

CRIAÇÃO DE GRAFOS DE TÓPICOS DO CONHECIMENTO BASEADA EM GENEALOGIA ACADÊMICA

Luciano Rossi

Universidade Federal do ABC, Santo André, SP, Brasil
luciano.rossi@ufabc.edu.br

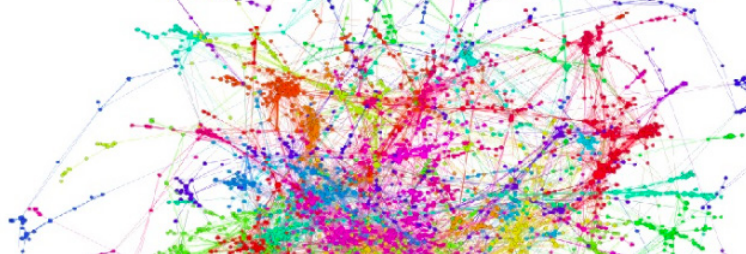
Jesús P. Mena-Chalco

Universidade Federal do ABC, Santo André, SP, Brasil
jesus.mena@ufabc.edu.br

1 INTRODUÇÃO

Tudo o que sabemos a respeito do surgimento e do desenvolvimento do conhecimento científico, na forma de disciplinas ou tópicos, é resultado predominantemente da análise de publicações. Neste contexto, a observação de possíveis influências entre tópicos do conhecimento envolve a identificação dos tópicos e das relações entre eles. O problema de identificar o fluxo do conhecimento passa pela obtenção de uma rede de tópicos que evidencie a estrutura topológica derivada dos seus relacionamentos. Assim, uma rede de tópicos do conhecimento é o objeto de estudo que viabiliza as análises referentes ao fluxo do conhecimento científico e constituem, respectivamente, o objetivo e o problema de pesquisa deste trabalho.

A criação de uma rede de tópicos do conhecimento é importante por possibilitar a realização de análises sobre as influências existentes entre esses tópicos e, conseqüentemente, sobre a interdisciplinaridade no desenvolvimento da ciência. Comumente, a representação de uma rede deste tipo é feita por meio de uma estrutura matemática denominada grafo direcionado. Um grafo é uma estrutura abstrata de dados que permite representar diferentes elementos e seus relacionamentos como vértices e arestas, respectivamente.



Neste trabalho, propomos um novo método cujo objetivo é a criação de um grafo de tópicos do conhecimento baseado em dados de genealogia acadêmica (GA). O método contempla (i) a prospecção de dados de GA, (ii) a estruturação hierárquica dos dados em grafo de GA, (iii) a identificação das áreas de atuação dos acadêmicos, (iv) a inferência das áreas de atuação por meio da propagação de tópicos no grafo de genealogia e (v) a criação do grafo de tópicos do conhecimento, em que os vértices e as arestas representam, respectivamente, os tópicos e suas influências que são herdadas dos relacionamentos de orientação acadêmica.

Como estudo de caso utilizamos a GA do proeminente matemático suíço Johann Bernoulli (1667-1748). A identificação das áreas de atuação dos matemáticos considerou as informações biográficas disponíveis no repositório *Wikipedia* e dois dicionários de tópicos baseados no sistema *Mathematics Subject Classification* e em um glossário de áreas da matemática, respectivamente. Como resultado preliminar, destacamos a criação do grafo de tópicos de pesquisa e sua análise descritiva, a qual considera os tópicos do conhecimento, as relações hierárquicas existentes entre eles e o processo de transição entre os tópicos.

