Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes

Rafael J. P. Damaceno[1], **Luciano Rossi**[1], **Jesús P. Mena-Chalco**[1]

[1]Universidade Federal do ABC (UFABC)
Santo André - SP - Brasil

{rafael.damaceno,luciano.rossi,jesus.mena}@ufabc.edu.br

Abstract. *Different academic-scientific knowledge areas have made efforts to create databases related to researchers as well as the advisor-advisee relationships. However, most of these databases present problems such as redundancy, absence, and imprecision of information. In this paper, we present an algorithm for the automatic identification of academic genealogical graph. The contribution of our work relies on the precision of the hierarchical structure of academic advisor-advisee relationships resulting from the algorithm (graph), which facilitates genealogical analysis. As a case study, we prospected more than 272 thousand Ph.D. researchers registered in the Lattes Platform. Novel characteristics on the Brazilian academic genealogy are discussed in detail.*

Resumo. *Diferentes áreas do conhecimento acadêmico-científico têm realizado esforços para a criação de bases de dados de pesquisadores e seus relacionamentos de orientação. No entanto, grande parte destas bases apresentam problemas como redundância, ausência e imprecisão de informações. Neste artigo, apresentamos um algoritmo para a identificação automática de registros acadêmicos, considerando estes problemas. A contribuição deste trabalho recai na precisão da estrutura hierárquica de orientação acadêmica resultante do algoritmo (grafo), que facilita análises genealógicas. Como estudo de caso, prospectamos mais de 272 mil doutores registrados na Plataforma Lattes e apresentamos características inéditas sobre a genealogia acadêmica brasileira.*

1. Introdução

O acesso e a estrutura de dados que estão disponíveis nos mais diversos contextos nem sempre são suficientes para a extração de conhecimento. Assim, muitas vezes informações de interesse não podem ser obtidas a partir de consultas simples que se limitam à estrutura estabelecida, sendo necessário reorganizar os dados previamente para facilitar a recuperação de informação [Korfhage 1997].

Este trabalho está relacionado com métodos computacionais para criação de grafos que permitam o estudo de parentesco acadêmico denominado Genealogia Acadêmica (GA). A GA é uma ferramenta importante no estudo da propagação do conhecimento acadêmico-científico por meio dos relacionamentos de orientação [Sugimoto 2014]. Bancos de dados com informações sobre genealogia, especificamente aqueles dedicados a academia, comumente apresentam uma estrutura bem definida. Dessa forma, é possível obter informações (atributos) diretamente relacionadas aos acadêmicos (registros).

Diferentes áreas do conhecimento acadêmico-científico têm realizado esforços na estruturação de bancos de dados com informações de parentesco acadêmico. Os matemáticos possuem um banco de dados com informações genealógicas desta comunidade acadêmica que é disponibilizado pelo *Mathematics Genealogy Project – MGP*[1]. Outras áreas do conhecimento, como os neurocientistas e os cientistas da computação, dentre outros, possuem seus registros genealógicos mantidos pela plataforma *The Academic FamilyTree*[2]. Ambos os exemplos, apesar da indiscutível relevância, não fornecem a possibilidade de consulta e análise do grafo de genealogia de forma ampla.

No contexto nacional, a Plataforma Lattes apresenta-se como o maior e mais relevante repositório curricular de acadêmicos relacionados com a Ciência Brasileira [Lane 2010]. Para todos os acadêmicos registrados na Plataforma Lattes são encontradas informações correspondentes à produção científica, técnica e artística, sendo que a informação sobre orientação/supervisão está disponível para acadêmicos com atuação na formação de alunos. Como apontado por [Mena-Chalco and Cesar Junior 2013], essa informação registrada não mantém um padrão e, em muitos casos, é incompleta e sujeita a erros (e.g., registros com nomes abreviados tanto do orientador quanto do orientado, e sem associação do ID identificador). Assim, a obtenção de uma estrutura hierárquica correta que represente os parentescos acadêmicos torna-se um problema não-trivial.

Neste artigo apresentamos um algoritmo desenvolvido com o objetivo de extrair, de forma automática e com alto grau de corretude, grafos de genealogia acadêmica a partir de bancos de dados que possuam informações explicitas e/ou implícitas sobre relacionamentos de orientação acadêmica. Consideramos os currículos de pesquisadores em nível de Doutorado registrados da Plataforma Lattes como estudo de caso. Os principais desafios encontrados na estruturação do grafo de genealogia foram: (i) devido à falta de declaração explícita da orientação, atribuir um relacionamento por meio do casamento exato e inexato de nomes, (ii) desambiguação de nomes grafados de formas diferentes, e (iii) tratamento de inconsistências das informações registradas.

