Universidade Federal do ABC

Identificando pesquisadores relevantes em grafos de genealogia acadêmica: Uma abordagem baseada em PageRank

Relatório Final - Iniciação Científica - FAPESP

Aluno: Arthur Veloso Kamienski
Bacharelado em Ciência & Tecnologia
a.kamienski@aluno.ufabc.edu.br

Orientador: Jesús P. Mena-Chalco
Centro de Matemática, Computação e Cognição
jesus.mena@ufabc.edu.br

Santo André, 10 de Outubro de 2018

Resumo

Identificar pesquisadores relevantes nas suas áreas de atuação é uma tarefa árdua, apesar de ser de grande relevância. Sua análise envolve diferentes indicadores não só na ciência, mas também no cenário político-social. Atualmente, existem múltiplos métodos quantitativos e qualitativos para avaliar pesquisadores nos diferentes contextos em que estão inseridos. Nos métodos quantitativos, a maior dificuldade para a associação desses indicadores é a avaliação do impacto de cada pesquisador não só na totalidade da comunidade a que pertencem, mas também na estrutura local que o envolve. Neste projeto de pesquisa usamos o algoritmo PageRank em um grafo de genealogia para analisar a relevância de cada membro que o compõe. Adicionalmente, modificamos esse algoritmo para obter uma informação local (frente à informação global), de modo a encontrar com maior confiabilidade os pesquisadores mais importantes. A contribuição deste projeto permite, além de identificar pesquisadores importantes desde o ponto de vista de formação de recursos humanos, a criação de novos métodos e métricas capazes de auxiliar futuros estudos sobre o tema da genealogia acadêmica.

Palavras-chave: grafos, genealogia acadêmica, PageRank, pesquisadores relevantes.

Sumário

1 Introdução

O desenvolvimento científico-tecnológico, observado principalmente nas últimas décadas, é resultado, dentre muitos fatores, da atuação de pesquisadores na busca da evolução e perpetuação do conhecimento acadêmico em nossa sociedade. Assim, buscar formas de avaliação destes pesquisadores é uma iniciativa importante para identificar os indivíduos de maior impacto na comunidade científica (Patton, 2005). Na atualidade existem muitos desafios no sentido de identificar atributos de um pesquisador (professor) de modo que seja possível caracterizá-lo com informações que vão além de sua produção acadêmica, na forma de artigos científicos, ou de habilidades de ensino, na forma de apresentação de conteúdo acadêmico em salas de aulas.

Os métodos clássicos para se identificar a relevância de um pesquisador, como a análise de suas premiações, atuações em eventos científicos e avaliação por meio de seus pares são comumente qualitativos e de difícil implementação, considerando que os métodos são aplicados individualmente e carregam consigo o juízo de quem o aplica, sendo suscetível à fatores pessoais que podem comprometer o resultado (Isaac & Michael, 1971). Há ainda métodos quantitativos baseados em análise de publicações (Tol, 2013), porém não consideram o desempenho do acadêmico na formação de recursos humanos.

Podemos avaliar um pesquisador/professor com base em sua genealogia acadêmica, que é definida como um estudo quantitativo da herança intelectual esquematizada por meio de cadeias de pesquisadores interligados através de seus relacionamentos de orientação acadêmica (Cronin & Sugimoto, 2014). A genealogia acadêmica possibilita obter informações relevantes sobre a configuração das diferentes áreas do conhecimento científico, bem como a identificação de seus principais membros (Rossi, 2015). Uma possível análise pode ser dada, por exemplo, pela identificação da influência que um pesquisador tem sobre sua descendência acadêmica, direta ou indireta. Estas relações de parentesco acadêmico e seus impactos constituem o objeto de pesquisa considerado nesta Iniciação Científica.

Seguindo as considerações do Manifesto Leiden (Hicks *et al.* , 2015), uma alternativa para tornar o processo de avaliação mais eficiente, assertivo e que considere a propagação do conhecimento científico, envolve métricas computacionais de base quantitativa que reflitam as

informações sobre os relacionamentos de orientação acadêmica do pesquisador. A busca de se extrair conhecimento relevante a partir de estruturas de genealogia (grafos) passa, comumente, pelo desenvolvimento de métricas que caracterizem sua topologia. No trabalho de David & Hayden (2012) a comunidade dos neurocientistas foi caracterizada por meio de medidas de distância, fecundidade e agrupamento. Gargiulo *et al.* (2016) consideraram medidas clássicas de distância, similaridade e agrupamento, enquanto Rossi *et al.* (2017) adaptaram o índice-h bibliométrico para o contexto genealógico. Finalmente, Rossi & Mena-Chalco (2014) apresentaram um conjunto de medidas topológicas, os quatro trabalhos consideraram a comunidade dos doutores em Matemática.

Existem diferentes iniciativas no sentido de registrar informações sobre a genealogia de comunidades acadêmicas (e.g., *Mathematics Genealogy Project*[1], *The Academic Family Tree*[2]) o que resulta em um grande volume de dados genealógicos que são disponibilizados. A utilização destes recursos envolve desafios como: (i) mineração de grande um volume de dados, (ii) desenvolvimento de métodos computacionais para a estruturação dos dados (grafo de genealogia) e para (iii) a subsequente extração da métrica desejada (Hey *et al.* , 2009). Estes métodos computacionais viabilizam a identificação de pesquisadores relevantes.

Segundo Sugimoto (2014), uma grande variedade de questões podem ser respondidas construindo e analisando uma genealogia acadêmica. Utilizando a ascendência e a descendência de um pesquisador, pode-se obter informações sobre o processo de formação da comunidade acadêmica de interesse, bem como sobre a dinâmica do fluxo de conhecimento científico e seu impacto na formação de novos pesquisadores (Malmgren *et al.* , 2010).

Este projeto de iniciação científica é pautado na utilização da genealogia acadêmica como abordagem para a extração de conhecimento em comunidades acadêmicas e para a proposição de uma nova métrica baseada na estrutura topológica do grafo, a qual pode ser considerada para auxiliar na avaliação dos pesquisadores.

[1] https://genealogy.math.ndsu.nodak.edu/ último acesso em 29 de Junho de 2017.
[2] https://academictree.org/ último acesso em 29 de Junho de 2017.

1.1 Objetivo geral

O objetivo geral deste projeto de Iniciação Científica é (i) a adaptação do algoritmo PageRank para a aplicação em grafos de genealogia acadêmica e (ii) o desenvolvimento de um método computacional que considere a hereditariedade local no cálculo da métrica, criando diferentes níveis de influência para os diferentes graus de parentesco acadêmico.

1.2 Objetivos específicos

- Criar um método de classificação de pesquisadores baseado em PageRank com influência local e global;
- Quantificar a relevância de pesquisadores em diferentes cenários de pesquisa (e.g., tipo de orientação, áreas do conhecimento, senioriedade acadêmica);
- Identificar grupos de pesquisadores relevantes em diferentes áreas do conhecimento utilizando o método obtido.

Além dos objetivos específicos de pesquisa, este projeto também tem os objetivos característicos de um projeto de Iniciação Científica padrão:

- Obter resultados inéditos e relevantes para os problemas estudados;
- Incentivar o gosto do aluno pela pesquisa e a criatividade para solução de problemas;
- Envolver o aluno em problemas de pesquisa na área de Ciência da Computação;
- Desenvolver a maturidade algorítmica do aluno; e
- Ampliar os conhecimentos e habilidades do aluno, especialmente aqueles relacionados aos assuntos de Ciência da Computação (desenvolvimento de algoritmos) e também aqueles relacionados ao desenvolvimento de pesquisa em geral, tais como escrita de artigos e relatórios, preparação e apresentação de seminários.

1.3 Metas

Como meta, esperamos que ao final das etapas de pesquisa previstas no projeto, possamos construir um método computacional que permita identificar, de forma automática, os pesquisadores mais influentes na formação de recursos humanos ao longo de suas gerações. Note que a proposta de Iniciação Científica, embora inicialmente destinada para uso em grafos de genealogia, pode ser aplicada para qualquer tipo de grafo direcionado.

2 Sobre grafos de genealogia acadêmica

A genealogia acadêmica pode ser definida como o estudo das relações entre professores e alunos (atuando por meio de orientações acadêmicas como orientadores e orientados) e do fluxo de conhecimento científico na forma de herança intelectual (Sugimoto, 2014). Por meio de uma abordagem quantitativa, utilizam-se diferentes métodos para a realização de estudos sobre a descendência ou ascendência de um indivíduo, como por exemplo: identificação de figuras históricas, extensão da transmissão de conhecimento e avaliação dos membros de comunidades acadêmicas.

Ao se falar em genealogia, é natural lembrar-se de uma árvore, onde a partir de um indivíduo surgem ramificações que descrevem suas relações de parentesco com aqueles de sua família. Essa representação pode ser modelada por um grafo onde cada indivíduo assume o papel de um vértice e suas relações correspondem às arestas que os associam. Este tipo de estruturação facilita a compreensão das intrincadas relações que se acumulam e se multiplicam com o passar do tempo (Gargiulo *et al.* , 2016). Para a genealogia acadêmica pode-se usar de uma ideia semelhante, apresentando professores e alunos como vértices interligados por meio de suas relações de orientação acadêmica, que são representados por arestas direcionadas.

A Figura 1 apresenta um exemplo de grafo de genealogia contendo todos os descendentes do pesquisador Jacob Palis. As informações utilizadas para compor o grafo foram obtidas a partir da plataforma Lattes[3], utilizando o método desenvolvido por Damaceno *et al.* (2017).