2 TRABALHOS CORRELATOS

A estruturação e análise de redes de conhecimento é objeto de estudo em diferentes áreas. Os estudos disponíveis na literatura podem considerar redes de publicações científicas que são relacionadas por critérios específicos como, por exemplo, os leitores destas publicações (MOHAMMADI; THELWALL, 2014). Por outro lado, as redes podem descrever elementos distintos, como publicações científicas e patentes, de modo a evidenciar a transferência do conhecimento científico teórico para aplicações tecnológicas (DING; HUNG; LEE; WANG, 2017). As redes de patentes podem ser úteis, também, para análises que envolvam a identificação do impacto que a geolocalização das fontes de conhecimento exercem em diferentes atores (SORENSEN; RIVKIN; FLEMING, 2006). A análise do fluxo da informação em redes que representa os diversos atores envolvidos em processos de negócios e/ou



desenvolvimento tecnológico pode contribuir para a gestão da inovação (INOMATA; RADOS, 2011). Por fim, redes de citações podem ser importantes como objeto de estudo para a propagação do conhecimento (BUFREM; SILVA; SOBRAL, 2017).

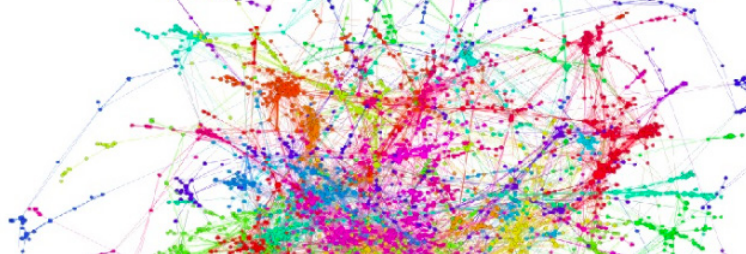
Os processos de identificação de tópicos que representem conhecimento específico também são amplamente documentados na literatura. Um diferencial importante entre as metodologias consideradas é a fonte de informações utilizada. A identificação de tópicos pode utilizar como fonte de informações fóruns *on-line* de discussão sobre tópicos de aprendizagem (TOBARRA; ROBLES-GÓMEZ; ROS; HERNÁNDEZ; CAMINERO, 2014) ou ainda a análise de expressões em títulos de publicações científicas (BOSCHMA; HEIMERIKS; BALLAND, 2014). Consideramos que, apesar de o estudo do conhecimento por meios de redes e a identificação de tópicos do conhecimento serem temas de pesquisa amplamente documentados, nosso trabalho difere dos demais quanto à utilização da GA para a estruturação hierárquica do conhecimento.

3 MÉTODO

O método está composto por quatro etapas em que a partir de um grafo de GA de um determinado grupo de pesquisadores são identificados todos seus tópicos de pesquisa. Caso não seja possível identificar os tópicos de pesquisa, é utilizado um processo de inferência de tópicos baseado na propagação de informação. Finalmente, com base no grafo de pesquisadores é criado outro grafo em que os vértices representam tópicos e as arestas as relações entre pares de tópicos.

3.1 CRIAÇÃO DO GRAFO DE GENEALOGIA ACADÊMICA

O método utilizado considera um conjunto de acadêmicos e seus relacionamentos de orientação/supervisão formal e concluída. A criação do grafo de genealogia utiliza, por exemplo, relacionamentos de orientação de mestrado e doutorado, aqui denominado grafo de GA, onde os vértices descrevem os acadêmicos e as arestas direcionadas descrevem todos os relacionamentos de orientação observados entre eles. Para um



determinado grupo de acadêmicos, um grafo de GA pode ser obtido por coleta direta em bancos de genealogia como, por exemplo, o *Mathematics Genealogy Project*, ou por métodos computacionais que inferem as relações de orientação. Neste último caso, os relacionamentos de orientação podem ser obtidos por meio de análises em publicações em coautoria ou em currículos acadêmicos.

3.2 IDENTIFICAÇÃO DOS TÓPICOS DE ATUAÇÃO

No processo de identificação dos tópicos de atuação do acadêmico consideramos seus respectivos registros biográficos obtidos por meio de busca em repositórios *Web*, utilizando o nome do acadêmico como elemento de busca. Após a prospecção dos registros na forma de textos, esses são submetidos a um processo de eliminação de elementos textuais que não sejam importantes para a identificação de seus tópicos de atuação, ou seja, são eliminados artigos, preposições, conjunções, dentre outros elementos similares. O resultado é um conjunto de termos, oriundos da descrição biográfica do acadêmico, que são utilizados para a inferência da área de atuação do acadêmico.

