Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos

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. *Documenting individuals and their relationships using the genealogy aims to obtain knowledge about the origin, evolution and characteristics of interrelated groups. This approach allows to understand the formation and future trends of groups. In this context, the characterization of the academic genealogy trees by topological metrics allows to categorize individuals screened by their academic lineage and enables to obtain important new knowledge for understanding the scientific scenario about an area. In this work, we present nine adapted and developed topological metrics to characterize academic genealogy trees. In order to show the feasibility of our characterization method by making use of topological metrics, we present an experiment focusing on the analysis of the genealogy of Johann Bernoulli (1667-1748), consisting of 81,768 mathematicians and 88,955 relationships of academic advising.*

Resumo. *Documentar indivíduos e seus relacionamentos utilizando a genealogia visa a obtenção de conhecimento sobre a origem, evolução e disseminação de grupos inter-relacionados. Essa tarefa de documentação auxilia o entendimento da formação e tendências futuras de grupos de pessoas. Nesse contexto, a caracterização de árvores de genealogia acadêmica, por meio de métricas topológicas, permite categorizar indivíduos através de sua linhagem acadêmica e possibilita a obtenção de novos conhecimentos importantes para a compreensão do cenário científico de uma área. Neste trabalho apresentamos nove métricas adaptadas e desenvolvidas para caracterizar árvores de genealogia acadêmica. A fim de demonstrar a viabilidade do nosso método de caracterização por meio da utilização de métricas topológicas, apresentamos testes preliminares voltados para a análise da genealogia de Johann Bernoulli (1667-1748), composto de 81.768 matemáticos e 88.955 relações de orientação acadêmica.*

1. Introdução

A genealogia é uma ciência auxiliar da história que estuda a origem, evolução e disseminação de grupos familiares (Malmgren *et al.*, 2010). O objeto de pesquisa da genealogia são os ascendentes e descendentes de um indivíduo. O processo de pesquisa envolvido na genealogia abrange a identificação de parentesco entre indivíduos através de registros históricos como certidões de nascimento, casamento, óbito, registro de propriedades e outros documentos que possam comprovar uma ligação entre indivíduos. Uma árvore genealógica é um grafo conexo acíclico que é comumente utilizado para documentar e facilitar o entendimento a respeitos de estudos de cunho genealógico (Hamberger *et al.*, 2011). Neste tipo de grafo, cada vértice representa um indivíduo na árvore e cada aresta indica a existência de algum tipo de relação entre os vértices. Neste contexto,

uma árvore de genealogia acadêmica é uma estrutura em que cada vértice é um orientador acadêmico e as arestas (direcionadas) representam as relações de orientação. Um conjunto de árvores pode ser denominado floresta.

A utilização da genealogia (Derrida *et al.*, 1999) como ferramenta para documentar e obter novos conhecimentos sobre grupos inter-relacionados é cada vez mais frequente em contextos acadêmico-científicos (Malmgren *et al.*, 2010; Chang, 2011; Hart & Cossuth, 2013; Mena-Chalco & Cesar-Jr., 2013). A estruturação de árvores de genealogia acadêmica, por meio de relações de orientação, pode ser de grande utilidade para o registro histórico de grupos atuantes em específicas áreas do conhecimento, onde os indivíduos de interesse (orientadores e orientados) são representados por vértices na árvore e os seus relacionamentos de orientação (e.g., orientações de doutorado ou supervisão de pós-doutorado) são representados por arestas. A obtenção da floresta de genealogia possibilita, também, a caracterização da área do conhecimento em questão por meio de métricas que permitem, através de análises estatísticas, *data mining* e técnicas de reconhecimento de padrões, extrair conhecimento relevante para a área que é objeto de estudo.

A importância deste tipo de análise se revela por meio da possibilidade de avaliar o impacto das orientações acadêmicas no desenvolvimento científico de específicas áreas do conhecimento e na identificação dos principais atores, ou grupos de maior relevância, que se destacaram por suas contribuições na proliferação do conhecimento através deste tipo de relação. A proposta deste trabalho é caracterizar árvores de genealogia por meio do desenvolvimento, adaptação e aplicação de métricas topológicas que permitam diferenciar os vértices das árvores, identificar grupos semelhantes e, de forma geral, promover um maior entendimento sobre este tipo de estrutura.

Este estudo está estruturado em cinco seções, além desta introdução. Na seção 2, apresentamos estudos que possuem temas correlatos ao aqui descrito. Na seção 3, apresentamos as métricas consideradas para a caracterização das árvores de genealogia. A estratégia utilizada para a obtenção dos dados e os resultados obtidos neste estudo são descritos nas seções 4 e 5, respectivamente. Finalmente, na seção 6 apresentamos os pontos de relevância obtidos com este estudo bem como possíveis trabalhos futuros.

2. Trabalhos correlatos

A análise de redes sociais é uma abordagem que origina-se em outras áreas do conhecimento (sociologia, psicologia social e antropologia) (Matheus *et al.*, 2006) e apresenta grande crescimento nos últimos anos devido ao (i) aumento da quantidade de dados disponíveis para análise, (ii) desenvolvimento das áreas de informática e processamento de dados e (iii) a ampliação dos assuntos de interesse e áreas do conhecimento que utilizam este tipo de análise. A utilização destas análises pode ser de grande valor para a obtenção de conhecimento sobre diversos grupos sociais e envolve quatro componentes principais: gerenciamento e estruturação de dados, descoberta de conhecimento, aprendizagem de máquina e técnicas de visualização (Freitas *et al.*, 2008).

