Índice-h genealógico expandido: Uma medida de impacto em grafos de orientação acadêmica

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

[1]Centro de Matemática, Computação e Cognição – Universidade Federal do ABC

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

Abstract. *Academic Genealogy is defined as the study of intellectual legacy perpetrated through the relationship between advisor and advisee. The set of these relationships over several generations is commonly represented by a social structure as a graph. In this paper, we present the definition of a new topological metric, called* "extended genealogical h-index"*, which can be used to evaluate the impact of an academic through their genealogical relationships. This metric is based on the h-index and expand its concept to measure the impact of an academic over different genealogical generations. For exemplification of our proposed concept, we present a case study about a genealogy graph composed by more than 178,000 mathematicians registered in the* Mathematics Genealogy Project.

Resumo. *A genealogia acadêmica é definida como o estudo da herança intelectual perpetrada por meio do relacionamento entre orientador e orientado. O conjunto desse tipo de relacionamentos, ao longo de várias gerações, é comumente abstraída por uma estrutura social que é representada por um grafo. Neste trabalho, apresentamos a definição de uma nova métrica, denominada* "índice-h genealógico expandido"*, que pode ser usada para avaliar o impacto de um acadêmico por meio de seus relacionamentos de orientação e tendo sua abrangência limitada somente pela topologia do grafo. Esta métrica baseia-se no índice-h bibliométrico e expande seu conceito para mensurar o impacto de um acadêmico ao longo de diferentes gerações. Para exemplificação da nova métrica, apresentamos um estudo de caso considerando um grafo de genealogia composto por mais de 178 mil doutores em matemática registrados no* Mathematics Genealogy Project.

1. Introdução

A genealogia acadêmica é definida como um estudo quantitativo da herança intelectual perpetrada por meio de relacionamentos de orientação entre estudantes e seus orientadores ao longo de diferentes gerações (Sugimoto, 2014). Os relacionamentos de orientação acadêmica promovem a propagação de conhecimento científico por meio da interação entre orientador, com diferentes desempenhos em orientação, e seus orientados, que são influenciados pelas características de seus orientadores (Malmgren *et al.*, 2010). Neste contexto, a genealogia acadêmica fornece meios para mensurar e analisar estas interações de forma quantitativa.

Diferentes estudos foram publicados sobre genealogia acadêmica com o objetivo de caracterizar áreas do conhecimento específicas, como a Neurociência (David & Hayden, 2012), a Química Orgânica (Andraos, 2005), a Matemática (Chang, 2011; Malmgren *et al.*, 2010), a Fisiologia (Bennett & Lowe, 2005; Jackson, 2011), a Metereologia

(Hart & Cossuth, 2013), entre outros. Há ainda, iniciativas para a captação e estruturação de dados genealógicos utilizando plataformas *Web*. O *Mathematics Genealogy Project* (`http://genealogy.math.ndsu.nodak.edu`) e o projeto *Neurotree* (`http://neurotree.org/neurotree`) são pautados na obtenção de dados genealógicos das respectivas áreas e na interpretação das estruturas genealógicas obtidas, a comunidade científica dos Físicos (`http://academictree.org/physics`) e, de forma mais específica, para os acadêmicos titulados com doutorado (`http://phdtree.org`). Tais projetos são, inicialmente, orientados para a obtenção e documentação de seus membros, não oferecendo análises dos conjuntos de dados. Porém, os dados registrados contribuem para a documentação histórica das comunidades acadêmicas e resulta em campo fértil para estudos futuros relacionados à influência ou impacto que tiveram acadêmicos desde o ponto de vista da formação de recursos humanos.

Segundo Sugimoto (2014), os estudos de genealogia acadêmica são principalmente utilizados no ambiente acadêmico por pesquisadores interessados em traçarem suas próprias raízes. Entretanto, estes estudos são pouco explorados por aqueles que estudam a ciência a partir de perspectivas históricas, filosóficas, sociológicas e científicas. A real importância da genealogia acadêmica deve-se ao fato de oferecer insumos quantitativos e qualitativos para mensurar as interações, em diferentes dimensões, dos orientadores e seus orientados/supervisados. Adicionalmente, este tipo de estudos permite analisar a ciência desde um ponto de vista de transferência de conhecimento científico entre diferentes gerações, assim como, seu impacto ou influência desta transferência.

Como apresentado no trabalho de Rossi & Mena-Chalco (2014), as estruturas de genealogia acadêmica podem ser analisadas por meio de métricas topológicas, que representam diferentes atributos destas estruturas e fornecem informações relevantes a respeito da formação da comunidade acadêmica bem como a identificação dos principais indivíduos que contribuíram para o desenvolvimento da área por meio dos relacionamentos de orientação. Dentre as diversas métricas utilizadas para a caracterização de estruturas de genealogia, o índice-h genealógico é uma medida com forte intuição semântica que fornece informações sobre a abrangência dos relacionamentos de orientação.