Assim, podemos definir formalmente um grafo de genealogia acadêmica $G = (V, E)$ como

[3] http://lattes.cnpq.br, último acesso em 18 de maio de 2017.

um conjunto de vértices (V) e arestas (E) em que cada vértice é a representação de um indivíduo da comunidade acadêmica, e cada aresta apresenta uma orientação entre dois indivíduos. Além disso, como a orientação é unidirecional (no sentido de que o professor orienta o aluno e não o oposto), é conveniente utilizar-se de arestas direcionadas, denotando não só o sentido da orientação mas também o fluxo de conhecimento. Apesar de assemelhar-se a uma árvore, o grafo de genealogia nem sempre é acíclico e pode ser desconexo, não caracterizando, portanto, uma árvore do ponto de vista formal/computacional.

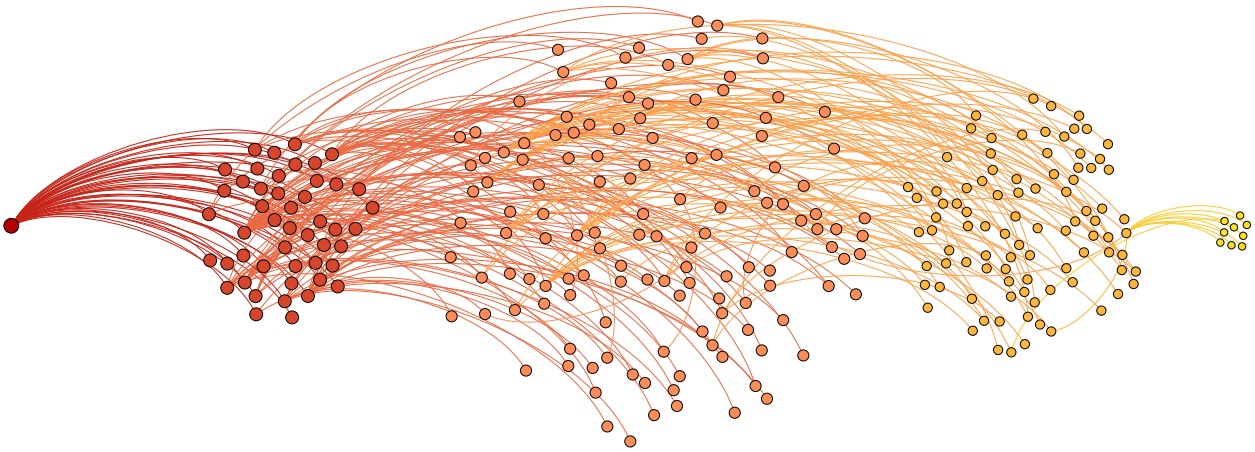


Figura 1: *Grafo de genealogia acadêmica apresentando todos os descendentes do pesquisador Jacob Palis, representado pelo vértice vermelho mais à esquerda. Cada geração a partir de Palis está representada por uma cor mais clara e tamanho menor dos vértices. A cor de cada aresta é a mesma do vértice de origem, representando sua unidirecionalidade. Note que existem diferentes gerações de pesquisadores influenciados pelo Prof. Palis.*

O crescente interesse da comunidade acadêmica em registrar e analisar seus dados genealógicos resulta em terreno fértil para a pesquisa nesta área. Existem diferentes maneiras de se obter um grafo de genealogia acadêmica, dado que existem diferentes fontes de dados genealógicos que necessitam métodos distintos de obtenção de informações (Andraos, 2005). Considerar os bancos de dados genealógicos, construídos por iniciativa das próprias comunidades acadêmicas, é a forma mais trivial. A estruturação deste tipo de banco de dados é orientada para a genealogia, i.e., as relações entre os acadêmicos são explicitamente declaradas.

Um exemplo importante, neste contexto, é o *Mathematics Genealogy Project*, onde doutores em matemática podem se cadastrar e juntar-se ao grafo de mais de 200 000 pesquisadores (Gargiulo *et al.* , 2016). Outra iniciativa, na mesma linha de atuação, refere-se à comuni-

dade dos astrônomos (Tenn, 2016) *Astronomy Genealogy Project* com objetivos similares ao anterior.

Por outro lado, é possível extrair um grafo a partir de conjuntos de dados que não são vocacionados para a genealogia. Neste caso, as relações devem ser obtidas por meio de processos mais sofisticados, i.e., as relações neste contexto não são explicitamente declaradas. São exemplos deste tipo de fonte de dados: publicações científicas em co-autoria e currículos acadêmicos (Damaceno *et al.* , 2017).

Com a aplicação das métricas clássicas descritas pela Teoria dos Grafos, pode-se aprofundar a análise das relações de genealogia acadêmica. Entretanto, devido à especificidade destas estruturas e do contexto hierárquico que elas representam, é necessário o desenvolvimento de métricas específicas para estudos genealógicos.

Os resultados obtidos a partir da aplicação de métricas genealógicas podem ser considerados como base para outros estudos com objetivos mais específicos, como a identificação de pesquisadores relevantes, descrita no trabalho realizado por Elias *et al.* (2016), em que foram identificados os pesquisadores pioneiros na área de Protozoologia do Brasil.

3 Sobre o algoritmo PageRank

No final dos anos 1990, Sergey Brin e Larry Page, dois estudantes de doutorado na Universidade de Stanford, na Califórnia, desenvolveram um algoritmo denominado PageRank com o objetivo de classificar páginas na web segundo a quantidade e a qualidade de seus *links.*

O algoritmo, que na época em que foi descrito no artigo *"The anatomy of large-scale hypertextual Web search engine"* (Brin & Page, 2012), autorado por seus dois criadores, causou grande impacto na comunidade científica, trouxe muitas melhorias para os antigos sistemas de pesquisa na web baseados em buscas textuais. Ao classificar as páginas, tornou-se possível uma busca mais eficiente, evidenciando aquelas de maior valor em detrimento daquelas que não são de grande interesse para o que está sendo procurado.

A atribuição da nota (quantitativa) para uma página no PageRank se dá com base em

suas conexões com outras páginas. Essas conexões são feitas por meio de *hyperlinks* que direcionam o usuário de uma para outra. A conexão é unidirecional, ou seja, liga a página de origem à página destino, mas não o contrário. Utilizando esse conceito, podemos quantificar a popularidade de cada página e criar um critério para definir aquelas que são mais conectadas e relevantes.

Considerando, como exemplo, uma rede social, onde a popularidade de uma pessoa é proporcional à quantidade de amigos (ou conexões) que ela possui, podemos dizer que uma página tem grande impacto na *web* se essa for muito referenciada por outras. Assim, podemos classificar uma página com base na quantidade de páginas que tem *hyperlinks* direcionando os usuários para ela.

Além disso, outro fator que influencia a relevância de uma página *web* é a qualidade daquele que a referencia. Uma indicação vinda de uma página que indica milhares de outras é de menor significância do que uma indicação de uma página que tem pouquíssimas outras referências, dado que assume-se que quanto mais indicações uma página faz, menor a relevância de cada uma. Portanto, o peso da conexão é inversamente proporcional ao número total de conexões que a página possui.

Utilizando essa lógica, calcula-se uma nota para todas as páginas existentes na web e, com uma classificação definida, pode-se dar prioridade àquelas que tem notas mais altas.

Apesar de simples, este algoritmo computacional revolucionou as ferramentas de pesquisa (Langville & Meyer, 2011). Seus dois criadores, após abandonarem seus cursos de doutorado, utilizaram o algoritmo como base para seu próprio site de buscas, ao qual deram o nome de Google. O sucesso dessa empresa, que está entre as mais reconhecidas e bem-sucedidas da área de tecnologia, está diretamente relacionado com a qualidade de seu algoritmo de buscas e classificação, o que exemplifica a superioridade em relação a outros. Até hoje, segundo Gleich (2015), o PageRank é utilizado em inúmeras aplicações nas mais diversas áreas, apesar de ter sido criado com o objetivo de facilitar buscas na *web*.

3.1 Descrição matemática simplificada

Considerando o conceito descrito anteriormente, o PageRank de um vértice em um grafo qualquer é a soma da razão entre os PageRanks dos vértices que apontam para esse e os seus graus de saída. Assim, o PageRank de um vértice a, $PR(a)$, que é apontado pelo conjunto de vértices V de tamanho n é definido como:

$$PR(a) = \sum_{v \in V} \frac{PR(v)}{S(v)}$$

em que $S(v)$ é o grau de saída do vértice v.

A partir dessa equação pode-se, então, aplicar um método iterativo que começa com valores iniciais de PageRank iguais a $\frac{1}{n+1}$ para todos os vértices do grafo, para que todos iniciem com um PageRank de valor igual e para que a soma de todos os PageRanks do grafo seja igual a 1. Os valores do PageRank para cada vértice convergem após um determinado número de iterações independentemente do valor inicial, o qual só altera a velocidade da convergência (Brin & Page, 2012).

3.2 Fator de amortecimento

Apesar do algoritmo apresentado ser simples e intuitivo, existem problemas quando o utilizamos da forma apresentada. Como o PageRank de um vértice depende de todos os outros vértices do grafo, o algoritmo pode não convergir, uma vez que se existirem ciclos o PageRank será continuamente transmitido entre os vértices, nunca estabilizando em um valor.

Além disso, em sua forma simplificada não existe nenhum mecanismo que impeça que os vértices que não tem nenhuma ligação de saída (*sink nodes*) acumulem PageRank, nem que vértices que não tem ligação de entrada percam PageRank, já que após cada iteração o PageRank dos vértices iniciais é transmitido mas não é reposto e o PageRank dos vértices finais é acrescido mas não é transmitido.