A representação de indivíduos e seus relacionamentos na forma de redes apresenta-se como forma eficaz para extrair conhecimento em contextos, por vezes, de difícil interpretação. Caracterizar a ciência, como por exemplo a área da Ciência da Informação, e a contribuição que a análise de redes sociais proporciona para a sua correta interpretação é ainda um desafio. Nesse contexto, utilizar árvores de genealogia como

ferramenta para o estudo e descoberta de conhecimento sobre um grupo de indivíduos é uma estratégia eficiente de ampla aplicação. Um desafio importante, neste projeto, recai sobre a forma pela qual pode-se extrair conhecimento relevante a respeito de tais estruturas. Um estudo, não muito recente, a respeito das propriedades estatísticas das árvores de genealogia foi conduzido por Derrida *et al.* (1999) onde se busca, a partir da reconstrução da genealogia de um indivíduo pertencente à um pequeno grupo, medir a distribuição de seus ancestrais que aparecem mais de uma vez na árvore construída.

Diferentes estudos foram dedicados à documentação, análise e classificação de árvores de genealogia acadêmica através de relacionamentos de orientação. O trabalho *'A Labor of Love: The Mathematics Genealogy Project'* (Jackson, 2007) descreve o projeto, idealizado e implementado por Harry Coonce, sobre os relacionamentos de orientação acadêmica entre os doutores em matemática, e tem como principal objetivo 'compilar informações a respeito de *todos* os matemáticos do mundo' (uma comunidade científica seleta e pequena). O projeto, que em Março de 2014 disponibiliza, via *Web*, consulta a mais de 178.000 matemáticos em diversos períodos, apresenta resultados históricos muito significantes no que tange à documentação da área da matemática, porém neste projeto não foi contemplada uma análise ampla do conjunto de dados. É importante destacar que o conjunto de dados gerado com o projeto de genealogia matemática (*Mathematics Genealogy Project*) é uma base ímpar que ainda não foi explorada completamente.

No estudo sobre o papel das relações de orientação acadêmica no desempenho dos orientados, Malmgren *et al.* (2010) utilizaram a genealogia dos matemáticos como base, estruturando-os por meio de suas relações de orientação acadêmica. As análises apresentadas foram referentes à um subconjunto de 7.259 matemáticos, com graduação ocorrida entre 1900 e 1960, e suas respectivas contagens de descendentes (fecundidade). O referido estudo apresenta resultados interessantes, utilizando análises estatísticas, para a compreensão, em escala temporal, do desenvolvimento do grupo pesquisado e correlações existentes entre fecundidade e outras medidas de desempenho acadêmico.

Por outro lado, a identificação do impacto que uma orientação acadêmica exerce sobre o orientado, a utilização dos registros do projeto de genealogia matemática e o entendimento de como a comunidade dos matemáticos se desenvolveu, são itens abordados por Narayan (2011). O conjunto de dados obtido (137.138 matemáticos e seus relacionamentos) foi modelado em diferentes tipos de grafos considerando os relacionamentos, primeiramente, como arestas direcionadas, posteriormente, como arestas não-direcionadas e os relacionamentos entre *irmãos* (quando dois ou mais indivíduos tiveram o mesmo orientador), de modo à possibilitar a análise dos grafos sob diferentes perspectivas.

A utilização de relacionamentos entre *irmãos* ou redes de parentesco (*kinship networks*) resulta em árvores de genealogia de composição mista. As arestas direcionadas (que indicam relacionamento *top-down* é utilizada comumente para interligar pais e filhos). As arestas não-direcionadas (que indicam relacionamento, como por exemplo casamento, onde não existe uma orientação de origem e destino) são menos frequentes neste tipo de abordagem. Essa forma de modelar às árvores é descrita no estudo de Hamberger *et al.* (2011) e demonstra as possibilidades de avaliação com diferentes estruturas.

A importância deste tipo de estudo também pode ser verificada no projeto *Neurotree* (David & Hayden, 2012). Em concordância com o projeto dos matemáticos, a

área da neurociência (outra comunidade científica seleta e pequena) também busca a compreensão da ciência através do estudo de sua genealogia. Uma dificuldade comum em ambos os projetos é a identificação dos orientadores e suas relações de orientação. O projeto *Neurotree* (`http://neurotree.org/neurotree`) foi pautado na obtenção das informações da área e, pela primeira vez em projetos deste tipo, na interpretação das árvores de genealogia acadêmica constituídas. A utilização de métricas de avaliação de árvores apresentou resultados interessantes na caracterização da área da Neurociência. Este projeto conta com, aproximadamente, 40.000 pesquisadores e 60.000 relacionamentos cadastrados. Iniciativas similares são observadas para a comunidade científica dos Físicos (`http://academictree.org/physics`) e, de forma mais ampla, para os acadêmicos titulados com doutorado (`http://phdtree.org`). Estes projetos são, inicialmente, pautados na obtenção e documentação de seus membros, não oferecendo análises destes conjuntos de dados.