Este trabalho apresenta uma nova métrica topológica denominada *índice-h genealógico expandido*, que pode ser considerada para identificar o impacto ou influência de acadêmicos em suas respectivas comunidades, considerando a amplitude de seus relacionamentos de orientação (número de orientados diretos) e expandindo a abrangência (ordem) da métrica a todas as gerações possíveis de serem identificadas (i.e., produtividade dos descendentes em termos de orientação). No nosso entendimento, esta abordagem é original e formaliza a adaptação do índice-h, originalmente concebida na área de Bibliometria para avaliação de citações bibliográficas, para analisar relações de orientação acadêmica. Essa medida abre uma nova perspectiva para estudar, de forma quantitativa, o grau de impacto ou influência de acadêmicos priorizando a formação de recursos humanos ao invés de considerar somente sua relevância na produção de ciência em termos de artigos acadêmicos ou participação em grandes projetos de pesquisa.

2. Grafos de genealogia acadêmica

A utilização de representações gráficas para estruturar os indivíduos que têm algum tipo de conexão facilita o estudo genealógico. A estrutura geralmente utilizada é denomi-

nada *árvore de genealogia*[1]. Neste trabalho é utilizado o termo *grafo de genealogia acadêmica* para nomear as estruturas de genealogia, sendo categorizadas como *grafo dirigido acíclico conexo*.

Formalmente, um grafo dirigido $\vec{G}$ é um par (V, E), onde V é um conjunto finito de vértices e E, as arestas, é uma relação binária ordenada em V. Para este trabalho, os acadêmicos e seus relacionamentos de orientação são estruturados na forma de *grafo de genealogia acadêmica*. Os vértices (V) representam os indivíduos (acadêmicos) e as arestas direcionadas (E) representam seus relacionamentos de supervisão ou orientação.

Neste trabalho, dado um acadêmico deseja-se analisar toda sua descendência. Assim, um conceito que naturalmente aparece é o do caminho existente entre o acadêmico e toda sua descendência. Formalmente, um *caminho* de *comprimento* k ($C^{(k)}$) de um vértice origem u a um vértice destino u' em um *grafo dirigido* $\vec{G}$ é uma sequência $(v_0, v_1, v_2, \ldots, v_k)$ de vértices tais que $u = v_0$, $u' = v_k$ e (v_{i-1}, v_i) para $i = 1, 2, 3, \ldots, k$. Em um *grafo dirigido*, um caminho $(v_0, v_1, v_2, \ldots, v_k)$ forma um *ciclo* se $v_0 = v_k$ e o caminho contém no mínimo uma aresta. Um grafo que não possui *ciclos* é *acíclico*. Adicionalmente, um grafo dirigido $\vec{G}$ é *conexo* se existe, no mínimo, um caminho ligando todos os vértices deste grafo.

3. Índice-h genealógico expandido

Na área de Bibliometria/Cientometria, o índice-h é uma medida de desempenho proposta por Hirsch (2005) que classifica pesquisadores em função do número de suas publicações e citações correspondentes. Apesar de existirem diferentes questionamentos quanto a eficiência do índice-h (Yong, 2014), esta medida é amplamente utilizada no meio acadêmico devido à sua característica de combinar quantidade (número de publicações) e qualidade relativa (número de citações) da produção acadêmica. Intuitivamente, o índice-h é definido como o maior número h de publicações que possuem, no mínimo, o mesmo número h de citações cada uma.

A adaptação do índice-h, com o objetivo de caracterizar grafos de genealogia acadêmica foi inicialmente desenvolvido por Rossi & Mena-Chalco (2014), entretanto não foi formalizada sua definição. Este *índice-h genealógioco* permite o estudo de acadêmicos orientadores em função do seu desempenho em formação de recursos humanos.

No contexto dos grafos de genealogia acadêmica, a descendência de um vértice é comumente chamada de *território* do vértice e é definida por:

$$T(v) = \{u \in V : \exists (v, u) - caminho\ em\ G\}. \tag{1}$$

Por outro lado, dado um grafo de genealogia $\vec{G}$ e um vértice de interesse $v \in V$, a descendência direta do vértice v em $\vec{G}$ pode ser definida por:

$$D(v) = \{u \in V : (v, u) \in E\}, \tag{2}$$

e a largura, $l(v)$, é dada por:

$$l(v) = |D(v)|. \tag{3}$$

[1] A rigor, as estruturas construídas a partir de dados de genealogia acadêmica não podem ser categorizadas como *árvores*, pois pode existir mais de um caminho entre dois vértices no grafo.

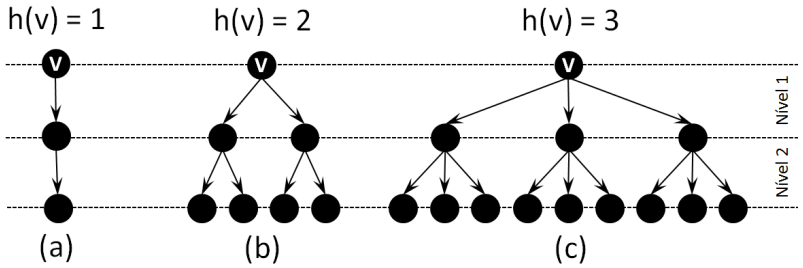


Figura 1. Grafos de genealogia que representam suas diferentes topologias em função do índice-h genealógico.

Esta medida representa o número de vértices adjacentes (vizinhos) a um vértice de interesse[2].