A aplicação do algoritmo desenvolvido possibilitou a estruturação do grafo de genealogia de todos os doutores registrados, de forma explícita ou implícita, na Plataforma Lattes. O grafo obtido é composto por mais de 334 mil vértices (doutores) e 300 mil arestas (relacionamentos de orientação), que foram quantificados em função das grandes áreas do conhecimento. Adicionalmente, foram aplicadas métricas genealógicas (descendência, fecundidade e índice genealógico) para caracterizar os indivíduos e os grupos em função de sua capacidade de orientação acadêmica. Os resultados apresentados evidenciam informações inéditas da Ciência Brasileira sob a perspectiva de formação de recursos humanos.

2. Trabalhos correlatos

O crescente interesse da comunidade científica, sobre trabalhos que considerem a GA, motiva o desenvolvimento de pesquisas sobre a extração de grafos de genealogia a partir de bancos de dados. O trabalho de [Dores et al. 2016] apresenta um método semi-automático que tem como objetivo específico a extração deste tipo de estrutura. Foi considerado como fonte de dados o *Networked Digital Library of Theses and Disser-*

[1] https://genealogy.math.ndsu.nodak.edu, último acesso em 27 de maio de 2017.
[2] https://academictree.org, último acesso em 27 de maio de 2017.

tations – NDLTD[3]. Do total de registros desta base (4 588 474), à época do artigo, os autores construíram o grafo considerando 638 812 (14%). Os autores apontam a ausência de informações e os erros de grafia no banco como as causas para este baixo percentual. Além da obtenção do grafo de genealogia, o referido trabalho apresenta uma caracterização por meio de métricas topológicas específicas.

Grande parte das pesquisas sobre GA se concentram na estruturação semi-automática do grafo e em sua caracterização. [Malmgren et al. 2010] é um dos pioneiros na análise deste tipo de estrutura. Este trabalho avalia o impacto da orientação acadêmica nos orientados, utilizando dados do *MGP* e caracterização por meio de métricas. No trabalho de [Gargiulo et al. 2016] o objetivo é a identificação da origem histórica dos matemáticos, também pela estruturação de grafos de genealogia. Este mesmo conjunto de dados foi utilizado por [Rossi et al. 2017] para caracterização de matemáticos através de uma nova métrica denominada índice genealógico. A plataforma *Web* denominada *The Academic Family Tree* é outra importante fonte de dados genealógicos que abrange diferentes áreas do conhecimento. Esta base de dados foi considerada no trabalho de [David and Hayden 2012] que estruturou o grafo de genealogia dos neurocientistas e promoveu a análise da estrutura com o uso de métricas.

No contexto nacional, a Plataforma Lattes foi utilizada como fonte de dados para identificar grafos de genealogia honoríficas de pesquisadores destacados ou de determinadas áreas do conhecimento [Elias et al. 2016]. É importante destacar que a coleta de dados nesses trabalhos foi realizada de forma semi-manual. Diferente dos trabalhos anteriormente descritos, este artigo apresenta como objetivo a proposição de um algoritmo que obtenha, de forma automática, grafos de genealogia acadêmica de fontes de dados curriculares, como a Plataforma Lattes. No nosso entendimento, trata-se de uma proposta inédita, tanto pela tratativa das inconsistências observadas, quanto pela precisão do grafo de genealogia acadêmica obtido.

3. Métodos

Neste trabalho definimos um algoritmo para identificação de grafo de genealogia acadêmica, obtido a partir de dados da Plataforma Lattes. Foram realizadas quatro etapas: (i) coleta, leitura e normalização dos dados, (ii) identificação de relações explícitas, (iii) identificação de relações implícitas, e (iv) redução de inconsistências e validação do grafo. A seguir, descrevemos estas etapas e, ao final, apresentamos o algoritmo que sumariza o processo de obtenção do grafo de genealogia.

3.1. Coleta, leitura e normalização dos dados

Os dados prospectados da Plataforma Lattes foram utilizados para a construção do grafo de genealogia acadêmica, onde os vértices representam os pesquisadores (com nível de Doutorado, Pós-Doutorado ou Livre-Docência) e as arestas direcionadas os relacionamentos de orientação entre eles. Os atributos disponíveis para cada registro são: identificador numérico exclusivo, nome completo, endereço, área de pesquisa, formação acadêmica, vínculo institucional, orientações em andamento e concluídas, publicações, dentre outros.

Após a coleta de todos os currículos disponíveis, os dados foram submetidos a um processo de normalização, que consiste em detectar e eliminar erros e/ou in-

[3]http://www.ndltd.org, último acesso em 27 de maio de 2017.

consistências nas informações, como definido nos trabalhos de [Fayyad et al. 1996] e [Rahm and Do 2000]. Todos os nomes de pesquisadores foram submetidos à eliminação de espaços em branco duplicados, acentos, pontuações e prefixos no início dos nomes completos (e.g., "Dr.", "Profa."), e à transformação dos caracteres para caixa-baixa. É importante frisar que nomes abreviados ou grafados de forma diferente podem acarretar em erros nas atribuições dos relacionamentos de orientação. O algoritmo apresentado neste trabalho permite lidar de modo eficiente com nomes grafados de forma similar.