Para solucionar tais problemas, um fator de amortecimento (*damping factor*) d é adicionado à equação.

$$PR(a) = \frac{1-d}{n} + d\sum_{v \in V} \frac{PR(v)}{S(v)}$$

Este fator, cuja escala varia entre zero e um, foi originalmente concebido para simular a forma aleatória como um usuário navega pela *web*: seguindo os *links* entre as páginas com chance d e, eventualmente (com probabilidade $1 - d$), saltando para uma página qualquer ao acessá diretamente pelos seus endereços. O PageRank pode ser pensado, segundo esta perspectiva, como uma distribuição de probabilidades de um passeio aleatório em um grafo.

Também considera-se que os vértices sem arestas de saída estão ligados a todos os outros vértices do grafo uma vez que, ao atingir uma página sem nenhuma ligação com outra, um usuário não tem outra opção a não ser acessar uma página qualquer da *web* requisitando-a através de seu endereço em vez de utilizando uma ligação.

Desta forma, garante-se que cada vértice sempre terá um valor mínimo de PageRank, $\frac{1-d}{N}$, obtido pela chance de ser "acessado" aleatoriamente, e que o valor acumulado por ciclos e por *sink nodes* seja redistribuído para todo o grafo.

Note que, para a versão simplificada do PageRank, o fator de amortecimento é igual a 1, por outro lado, conforme descrito no artigo original de Sergey Brin e Larry Page, o valor padrão para o fator de amortecimento é de $0,85$, o qual foi considerado neste projeto de pesquisa (Page *et al.* , 1999).

4 Alterações no algoritmo PageRank

Conforme definido anteriormente, um grafo de genealogia acadêmica é uma estrutura onde as arestas direcionadas interligam o orientador ao orientado. Consequentemente, a estrutura do grafo se assemelha a estrutura de uma árvore (apesar de não ter todos os requisitos para ser classificado como uma) na qual existem poucos vértices que agem como raízes. Portanto, o PageRank obtido por meio do algoritmo original, aplicado a estes grafos, resultará em valores de PageRank altos para orientados e baixos para orientadores, uma vez que o PageRank será repassado entre os vértices no sentido da aresta e do fluxo de conhecimento que ela representa. Ainda que este resultado seja útil para identificar a influência dos orientadores

em seus orientados, exaltando deste modo aqueles que mais se beneficiam de suas orientações, ele é o oposto do que procuramos.

Para alcançar os objetivos e resultados desejados, o algoritmo do PageRank precisou ser alterado de modo a oferecer uma métrica mais adequada. Apesar dos resultados obtidos pelo algoritmo original já demonstrarem sua capacidade de atribuir um determinado ranking de acordo com a relevância do pesquisador, ainda buscamos uma forma de torná-lo mais justo. Assim, podemos obter resultados que melhor refletem a realidade, facilitando a identificação dos pesquisadores relevantes.

A seguir estão descritas as adaptações propostas para a obtenção de valores e, posteriormente, uma ordenação de pesquisadores mais satisfatórios. Para ilustrar o raciocínio seguido no desenvolvimento das modificações e exemplificar com maior clareza o resultado de tais alterações, o grafo da Figura 2 será utilizado. O resultado da aplicação de cada tipo de PageRank aqui discutido ao grafo está exposto na Tabela 1.

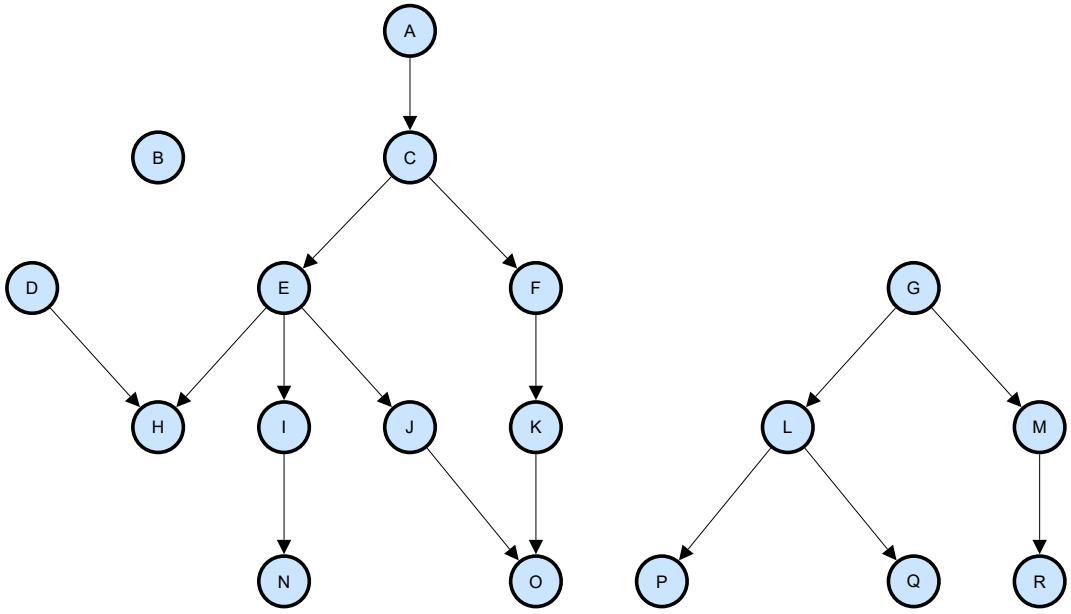


Figura 2: *Exemplo de um grafo de genealogia com 18 vértices e 16 arestas.*

4.1 PageRank Invertido

Para obter um resultado onde os orientadores (em oposição aos alunos, no algoritmo original do PageRank) sejam exaltados, o sentido de fluxo de PageRank deve ser invertido, sendo direcionado dos orientados para os orientadores. Para isso, pode-se inverter o sentido

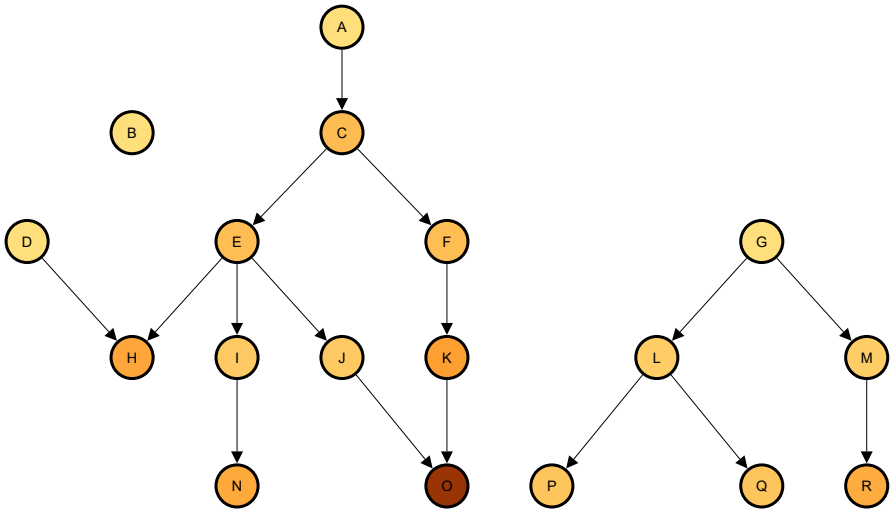
Tabela 1: *Tabela com os valores de diversos tipos de PageRank para o grafo da Figura 2 com valores truncados à quatro casas decimais, onde PR = PageRank, PRI = PageRank Invertido, PRIL = PageRank Invertido Local, PRILN = PageRank Invertido Local Normalizado e V(S) = número de vértices contidos no sub-grafo extraído.*

Vértice	PR	PRI	PRIL	1-PRIL	PRILN	V(S_2)
A	0,0309	0,1548	0,4121	0,5878	0,5878	4
B	0,0309	0,0238	1,0000	0,0000	0,0000	1
C	0,0573	0,1540	0,3729	0,6270	0,6270	7
D	0,0309	0,0340	0,6491	0,3508	0,3508	2
E	0,0553	0,1004	0,4271	0,5728	0,5728	6
F	0,0553	0,0527	0,4744	0,5255	0,3514	3
G	0,0309	0,1161	0,3919	0,6080	0,5257	6
H	0,0729	0,0238	1,0000	0,0000	0,0000	1
I	0,0466	0,0441	0,6491	0,3508	0,3508	2
J	0,0466	0,0340	0,6491	0,3508	0,3508	2
K	0,0780	0,0340	0,6491	0,3508	0,3508	2
L	0,0441	0,0644	0,5744	0,4255	0,4255	3
M	0,0441	0,0441	0,6491	0,3508	0,3508	2
N	0,0706	0,0238	1,0000	0,0000	0,0000	1
O	0,1369	0,0238	1,0000	0,0000	0,0000	1
P	0,0497	0,0238	1,0000	0,0000	0,0000	1
Q	0,0497	0,0238	1,0000	0,0000	0,0000	1
R	0,0684	0,0238	1,0000	0,0000	0,0000	1

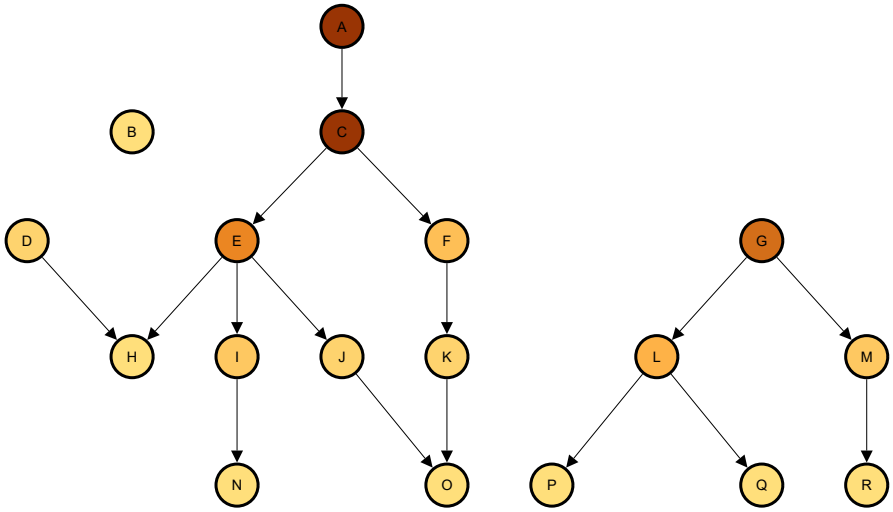
de todas as arestas do grafo, de modo que seus sentidos sejam contrários ao sentido original e, assim, inverter o fluxo de PageRank. Também é possível inverter o sentido do PageRank durante seu cálculo (sem alterar o grafo), ao fazer com que os vértices recebam o PageRank de seus filhos ao invés do PageRank de seus pais.