Documentar a história e compreender a expansão de grupos com interesses comuns, destacando principalmente as comunidades acadêmicas, passa obrigatoriamente pela utilização da genealogia e, consequentemente, pela construção de árvores genealógicas. A utilização da genealogia foi o caminho para o estudo de um seleto grupo de meteorologistas tropicais, apresentado por Hart & Cossuth (2013). Os resultados desse estudo motivaram, devido às características de interdisciplinaridade dos indivíduos pertencentes à árvore, a ampliação da busca por indivíduos fora dos limites da área.

O desenvolvimento de métodos para caracterizar árvores genealógicas é parte importante do trabalho de gerar conhecimento por meio destas estruturas. Estudos neste sentido, como o de Griffiths (1987), demonstram a viabilidade da caracterização de árvores de genealogia por meio de métricas específicas. No nosso trabalho, exploramos nove métricas topológicas adaptadas e desenvolvidas para caracterizar árvores de genealogia acadêmica.

3. Métricas em grafos para a caracterização de árvores de genealogia

As árvores de genealogia podem ser caracterizadas por meio de métricas de avaliação de grafos. Estas métricas têm como objetivo caracterizar o indivíduo, ou seja, atribuir um valor numérico que possa ser utilizado para qualificar este indivíduo pela topologia de sua árvore, de forma a descobrir informações ou padrões que possam auxiliar à uma compreensão a respeito de sua formação, capacidade de propagação e diferenciação entre as outras árvores da floresta. As métricas consideradas neste estudo são descritas a seguir.

- **Fecundidade**[1]**.** O objetivo desta métrica é dimensionar a árvore por meio do número de vértices que ela apresenta. É uma métrica importante para a classificação de um vértice raiz com base na quantidade de descendentes que ele influenciou. A fecundidade (f) é estimada considerando a somatória do número de vértices existentes em cada nível, m_i, da árvore $f = \sum_{i=1}^{m} (n_i)$ onde n_i é o número de vértices no nível m_i.
- **Fecundidade ponderada.** Esta métrica tem objetivo similar ao da fecundidade, sua principal característica é a atribuição de um peso maior para os vértices que estão mais próximos do vértice raiz. Os relacionamentos diretos têm maior peso

[1] As métricas *fecundidade e fecundidade ponderada* foram adaptadas do trabalho de árvores de genealogia dos neurocientistas descrito por David & Hayden (2012).

no cálculo do valor da métrica. A fecundidade ponderada (fp) reflete o potencial de um vértice em se relacionar com outros vértices (orientação acadêmica) e sua influência na propagação de relacionamentos. Neste trabalho, a fp utiliza como fator de ponderação a distância existente entre o vértice raiz e seus descendentes, $fp = \sum_{i=1}^{m} \left(\frac{n_i}{i^2}\right)$, onde n_i é o número de vértices no nível i da árvore. Esta métrica reduz o impacto da quantidade de vértices pertencentes à linhagem de um vértice raiz a medida que estes se distanciam.

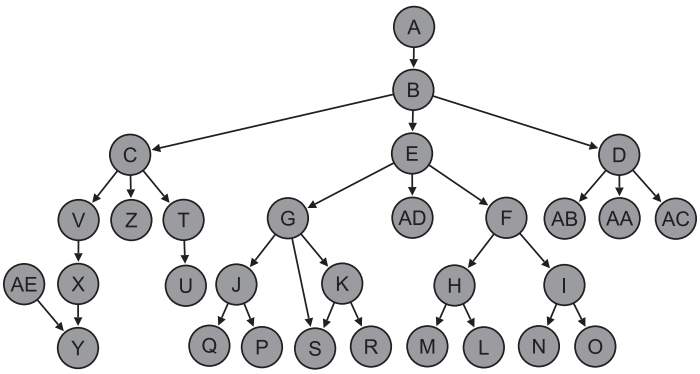
- **Número de folhas.** Definida como a quantidade de vértices não fecundos, ou seja, aqueles que não têm nenhum filho. O número de folhas (nf) totaliza a quantidade de vértices, na árvore, que não orientaram alunos. Comumente, isso acontece quando um pesquisador não segue a vida acadêmica ou quando estiver no início da vida acadêmica. Por outro lado, a quantidade de folhas existente no último nível da árvore, poderia nos indicar que esta estrutura tem potencial de crescimento, visto que estas folhas tendem a se propagar. Já a quantidade de folhas observadas nos níveis intermediários indicariam vértices com potencial esterilidade, pois não procriaram em tempo hábil.
- **Profundidade[2].** O objetivo da métrica profundidade (p) é fornecer o grau de maturidade da árvore genealógica formada a partir de um vértice raiz. Ela mede a quantidade de arestas existentes entre o vértice raiz e um vértice mais distante que possa ser alcançado. A métrica profundidade (p) pode ser definida por: $p = max(d(i,j))$, onde $d(i,j)$ corresponde à distância geodésica entre os vértices i e j.
- **Largura.** A métrica largura tem como objetivo medir a quantidade de relacionamentos diretos que um vértice raiz possui. Representa a quantidade de orientados existente no nível imediatamente posterior ao nível do orientador (vértice raiz) e reflete a produtividade (em termos de orientação) direta deste. A largura (l) é uma medida simples usada para classificar um orientador. Trata-se de uma análise quantitativa importante, porém, pouco diz sobre a qualidade da orientação.
- **Maior largura[3].** A métrica maior largura (ml) tem como objetivo identificar o maior número de relacionamentos em um mesmo nível da árvore genealógica. Este valor demonstra o quão ampla foi a influência de um vértice raiz na propagação dos relacionamentos em sua árvore.
- **Distância média[4].** A definição de proximidade entre um vértice raiz e todos os vértices pertencentes à sua ascendência é o objetivo da métrica distância média (dm). Neste contexto, quanto menor for o valor da dm, maior é a proximidade existente entre os vértices de uma árvore. A dm é a média dos comprimentos dos caminhos possíveis entre um vértice raiz e os vértices pertencentes à sua linhagem, e é definida por $\frac{1}{n}\sum_{i \neq j} d(i,j)$, onde $d(i,j)$ é a distância (quantidade de arestas existentes) entre os vértices i e j, n é a quantidade de caminhos contabilizados.
- **Média dos menores caminhos.** Esta métrica apresenta um objetivo similar ao da distância média, sua principal diferença está nos caminhos utilizados para o cálculo. Objetiva-se com esta métrica ponderar o valor obtido. Assim, esta medida representa a distância média entre os indivíduos pertencentes à árvore. A média dos menores caminhos (mmc) é definida por $\frac{1}{n(n-1)}\sum_{i \neq j} d(i,j)$, onde $d(i,j)$ é a