3.2. Identificação de relações explícitas

As relações explícitas de orientação são aquelas onde o nome de um pesquisador e seu respectivo número identificador[4] estão presentes nos campos de "Formação Acadêmica" e "Orientações Concluídas" de cada currículo. No primeiro campo, há informações sobre o nível de formação do pesquisador, bem como o registro de seu orientador e coorientador, que são identificados por meio dos nomes completos e os números identificadores. Usamos estas informações para identificar a ascendência explícita de cada pesquisador.

O campo "Orientações Concluídas" contém os registros das orientações finalizadas pelo pesquisador. Podem ser consideradas as orientações e coorientações concluídas em nível de Iniciação Científica, Mestrado, Doutorado e Pós-doutorado. As informações disponíveis no registro de orientação são: nome completo do orientado, anos de início e conclusão da orientação, e número identificador do orientado. Extraímos os dados somente das orientações concluídas em nível de Doutorado.

As relações identificadas (por ascendência ou descendência) representam arestas no grafo de genealogia, e seus atributos disponíveis são: número identificador do vértice orientador, número identificador do vértice orientado, ano de início, ano de conclusão, nome completo do orientador e nome completo do orientado. Cada vértice representa um pesquisador com os seguintes atributos: número identificador, nome completo e nome da Grande Área do Conhecimento.

3.3. Identificação de relações implícitas

Para os registros nos quais o número identificador de um pesquisador está ausente (nos campos "Formação Acadêmica" e/ou "Orientações Concluídas"), utilizamos os nomes completos do orientador e orientado para fazer a identificação do relacionamento (i.e., deduplicação de instâncias[5]). Consideramos casamentos exatos, i.e., aqueles nos quais os nomes não apresentam diferenças em nenhum caractere e inexatos, ou seja, por até um caractere de distância de edição, conforme método proposto por [Levenshtein 1966].

Estes casamentos foram aplicados em dois escopos diferentes do grafo, primeiro de forma global, usando os nomes completos (como registrados na Plataforma Lattes) dos pesquisadores e por último, de forma local usando o nome e sobrenome (ignorando os termos intermediários) dos pesquisadores. Para os casamentos em escopo global, cada nome completo foi comparado com registros de um dicionário de nomes e identificadores numéricos, criado previamente com informações obtidas da Plataforma Lattes. Nos casamentos em escopo local, a comparação de nomes se deu entre irmãos acadêmicos.

[4]Um número identificar na Plataforma Lattes está composto de 16 algarismos e é um valor único associado a cada pesquisador registrado na Plataforma.

[5]O tratamento de ambiguidades é um problema importante que afeta a qualidade dos dados coletados [Kim et al. 2014, Ferreira et al. 2014].

Nos casos em que o casamento de nomes não fosse efetivo (tanto por casamento exato quanto por inexato) ou por haver homônimos nestes casamentos, novos vértices, representando pesquisadores não encontrados na Plataforma Lattes, foram construídos de modo artificial. Isto foi feito com o intuito de manter de forma fiel o número de orientações registradas por cada pesquisador. Nestes casos, os vértices novos não contém atributos como Grande Área e dados de formação acadêmica, bem como o próprio número identificador (sendo atribuído um valor artificial). Nas relações implícitas as orientações identificadas também representam arestas no grafo de genealogia.

3.4. Redução de inconsistências e validação do grafo

Esta etapa consiste na eliminação de vértices artificiais com nomes inconsistentes, na unificação de arestas duplicadas e na validação amostral do grafo. Primeiramente, submetemos os vértices artificiais a dois filtros: (a) nomes de tamanho menor que seis caracteres ou compostos por apenas um termo e (b) nomes presentes em uma lista de termos pré-definida (e.g., "a informar", "sem orientador", "não tinha na época"). Removemos os vértices artificiais e suas arestas, caso os nomes se enquadrassem nestes filtros.

Como o algoritmo coleta tanto as relações apontadas na ascendência quanto na descendência, ocorre de uma mesma relação ser apontada duas vezes, uma pelo orientador (em "Orientações Concluídas" do currículo do orientador) e outra pelo orientado (em "Formação Acadêmica" do currículo do orientado). Ainda, pode ocorrer de um pesquisador relacionar-se com outro mais de uma vez, i.e., ter mais de um Doutorado concluído com o mesmo orientador, gerando também arestas duplicadas. Nestes casos, as arestas foram unificadas considerando-se a mais antiga (ano de conclusão da relação), pois representam a mesma entidade no grafo de genealogia.

Por fim, fizemos uma validação amostral do grafo, junto aos currículos de 10 pesquisadores. Essa validação consistiu em verificar manualmente de qual currículo surgiu cada relação (ou de "Orientações Concluídas" do próprio currículo ou de "Formação Acadêmica" do currículo de outro pesquisador) e quantas são implícitas, explícitas ou transformadas, por meio de casamento, de implícitas para explícitas (ver Seção 4).