A partir dessa adaptação no cálculo do PageRank, a qual chamamos de PageRank Invertido (PRI), foi possível encontrar uma ordenação de pesquisadores na qual indivíduos são recompensados por realizarem orientações (ação de caráter ativo), ao invés de serem recompensados por receberem orientações (ação de caráter passivo).

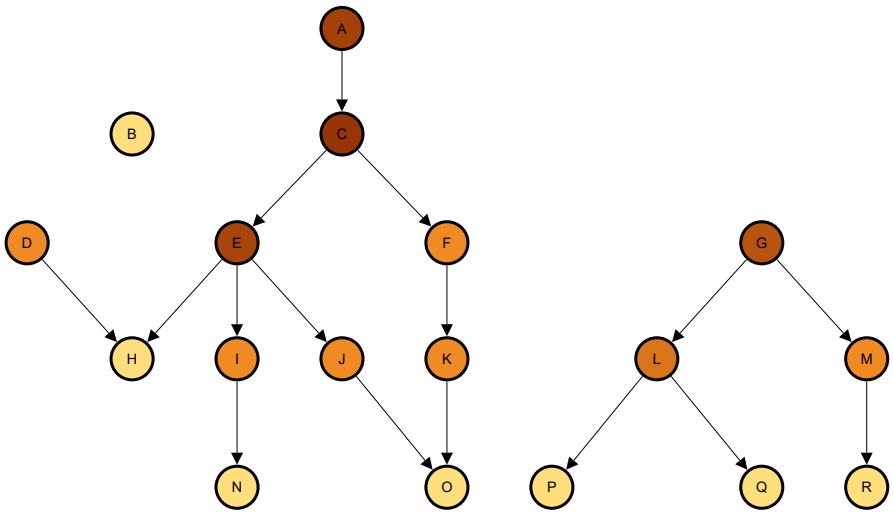
O efeito da inversão do PageRank pode ser observado pelo resultado obtido após o cálculo do PageRank e do PageRank Invertido para o grafo da Figura 2, mostrados na segunda e terceira coluna da Tabela 1. A figura 3 evidencia esse fato ao mostrar valores de PageRank mais altos de acordo com a intensidade das cores. Para o PageRank original (a), os vértices que mais se destacaram foram aqueles encontrados na base da árvore, mais especificamente os vértices O, K, H e N. Já os resultados do PageRank Invertido (b) exaltam vértices no topo da árvore, como os vértices A, C, G e E (em ordem de PageRank).



(a) PageRank



(b) PageRank Invertido



(c) PageRank Invertido Local Normalizado

Figura 3: *Mapa de calor mostrando as diferenças entre as medidas do PageRank. Cores mais escuras indicam um valor mais alto de PageRank.*

4.2 PageRank Invertido Local

Ainda que o PageRank Invertido tenha colaborado para o alcance de uma métrica mais satisfatória, ainda existe um problema com relação à justiça: a aplicação do PageRank no grafo de genealogia completo, ou seja, de forma global, pode produzir distorções se compararmos um pesquisador mais antigo com um mais jovem, visto que os descendentes do primeiro são em maior número. Assim, um pesquisador mais antigo possui maior probabilidade de se destacar com o ranking gerado a partir dessa análise (em número de filhos acadêmicos assim como em quantidade de orientações).

O maior número de descendentes de um pesquisador pode, de fato, ser um indício de sua relevância quando consideramos um número pequeno de gerações descendentes de um pesquisador. Contudo, existem pesquisadores para os quais o número de descendentes é grande apenas para gerações distantes (trinetos, tetranetos, etc.) dado pelo surgimento constante de novas orientações. Consideramos, deste modo, que o pesquisador não teve grande influência na formação destas gerações e, portanto, não seria justo se fosse recompensado por isso.

Existem ainda pesquisadores que não pertencem a grandes componentes do grafo, como é o caso dos vértices G, L, M, P, Q e R para o grafo da Figura 2. Esses pesquisadores, ainda que tenham uma grande relevância em sua própria comunidade, acabam sendo prejudicados por não terem elos com outros pesquisadores, enquanto pesquisadores de pequena relevância se beneficiam por terem apenas uma ligação com grandes componentes (como é o caso do vértice A).

Neste contexto, consideramos limitar a abrangência do PageRank de modo que cada pesquisador seja analisado de forma local. Para tanto, restringimos o tamanho do grafo que será utilizado para o cálculo do PageRank do pesquisador, ao gerar um sub-grafo contendo apenas o pesquisador em questão e um número k de suas gerações descendentes.

Os sub-grafos gerados consideram um número k arbitrário de gerações de descendentes do vértice ao qual diz respeito, sendo esse a raiz do sub-grafo. Esta escolha foi feita com base na ideia de que queremos analisar o vértice em sua maior relevância possível e, portanto, deve-se escolher uma comunidade que é centrada nele. Como todos os vértices do sub-grafo são descendentes desse vértice, espera-se que ele acumule um maior PageRank que os outros.

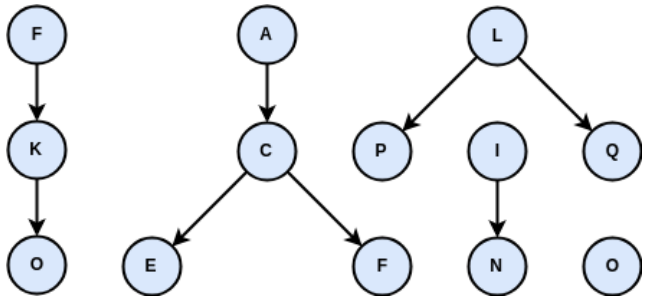


Figura 4: *Exemplos de sub-grafos retirados do grafo da Figura 2, para os vértices A, F, L, I e O.*

O PageRank Invertido pode, então, ser aplicado para cada sub-grafo e um valor de PageRank Invertido Local (PRIL) será obtido. O valor do PRIL de um pesquisador será o PageRank Invertido do pesquisador dentro do sub-grafo extraído a partir dele. O resultado do cálculo do PRIL para sub-grafos extraídos do grafo da Figura 2 e com um número de gerações (k) igual a 2 é mostrado na quarta coluna da Tabela 1.

Essa nova métrica, além levar em consideração o que foi discutido na seção 4.1, permite que analisemos um pesquisador para um grupo de orientações arbitrariamente próximo, desconsiderando a influência de vértices que estão fora do sub-grafo. Dessa maneira, tornamos a métrica mais justa e mais condizente com a realidade, uma vez que todos os pesquisadores devem se destacar dentro de suas próprias comunidades, sem se beneficiar da relevância de outros pesquisadores com baixo grau de parentesco.

Normalização de PageRank Invertido Local

Apesar de ter sido concebido para ser uma métrica com características desejadas, o PageRank Invertido Local trás um novo problema. Como leva-se em conta apenas o número de gerações durante a criação de cada sub-grafo (e não o número de vértices), existem sub-grafos de tamanhos muito distintos, o que impede que a métrica obtida para pesquisadores de sub-grafos diferentes seja comparada.

Dado que o valor da soma de todos os PageRanks de um grafo é sempre igual a 1, quanto maior o número de vértices menor a quantidade de PageRank disponível para cada vértice acumular. Por exemplo, em um grafo com 7 vértices, a quantidade de PageRank por vértice é igual a $1/7 = 0,14$, enquanto para um grafo com quatro vértices esse valor sobe para $1/4 = 0,25$, e para um grafo com dois vértices sobe para $1/2 = 0,5$.

Portanto, tomando dois pesquisadores com sub-grafos distintos, caso o tamanho destes sub-grafos seja muito diferente, o pesquisador com maior sub-grafo terá um valor de PageRank menor que o do outro pesquisador. Esse fato pode ser observado nos resultados obtidos pelo cálculo do PRIL para o grafo da Figura 2. Perceba que enquanto C têm um valor de PRIL igual a 0.3729, vértices que têm sub-grafos de tamanho 2 têm PageRank igual a 0.6491 e sub-grafos de tamanho 1 (vértices folha) têm PageRank igual a 1.0000. Isso vai contra o desejado, uma vez que o pesquisador com maior sub-grafo deve ser mais influente (dado que tem maior quantidade de orientações para um mesmo número de gerações).

