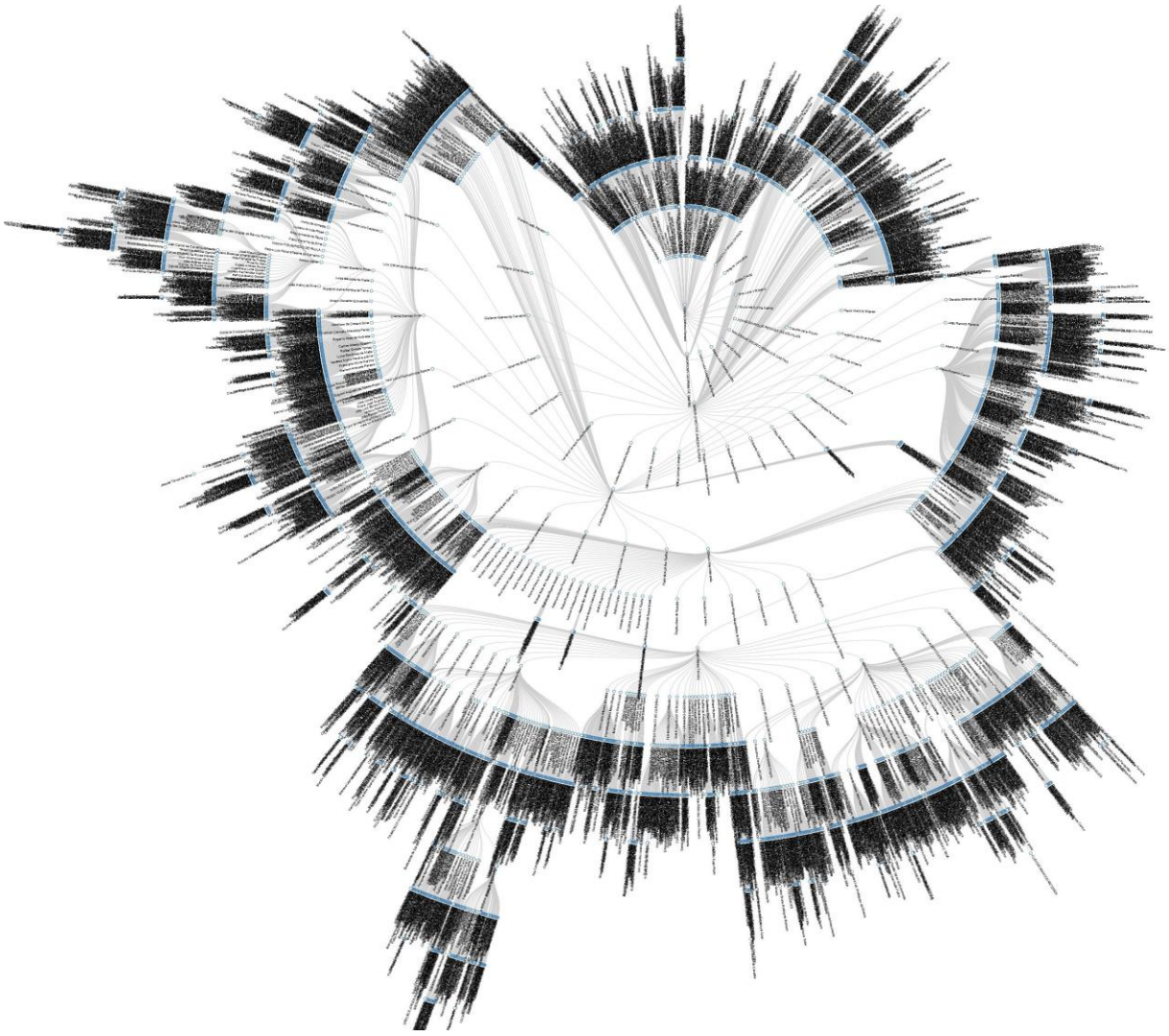
Fonte: Próprio Autor.

A árvore da Figura 12 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).

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.

Figura 2: Árvore genealógica do orientador com a maior quantidade de gerações e descendentes.