A identificação da área de atuação do acadêmico é guiada por um dicionário de tópicos do conhecimento que é derivado de algum sistema de classificação do conhecimento. Os sistemas de classificação do conhecimento são esquemas hierárquicos que organizam o conhecimento formal em classes. Esses sistemas podem ser utilizados para a composição do dicionário de tópicos do conhecimento, que consiste de um conjunto de termos. Este processo considera os termos pertencentes ao dicionário de tópicos como elementos de busca no registro, e tem por objetivo identificar a frequência de cada tópico no registro. Assim, um tópico é k -frequente se ele aparece k vezes no registro do acadêmico. Maiores valores de k tornam a identificação mais assertiva, por outro lado, valores menores de k possibilitam a identificação com menor restrição e menor assertividade. Ao final desse processo, os vértices do grafo de GA, para os quais existe registro biográfico identificado, são rotulados com um ou mais tópicos, sendo que este atributo associado descreve sua área de atuação.



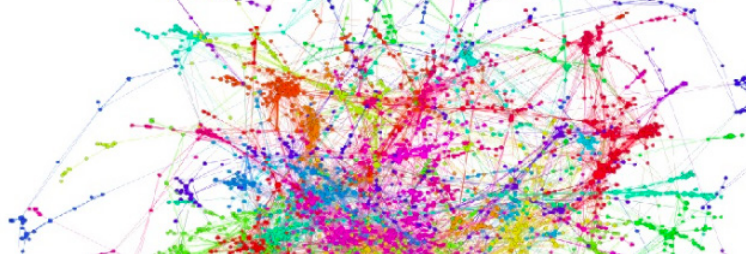
3.3 PROPAGAÇÃO DE TÓPICOS

Os vértices não rotulados (acadêmicos cujos tópicos não foram identificados) receberam tópicos inferidos por meio da propagação dos tópicos dos vértices rotulados (acadêmicos cujos tópicos foram identificados), considerando a estrutura do grafo de GA.

A propagação é um processo iterativo baseado na adjacência rotulada de um vértice não rotulado: a primeira iteração considera a descendência do vértice não rotulado, ou seja, os acadêmicos que foram orientados por ele. A atribuição de tópicos é feita por meio da interseção dos conjuntos de tópicos de seus descendentes. Caso haja apenas um descendente rotulado, seus tópicos serão atribuídos diretamente ao vértice não rotulado. Nesta iteração são submetidos ao processo todos os vértices não rotulados do grafo. Se um vértice não apresenta descendência rotulada ou se a interseção for um conjunto vazio, esse vértice mantém a condição de não rotulado. Na iteração seguinte, o processo é repetido com a interseção da ascendência rotulada, ou seja, consideram-se os orientadores do acadêmico em questão. A descendência e ascendência rotuladas são utilizadas de maneira alternada até um determinado número de iterações que é definido por um limiar de rotulação de vértices. A definição de um limiar é importante visto que a estrutura do grafo não garante a rotulação total de seus vértices.

3.4 CRIAÇÃO DO GRAFO DE TÓPICOS

Finalmente, o grafo de GA rotulado é utilizado para o processo de criação de um grafo de tópicos do conhecimento. O grafo resultante sumariza a interação entre tópicos. Cada tópico é representado por um vértice e dado o relacionamento AàB todos os tópicos associados a um orientador (A) são conectados a todos os tópicos associados a um orientado (B), ou seja, as arestas do grafo são resultantes do produto cartesiano entre os conjuntos de tópicos do orientador e do orientado ($A \times B$). Em outras palavras, todos os tópicos associados a um vértice orientador são conectados a todos os tópicos associados a um vértice orientado. Esta abordagem permite realizar a rotulação dos vértices que



não possuem registro biográfico associado evidenciando a transferência dos tópicos de um acadêmico a outro.