Em uma abordagem inicial e ingênua, tentou-se inverter todos os valores de PageRank obtidos. Seguindo esse método, pesquisadores com maior PRIL (que neste caso são aqueles com menores sub-grafos) obteriam os menores PageRanks, enquanto aqueles com menor PRIL obteriam valores maiores. Dado que o domínio do PageRank para todos os vértices é $]0, 1]$, pôde-se redefinir o valor do PageRank Invertido Local ($PRIL$) para $PRIL = 1 - PRIL$. Este método resolve, de fato, o problema discutido anteriormente, fazendo com que pesquisadores com menores PageRanks (e maiores sub-grafos) obtivessem maiores PageRanks ao final do cálculo.

Observou-se, contudo, que apenas fazer uma inversão direta dos PageRanks provocou uma distorção dos resultados. Apesar do PageRank de grafos menores ser maior que o PageRank de grafos maiores, não indicando uma relevância maior, um PageRank maior para sub-grafos de tamanho igual continua indicando maior relevância. Por exemplo, tomando os vértices E e G do grafo da Figura 2, ambos têm um sub-grafo de tamanho 6 mas, por outro lado, o PageRank de E é maior que o PageRank de G dado à morfologia do sub-grafo. Inverter o PRIL desses vértices implica na inversão da situação e do resultado desejado, tornando G mais relevante que E, como mostrado na Tabela 1.

Por conta disso, fez-se necessário o uso de um novo método de normalização de valores. Dado que o problema se origina devido ao tamanho distinto dos sub-grafos, é natural que uma normalização baseada nesse valor seja aplicada.

A normalização encontrada para solucionar este problema se baseou na ideia de que uma vez que vértices com o mesmo tamanho de sub-grafo seguem o ranking desejado, mas os de tamanho distinto não. Por isso, devemos fazer a inversão apenas para o segundo caso,

mantendo o ranking para os vértices do primeiro grupo. Assim, elegemos um dos vértices de cada tamanho de sub-grafo como um representante para a realizar a inversão, enquanto o ranking dos demais vértices é calculado utilizando o valor obtido pelo representante após a inversão, de maneira a manter a proporção de PRIL anterior à inversão. A fórmula resultante para o cálculo do PageRank Invertido Local Normalizado (PRILN) é

$$PRILN(v) = \frac{PRIL(v)}{PRIL(w)} \times (1 - PRIL(w)) = \frac{PRIL(v)}{PRIL(w)} - PRIL(v)$$

em que v e w tem o mesmo tamanho de sub-grafo e w é o representante deste conjunto de vértices. Decidiu-se por eleger como representante o vértice com maior valor de PRIL para cada tamanho de sub-grafo.

A métrica obtida pelo cálculo utilizando a formula mantém o ranking entre vértices com mesmo tamanho de sub-grafo antes da inversão. Por exemplo, para F e L, temos que a proporção de PRIL é $\frac{0{,}4744}{0{,}5744} = 0,8259$, enquanto para PRILN $\frac{0{,}3514}{0{,}4255} = 0,8259$. Além disso, observe que para vértices que não tem nenhuma orientação passam a ter PageRank $0,0000$, indicando sua relevância nula dado que não têm orientações e, portanto, não influenciaram a formação de nenhum pesquisador.

O PRILN mostrou sua capacidade de identificação de pesquisadores de maneira justa. O resultado obtido para o grafo (Figura 3 (c)) mostra que o pesquisador mais relevante do grafo é aquele representado pelo vértice C, seguido por A, E e G. De fato, percebemos que esses são os vértices mais conectados e mais ao topo de suas árvores, indicando sua alta relevância. Ademais, pode-se ver que a métrica é mais justa que o PRI, onde A é o vértice mais relevante apenas por estar no topo do grupo, apesar de ter apenas uma orientação. Finalmente, o PRILN se mostra melhor distribuído pelo grafo, dado que vértices mais próximos à base recebem uma parcela maior de relevância, essa não estando centralizada em vértices mais altos.

Esta abordagem pode aparentar ser altamente correlacionada ao tamanho do sub-grafo de cada vértice. Entretanto, após testes em grafos de genealogia grandes, com número de vértices na casa de dezenas e centenas de milhares, a correlação entre ambos mostrou-se baixa (ao redor de 0,6). Isso ocorre pois, diferentemente do exemplo aqui apresentado, onde muitos sub-

grafos têm tamanho e estrutura semelhantes, em situações reais existe uma grande variedade de formas que um grafo pode assumir, especialmente para valores de k (número de gerações) maiores. Como o PageRank leva em consideração a maneira em que os vértices estão ligados em um grafo, a morfologia distinta de cada um é levado em consideração. Contudo, ainda existem sub-grafos triviais (como aqueles com 1, 2 ou 3 vértices) que aglomeram um grande número de ocorrências.

Por fim, vale notar que o grafo utilizado como exemplo nesta seção é muito pequeno quando comparado com grafos de genealogia reais. Assim, os problemas aqui discutidos são muito mais evidentes, apresentando diferenças entre os valores obtidos pelas métricas que têm um impacto realmente grande no ranking final de pesquisadores.

5 Procedimento metodológico

Este projeto de Iniciação Científica foi dividido em sete etapas ou módulos. Na Figura 5 apresentamos um fluxograma que ilustra as etapas e o respectivo sequenciamento. Cada bloco representa um processo e as setas representam o fluxo de informação entre os módulos.

Como elementos de entrada consideramos um grafo de genealogia acadêmica, o número de gerações a considerarmos no cálculo do PageRank local, e um fator de amortecimento. A saída é um lista de pesquisadores com seus respectivos valores de PageRank (global e local).

5.1 Leitura e representação do grafo

A identificação de pesquisadores mais relevantes com o uso do algoritmo PageRank se inicia com a leitura do grafo. Para nosso projeto consideraremos grafos de genealogia de pesquisadores que o Grupo de Pesquisa em Cientometria da UFABC[4] construiu no contexto de pesquisas de mestrado e doutorado em Ciência da Computação. Em geral cada grafo contem atributos relacionados com: (i) o nome do pesquisador, (ii) a instituição e (iii) a área nas quais o acadêmico atua e (iv) a lista de seus orientados. Todas essas informações são armazenadas pelo programa em objetos específicos que representam cada indivíduo.

[4]Os projetos do grupo de cientometria da UFABC estão disponíveis em http://pesquisa.ufabc.edu.br/cientometria, último acesso em 10 de março de 2018.

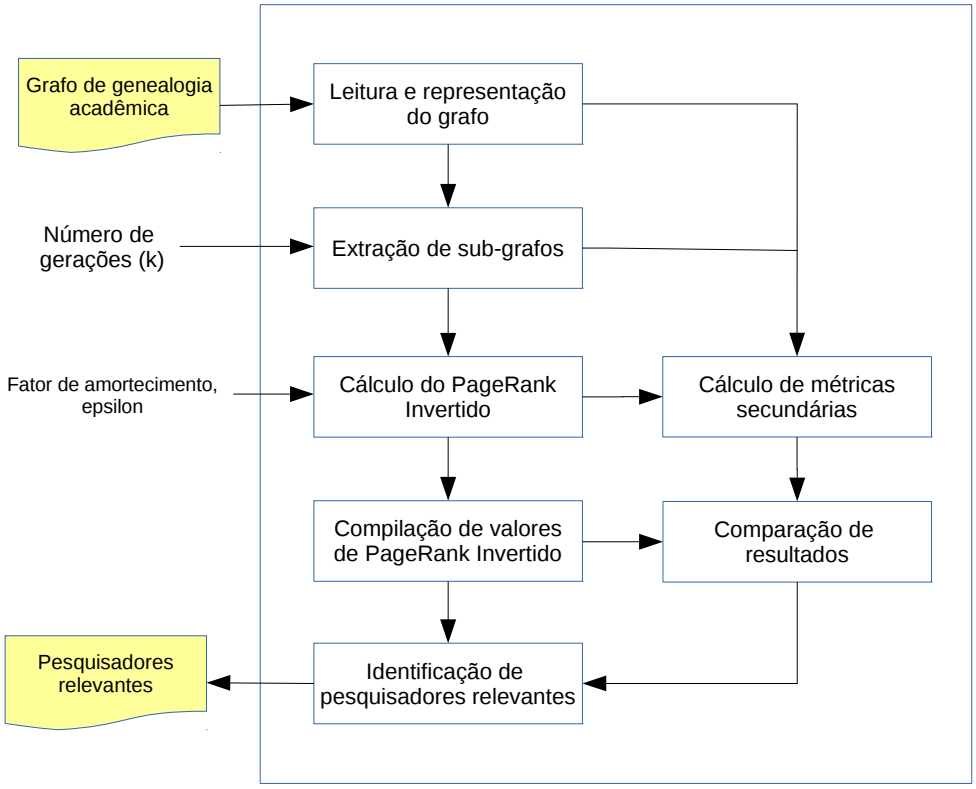


Figura 5: *Diagrama de Fluxo da proposta de Iniciação Científica para a identificação de pesquisadores relevantes utilizando o PageRank.*

O grafo é criado e representado utilizando uma estrutura de lista encadeada, ligando cada objeto com as informações de um pesquisador ao objeto contendo as informações de cada um de seus alunos.

5.2 Extração de sub-grafos

Para o cálculo do PageRank Local descrito na seção 4.2, um sub-grafo é extraído para cada vértice do grafo original, esse composto apenas pelos descendentes do vértice de interesse, cuja distância entre eles não seja superior a um determinado parâmetro (por exemplo, um valor inteiro k que represente o número máximo de gerações posteriores a cada vértice).

5.3 Cálculo do PageRank invertido

Com os sub-grafos definidos, aplica-se o algoritmo PageRank Invertido a cada um deles e para o grafo completo. Nesta etapa utilizam-se o fator de amortecimento para o cálculo do PageRank e um critério de parada para as iterações (e.g., um *epsilon* ϵ).