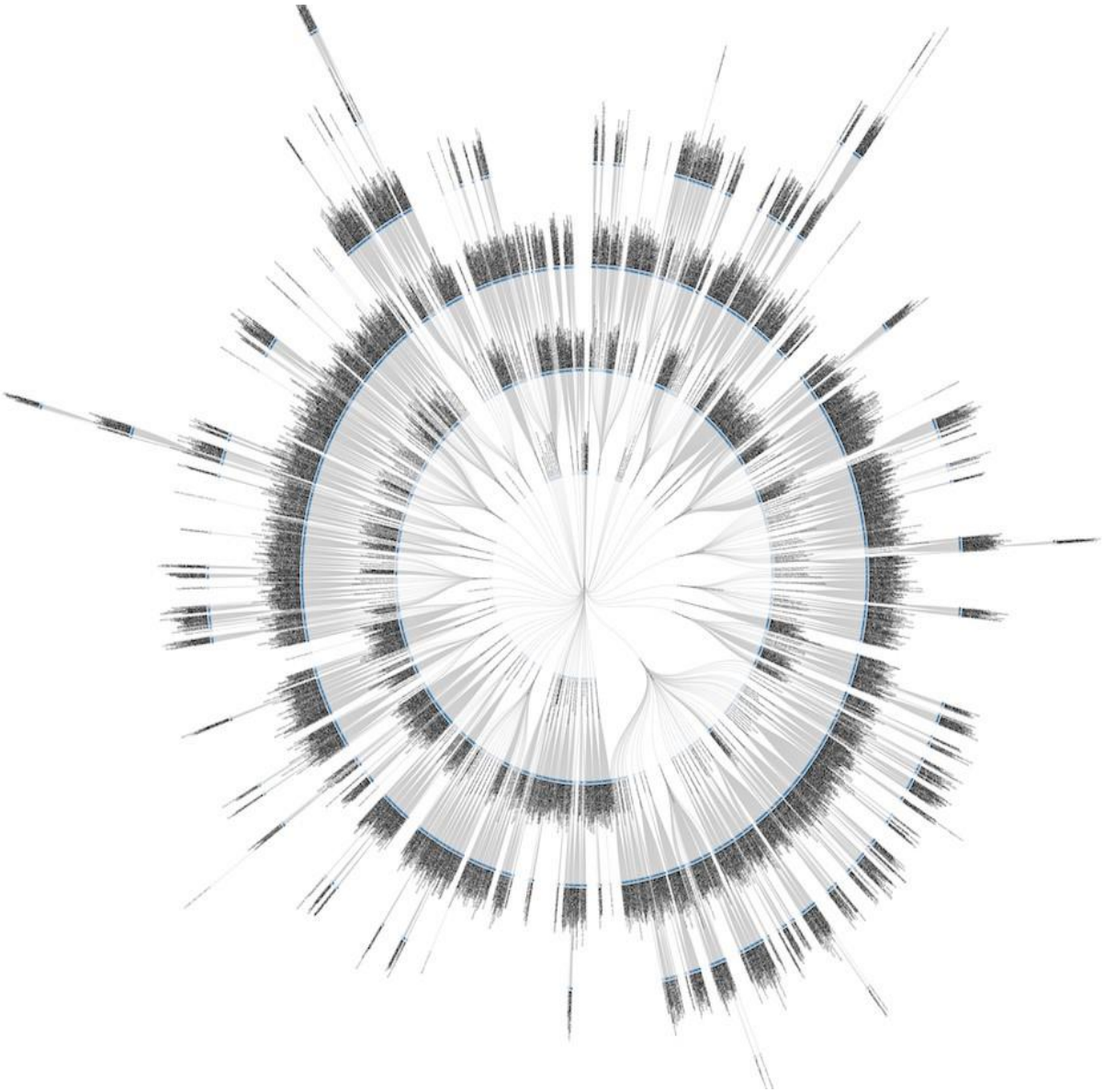


Fonte: Próprio Autor.

Já a árvore da Figura 13 possui a maior quantidade de descendentes, apesar de possuir apenas 6 gerações (muito abaixo das 15 encontradas na árvore da Figura 23) 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 13: 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.

5 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 bibliométricas e baseadas em análises de redes realizadas sobre dados de registros de orientações acadêmicas dos currículos cadastrados na Plataforma Lattes.

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.

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. Na Plataforma Lattes existe, atualmente, uma grande dificuldade de identificar indivíduos nos currículos, já que muitas vezes não há vínculo entre identificadores e nomes, sendo necessária a aplicação de técnicas de desambiguação dos próprios nomes. Com a aplicação do método proposto neste trabalho, foi possível identificar orientados não vinculados a seus orientadores, permitindo, consequentemente, a caracterização de redes com maior precisão.

REFERÊNCIAS

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

CHRISTEN, P. **Data matching:** Concepts and techniques for record linkage, entity resolution, and duplicate detection. [S.l.]: Springer Data-Centric Systems and Applications, 2012. p. 1-279.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). **Sobre a plataforma Lattes**. 2017. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 01 mar. 2018.

SILVA, L. C. R. da; NASCIMENTO, H. A. D. do. **Visualizando bases curriculares**. 2006.

DIAS, T. M. R. **Um Estudo Sobre a Produção Científica Brasileira a partir de Dados da Plataforma Lattes**. 2016. 181 f. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

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. Disponível em: <http://dl.acm.org/citation.cfm? doid=2910896.2910916>. Acesso em: 01 mar. 2018.

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.

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

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**, v. 46, n. esp., p. 99-109, 2006.

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

MUGNAINI, R.; JANNUZZI, P. d. M.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. **Ciência da Informação**, 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.

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: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 4., 2015. [S.l.: s.n.], 2015. p. 12.

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

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.