O índice-h genealógico, de ordem 1, de um vértice é definido como o maior número h de conexões existentes entre ele e seus vértices adjacentes (orientados diretos) que possuem, pelo menos, o mesmo número h de conexões cada um, ou seja, dado um grafo de genealogia $\vec{G}$, um vértice de interesse $v \in V$ é k-fértil se $l(v) \geq k$. Assim, a descendência direta k-fértil de um vértice $u \in V$ é o conjunto:

$$D^{(k)}(u) = \{v \in D(u) : l(v) \geq k\}, \tag{4}$$

e a largura k-fértil de u é:

$$l^{(k)}(u) = |D^{(k)}(u)|. \tag{5}$$

Neste contexto, o índice-h genealógico, de ordem 1, de um vértice u é definido por:

$$h(u) = max\{k \in \mathbb{N} : l(u), l^{(k)}(u) \geq k\}. \tag{6}$$

Contextualizando o valor do índice-h genealógico para a caracterização de grafos de genealogia acadêmica, pode-se dizer que um vértice de interesse $v \in \vec{G}$ para o qual observa-se $h(v) = x$, com $x = (1, 2, 3, \ldots, n)$, os grafos de genealogia, obtidos a partir do vértice v, possuem, no mínimo, um sub-grafo unário completo (para $x = 1$), um sub-grafo binário completo (para $x = 2$), um sub-grafo ternário completo (para $x = 3$) e assim sucessivamente, todos com 2 níveis de profundidade, conforme representado nas Figuras 1(a), 1(b) e 1(c), respectivamente.

Intuitivamente, o índice-h genealógico define uma progressão geométrica de razão $q = h(v)$ com 3 termos, onde o primeiro termo representa o vértice de interesse, os demais indicam o número de vértices encontrados em cada nível. O índice-h genealógico captura o impacto que um vértice de interesse v exerce sobre o grafo de genealogia $\vec{G}$ com abrangência de até dois níveis. Dessa forma, o total de vértices pertencentes ao sub-grafo n-ário completo é $\sum_{i=0}^{2}[h(v)]^i$, onde d é um fator de expansão da métrica e indica o número de níveis $(d + 1)$ considerados.

Claramente, o valor obtido para $h(v)$ representa uma cota inferior, visto que existe, no mínimo, um sub-grafo n-ário completo e não existe um sub-grafo $(n+1)$-ário completo para abrangência até o segundo nível do grafo, considerando o território de v.

[2]A largura é uma medida usada para classificar um vértice com base em sua capacidade de conexão.

É importante notar que, a co-orientação é uma atividade comum no contexto acadêmico (um aluno pode ser orientado por mais do que um acadêmico). Assim, para os casos onde se observa um vértice com grau de entrada[3] maior que 1, segundo a métrica apresentada, este vértice será considerado (contabilizado) para todos os adjacentes no nível anterior.

O índice-h genealógico apresenta-se como uma medida interessante para a identificação do impacto de um orientador sobre a comunidade acadêmica, em termos de relacionamentos de orientação, porém há uma limitação na ordem desta métrica, ficando a análise restrita aos dois primeiros níveis do seu território no grafo de genealogia acadêmica.

Para aumentar a abrangência na análise, se faz necessário recalcular a medida substituindo o parâmetro de entrada *largura* pelos valores de *índice-h* obtidos. Trata-se de um processamento recursivo. Para um vértice v suponha $h(v) = 2$, conforme discutido anteriormente, o grafo proveniente de v possui, no mínimo, um sub-grafo binário completo de dois níveis. Caso pelo menos dois dos vértices adjacentes a v apresentem o mesmo valor (i.e., $h = 2$) podemos concluir que existe, no mínimo, um sub-grafo binário completo com *três níveis* de profundidade a partir do vértice v.

Dado um grafo de genealogia $\vec{G}(V, E)$ e um vértice de interesse $v \in V$, o conjunto A dos índices-h dos vértices u adjacentes a v com $h(u) \geq k$ é:

$$A^{(k)}(v) = \{h(u) : (v, u) \in E, h(u) \geq k\}. \tag{7}$$

Com essa definição, o número de vértices adjacentes a v com índice-h maior ou igual a k é $|A^{(k)}(v)|$.

O índice-h genealógico pode ser definido de forma recursiva para considerar mais do que dois níveis, i.e., para analisar o impacto de um acadêmico, considerando diferentes ordens:

$$h_{(d)}(v) = max\{k \in \mathbb{N} : h_{(d-1)}(v), |A^{(k)}(v)| \geq k\}. \tag{8}$$

onde d é a ordem a ser considerada na análise, para $d \geq 1$. No caso $d = 0$, considera-se $h_{(0)} = l$, i.e. o número de descendentes diretos. Note que a definição do índice-h genealógico apresentada na Equação 6 corresponde a ordem 1 (i.e., $h_{(1)}$).

O índice-h genealógico expandido pode ser utilizado para análises de impacto com ordem limitada somente pela topologia do grafo de genealogia, ou seja, é possível se *aprofundar* no cálculo da métrica até o último nível do grafo.

Para ilustrar a proposta, na Figura 2 apresentamos três resultados do cálculo do índice-h expandido para um mesmo grafo de genealogia de profundidade igual a quatro. O cálculo da métrica foi realizado considerando o limite topológico do grafo.