O critério de parada utilizado para o cálculo de PageRank pode ser tanto um número determinado de iterações desejadas, quanto um valor a ser comparado com alguma métrica de similaridade entre o PageRank da iteração atual e da iteração anterior. Para este projeto calculou-se a distância euclidiana entre os vetores de PageRanks do grafo para cada iteração. Assim, quando a distância entre os vetores for menor que ϵ o algoritmo é finalizado. O valor de ϵ foi definido como 10^{-10}.

Após o cálculo do PageRank para todos os sub-grafos, os valores obtidos são normalizados como indicado na seção 4.2

5.4 Cálculo de métricas secundárias

Utilizando o grafo original e os sub-grafos, métricas secundárias são calculadas como o tamanho do grafo, número total de arestas, tamanho de cada sub-grafo, número total de gerações do grafo, número de componentes conexas, PageRank invertido global, número de alunos e orientadores (grai de entrada e de saída) de cada vértice, entre outros.

Essas métricas são utilizadas para analisar o grafo, colaborando para a verificação de sua viabilidade para o uso nesta pesquisa e de que não existem problemas em sua estrutura, além de servirem para a validação dos resultados, servindo como bases de comparação.

5.5 Compilação de valores de PageRank invertido

Os valores obtidos pelo cálculo do PageRank invertido são compilados em uma lista a qual também contem as outras informações dos indivíduos. Essa lista pode ser exportada para uma estrutura que permita a sua fácil exploração em planilhas ou documentos tabulados.

5.6 Comparação dos resultados

Com o PageRank invertido e as métricas secundárias calculadas, os resultados são analisados. As métricas obtidas (PageRank invertido global, local para vários valores de k, métricas secundárias) são comparadas para evidenciar suas diferenças e similaridades e colaborar para a escolha de uma métrica mais satisfatória e viável de identificação de pesquisadores.

5.7 Identificação de pesquisadores relevantes

Com base nos resultados das últimas etapas, identificam-se os pesquisadores com maior valor de PageRank invertido (global e local). Para isso, utiliza-se tanto o valor bruto (sem nenhum tratamento) das métricas calculadas, tanto como o ranking calculado com o uso destes valores.

O ranking pode ser calculado ao ordenar-se os pesquisadores por valor de PageRank em ordem decrescente, e definindo sua posição no rank como sendo a sua posição na sua lista. Contudo, essa abordagem pode ser injusta pois pesquisadores com PageRank igual terão valores distintos. Desta forma, definiu-se que pesquisadores com o mesmo valor de PageRank terão o mesmo rank.

Com base nos valores e rankings obtidos, os pesquisadores com maior desempenho nas métricas ou com maior colocação nos rankings são escolhidos. Estes correspondem aos pesquisadores que seriam mais relevantes de acordo com a analise de cada métrica.

As listas de pesquisadores obtidas são utilizadas então para a a realização de uma análise mais profunda, ao se calcular a distribuição de valores obtidos, correlação entre valores, valores máximos e mínimos, entre outras. Uma análise subjetiva feita através da verificação dos nomes dos pesquisadores mais relevantes também é realizado para que se verifique que os pesquisadores identificados são de fato renomados e podem ser considerados relevantes de forma subjetiva.

Como resultado de todo o processo apresentado obtêm-se, por fim, um método para se identificar aqueles de maior relevância na comunidade representada pelo grafo.

6 Implementação computacional

6.1 Algoritmo de extração de sub-grafos

O algoritmo recursivo de extração de sub-grafos está descrito em Algoritmo 1. Esse recebe como entrada o grafo original G, um sub-grafo H de G que será expandido, uma lista de vértices P cujos filhos serão inseridos no sub-grafo, e o número de gerações, k, a serem

adicionadas a partir do vértice de interesse.

Cada vez que o algoritmo é executado, todos os vértices que estão em P são retirados e seus filhos são adicionados ao sub-grafo e à lista, desde que já não façam parte de H. Quando o algoritmo é chamado pela primeira vez, tanto H quanto P contém apenas o vértice a partir do qual o sub-grafo será gerado. Após cada chamada recursiva, o valor de k é diminuído para indicar que uma nova geração foi adicionada.

Quando $k = 1$ ou quando $P = \emptyset$, não é preciso ou possível adicionar novos vértices ao sub-grafo e a recursão é parada. Observe que, pelo modo como o algoritmo foi escrito, grafos com um número de gerações menor que k ainda serão considerados. Para que isso não aconteça, o algoritmo pode ser facilmente alterado para rejeitar tais casos se necessário.

Algoritmo 1: ExtrairSub-grafo

Entrada: Grafo $G = (V, E)$, Sub-grafo H de G, lista de vértices P, número de gerações do sub-grafo k

```
1  início
2      Para cada vértice v em P faça
3          P ← P \ {v}
4          Para toda aresta e = (v, w) ∈ E(G) faça
5              Se w ∉ V(H) então
6                  P ← P ∪ {w}
7                  V(H) ← V(H) ∪ {w}
8                  E(H) ← E(H) ∪ {e}
9              fim-Se
10         fim-Para
11     fim-Para
12     Se k > 1 e P ≠ ∅ então
13         ExtrairSub-grafo(G, H, P, k − 1)
14     fim-Se
15 fim
```

6.2 Algoritmo de cálculo de PageRank invertido

Para calcular o PageRank invertido nos grafos de genealogia descritos, foi utilizado o algoritmo exposto em Algoritmo 2. A entrada é composta por um grafo G, um critério de parada ϵ (epsilon) e um coeficiente de amortecimento d.

Primeiro, a variável utilizada para calcular o critério de parada ϵ_i é inicializada com um valor que permitirá a execução da primeira iteração. Em seguida, um PageRank inicial é

atribuído para todos os vértices. Esse valor é calculado de modo que todos tenham um mesmo PageRank. Como discutido por Brin & Page (2012), o valor inicial não altera o resultado final mas apenas a taxa de convergência.

Ao começo de uma iteração, cada vértice recebe um valor de PageRank igual a $\frac{1-d}{|V(G)|}$. A esse são somadas as parcelas de PageRank obtidos na última iteração ($\frac{PageRank_{i-1}(v)}{S(v)}$) dos filhos de cada vértice, caracterizando o cálculo invertido do PageRank. Essa parcela é multiplicada pelo fator de amortecimento e leva em conta a contribuição dos *sink nodes*, como discutido na Seção 3.2.

Após o término do cálculo de PageRank na iteração, ϵ_i recebe a distância Euclidiana entre $PageRank_i$ e $PageRank_{i-1}$. O algoritmo para quando a distância encontrada é menor que o parâmetro ϵ indicado na entrada. Por fim, o PageRank de cada vértice na iteração é armazenado em $PageRank_{i-1}$ para ser utilizado na iteração seguinte.

Para obter o PageRank local de cada pesquisador, o algoritmo é executado tendo o sub-grafo como o parâmetro de entrada. Ao final das iterações apenas o PageRank obtido para o pesquisador do qual o sub-grafo foi extraído é utilizado. Note que o calculo do PageRank local, para um dado vértice, é muito mais rápido do que o PageRank global, no entanto para todo o grafo, deverá de ser estimado o PageRank local para todos os vértices.

Uma análise inicial grosseira mostra que o algoritmo tem complexidade igual a $O(\alpha NM)$, onde α é o número de iterações, N o número de vértices e M o número de arestas. Entretanto, durante cada iteração os vértices e arestas são acessados uma única vez e a complexidade se reduz a $O(\alpha(N + M))$. Além disso, os grafos de genealogia utilizados nessa pesquisa são esparsos e cada vértice tem poucas arestas. Logo, o número de arestas é próximo ao número de vértices ($N \sim M$) e a complexidade se torna $O(\alpha N)$.

Ademais, para grafos muito grandes (como no caso dos grafos aqui utilizados) o número de iterações é muito pequeno comparado ao número de vértices ($N \gg \alpha$) e, portanto, pode ser desconsiderado[5]. Por fim, a complexidade final do algoritmo é $O(N)$.

[5]Note que essa premissa não é válida para o cálculo de PageRank em sub-grafos, dado que $N \sim \alpha$ na grande maioria dos casos.

Algoritmo 2: PageRank

Entrada: Grafo $G = (V, E)$, critério de parada ϵ, coeficiente de amortecimento d

início
  $\epsilon_i \leftarrow inf$
  Para *cada vértice* $v \in V$ **faça**
    $PageRank_{i-1}(v) \leftarrow \frac{1}{|V(G)|}$
  fim-Para
  Enquanto $\epsilon_i > \epsilon$ **faça**
    Para *cada vértice* $v \in V$ **faça**
      $PageRank_i(v) \leftarrow \frac{1-d}{|V(G)|}$
      Para *cada aresta* $e = (v, w) \in E$ **faça**
        $PageRank_i(v) \leftarrow PageRank_i(v) + d * \frac{PageRank_{i-1}(w)}{S(w)}$
      fim-Para
    fim-Para
    $\epsilon_i \leftarrow$ distEuclidiana($PageRank_i$, $PageRank_{i-1}$)
    Para *cada vértice* $v \in V$ **faça**
      $PageRank_{i-1}(v) \leftarrow PageRank_i(v)$
    fim-Para
  fim-enquanto
fim

7 Conjuntos de dados utilizados nos experimentos

A fim de testar o funcionamento do algoritmo em um caso real e obter resultados que podem ser validados, dados foram coletados de duas plataformas distintas: (i) *Academic Family Tree*[6] e (ii) Plataforma Lattes[7]. Ambos os grafos gerados a partir desses conjuntos de dados são grafos de genealogia acadêmica que contêm as informações dos pesquisadores cadastrados e relações de orientações de diversas áreas do conhecimento.