4 CONJUNTOS DE DADOS

Consideramos, como estudo de caso, a descendência acadêmica do matemático suíço Johann Bernoulli (Rossi & Mena-Chalco, 2014), a qual foi obtida junto à plataforma do *Mathematics Genealogy Project*¹. Essa plataforma reúne informações sobre a comunidade dos doutores em matemática e áreas correlatas. Foram obtidos todos os registros dos matemáticos da GA de Bernoulli. Esta estrutura é composta por 100.221 matemáticos e 110.061 relacionamentos de orientação concluída. Todos os matemáticos pertencentes à descendência acadêmica de J. Bernoulli foram submetidos ao processo de busca pelas respectivas biografias. O processo de busca considerou o nome completo dos matemáticos e foi feita junto ao repositório *Wikipedia*, que consideramos como fonte de dados para a composição dos registros biográficos associados aos matemáticos do grafo de GA. Ao todo, foram obtidos 13.477 registros descritivos ou biográficos (13,48% dos descendentes de Bernoulli). A opção por essa fonte de informações considerou que a existência de um registro no *Wikipedia* pressupõe a importância e a contribuição do matemático para o desenvolvimento da área.

Para a composição dos dicionários de tópicos utilizamos dois sistemas de classificação de áreas da Matemática: (i) o sistema *Mathematics Subject Classification*² (MSC), que disponibiliza 62 classes hierárquicas, e (ii) um glossário³ composto por 446 termos, ambos referenciando tópicos para a área da Matemática. O MSC é um sistema de classificação formal disponibilizado pela *American Mathematical Society*. Por outro lado, o

1 Disponível em: <genealogy.math.ndsu.nodak.edu>. Acesso em: 05 jan. 2018.

2 Disponível em: <mathscinet.ams.org/msc/msc2010.html>. Acesso em: 31 jan. 2018.

3 Disponível em: <en.wikipedia.org/wiki/Glossary_of_areas_of_mathematics>. Acesso em: 31 jan. 2018.



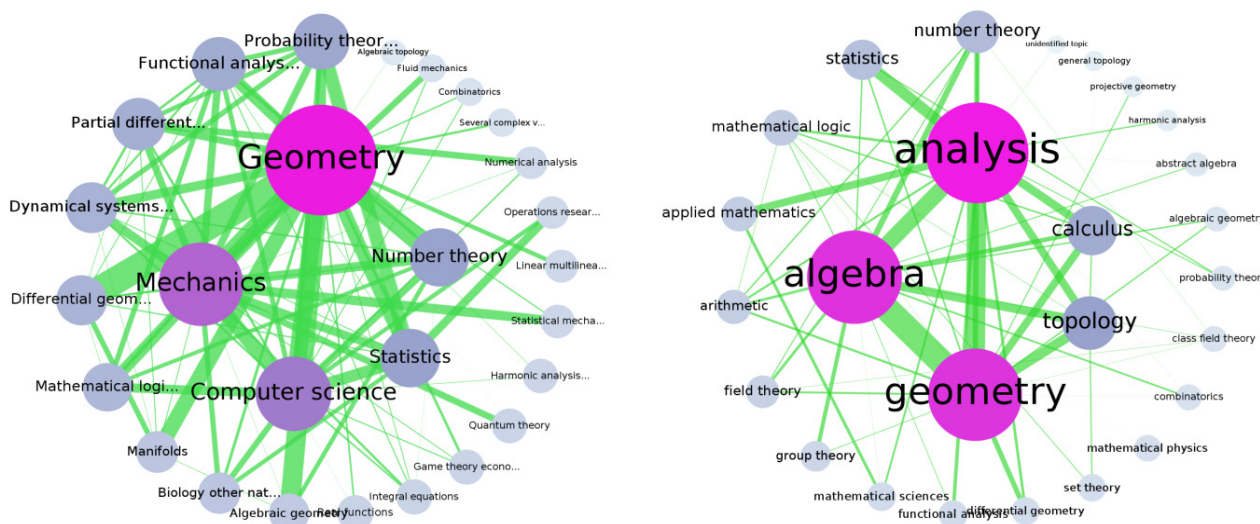
glossário de tópicos é um sistema informal não hierárquico que consiste em uma lista de termos que definem disciplinas ou áreas da Matemática.