No primeiro grafo, os vértices estão rotulados com os respectivos índices-h de ordem 1. O vértice da raiz do grafo (vértice de interesse) apresenta $h_{(1)} = 4$, conforme discutido anteriormente o território deste vértice contém no mínimo um *sub-grafo quaternário completo com 2 níveis de abrangência* a partir do vértice de interesse (destacado na figura). No contexto deste trabalho, um grafo quaternário completo é aquele em que

[3]O grau de entrada é o número de arestas que *incidem no* vértice de interesse.

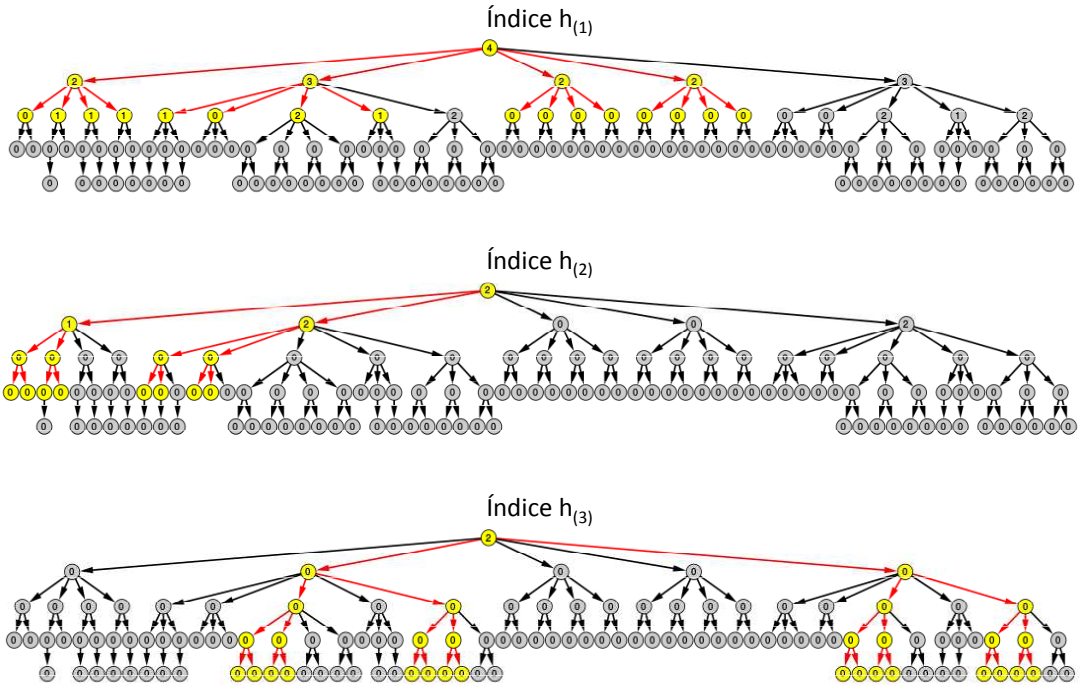


Figura 2. Exemplos de grafos de genealogia com seus vértices rotulados com os índices-h: $h_{(1)}$, $h_{(2)}$ e $h_{(3)}$. Os sub-grafos n-ários completos, identificados as ordens 1, 2 e 3, são destacadas na cor vermelha.

seus vértices possuem grau de saída[4] igual a 4, exceto os vértices pertencentes ao último nível considerado.

O segundo grafo tem seus vértices rotulados com o índice-h de ordem 2 e, para o vértice de interesse, seu valor é $h_{(2)} = 2$. Isto significa que no território a partir do vértice de interesse, existe, pelo menos, um *sub-grafo binário completo com 3 níveis de abrangência*. É importante notar que pode-se encontrar outro exemplo de sub-grafo binário completo válido no grafo em questão, porém não existe um outro sub-grafo ternário para este caso. Isto se deve ao fato de que $h_{(d)}$ é uma cota de limite inferior.

Para o terceiro grafo disponível na Figura 2, os índices-h apresentados nos vértices referem-se a ordem 3, que para o vértice de interesse tem valor $h_{(3)} = 2$, o que sugere, no mínimo, um *sub-grafo binário completo com 4 níveis de abrangência* contido no grafo em questão. É importante frisar que, a recursão pode ser aplicada até que o último nível do grafo seja igual a $d + 1$.

Algoritmo para o cálculo do índice-h genealógico expandido

Como apresentado na Equação 8, o índice-h de ordem d pode ser implementado com uma abordagem recursiva. O pseudocódigo apresentado a seguir foi projetado para calcular do índice-h genealógico expandido (IHE). O procedimento IHE recebe como entrada três parâmetros: o grafo de genealogia $\vec{G}(V, E)$, um vértice de interesse (v) e a ordem (d).

[4]O grau de saída é o número de arestas que *incidem do* (saem) vértice de interesse.

```
IHE(G⃗, v, d)
1   for i ← 0 to d
2       IH(G⃗, v, i)
3       for each u ∈ G⃗.adj[v]
4           IHE(G⃗, u, i)
5   return v.h_d

IH(G⃗, v, i)
1   if i = 0
2       v.h_i ← |G⃗.adj[v]|
3       for each u ∈ G⃗.adj[v]
4           u.h_i ← |G⃗.adj[u]|
5   c ← 0
6   while v.h_i > 0 and v.h_i > c
7       for each u ∈ G⃗.adj[v]
8           if v.h_i ⩽ u.h_i
9               c ← c + 1
10      v.h_i ← v.h_i − 1
11  v.h_{i+1} ← v.h_i
```

No procedimento IHE, o *laço* da linha 1 é executado d vezes. Para cada execução o vértice de interesse v é considerado como parâmetro de entrada para o procedimento IH, juntamente com o grafo $\vec{G}$ e a ordem i que será calculada (linha 2). O procedimento é repetido recursivamente para cada vértice adjacente de v. No procedimento IH verifica-se se o cálculo é referente a $h_{(0)}$ (linha 1) e, caso verdadeiro, é utilizado como elemento de comparação a *largura* do vértice de interesse e de seus adjacentes (linhas 2 – 4).