Dado que estes grafos contém dados de pesquisadores e orientações reais, e tendo em vista que muitos destes pesquisadores são de fato relevantes e já carregam consigo certo renome, o seu uso permite que os resultados sejam verificados através de avaliações quantitativas, através outros dados sobre sua carreira acadêmica (como premiações, publicações, bolsas recebidas, outras métricas), e também por meio de avaliações qualitativas e subjetivas, através da análise de seu currículo, área e local de atuação, entre outros.

Apesar de serem fontes reconhecidas e serem amplamente utilizadas, deve-se ter em mente que essas bases de dados podem conter informações incoerentes e enviesadas: como os ca-

[6] https://academictree.org/ último acesso em 29 de Junho de 2017.
[7] http://lattes.cnpq.br, último acesso em 18 de maio de 2017.

dastros são feitos pelos próprios pesquisadores, não existe garantia da qualidade dessas informações (podem existir auto-laços, e orientações reciprocas, i.e., A orienta B, e B orienta A).

O cálculo das gerações para esses grafos (expostos nas Tabelas 2 e 3) foi feito considerando que pesquisadores que não tem ascendentes (cadastrados nas bases de dados utilizadas) são os primeiros de suas linhagens e, consequentemente, pertencem a geração 0, como é o caso para os vértices A e D da Figura 2. A partir desses, podemos designar uma geração para cada vértice fazendo Geração(filho) $\leftarrow$ Geração(pai)+1. Em casos de conflito, como quando um pesquisador tem orientações de pais de gerações diferentes, a geração mais baixa foi considerada. Assim, os vértices E, H, I e N tem gerações 2, 1, 3 e 4, respectivamente.

Como as gerações são definidas apenas a partir das orientações e não do ano de atuação do pesquisador, pesquisadores da mesma geração não são necessariamente contemporâneos. Além disso, os pesquisadores cadastrados na primeira geração são aqueles que não tem ascendentes cadastrados nas plataformas e não devem ser pensados como pioneiros em suas áreas de pesquisa. Consequentemente, o conceito de geração aqui utilizado é melhor definido como o tamanho da ascendência (cadastrada) do pesquisador.

É importante ressaltar que, apesar de todos os vértices de uma geração (com a exceção da primeira) serem necessariamente filhos de ao menos um vértices da geração anterior, ainda podem haver orientações para a mesma geração ou para gerações anteriores. Por este motivo, o número de orientações recebidas de cada geração não é necessariamente igual ao número de orientações realizadas pela geração anterior.

7.1 Grafo dos doutores registrados na *Academic Family Tree*

O primeiro grafo utilizado foi gerado a partir das informações coletadas pelo projeto *Academic Family Tree*. Esse, originado a partir da expansão da iniciativa *NeuroTree*, tem por objetivo reunir e compartilhar informações sobre a genealogia acadêmica de pesquisadores contemporâneos e históricos.

A base de dados é sustentada pela contribuição voluntária de pesquisadores que desejam fazer parte da árvore genealógica, estando constantemente crescendo devido à inclusão de

Tabela 2: *Número de pesquisadores (vértices), orientações recebidas (arestas incidentes) e realizadas (arestas emergentes) para cada geração do grafo de genealogia extraído da Plataforma* Academic Family Tree.

Geração	**Número de Vértices**	**Arestas Incidentes**		**Arestas Emergentes**	
		Total	Média	Total	Média
0	152522	0	0,00	429557	2,82
1	413937	549809	1,33	135755	0,33
2	39494	134941	3,42	100858	2,55
3	26026	75326	2,89	100235	3,85
4	25079	67725	2,70	77644	3,10
5	13697	32727	2,39	20572	1,50
6	3424	6795	1,98	3484	1,02
7	549	987	1,80	314	0,57
8	68	137	2,01	34	0,50
9	11	22	2,00	11	1,00
10	4	8	2,00	8	2,00
11	4	8	2,00	15	3,75
12	7	14	2,00	11	1,57
13	5	10	2,00	8	1,60
14	2	4	2,00	4	2,00
15	2	4	2,00	4	2,00
16	2	4	2,00	4	2,00
17	2	4	2,00	4	2,00
18	2	3	1,50	4	2,00
19	4	6	1,50	7	1,75
20	4	8	2,00	8	2,00
21	4	6	1,50	4	1,00
22	2	3	1,50	6	3,00
23	2	2	1,00	2	1,00
Total:	674853	868553		868553	

novos cadastros. As informações são estruturadas em 80 árvores, uma para cada área do conhecimento, interligadas pela participação de pesquisadores em mais de um campo de pesquisa. O grafo obtido pela junção de todas as árvores tem um total de 674 853 pesquisadores e 868 553 orientações, contendo informações como o nome do pesquisador, suas orientações e a árvore a qual pertence.

Percebe-se que a grande maioria dos vértices (61%) se concentra na geração 1 do grafo. Esse fenômeno é consequência do alto número de pesquisadores que estão cadastrados na plataforma mas não indicaram seu orientador. Por consequência, também é nessa geração que se concentram o maior número de orientações recebidas e o maior número de orientações realizadas acontece na primeira geração. O número de pesquisadores passa a diminuir após atingir seu pico na geração 1. Após a sétima já são pouquíssimos os vértices em cada geração.

Esses vértices pertencem às poucas grandes linhagens do grafo e mostram que existe um limite de viabilidade para o número máximo de gerações no cálculo do PageRank local: mesmo havendo um grande número de gerações, o número de vértices que realmente se beneficia disso é muito pequeno.

7.2 Grafo dos pesquisadores registrados na Plataforma Lattes

O segundo grafo de genealogia foi obtido no contexto dos projetos de pesquisa em Cientometria realizados na UFABC ao qual este projeto de Iniciação Científica está inserido. O grafo foi gerado por meio da aplicação de métodos de *data mining* aos dados da plataforma Lattes, de onde todas as informações de cada pesquisador foram retiradas. Informações mais detalhadas sobre a metodologia utilizada para a mineração das informações e construção do grafo podem ser encontradas no trabalho de Damaceno *et al.* (2017).

Atualmente a Plataforma Lattes consta com mais de 5 milhões de currículos registrados. Contudo, para a geração do grafo e a sua análise nesta pesquisa foram consideradas apenas as informações de doutores, mestres e suas orientações de doutorado e mestrado. Além do mais, foram incluídas também pesquisadores que foram referenciados em currículos mas que não têm currículos na plataforma (por já terem encerrado sua carreira acadêmica antes de seu surgimento, por exemplo). Com isso, o grafo resultante tem 1 111 544 pesquisadores e 1 208 398 orientações.

Tabela 3: *Número de pesquisadores (vértices), orientações recebidas (arestas incidentes) e realizadas (arestas emergentes) para cada geração do grafo de genealogia extraído da Plataforma Lattes*

Geração	**Número de Vértices**	**Arestas Incidentes**		**Arestas Emergentes**	
		Total	Média	Total	Média
0	109791	0	0,00	147424	1,34
1	129727	205016	1,58	561729	4,33
2	477346	582412	1,22	407087	0,85
3	322943	347864	1,07	81592	0,25
4	63526	64840	1,02	9048	0,14
5	7047	7099	1,00	1389	0,19
6	1061	1064	1,00	129	0,12
7	103	103	1,00	0	0,00
Total:	1111544	1208398		1208398	

O número de gerações deste grafo é muito inferior ao do grafo discutido anteriormente. Além disso, observa-se que o maior número de vértices está localizado na geração 2, indicando que boa parte dos vértices tem pelo menos seus avôs acadêmicos cadastrados na plataforma. Note que o número de gerações evidencia a idade jovem da Ciência Brasileira em conjunto com o surgimento recente da Plataforma Lattes e de seus registros.

De maneira similar ao grafo da *Academic Family Tree*, o maior número de orientações recebidas acontece na geração de maior tamanho e o maior número de orientações realizadas acontece na geração anterior a essa.

Outro fenômeno que se repete é o da diminuição do número de pesquisadores em gerações de maior número, o que reafirma o limite já discutido para o cálculo de PageRank local.

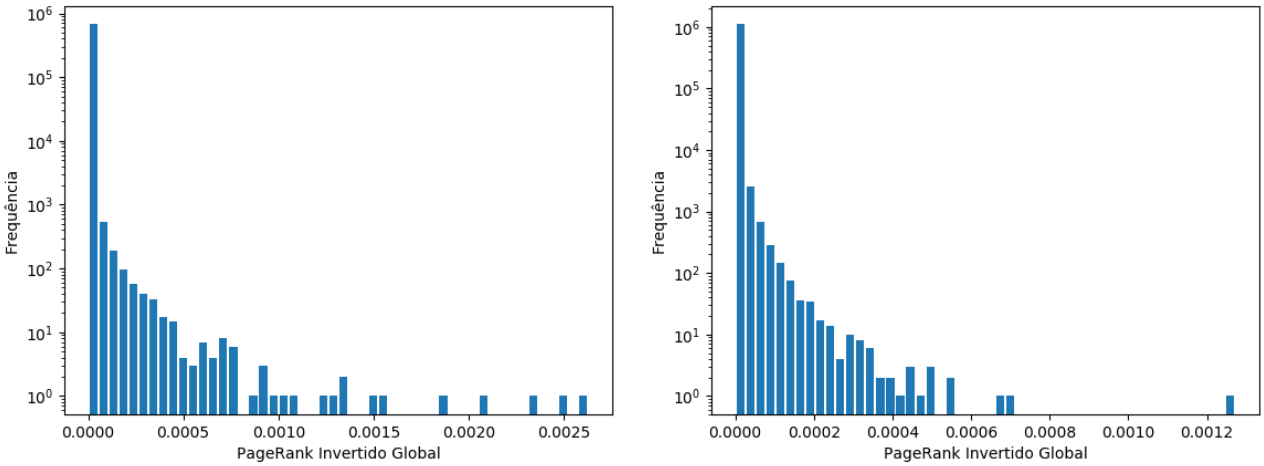
8 Resultados experimentais

8.1 PageRank invertido global

Utilizando os módulos 5.1 e 5.5 descritos na Seção 5, o algoritmo do PageRank foi aplicado aos grafos de genealogia acadêmica descritos acima para a obtenção do PageRank invertido global. Durante o processamento, foram utilizados como parâmetros de entrada um fator de amortecimento padrão igual a $0,85$ e um *epsilon* (ϵ), utilizado como critério de parada para as iterações, igual a 10^{-10}.