[2] As métricas *profundidade, largura e número de folhas* foram adaptadas da Teoria dos Grafos.
[3] A métrica *maior largura* foi desenvolvida para este trabalho.
[4] As métricas *distância média e média dos menores caminhos* foram adaptadas da Teoria dos Grafos.

distância, quantidade de arestas existentes, entre os vértices i e j, n é a quantidade de caminhos contabilizados.

- **Índice H.** O índice H genealógico (h) de um vértice é definido como o maior número h de relações que este vértice possui com outros vértices que tenham, pelo menos, o mesmo número h de relacionamentos cada um[5]. O objetivo desta métrica é considerar a quantidade e a qualidade genealógica (no sentido de perpetuidade) dos relacionamentos dos vértices da árvore.

Para exemplificar as métricas, apresentamos na Figura 1 uma árvore de genealogia e os resultados dos cálculos das respectivas métricas para os vértices mais representativos da árvore.



Vértice	f	fp	nf	p	l	ml	dm	mmc	h
A	28	3,51	15	5	1	9	3,52	0,13	1
B	27	6,53	15	4	3	9	2,64	0,10	3
C	6	3,61	3	3	3	3	1,43	0,24	1
D	3	3,00	3	1	3	3	0,75	0,25	0
E	15	5,03	9	3	3	7	2,13	0,14	2

Figura 1. Exemplo de árvore de genealogia com os respectivos resultados das métricas calculadas para seus principais vértices.

4. Conjunto de dados utilizados

Utilizamos em nosso estudo o conjunto de indivíduos pertencentes à linhagem de Johann Bernoulli, importante matemático de Basel (Basiléia) que, juntamente com Newton e Leibniz, é considerado um dos fundadores do cálculo. Os dados foram obtidos por meio de consultas recursivas ao *website* do projeto de genealogia de Matemáticos (*Mathematics Genealogy Project*, `http://genealogy.math.ndsu.nodak.edu`), onde, através do fornecimento de um identificador numérico exclusivo (ID), temos acesso a uma página *html* com informações sobre o matemático em questão. Em cada consulta foram obtidos: (i) os ID's referentes aos matemáticos orientados pelo indivíduo em questão e (ii) seu nome completo. As consultas recursivas foram realizadas em fevereiro de 2014 e totalizaram 81.768 matemáticos e 88.955 relacionamentos. É importante ressaltar que os resultados apresentados pelo projeto dos matemáticos é de grande relevância, a motivação deste trabalho é baseada na assertividade destes resultados e na possibilidade de, por meio

[5] O índice H, proposta por Hirsch (2005), é uma métrica que combina quantidade (número de publicações) e qualidade (número de citações) da produção acadêmica.

das métricas topológicas, aprofundar as análises e, consequentemente, o conhecimento a respeito dos indivíduos envolvidos e da estrutura resultante de seus relacionamentos.

O conjunto de dados foi utilizado para povoar um banco de dados em estrutura de grafo, por meio da plataforma Neo4j (banco de dados orientado à grafos). A escolha deste tipo de estrutura se justifica pelo ganho de desempenho que pode ser obtido quando comparado à outras estruturas relacionais. Cada matemático obtido é representado, no banco de dados, como um vértice da árvore e para cada relação de orientação acadêmica existente é adicionado uma aresta (direcionada) ligando o orientador ao orientado.

5. Resultados

A árvore resultante da estruturação dos descendentes de Johann Bernoulli e seus relacionamentos apresenta a profundidade de 20 e a maior largura de 20.242. A fecundidade e fecundidade ponderada do vértice raiz são 81.767 e 623,63, respectivamente. Apesar dos valores expressivos apresentados, Bernoulli orientou somente quatro matemáticos ($l = 4$), destes apenas dois tiveram alunos ($h = 2$) e 80,69% dos indivíduos pesquisados não orientaram alunos ($nf = 65.977$).

A Figura 2 ilustra a árvore composta pela linhagem de Johann Bernoulli a título de visualizar sua magnitude e estrutura. O vértice existente na parte superior da figura representa a raiz da árvore (Johann Bernoulli) e sua descendência é apresentada nos 20 níveis inferiores. Vértices e arestas com maior contraste indicam uma sobreposição destes elementos.