Um *laço* (linha 6), em IH, será executado enquanto o valor do atributo em questão do vértice de interesse for maior que zero e maior que a contagem dos seus vértices adjacentes O *laço aninhado* (linha 7) é utilizado para comparar os atributos do vértice de interesse com todos os seus adjacentes, contabilizando o número de adjacentes que possuem seus atributos maior ou igual ao valor do atributo dos adjacentes (linhas 8 – 9). Caso o atributo do vértice de interesse seja menor ou igual ao total da contagem, o valor deste atributo é assumido para $h_{(i+1)}$ (linha 11). Caso contrário, o atributo é decrementado em uma unidade.

4. Conjunto de dados utilizado

A aplicabilidade do índice-h genealógico expandido foi testada utilizando-se o conjunto dos doutores em matemática e seus relacionamentos de orientação acadêmica. Estes dados são livremente disponibilizados pelo projeto de genealogia dos matemáticos (*Mathematic Genealogy Project – MGP*, disponível em: `http://genealogy.math.ndsu.nodak.edu/`).

O *MGP* foi idealizado por Harry Coonce, um professor na *North Dakota State University*, no início da década de 1990 (Jackson, 2007). O projeto tem como objetivo compilar informações sobre todos os matemáticos do mundo, por meio do registro histórico, via *Web*, dos indíviduos que obtiveram o título de doutor em matemática (ou título semelhante) e seus respectivos alunos/doutores com formação concluída. O *site* do

MGP é apresentado como ferramenta para a captação e documentação de novos registros genealógicos deste seleto grupo de acadêmicos. As informações que são possíveis de se obter, através do *site* do projeto, são listadas a seguir:

- O *nome* completo do matemático;
- A *instituição* e o *país* onde foi obtida a titulação;
- O *ano* no qual o grau foi obtido;
- O *título* da tese;
- O número de *classificação* da área de atuação (*Mathematics Subject Classification*[5]);
- Seu(s) *orientador(es)* e *orientado(s)*;
- A quantidade total de *descendentes*.

Os registros do *MGP* são identificados por meio de um número exclusivo (id) para cada matemático. Os dados, que são objeto de estudo neste trabalho, foram obtidos por meio de consultas recursivas ao *site* do *MGP* (*web crawling*).

Em Abril de 2014 foram obtidos 178.698 registros de matemáticos e identificados 187.199 relacionamentos de orientação acadêmica. Estes indivíduos estão distribuídos em 185 países ou combinação destes (isso ocorre devido à declaração de dois países como local de titulação) e 2.671 instituições ou combinações destas.

O grafo de genealogia, resultante da representação dos matemáticos como vértices e seus relacionamentos de orientação acadêmica como arestas direcionadas, possui 10.048 componentes conexas. A maior componente conexa contém aproximadamente 88,72% dos vértices totais (158.548 vértices), por outro lado, a segunda componente conexa, em relação ao número de vértices, apresenta apenas 0,08% dos vértices totais (141 vértices). As últimas 7.542 componentes conexas referem-se a vértices isolados, ou seja, não possuem ascendentes ou descendentes. Ao todo, em média cada vértice do grafo possue 2,094 vizinhos.

5. Estudo de caso

O índice-h genealógico expandido foi aplicado ao conjunto de dados extraído do *MGP*. A classificação dos vértices do conjunto de dados foi realizada considerando as duas dimensões do $h_{(d)}$. A primeira dimensão é o resultado da métrica que apresenta $h_{(d)} = n$, para $n = (0, 1, 2, 3, \dots)$. Esta dimensão representa a amplitude do grafo n-ário completo, ou seja, o número de descendentes diretos para cada vértice do grafo, exceto os vértices do último nível.

A segunda dimensão considerada representa a ordem d, indicando os $d + 1$ níveis ou gerações a partir do vértice de interesse. O calculo do índice-h foi realizado até a ordem 10 (11 níveis), este limite foi escolhido pois, a partir da ordem 6 ($d = 6$) observa-se apenas grafos unários completos (caminhos), ou seja, o máximo resultado obtido para $d > 6$ é $h_{(d)} = 1$. Vale ressaltar que, o maior caminho existente neste conjunto de dados é de 41.