Algoritmo para identificação do grafo de genealogia acadêmica

O principal procedimento BUILD-ACADEMIC-GENEALOGICAL-GRAPH requer como parâmetro inicial uma lista S composta por números identificadores de currículos de pesquisadores, e chama procedimentos auxiliares para tratar cada currículo coletado. Como saída, produz uma lista de vértices (V) e arestas (E) que representam o grafo de genealogia acadêmica identificado. Um dos procedimentos auxiliares é GET-RELATIONSHIP, cuja função básica é extrair do currículo coletado os relacionamentos ($i_{relationships}$) de ascendência e descendência descritos nos campos relacionados à "Formação Acadêmica" e a "Orientações Concluídas", respectivamente.

Para cada relacionamento extraído, outro procedimento auxiliar (GET-IDENTIFICATION) busca o número identificador ($r_{identification}$) do orientado (descendência) ou do orientador (ascendência). Caso sejam relações explícitas, o número identificador do pesquisador está descrito no currículo, e portanto, a sua obtenção é trivial. Caso sejam relações implícitas, este número está ausente e o procedimento usa o nome completo do pesquisador para tentar fazer casamento junto a um dicionário (D) de nomes completos e identificadores numéricos, criado previamente.

```
BUILD-ACADEMIC-GENEALOGICAL-GRAPH(S)
1   V = E = {}
2   while |S| ≠ 0
3     for each i ∈ S
4       if i ∉ V
5         INSERT(V, i)
6         i_relationships ← GET-RELATIONSHIPS(i)
7         for each r ∈ i_relationships
8           r_identification ← GET-IDENTIFICATION(r)
9           if r_identification ∉ V
10            INSERT(V, r_identification)
11            if r_identification not artificial
12              INSERT(S, r_identification)
13          if r ∈ E
14            r_older ← GET-OLDER-EDGE(r, E)
15            INSERT(E, r_older)
16          else
17            INSERT(E, r)
18    REMOVE(S, i)
19  LOCAL-MATCHING(V, E)
20  FILTER-NODES(V, E)
21  return (V, E)

GET-IDENTIFICATION(r, D)
1   if r is implicit
2     r_identification ← PERFECT-MATCHING(r, D)
3     if r_identification = NIL
4       r_identification ← LEVENSHTEIN-MATCHING(r, D)
5     if r_identification = NIL
6       r_identification ← new artificial identifier
7   return r_identification
```

Este casamento pode ser de duas formas, ou PERFECT-MATCHING ou LEVENSHTEIN-MATCHING. Primeiramente, tenta-se fazer o casamento exato, e em caso de falha, procede-se ao inexato, usando, em ambos os casos, o nome completo do pesquisador. Caso não ocorra casamento, um vértice artificial é criado para o nome, evitando assim que o número de orientações realizadas seja menor do que a descrita nos currículos.

O procedimento GET-OLDER-EDGE certifica de manter em E a aresta mais antiga entre dois pesquisadores, comparando, para isso, o ano de conclusão dos registros. Durante a execução do procedimento principal, novos números identificadores são adicionados à lista S (somente os não artificiais) e outros são excluídos, conforme são tratados os relacionamentos. O laço principal do algoritmo só encerra após não haver mais números identificadores para serem verificados. Após, o procedimento LOCAL-MATCHING, tenta mesclar vértices artificiais com seus irmãos, comparando neste caso, o nome e sobrenome (os termos intermediários são ignorados) dos vértices em questão. Isso é feito com o intuito de reduzir o número de vértices artificiais duplicados e que representam, possivelmente, o mesmo pesquisador (nome abreviado). Por fim, o procedimento FILTER-NODES exclui vértices artificiais com nomes inconsistentes e as arestas a eles relacionadas.

A principal vantagem da utilização deste algoritmo é a capacidade de transformar relações implícitas de orientação em explícitas, por meio de deduplicação de nomes. Com isso, obtém-se um grafo mais completo, em que há mais relacionamentos informativos quando comparado à estrutura obtida sem essa transformação.

4. Resultados

O processo anteriormente descrito foi aplicado no conjunto de currículos disponíveis na Plataforma Lattes. Apresentamos, a seguir, os resultados obtidos, onde são detalhados os dados considerados, a validação manual e a caracterização do grafo conforme subconjuntos que representam as Grandes Áreas do Conhecimento. Adicionalmente, com objetivo estritamente ilustrativo, apresentamos uma análise genealógica do grafo obtido.

4.1. Conjunto de dados e validação das relações de orientação

Os dados de 272 165 currículos foram coletados da Plataforma Lattes em Maio de 2017[6]. A Tabela 1 apresenta as principais estatísticas que descrevem tanto os dados obtidos quanto o grafo resultante. A Tabela 2 descreve os resultados referentes ao processo de casamento de nomes nos escopos global e local, aplicados nos vértices cujos identificadores numéricos estavam ausentes.

Foi possível inferir, ou por casamento exato ou por inexato, registros da Plataforma Lattes para 57,3% do total de nomes sem número identificador (148 980), o que reduziu a quantidade de vértices artificiais e ampliou a qualidade e quantidade das informações da estrutura obtida.