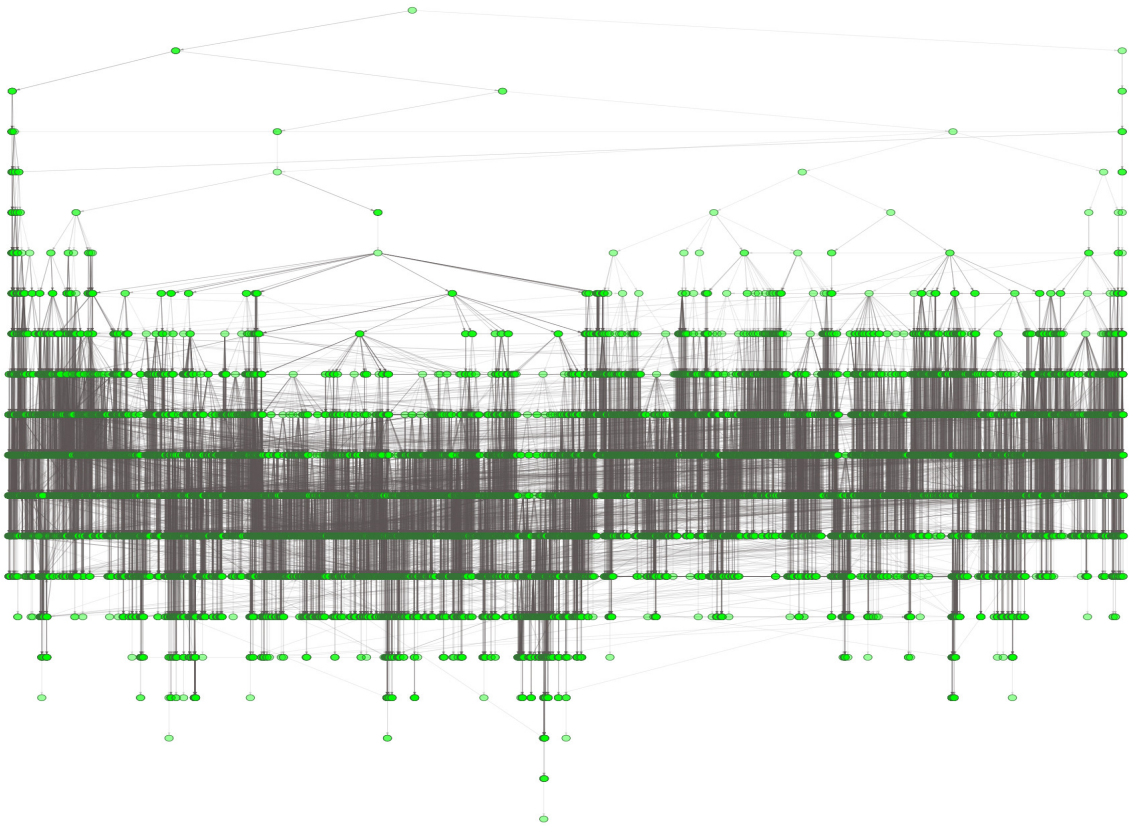


Figura 2. Árvore genealógica de Johann Bernoulli.

A proposta de caracterização de árvores de genealogia, por meio das relações de orientação acadêmica, foi implementada utilizando-se a árvore acima descrita, os resultados observados foram divididos em três tópicos principais: (i) classificar os matemáticos (*ranking*) por meio de seus resultados, (ii) identificar grupos (*clusters*) que compartilhem características ou atributos semelhantes, e (iii) análise da distribuição das frequências observadas para intervalos pré-estabelecidos.

5.1. Classificação dos matemáticos

As medidas utilizadas neste estudo representam diferentes aspectos das árvores analisadas. Métricas que utilizam apenas contagens, sem ponderação, para o cálculo de seus valores são de representatividade exclusivamente quantitativa e importantes para dimensionar a árvore derivada da linhagem de um vértice raiz. Dimensionar uma árvore utilizando apenas medidas baseadas em contagens, pode produzir classificações inconsistentes, visto que estamos atribuindo um valor numérico individual para um determinado vértice e, este valor, não é resultado apenas dos relacionamentos diretos do vértice em questão, mas também dos relacionamentos de seus descendentes. Para minimizar este tipo de inconsistência na classificação de vértices, medidas como fp e h, que no cálculo de seus resultados apresentam alguma ponderação, atribuem um maior grau de classificação para os relacionamentos diretos do vértice analisado, refletindo com maior assertividade o desempenho do próprio indivíduo em questão e, consequentemente, sua qualidade em termos de orientação acadêmica. Para avaliar a densidade de uma árvore, ou seja, a proximidade existente entre os vértices que a compõem, utilizamos as medidas dm e mmc.