Na Tabela 1 é apresentada a classificação dos grafos de genealogia dos matemáticos em função de $h_{(d)}$. As linhas estão associadas aos valores do $h_{(d)}$. Já as colunas estão associadas à ordem d. Para cada célula, linha x, coluna d, é apresentado,

[5]Classificador alfanumérico formulado pela *American Mathematical Society* utilizado para categorizar temas da matemática, disponível em: `http://www.ams.org/msc/msc2010.html`

Tabela 1. Índices-h obtidos para o conjunto de dados dos matemáticos do *MGP*. Cada célula contém os resultados considerando $h_{(d)} = x$, para $x = 0, \ldots, 12$, e ordens $d = 1, \ldots, 10$. As células em cinza correspondem à existência de acadêmicos com estas características no conjunto de dados.

x	$h_{(1)}$	$h_{(2)}$	$h_{(3)}$	$h_{(4)}$	$h_{(5)}$	$h_{(6)}$	$h_{(7)}$	$h_{(8)}$	$h_{(9)}$	$h_{(10)}$
0	162.647	171.072	174.519	176.157	176.991	177.454	177.727	177.896	178.023	178.111
	1	1	1	1	1	1	1	1	1	1
1	11.371	6.676	3.987	2.506	1.700	1.244	971	802	675	587
	3	4	5	6	7	8	9	10	11	12
2	2.753	767	176	35	7					
	7	15	31	63	127	255	511	1.023	2.047	4.095
3	1.013	149	16							
	13	40	121	364	1.093	3.280	9.841	29.524	88.573	265.720
4	463	28								
	21	85	341	1.365	5.461	21.845	87.381	349.525	$1,4\times10^6$	$5,6\times10^6$
5	238	5								
	31	156	781	3.906	19.531	97.656	488.281	$2,4\times10^6$	$1,2\times10^7$	$6,1\times10^7$
6	94	**1**								
	43	**259**	1.555	9.331	55.987	335.923	$2,0\times10^6$	$1,2\times10^7$	$7,3\times10^7$	$4,4\times10^8$
7	45									
	57	400	2.801	19.608	137.257	960.800	$6,7\times10^6$	$4,7\times10^7$	$3,3\times10^8$	$2,3\times10^9$
8	31									
	73	585	4.681	37.449	299.593	$2,4\times10^6$	$1,9\times10^7$	$1,5\times10^8$	$1,2\times10^9$	$9,8\times10^9$
9	26									
	91	820	7.381	66.430	597.871	$5,4\times10^6$	$4,8\times10^7$	$4,4\times10^8$	$3,9\times10^9$	$3,5\times10^{10}$
10	11									
	111	1.111	11.111	111.111	$1,1\times10^6$	$1,1\times10^7$	$1,1\times10^8$	$1,1\times10^9$	$1,1\times10^{10}$	$1,1\times10^{11}$
11	5									
	133	1.464	16.105	177.156	$1,9\times10^6$	$2,1\times10^7$	$2,4\times10^8$	$2,6\times10^9$	$2,9\times10^{10}$	$3,1\times10^{11}$
12	1									
	157	1.885	22.621	271.453	$3,2\times10^6$	$3,9\times10^7$	$4,7\times10^8$	$5,6\times10^9$	$6,8\times10^{10}$	$8,1\times10^{11}$

na parte superior, o número total de acadêmicos com $h_{(d)} = x$. Já na parte inferior da célula é apresentado, o número total de descendentes que um acadêmico teria se $h_{(d)} = x$. Por exemplo, $h_{(5)} = 2$ indica um grafo binário completo com 6 níveis de profundidade, este tipo de sub-árvore contém 127 vértices e existem, no conjunto de dados do *MGP*, 7 acadêmicos com estas características.

A identificação dos acadêmicos mais representativos em função de sua capacidade de propagação pode ser feita buscando-se os maiores ordens d e, simultâneamente, os maiores valores de x. Para este conjunto de dados, um sub-grafo de genealogia representativo é originado a partir do vértice que representa o matemático alemão Heinz Hopf (seus valores são destacados em negrito na tabela), que tem $h_{(2)} = 6$ e é o único sub-grafo com estas dimensões, sendo que há 259 vértices neste sub-grafo hexanário completo.

Na Figura 3 ilustra-se o sub-grafo de genealogia, originado a partir de Heinz Hopf, identificado pelo maior índice-h genealógico expandido para a ordem 2. Pode-se verificar que existem seis descendentes diretos de Hopf onde cada um deles possue, também, seis descendentes com o mesmo grau de produtividade em termos de orientação acadêmica. Trata-se de indivíduos com desempenho similar em orientação acadêmica, considerando a descendência direta de cada um.