É importante observar que mantivemos no grafo os vértices referentes a nomes que não foram identificados explicitamente. Apresentamos na Figura 1 a quantidade destes vértices por graus de entrada e saída. Nota-se que a maior parte (93,2%) deles possui grau de saída de no máximo 1. Estes casos foram gerados pelo campo "Formação Acadêmica" dos currículos, onde os pesquisadores indicaram seus orientadores, mas não os vincularam a um identificador numérico (ou pelo orientador não possuir currículo Lattes ou pelo nome grafado ser diferente do cadastrado na plataforma).

A Tabela 3 apresenta a análise amostral de descendência, dos currículos de 10 pesquisadores comparando-os com o grafo identificado. Para estes pesquisadores boa parte dos relacionamentos implícitos foram convertidos em explícitos. O número final de descendentes no grafo foi maior para os 10 pesquisadores, quando comparado ao número registrado nos currículos. Isto ocorreu porque estes pesquisadores não registraram em seus currículos alguns descendentes, que os registraram como orientadores.

[6]A extração de dados foi automática seguindo as informações de acesso oferecidas pela Plataforma Lattes: http://memoria.cnpq.br/web/portal-lattes/extracoes-de-dados

Tabela 1. Estatísticas referentes aos dados considerados e ao grafo resultante

Currículos disponíveis na Plataforma Lattes	5 102 445	100,00%
Currículos considerados neste trabalho (doutores)	272 165	5,33%
Vértices no grafo obtido	334 952	–
Arestas no grafo obtido	300 696	–
Componentes conexas no grafo obtido	37 444	–
Vértices na maior componente conexa	178 767	53,4%
Arestas na maior componente conexa	181 731	60,4%
Vértices isolados (doutores sem orientador nem orientados)	4 842	1,45%
Densidade do grafo	$2{,}68 \times 10^{-6}$	–
Grau médio do grafo	1,79	–
Grau máximo do grafo	130	–

Tabela 2. Estatísticas referentes ao processo de casamento de nomes

Nomes sem número identificador	148 980	100, 00%
Nomes identificados por casamento global exato	71 307	47, 86%
Nomes identificados por casamento global inexato	7 024	4, 71%
Nomes identificados por casamento local	7 033	4, 72%
Nomes inconsistentes excluídos	426	0, 29%
Nomes não identificados (vértices artificiais)	63 190	42, 41%

4.2. Caracterização do grafo em termos de Grande Área

O grafo obtido pode ser utilizado como insumo para estudo da multidisciplinaridade na orientação acadêmica. Para isso, associamos as grandes áreas de conhecimento para cada pesquisador. A Figura 2 mostra a proporção de pesquisadores e de orientações realizadas e recebidas, por Grande Área. Há ao todo nove grandes áreas, conforme descrito pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[7]): Ciências Agrárias (AGR), Ciências Biológicas (BIO), Ciências da Saúde (SAU), Ciências Exatas e da Terra (EXA), Ciências Humanas (HUM), Ciências Sociais Aplicadas (SOC), Engenharias (ENG), Linguística, Letras e Artes (LIN) e Outros (OUT). Acrescentamos a Grande Área Indefinido (IND) para rotular os casos em que esta informação estava ausente nos currículos e nos vértices artificiais.

Excluindo-se a categoria Indefinido, nota-se que as grandes áreas com maior número absoluto de orientações realizadas são HUM com 37 685 orientações (17,98%) e SAU com 34 812 orientações (16,61%). As que possuem maior número de orientações recebidas são SAU com 39 005 orientações (18,61%) e HUM com 34 557 orientações (16,49%). As grandes áreas que possuem maior número de pesquisadores doutores são SAU com 43 831 pesquisadores (17,36%) e HUM com 42 417 pesquisadores (16,80%).

A Figura 3 mostra a influência recebida e exercida por Grande Área. As linhas vermelhas indicam a proporção de orientações recebidas, ou seja, indicam o quanto cada Grande Área influenciou a área em questão. De forma similar, as linhas verdes representam o quanto a Grande Área em questão influenciou as demais. Cabe destacar que,

[7]Tabela de áreas do conhecimento da CAPES, disponível em http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-conhecimento-avaliacao, último acesso em 27 Mai 2017.