A Tabela 1 apresenta os dez matemáticos melhores colocados em cada uma da métricas avaliadas. Em concordância com o objetivo das métricas utilizadas, observamos uma tendência de um indivíduo que figura nas primeiras posições de um *ranking* de medida com base quantitativa (e.g., *ranking* f - Johann Bernoulli) figurar, também, no topo de outras medidas de mesma base (e.g., *ranking* fo - Johann Bernoulli). Por outro lado, as medidas ponderadas ou normalizadas (e.g., fp) apresentam matemáticos diferentes nas primeiras posições, sugerindo que o desempenho destes indivíduos foi mais relevante em termos de contribuição direta com orientação acadêmica. Como exemplo da importância do trabalho realizado pelo matemático, consideremos o primeiro colocado no *ranking* h igual a 12, Heinz Hopf. Isso indica que este matemático orientou, no mínimo, 12 alunos que, por sua vez, orientaram, no mínimo, outros 12 alunos cada um. Um desempenho impressionante, não sendo possível encontrar outro igual na linhagem de Bernoulli. Os resultados ligados à densidade das árvores (e.g., dm), para este conjunto de dados, apresentou uma alta correlação com a magnitude de sua árvore, ou seja, a densidade é diretamente proporcional ao tamanho da árvore em questão.

A correta classificação dos matemáticos, identificando os indivíduos mais relevantes quanto à realização e proliferação da atividade de orientação acadêmica, pode ser feita analisando as medidas calculadas de maneira individual. Conforme discutido anteriormente, cada grupo de métricas de avaliação reflete uma característica importante a respeito da árvore de genealogia (i.e., quantidade e qualidade das relações e densidade da árvore), porém, considerando o conjunto das métricas pode-se identificar os indivíduos mais prolíficos em forma de grupos (*clusters*) com características similares.

Tabela 1. *Ranking* dos matemáticos pertencentes à árvore genealógica de Bernoulli para cada métrica calculada.

Fecundidade		Fecundidade ponderada		Número de folhas	
J. Bernoulli	81767	C. F. Klein	1326,63	J. Bernoulli	65977
L. Euler	81578	S. Poisson	1099,42	L. Euler	65843
J. Lagrange	78218	D. Hilbert	1093,95	J. Lagrange	63216
S. Poisson	78215	C. L. F. Lindemann	1082,23	S. Poisson	63215
J. B. Fourier	45929	R.Lipschitz	901,72	J. B. Fourier	37713
G. Dirichlet	45927	J. Lagrange	868,36	G. Dirichlet	37712
R. Lipschitz	43954	G. Dirichlet	754,42	C. F. Klein	36135
C. F. Klein	43953	L. Euler	751,92	R. Lipschitz	36135
C. L. F. Lindemann	32069	E. H. Moore	710,70	C. L. F. Lindemann	26415
M. Chasles	31734	M.Chasles	633,01	M. Chasles	25302
Profundidade		**Largura**		**Maior largura**	
J. Bernoulli	20	C. C. J. Kuo	120	J. Bernoulli	20242
L. Euler	19	R. Temam	111	L. Euler	20226
J. Lagrange	18	L. Ornstein	95	J. Lagrange	19361
S. Poisson	17	W. Jager	91	S. Poisson	19361
J. B. Fourier	17	L. Prandtl	88	C. F. Klein	11878
G. Dirichlet	16	A. Kolmogorov	82	R. Lipschitz	11878
R. Lipschitz	15	R. Eden	80	G. Dirichlet	11515
M. Chasles	15	C. Ehresmann	78	J. B. Fourier	11515
J. Hennert	15	B. De Moor	77	C. L. F. Lindemann	9911
C. F. Klein	14	E. Krause	76	M. Chasles	8401
Distância média		**Média menores caminhos**		**Índice h**	
J. Bernoulli	11,90	A. V. Perez	0,5	H. Hopf	12
L. Euler	10,91	R. Mazet	0,5	E. Schmidt	11
J. Lagrange	9,95	R. Oldenburger	0,5	H. Behnke	11
S. Konig	9,57	G. Glaeser	0,5	R. Baer	11
J. Hennert	9,54	W. Krolikowski	0,5	C. F. Klein	10
J. B. Fourier	9,43	A. Chaudoir	0,5	R. L. Moore	10
S. Poisson	8,95	F. Pfeiffer	0,5	S. Bochner	10
P. Nieuwland	8,70	U. N. de Alba	0,5	H. Kneser	10
C. Damen	8,61	J. L. Chaboche	0,5	A. Kolmogorov	10
A. Brugmans	8,59	A. Vacroux	0,5	J. L. Lions	10

5.2. Identificação de grupos similares

A identificação dos matemáticos mais relevantes em cada medida apresentada não permite uma avaliação global, com a utilização das medidas em conjunto. Para realizar essa classificação as dimensões obtidas, ou seja, as nove métricas, foram reduzidas para apenas duas dimensões por meio da análise de componentes principais (PCA). O método PCA consiste da utilização de combinações lineares dos dados originais com o objetivo de reduzir suas dimensões para obter formas representativas destes dados. A PCA é considerada uma 'transformação linear ótima' e apresenta-se como uma ferramenta muito útil para os processos ligados a reconhecimento de padrões.

A Figura 3a apresenta o gráfico com os matemáticos diagramados nas duas dimensões ou componentes principais obtidas. Ambas componentes concentram cerca de 82% da variância total. Podemos identificar três grupos distintos (A, B e C). O grupo A reúne somente 4 indivíduos com destaque evidente, *Simeon Poisson*, *Leonhard Euler*, *Johann Bernoulli* e *Joseph Lagrange* nesta ordem. Trata-se de um grupo de elite, matemáticos com relevância histórica, que apresentam, em suas biografias, diversas contribuições na evolução da matemática (Chang, 2011).