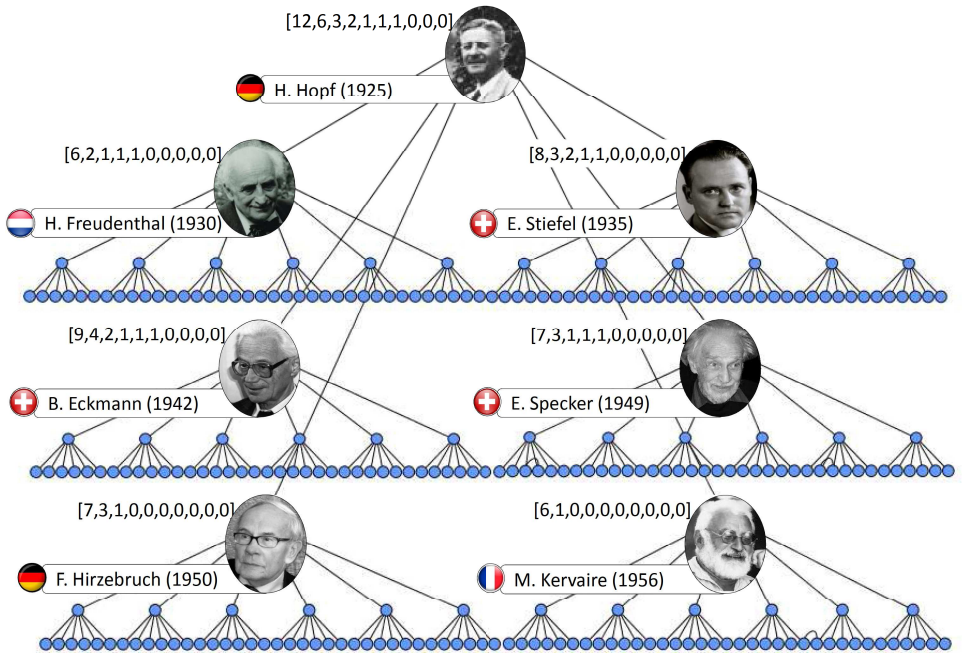


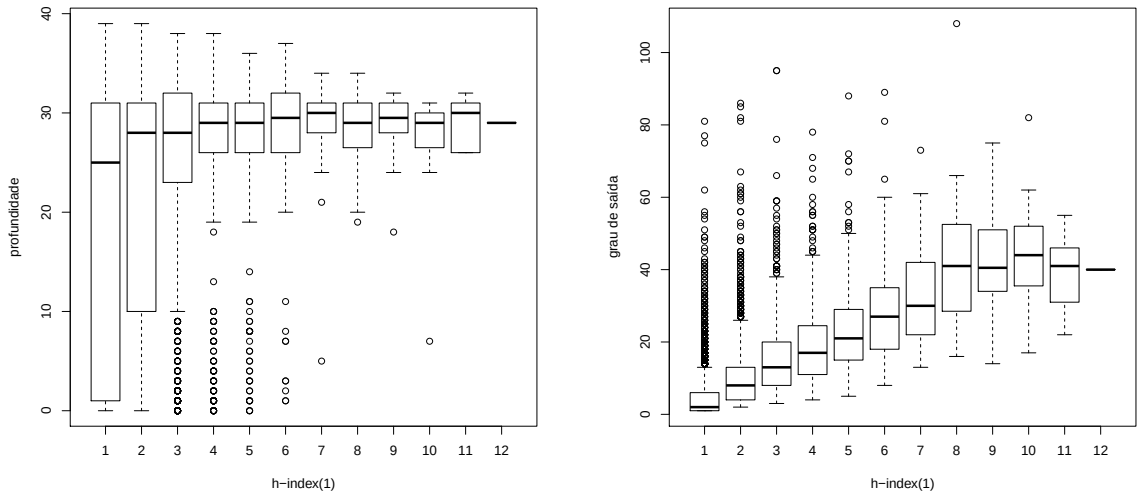
Figura 3. Grafo de genealogia de Heinz Hopf, identificado pelo índice-h genealógico expandido com 3 gerações de abrangência (ordem 2) e $h_{(2)} = 6$. Para cada matemático destacado é apresentado seu nome, o ano de titulação, o país de origem e seu respectivo vetor de índices-h para as 10 primeiras gerações.

A relevância da atividade de orientação acadêmica de Heinz Hopf pode ser verificada por meio do seu vetor de índices-h para outras ordens $h_{(d)} = [12, 6, 3, 2, 1, 1, 1]$, para d variando de 1 a 7. É importante notar que a comparação entre diferentes indivíduos é efetiva somente quando se utiliza a mesma ordem d para comparação ou o vetor completo aplicando algum método de classificação estatística. Apesar de Hopf ser o único matemático com $h_{(1)} = 12$ e $h_{(2)} = 6$ existem 16 indivíduos com $h_{(3)} = 3$. Para $h_{(4)} = 2$ são 35 no total.

A fim de estudarmos o grupo dos matemáticos sob a perspectiva do índice-h e do número de gerações posteriores ao matemático em questão (profundidade – maior caminho existente entre o vértice de interesse e outro sem descendente), na Figura 4(a), apresentamos as distribuições correspondentes índice-h de ordem 1. Para os resultados de $h_{(1)}$ variando de 1 a 12, observa-se que as medianas tendem a ser uniformes, indicando que o número de gerações posteriores para a maior parte dos matemáticos que apresentam valores de $h_{(1)}$ no intervalo especificado é em torno de 29. A dispersão nas distribuições diminui à medida que os resultados de $h_{(1)}$ aumentam.

Um grupo de matemáticos com especial desempenho pode ser encontrado por meio da identificação dos *outlier's*. Considerando que o número de gerações posteriores indica o quão remoto é o matemático, pode-se utilizar este parâmetro como complemento para identificação de desempenho. Analisando, por exemplo, a distribuição do número de gerações posteriores dos matemáticos com $h_{(1)} = 10$ identifica-se um único indivíduo (*outlier*) com este resultado, apresentando somente 7 gerações posteriores, enquanto seus pares apresentam de 24 a 31 gerações, sendo potencialmente indivíduos mais antigos.