5 RESULTADOS

Para J. Bernoulli, com o método proposto, foram coobtidos dois grafos de tópicos que correspondem a cada um dos dicionários utilizados. O processo de propagação dos tópicos considerou um limiar de 95% de vértices rotulados para finalização (isto é, a propagação foi realizada iterativamente até que 95% dos vértices sejam rotulados). A Figura 1 ilustra ambos os grafos cujos vértices representamos tópicos derivados do dicionário MSC, e do glossário de tópicos. O dimensionamento dos vértices é proporcional ao número de arestas que nele incidem ou que dele emergem (grau). A espessura das arestas é proporcional à sua frequência. Para facilitar a visualização das influências entre os vértices, são representados somente aqueles cujas arestas possuem frequência maior ou igual a 1.000. O grafo derivado do MSC, para o qual consideramos uma k -frequência igual no processo de identificação dos tópicos, apresenta 27 vértices. O tópico *Geometry* apresenta destaque tanto em número de relacionamentos com outros tópicos quanto em número de relacionamentos totais. Por outro lado, o grafo derivado do glossário de tópicos, para o qual consideramos uma k -frequência igual a 3, possui 26 vértices, de acordo com os mesmos critérios do grafo anterior, e o tópico em destaque neste grafo é o *Analysis*.



FIGURA 1 - GRAFOS DE TÓPICOS. OS VÉRTICES E ARESTAS TÊM DIMENSÃO PROPORCIONAL AO GRAU E A FREQUÊNCIA, RESPECTIVAMENTE. SÃO REPRESENTADAS SOMENTE AS ARESTAS COM FREQUÊNCIA ≥ 1.000 .



Fonte: Elaborado pelos autores, 2018.

A Tabela 1 apresenta os cinco tópicos mais influentes e os cinco tópicos mais influenciados por cada tópico influente, com os respectivos números de arestas. Note que os grafos de tópicos do MSC e do glossário de tópicos apresentam originalmente 860.097 e 888.225 arestas, respectivamente.



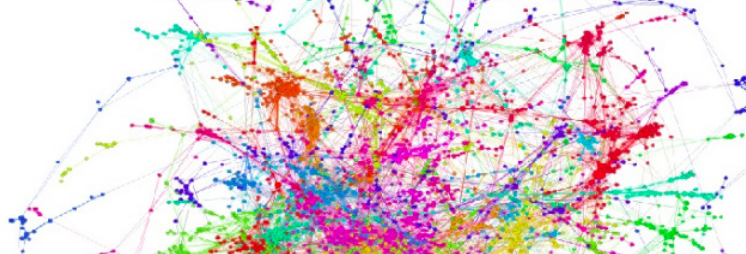
TABELA 1 - IDENTIFICAÇÃO DOS CINCO TÓPICOS MAIS INFLUENTES COM SEUS NÚMEROS DE ARESTASE OS RESPECTIVOS TÓPICOS INFLUENCIADOS PARA CADA DICIONÁRIO CONSIDERADO

Dicionário	Tópicos (número de relacionamentos)					
	Influentes	Influenciados				
MSC	Geometry (100.598)	Differential geometry (7.054)	Number theory (5.663)	Mechanics (5.363)	Algebraic geometry (5.275)	Computer science (4.836)
	Computer science (69.181)	Geometry (4.894)	Mechanics (4.108)	Statistics (3.653)	Mathematical logic (2.953)	Number theory (2.600)
	Mechanics (68.572)	Geometry (5.539)	Computer science (4.148)	Number theory (3.068)	Statistical mechanics (3.051)	Statistics (2.959)
	Statistics (56.285)	Geometry (3.789)	Probability theory (3.743)	Computer science (3.491)	Mechanics (2.754)	Operations research (2.695)
	Number theory (41.863)	Geometry (5.841)	Mechanics (3.058)	Computer science (2.605)	Algebraic geometry (1.982)	Mathematical logic (1.861)
Glossário de Tópicos	Analysis (122.018)	Geometry (9.688)	Algebra (8.370)	Statistics (6.178)	Calculus (4.758)	Topology (4.498)
	Algebra (98.710)	Geometry (12.439)	Analysis (8.272)	Topology (4.998)	Number theory (3.634)	Calculus (3.043)
	Geometry (95.250)	Algebra (12.556)	Analysis (9.644)	Topology (5.394)	Calculus (4.406)	Number theory (3.348)
	Topology (47.441)	Geometry (5.505)	Algebra (5.157)	Analysis (4.590)	Calculus (1.531)	Set theory (1.516)
	Statistics (40.757)	Analysis (6.079)	Geometry (1.666)	Algebra (1.560)	Probability theory (1.558)	Topology (958)