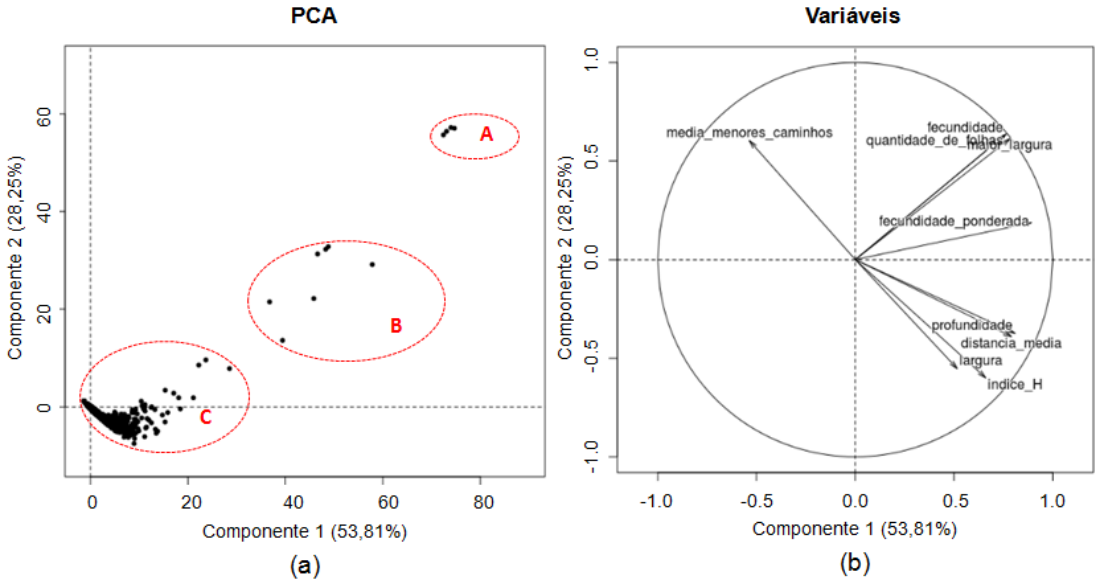


Figura 3. Análise de componentes principais: (a) conjunto de dados representados nas duas primeiras componentes principais. (b) orientação das variáveis (métricas) consideradas.

No grupo B, podemos observar a presença de 7 matemáticos importantes (*C. Felix Klein*, *Rudolf Lipschitz*, *Gustav Dirichlet*, *Jean-Baptiste Fourier*, *C. L. Ferdinand Lindemann*, *David Hilbert* e *Michel Chasles*) porém com relevância inferior aos anteriores. O último grupo reúne o restante dos indivíduos analisados, podendo ainda serem diferenciados entre si.

A Figura 3b apresenta a orientação obtida para as métricas analisadas. As métricas f, nf e ml são, praticamente, de mesma orientação. Já a métrica fp apresenta uma orientação diferenciada das anteriores. Medidas com orientações muito semelhantes indicam que a informação fornecida por elas também é semelhante ou até redundante. Há uma forte correlação, também, nas métricas p e dm que apresentam uma tendência próxima às métricas l e h, que também se correlacionam bem. Por fim, a mmc mostra-se com uma orientação oposta às últimas citadas, isso se deve pelas próprias características da medida. Aqui é importante notar que, com a utilização do método PCA, é possível diferenciar a maioria dos vértices da árvore genealógica.

5.3. Distribuição das frequências

Os resultados das métricas calculadas, para a árvore genealógica de Bernoulli, foram divididos em dez intervalos, onde cada um deles representa 10% do intervalo completo, e realizado uma contagem para identificar a frequência de valores obtidos para cada um deles. A Figura 4 apresenta os gráficos dos intervalos e suas respectivas frequências para as métricas utilizadas.

Para todos os casos, observamos que os valores de métricas que estão entre os 10% menores resultados, concentram a grande maioria das ocorrências, confirmando, para o conjunto de dados analisados, o princípio de Pareto ou Lei da Potência (Malmgren *et al.*, 2010). Na Figura 4a, observamos que existem poucos matemáticos com valores de f compreendidos na faixa de 90% do intervalo considerado. Este tipo de representação é devido à estrutura que as árvores apresentam (a propagação dos vértices) à medida que descemos