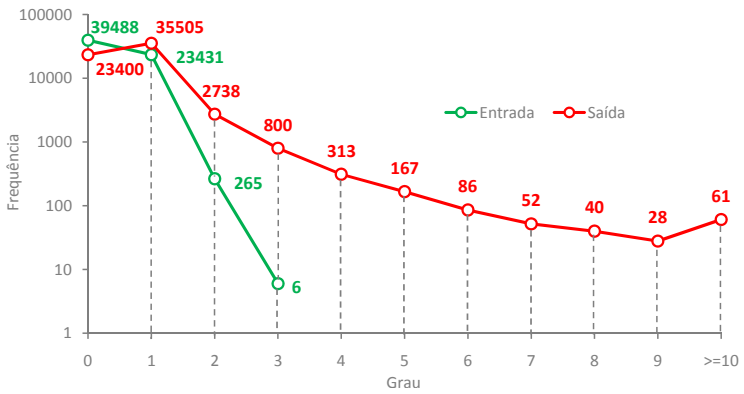

Figura 1. Quantidade de vértices artificiais quanto aos graus de entrada e saída

Tabela 3. Comparação de descendentes identificados nos currículos e no grafo

	Currículo Lattes					Grafo obtido				
	Total	Explícito		Implícito		Total	Explícito		Implícito	
F. A. P. Fialho	119	62	52,10%	57	47,90%	129	124	96,12%	5	3,88%
N. F. F. Ebecken	124	67	54,03%	57	45,97%	129	120	93,02%	9	6,98%
M. L. S. Braga	114	78	68,42%	36	31,58%	119	106	89,08%	13	10,92%
C. A. N. Cosenza	104	29	27,88%	75	72,12%	111	90	81,08%	21	18,92%
O. R. Gottilieb	66	1	1,52%	65	98,48%	98	35	35,71%	63	64,29%
P. de B. Carvalho	85	61	71,76%	24	28,24%	93	88	94,62%	5	5,38%
G. P. Witter	82	16	19,51%	66	80,49%	88	59	67,05%	29	32,95%
J. L. de Azevedo	79	36	45,57%	43	54,43%	85	65	76,47%	20	23,53%
E. L. da Silva	67	44	65,67%	23	34,33%	82	73	89,02%	9	10,98%
R. M. Filho	76	61	80,26%	15	19,74%	78	74	94,87%	4	5,13%

apesar dos maiores números de orientados estar presente entre as mesmas grandes áreas, há considerável número de relações entre grandes áreas distintas, evidenciando assim a multidisciplinaridade na orientação acadêmica. Exemplos desta multidisciplinaridade ocorrem entre BIO e SAU (4 603 orientações), e entre HUM e SOC (3 658 orientações).

4.3. Caracterização do grafo em termos genealógicos

Para ilustrar as possibilidades analíticas deste tipo de estrutura, consideramos um conjunto de métricas genealógicas desenvolvidas para este objetivo. Neste exemplo, utilizamos três métricas para a caracterização genealógica do grafo [Rossi and Mena-Chalco 2014]. A métrica *fecundidade* representa o total de descendentes, em todas as gerações, que um acadêmico de interesse possui, a métrica *descendência* captura o número de alunos que um acadêmico de interesse orientou no decorrer de sua carreira e a métrica *índice genealógico* [Rossi et al. 2017] é uma adaptação da métrica bibliométrica índice-h, e considera a quantidade (descendentes) e qualidade (abrangência das gerações) dos relacionamentos. O índice genealógico g de um acadêmico é o maior número g de descendentes que, por sua vez, possuem, no mínimo, g descendentes cada um.

As métricas foram calculadas para os vértices do grafo obtido pelo algoritmo proposto. Considerando o conjunto completo, os acadêmicos apresentam, em média, 9,6 descendentes em sua linhagem (fecundidade), 3,7 descendentes diretos (descendência)