Por outro lado, na Figura 4(b) apresentamos as distribuições dos valores de grau



(a) (b)

Figura 4. Distribuição dos índices-h sob a perspectiva: (a) da profundidade dos vértices (número de gerações posteriores), (b) do grau de saída dos vértices (número de orientados diretos).

de entrada para os matemáticos com o mesmo resultado de $h_{(1)}$. Existe um comportamento crescente, do grau de entrada, para os valores de 1 a 7. Este comportamento não é observado para valores maiores a 8. É importante destacar que, um comportamento semelhante é obtido para os índices-h com maiores ordens, i.e., para $d = 2, 3, 4, 5, 6$.

6. Conclusões e direcionamentos futuros

A genealogia acadêmica apresenta-se como uma importante opção à análise de publicações e citações, que atualmente é responsável por tudo que se sabe sobre o surgimento e desenvolvimento das disciplinas, a difusão do conhecimento e a evolução da ciência. O índice-h genealógico expandido, apresentado neste trabalho, utiliza o número de orientações para classificar um indivíduo e possibilita uma expansão do número de níveis (gerações) considerados. O desenvolvimento de métricas topológicas, como o índice-h genealógico expandido, e sua aplicação em grafos de genealogia acadêmica pode ser considerado como um meio efetivo de se mensurar e analisar a influência de orientadores acadêmicos em suas respectivas comunidades ao longo de diferentes gerações.

A estruturação de conjuntos de dados genealógicos mais heterogêneos, como os currículos disponíveis na Plataforma Lattes (Mena-Chalco *et al.*, 2014), em grafos de genealogia e a utilização de métricas para sua caracterização, e.g., (Tuesta *et al.*, 2015), pode resultar em importantes informações a respeito da formação, expansão e abrangência da comunidade acadêmico-científica do Brasil. Além de possibilitar analises sobre a interdisciplinaridade entre acadêmicos. Neste contexto, como trabalhos futuros pretendemos analisar os registros curriculares do banco de dados da plataforma Lattes e fazer seu mapeamento com o intuito de estudar a interdisciplinaridade na formação de recursos humanos (Rafols & Meyer, 2010).

Finalmente, é importante frisar que, este trabalho está alinhado com a epistemologia da análise de grande volume de dados (*Big Data*), sob a forma de ciência orientada

a dados, e às questões referentes à possibilidade de descoberta e/ou avaliação de teorias científicas universais, ferramentas instrumentistas, ou inferências indutivas como relatado por Frické (2015).

Agradecimentos

Os autores agradecem ao CNPq (Projeto Universal 461757/2014-1) e à CAPES pelo apoio financeiro concedido para a realização deste trabalho. Os autores agradecem também aos pareceristas anônimos pelas sugestões e comentários que contribuíram com o trabalho.

Referências Bibliográficas

J. ANDRAOS (2005). **Scientific genealogies of physical and mechanistic organic chemists**. *Canadian journal of chemistry* **83**(9), 1400–1414.

A. F. BENNETT & C. LOWE (2005). **The academic genealogy of George A. Bartholomew**. *Integrative and comparative biology* **45**(2), 231–233.

S. CHANG (2011). ***Academic Genealogy of Mathematicians***. World Scientific.

S. V. DAVID & B. Y. HAYDEN (2012). **Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience**. *PloS one* **7**(10), e46 608.

M. FRICKÉ (2015). **Big data and its epistemology**. *Journal of the Association for Information Science and Technology* **66**(4), 651–661.

R. E. HART & J. H. COSSUTH (2013). **A Family Tree of Tropical Meteorology's Academic Community and its Proposed Expansion**. *Bulletin of the American Meteorological Society* **94**(12), 1837–1848.

J. HIRSCH (2005). **An index to quantify an individual's scientific research output**. *Proceedings of the National academy of Sciences of the United States of America* **102**(46), 16 569–16 572.

A. JACKSON (2007). **A labor of love: the mathematics genealogy project**. *Notices of the AMS* **54**(8), 1002–1003.

D. C. JACKSON (2011). **Academic genealogy and direct calorimetry: a personal account**. *Advances in physiology education* **35**(2), 120–127.

R.D. MALMGREN, J.M. OTTINO & L.A.N. AMARAL (2010). **The role of mentorship in protégé performance**. *Nature* **465**(7298), 622–626.

J. P. MENA-CHALCO, L. A. DIGIAMPIETRI, F. M. LOPES & R. M. CESAR-JR. (2014). **Brazilian bibliometric coauthorship networks**. *Journal of the Association for Information Science and Technology* **65**(7), 1424–1445.

I. RAFOLS & M. MEYER (2010). **Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience**. *Scientometrics* **82**(2), 263–287.

L. ROSSI & J. P. MENA-CHALCO (2014). **Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos**. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 1–12. Brasília, DF, Brazil.

C. R. SUGIMOTO (2014). **Academic Genealogy**. In *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, B. CRONIN & C. R. SUGIMOTO, editors, 365–382. MIT Press, 1st edition.

E. F. TUESTA, K. V. DELGADO, R. MUGNAINI, L. A. DIGIAMPIETRI, J. P. MENA-CHALCO & J. J. PÉREZ-ALCÁZAR (2015). **Analysis of an Advisor-Advisee Relationship: An Exploratory Study of the Area of Exact and Earth Sciences in Brazil**. *PLoS ONE* **10**(5), e0129 065.

A. YONG (2014). **Critique of Hirsch's Citation Index: A Combinatorial Fermi Problem**. *Notices of the American Mathematical Society* **61**(9), 1040–1050.