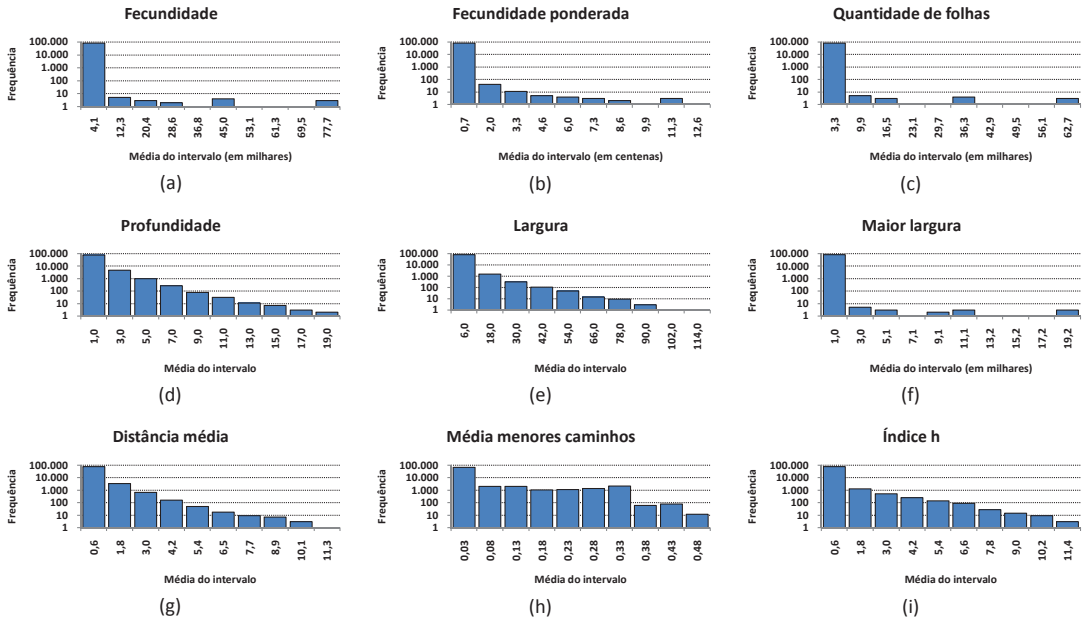


Figura 4. Distribuição das frequências observadas para cada intervalo de valores das métricas. O eixo das frequências é apresentado em escala logarítmica.

aos níveis inferiores da árvore, identificamos um crescimento geométrico, resultando em uma quantidade de vértices maior nos níveis inferiores e pequenas quantidades no topo da árvore. Podemos verificar que esta mesma configuração é repetida nos gráficos apresentados nas Figuras 4c e 4f, estas métricas são, predominantemente, resultado de contagens sem a aplicação de nenhum método de normalização ou ponderação. Quando utilizamos métricas que são ponderadas, as distribuições mantém as características das estruturas das árvores, porém, com uma maior uniformidade na distribuição das frequências dos intervalos. As métricas que apresentam essa uniformidade são representadas nos gráficos das Figuras 4b, 4d, 4e, 4g e 4i. Finalmente, a métrica mmc, Figura 4h, apresenta um padrão mais linear, quando comparado às métricas anteriores, com exceção feita aos 10% menores valores que apresentam conformidade com as demais métricas.

6. Conclusões

A caracterização de redes sociais, especificamente redes estruturadas em forma de árvores genealógicas, é uma importante forma de se obter conhecimento a respeito destas estruturas. Neste contexto, neste trabalho foi apresentada uma proposta de caracterização de árvores de genealogia considerando métricas de avaliação de grafos. A classificação dos indivíduos e a identificação de grupos com características comuns foram consideradas e podem contribuir para a compreensão de grupos inter-relacionados, sejam estas relações de orientação acadêmica ou outro tipo de relacionamento.

Foi considerada a árvore de genealogia de J. Bernoulli como estudo de caso. Embora a linhagem de Bernoulli seja um conjunto de dados médio e não considerarmos atributos dos indivíduos (e.g., país de origem) nem de seus relacionamentos (e.g., ano da formação), os resultados aqui apresentados são relevantes e difíceis de serem obtidos apenas com a utilização de abordagens convencionais. Estes resultados correspondem a informações que até agora não foram tratadas por outras pesquisas.

O projeto e aplicação de novas métricas, a consideração de atributos para os vértices e as arestas e o aprofundamento das análises sobre as estruturas obtidas podem enriquecer as análises e a descoberta de conhecimento. Nosso trabalho considera como direcionamentos futuros (i) aplicação do método em conjuntos de dados heterogéneos e de grande magnitude (e.g., CVs extraídos da Plataforma Lattes), (ii) utilização de diferentes atributos associados aos vértices e arestas, e (iii) identificação de subgrafos mais representativos nas árvores de genealogia (e.g., *motifs* (Milo *et al.*, 2002)).

Agradecimentos

Os autores agradecem à Fundação UFABC 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

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.

B. DERRIDA, S. C. MANRUBIA & D. H. ZANETTE (1999). Statistical Properties of Genealogical Trees. *Physicca Review Letters.* **82**, 1987–1990.

C. M. D. S. FREITAS, L. P. NEDEL, R. GALANTE, L. C. LAMB, A. S. SPRITZER, S. FUJII, J. P. M. DE OLIVEIRA, R. M. ARAUJO & M. M. MORO (2008). Extração de conhecimento e análise visual de redes sociais. *SEMISH-SBC* 106–120.

R. C. GRIFFITHS (1987). Counting genealogical trees. *Journal of mathematical biology* **25**(4), 423–431.

K. HAMBERGER, M. HOUSEMAN & R.W. DOUGLAS (2011). Kinship network analysis. *The Sage Handbook of Social Network Analysis* 533–549.

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).

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.

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

R. F. MATHEUS, F. S. PARREIRAS & T. A. S. PARREIRAS (2006). Análise de redes sociais como metodologia de apoio para a discussão da interdisciplinaridade na ciência da informação. *Ciência da Informação* **35**(1), 72–93.

J.P. MENA-CHALCO & R.M. CESAR-JR. (2013). *Bibliometria e Cientometria: reflexões teóricas e interfaces*, chapter Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes, 109–128. São Carlos: Pedro & João Editores.

R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII & U. ALON (2002). Network motifs: simple building blocks of complex networks. *Science* **298**(5594), 824–827.

P. NARAYAN (2011). *Mathematics Genealogy Networks*. Master's thesis, University of Oxford, United Kingdom.