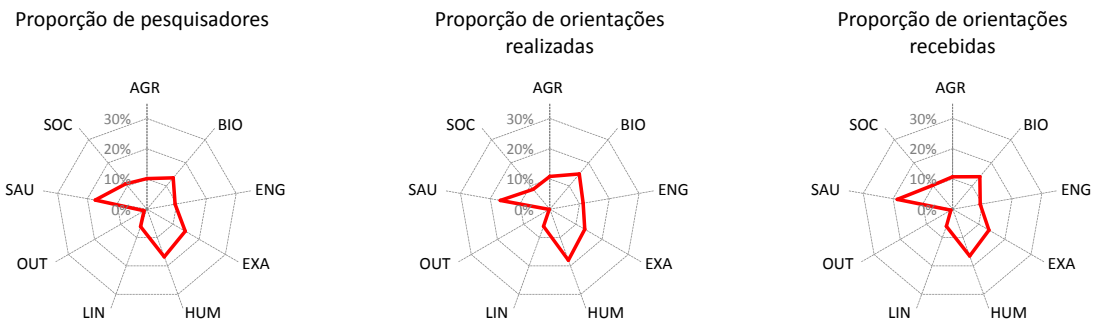

Figura 2. Proporções dos pesquisadores e orientações para cada Grande Área

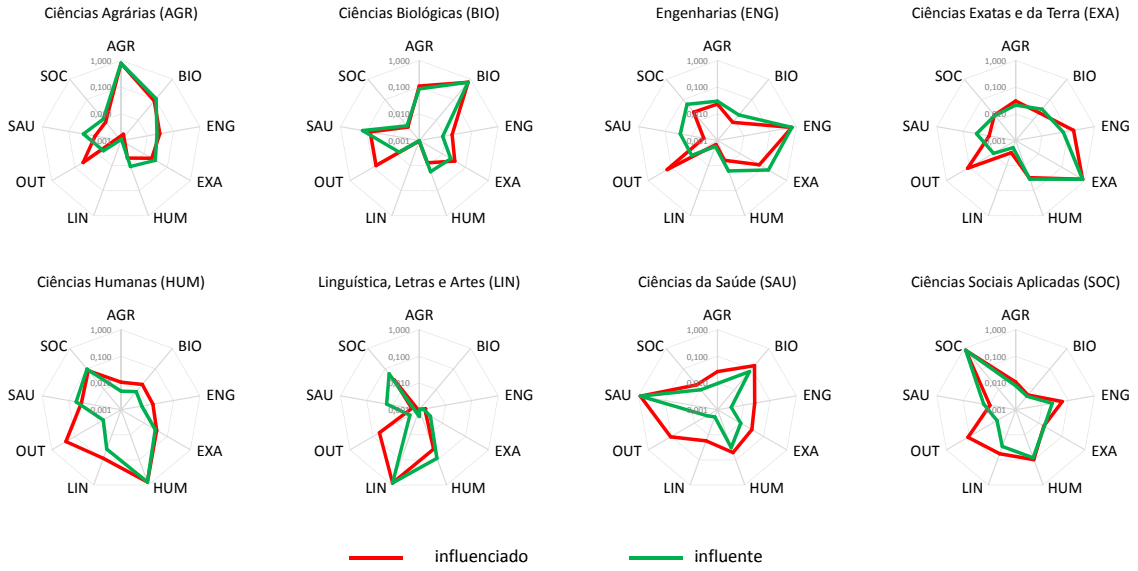

Figura 3. Interdisciplinaridade segundo as orientações observadas entre as Grandes Áreas. Os eixos estão em escala logarítmica.

e o índice genealógico médio é igual a 0,4. Estes resultados consideram somente os acadêmicos férteis, ou seja, não consideramos os vértices com descendência igual a zero; o conjunto resultante totaliza 80 479 (24,0%) acadêmicos. Na Tabela 4 apresentamos os resultados obtidos, estratificados por Grande Área.

Considerando os resultados individuais podemos destacar o acadêmico Joel Martins (BIO), que apresenta 1 747 acadêmicos em sua linhagem e o acadêmico Francisco Antonio Pereira Fialho (HUM), que orientou 129 alunos. Os professores Geraldina Porto Witter (HUM), Cidmar Teodoro Pais (LIN) e Carolina Martuscelli Bori (IND) são os acadêmicos com maior índice genealógico, cujo valor é igual a 12.

Finalmente, na Tabela 5 apresentamos os 10 acadêmicos mais relevantes, de acordo com os resultados das métricas, aplicadas ao grafo de genealogia. Para o índice genealógico, os seguintes acadêmicos também apresentam resultado igual a 10: Eduardo

Tabela 4. Métricas estratificadas por Grande Área (segundo a Plataforma Lattes)

	Fecundidade			Descendência			Índice genealógico			Acadêmicos
	média	mediana	máximo	média	mediana	máximo	média	mediana	máximo	considerados
AGR	11,81	6	388	7,07	5	63	0,44	0	8	14,37%
BIO	14,58	5	1747	6,59	4	85	0,53	0	11	16,57%
ENG	11,07	5	474	6,62	4	129	0,43	0	9	17,87%
EXA	9,69	4	865	5,47	3	98	0,42	0	11	17,12%
HUM	13,02	5	989	6,88	4	129	0,48	0	12	14,85%
IND	7,52	1	1536	1,30	1	48	0,35	0	12	49,98%
LIN	12,84	5	975	6,71	5	119	0,47	0	12	14,16%
OUT	4,09	2	54	2,47	1	31	0,18	0	4	8,58%
SAU	11,54	4	1003	5,98	4	70	0,48	0	10	15,75%
SOC	10,75	4	877	5,70	3	93	0,40	0	8	13,50%

Tabela 5. Acadêmicos com os 10 melhores resultados em cada métrica

Fecundidade			Descendência			Índice genealógico		
Acadêmico	Área	Valor	Acadêmico	Área	Valor	Acadêmico	Área	Valor
J. Martins	BIO	1747	F. A. P. Fialho	HUM	129	G. P. Witter	HUM	12
F. G. Brieger	IND	1536	N. F. F. Ebecken	ENG	129	C. T. Pais	LIN	12
A. Dreyfus	IND	1337	M. L. S. Braga	LIN	119	C. M. Bori	IND	12
C. Pavan	BIO	1336	C. A. N. Cosenza	ENG	111	O. R. Gottlieb	EXA	11
C. M. Bori	IND	1184	O. R. Gottlieb	EXA	98	J. L. de Azevedo	BIO	11
A. Coutinho	IND	1031	P. de B.Carvalho	SOC	93	D. Saviani	HUM	11
B. Pottier	IND	1020	G. P. Witter	HUM	88	M. de S. Chaui	HUM	11
J. P. Lima	IND	1004	J. L. de Azevedo	BIO	85	J. Martins	BIO	11
G. M. Böhm	SAU	1003	E. L. da Silva	EXA	82	E. M. Portella	LIN	11
L. Pereira	HUM	989	R. M. Filho	ENG	78	E. F. de A. Neves	EXA	10