Fonte: Elaborado pelos autores, 2018.

6. CONCLUSÕES

Mapear os tópicos do conhecimento e organizá-los, por meio da hierarquia fornecida pela GA, resulta em uma estrutura rica em possi-



bilidades de exploração. Dentre essas possibilidades, podemos destacar a identificação do processo de transição entre os tópicos de pesquisa. A identificação desse processo é importante para o entendimento do padrão de desenvolvimento do conhecimento científico. Neste trabalho, apresentamos um método que permite construir um grafo de tópicos do conhecimento que pode servir de base para descrever o processo de desenvolvimento do conhecimento científico em grupos específicos ou de forma global. Adicionalmente, analisamos as influências observadas entre os tópicos associados a J. Bernoulli. Acreditamos que a pesquisa de modelos baseados em GA contribui de forma importante para a obtenção de novos conhecimentos sobre a formação da ciência.

Finalmente, é necessário indicar que, o novo método aqui proposto é sensível ao sistema de classificação utilizado, ou seja, a assertividade da inferência de tópicos aos acadêmicos depende do quão específico for a classificação. Além disso, destaca-se que o processo de propagação de tópicos é de fundamental importância para possibilitar a correta inferência aos vértices não rotulados. Nesse sentido, como desdobramento futuro, pretendemos validar os processos com respeito ao grau de assertividade de seus resultados, assim como investigar novas formas de análise do fluxo de conhecimento que considere como base a genealogia acadêmica e científica.

REFERÊNCIAS

- BOSCHMA, R.; HEIMERIKS, G.; BALLAND, P.-A. Scientific knowledge dynamics and relatedness in biotech cities. **Research Policy**, v. 43, n. 1, p. 107-114, 2014.
- BUFREM, L. S.; SILVA, F. M.; SOBRAL, N. V. Análise das influências intelectuais na produção científica da área de Ciência da Informação: um estudo sobre os bolsistas de produtividade em pesquisa (PQ-CNPq). **Em Questão**, v. 23, p. 115-141, 2017.
- DING, C. G. et al. Exploring paper characteristics that facilitate the knowledge flow from science to technology. **Journal of Informetrics**, v. 11, n. 1, p. 244-256, 2017.



INOMATA, D. O.; RADOS, G. J. V. **O fluxo da informação tecnológica no processo de desenvolvimento de produtos biotecnológicos: uma construção teórico conceitual**. ENANCIB - Encontro Nacional de Pesquisa em Ciência da Informação. Brasília - DF: [s.n.]. 2011.

MOHAMMADI, E.; THELWALL, M. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. **Journal of the Association for Information Science and Technology**, v. 65, n. 8, p. 1627-1638, 2014.

ROSSI, L.; MENA-CHALCO, J. P. **Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos**. BRASNAM – Brazilian Workshop on Social Network Analysis and Mining. Brasília - DF: [s.n.]. 2014.

SORENSEN, O.; RIVKIN, J. W.; FLEMING, L. Complexity, networks and knowledge flow. **Research policy**, v. 35, n. 7, p. 994-1017, 2006.

TOBARRA, L. et al. Analyzing the students' behavior and relevant topics in virtual learning communities. **Computers in Human Behavior**, v. 31, n. 1, p. 659-669, 2014.