Moacyr Krieger, Gabriel Cohn, Ivan Antônio Izquierdo e Massayoshi Yoshida[8]

5. Conclusão

A evolução do conhecimento científico é a base para o desenvolvimento sócio-econômico de uma comunidade. A ciência evolui por meio da interação entre acadêmicos e pela transferência de conhecimento que é realizada nestas interações. Bancos de dados inseridos neste contexto são importantes fontes para identificação do impacto de pesquisadores na comunidade acadêmica, bem como sobre a formação de comunidades e a evolução do conhecimento de modo abrangente. Considerando uma base de dados curriculares de pesquisadores, desenvolvemos um algoritmo que identifica um grafo de forma automática. As relações de ascendência e descendência descritas nestes currículos, por vezes não continham identificadores únicos, e o algoritmo, por meio de deduplicação, permitiu expandir o grafo para uma maior quantidade de pesquisadores e relações explícitas, melhorando a qualidade da estrutura obtida.

Analisamos este grafo sob a ótica da GA analítica [Sugimoto et al. 2011], considerando as grandes áreas do conhecimento, em função de seu desempenho individual e das interações observadas entre elas. Observamos que as grandes áreas apresentam um equilíbrio de desempenho, capturado pelas métricas, na formação de recursos humanos. Identificamos, de forma preliminar, as interações existentes entre as áreas, que reflete na multidisciplinaridade da jovem comunidade acadêmica no Brasil.

Diferentes são as frentes a serem consideradas como trabalhos futuros. Dentre elas destacamos: (i) a ampliação da base de dados para pesquisadores com nível de Mestrado, (ii) a utilização de técnicas de deduplicação mais sofisticadas, e (iii) a geração de redes mesoescala para estudar longitudinalmente a evolução, no tempo, das relações de orientação, conforme outras informações dos pesquisadores, como instituição e gênero.

Agradecimentos

Os autores agradecem à UFABC, ao CNPq e à CAPES pelo apoio financeiro ao projeto de pesquisa. Os autores também agradecem aos revisores anônimos, cujas sugestões foram muito relevantes para a melhora do artigo.

[8]A grafia dos nomes foi mantida da mesma forma disponível na Plataforma Lattes.

Referências

[David and Hayden 2012] David, S. V. and Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PloS one*, 7(10):e46608.

[Dores et al. 2016] Dores, W., Benevenuto, F., and Laender, A. H. F. (2016). Extracting academic genealogy trees from the networked digital library of theses and dissertations. In *Digital Libraries, 2016 IEEE/ACM Joint Conference on*, pages 163–166.

[Elias et al. 2016] Elias, M., Floeter-Winter, L. M., and Mena-Chalco, J. P. (2016). The dynamics of brazilian protozoology over the past century. *Memórias do Instituto Oswaldo Cruz*, 111(1):67–74.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[Ferreira et al. 2014] Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2014). Disambiguating author names using minimum bibliographic information. *World Digital Libraries*, 7(1):71.

[Gargiulo et al. 2016] Gargiulo, F., Caen, A., and Carletti, R. L. T. (2016). The classical origin of modern mathematics. *EPJ Data Science*, 5(1):26.

[Kim et al. 2014] Kim, J., Diesner, J., Aleyasen, A., Heejun, K., and Hwan-Min, K. (2014). Why name ambiguity resolution matters for scholarly big data research. In *2014 IEEE International Conference on Big Data*, pages 1–6.

[Korfhage 1997] Korfhage, R. R. (1997). *Information Storage and Retrieval*. Wiley, 1st edition.

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

[Levenshtein 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707.

[Malmgren et al. 2010] Malmgren, R. D., Ottino, J. M., and Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, 465(7298):622–626.

[Mena-Chalco and Cesar Junior 2013] Mena-Chalco, J. P. and Cesar Junior, R. (2013). Prospecção de dados acadêmicos de currículos lattes através de scriptlattes. In *Bibliometria e Cientometria: reflexões teóricas e interfaces*, pages 109–128.

[Rahm and Do 2000] Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.

[Rossi et al. 2017] Rossi, L., Freire, I. L., and Mena-Chalco, J. P. (2017). Genealogical index: A metric to analyze advisor-advisee relationships. *Journal of Informetrics*, 11(2):564–582.

[Rossi and Mena-Chalco 2014] Rossi, L. and Mena-Chalco, J. P. (2014). Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 1–12, Brasília, DF, Brazil.

[Sugimoto 2014] Sugimoto, C. R. (2014). Academic genealogy. In Cronin, B. and Sugimoto, C. R., editors, *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, pages 365–382. first edition.

[Sugimoto et al. 2011] Sugimoto, C. R., Ni, C., Russell, T. G., and Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science. *Journal of the American Society for Information Science and Technology*, 62(9):1808–1828.