Após a execução do programa escrito para ler e calcular o PageRank invertido do grafo dado, obtivemos como saída uma relação de pesquisadores e seus respectivos PageRanks invertidos. Devido ao grande número de indivíduos pertencentes ao grafo e pelo fato de o PageRank se basear na divisão de 1 pelo total de vértices do grafo, os valores obtidos são da ordem de 10^{-6}.

Como é possível observar nos histograma apresentados na Figura 6, os PageRanks invertidos seguem uma distribuição semelhante a uma Lei da Potência, havendo muitas ocorrências de valores baixos (próximos da casa de 10^{-6}) e alguns poucos valores que se destacam na vizinhança de 10^{-3}.



(a) (b)

Figura 6: *Histograma de PageRanks calculados para os grafos de genealogia: (a) Academic Family Tree, (b) Pesquisadores da Plataforma Lattes.*

Para o grafo obtido a partir da *Academic Family Tree*, após um total de 71 iterações o valor máximo de PageRank encontrado foi de $2,62 \times 10^{-3}$ e o mínimo foi $6,75 \times 10^{-7}$. Já para o grafo de pesquisadores da Plataforma Lattes, os valores máximo e mínimo encontrados foram $1,27 \times 10^{-3}$ e $3,24 \times 10^{-7}$, respectivamente.

O valor mais frequente para ambos os grafos corresponde ao valor inicial atribuído a todos os vértices do grafo antes do início do processamento. Esse valor é mantido durante o processamento para aqueles acadêmicos que não possuem orientações (arestas apontando para si) e indica os indivíduos que não orientaram nenhum aluno. O mesmo fenômeno pode ser observado para outros valores de alta frequência, que correspondem a PageRanks de pesquisadores com poucos alunos. Os valores menos frequentes, com ocorrência de 1, indicam os indivíduos mais influentes do grafo, com maior número de filhos acadêmicos e grande descendência.

Apesar da distribuição de PageRanks de fato identificar aqueles pesquisadores mais influentes, nota-se que existe uma desproporção com relação ao número de pesquisadores identificados, dado que são pouquíssimos os pesquisadores que de fato se destacam. Existe ainda um senso de injustiça pois, como discutido na Seção 8.2, os pesquisadores destacados são aqueles mais antigos e que se encontram no topo da árvore genealógica. Estes resultados, embora preliminares, permitiram evidenciar a real exequibilidade da proposta e abriram uma oportunidade para uma análise mais aprofundada onde um número de gerações (k) foi considerado.

Identificação de pesquisadores brasileiros com maiores PageRanks globais

Utilizando o grafo de pesquisadores da Plataforma Lattes, o qual contempla pesquisadores de todas as áreas cadastradas, foi possível selecionar aqueles com os maiores PageRanks invertidos. A Tabela 4 mostra os 30 pesquisadores com maior PageRank invertido global.

Com a análise da lista apresentada na Tabela 4, vemos que existe uma grande diferença entre o pesquisador com maior PageRank e aquele com menor PageRank, evidenciando a distribuição desigual de valores pelos vértices. De forma semelhante, os valores de PageRank ficam cada vez mais próximos e se aglomeram cada vez mais conforme descemos o ranking de pesquisadores. Isso pode ser observado pela grande quantidade de pesquisadores com PageRank na casa de 3×10^{-4}, enquanto o primeiro tem PageRank de 1×10^{-3}.

Também foram selecionados para análise os maiores pesquisadores separados de acordo com a Grande Área do conhecimento em que atuam segundo seus currículos. A lista obtida é composta apenas pelos pesquisadores com currículos cadastrados na plataforma (e não que foram apenas referenciados) e que tem, portanto, sua área explicitamente declarada. Esses

Tabela 4: *Listas dos 30 pesquisadores com maior valor de PageRank invertido global.*

Nome	Rank	PRI
Joel Martins	1	1,2727E-03
Dermeval Saviani	2	7,0162E-04
Carolina Martuscelli Bori	3	6,8457E-04
Eduardo Doliveira Franca	4	5,4705E-04
Crodowaldo Pavan	5	5,3845E-04
Cidmar Teodoro Pais	6	4,9876E-04
Massaud Moisés	7	4,9796E-04
Lucrecia D'alessio Ferrara	8	4,8881E-04
Geraldina Porto Witter	9	4,7095E-04
Andre Dreyfus	10	4,5801E-04
João Lúcio de Azevedo	11	4,5503E-04
Arrigo Leonardo Angelini	12	4,3771E-04
Friedrich Gustav Brieger	13	4,2955E-04
Warwick Estevam Kerr	14	3,9279E-04
José Manoel de Arruda Alvim Netto	15	3,8734E-04
Elisaldo Luiz de Araujo Carlini	16	3,7162E-04
Maria Lucia Santaella Braga	17	3,6748E-04
Maria Antonieta Alba Celani	18	3,5178E-04
Sérgio de Iudícibus	19	3,4866E-04
György Miklós Böhm	20	3,4353E-04
Carlos Guilherme Santos Serôa da Mota	21	3,3564E-04
Miguel Rolando Covian	22	3,3451E-04
Ivan Antônio Izquierdo	23	3,3115E-04
Marilena de Souza Chaui	24	3,3007E-04
Sérgio Buarque de Holanda	25	3,2811E-04
Luiz Pereira	26	3,2115E-04
Jack Peter Green	27	3,1620E-04
Otto Richard Gottlieb	28	3,1399E-04
Octavio Ianni	29	3,1134E-04
Winifred Kera Stevens	30	3,1095E-04

pesquisadores não são necessariamente os pesquisadores com maiores PageRanks invertidos do grafo, dado que existem muitos pesquisadores relevantes que não foram considerados por não terem a informação de suas áreas explicitada. Obteve-se, assim, dez pesquisadores de cada uma das oito diferentes grandes áreas, sendo essas: Ciências Agrárias; Ciências da Saúde; Ciências Biológicas; Ciências Exatas e da Terra; Engenharias; Linguística, Letras e Arte; Ciências Humanas e Ciências Sociais.

Percebe-se pelas listas geradas (Tabela 5) que há certa variação entre os valores selecionados, sendo o menor de $1,0166 \times 10^{-4}$ (rank 334) e o maior de $7,0162 \times 10^{-4}$ (rank 2). Apesar de todos os valores estarem na casa de 10^{-4}, a variação observada do valor de PageRank é

Tabela 5: *Listas de pesquisadores com maior valor de PageRank invertido global separados em suas respectivas áreas.*

Ciências Agrarias		
Nome	Rank	PRI
Octavio Nakano	113	1,7292E-04
Franco Maria Lajolo	126	1,6559E-04
Vicente Wagner Dias Casali	128	1,6157E-04
Cyro Paulino da Costa	176	1,4113E-04
Paulo Leonel Libardi	184	1,3838E-04
José Fernando Coelho da Silva	211	1,2807E-04
Yong Kun Park	285	1,1102E-04
Norberto Mario Rodriguez	305	1,0580E-04
Otto Jesu Crocomo	307	1,0563E-04
Tuneo Sediyama	334	1,0166E-04

Ciências da Saúde		
Nome	Rank	PRI
György Miklós Böhm	20	3,4353E-04
Eduardo Moacyr Krieger	71	2,1010E-04
Ruy Laurenti	106	1,7963E-04
Paulo Hilario Nascimento Saldiva	133	1,5868E-04
Antonio Ruffino Netto	146	1,5172E-04
José Eduardo Dutra de Oliveira	163	1,4618E-04
Emilia Luigia Saporiti Angerami	191	1,3589E-04
Guilherme Rodrigues da Silva	257	1,1770E-04
Vicente Amato Neto	262	1,1564E-04
Mauricio Malavasi Ganança	284	1,1113E-04

Ciências Biológicas		
Nome	Rank	PRI
Crodowaldo Pavan	5	5,3845E-04
João Lúcio de Azevedo	11	4,5503E-04
Warwick Estevam Kerr	14	3,9279E-04
Elisaldo Luiz de Araujo Carlini	16	3,7162E-04
Ivan Antônio Izquierdo	23	3,3115E-04
Francisco Mauro Salzano	34	2,9858E-04
Leopoldo de Meis	46	2,5377E-04
Almiro Blumenschein	60	2,2482E-04
Roland Vencovsky	62	2,2437E-04
Isaias Raw	92	1,8907E-04

Ciências Exatas e da Terra		
Nome	Rank	PRI
Otto Richard Gottlieb	28	3,1399E-04
Carlos José Pereira de Lucena	54	2,3800E-04
Sergio Mascarenhas Oliveira	55	2,3694E-04
Eduardo Fausto de Almeida Neves	57	2,3385E-04
Klaus Reichardt	64	2,2403E-04
Shigueo Watanabe	147	1,5125E-04
Manfredo Perdigao do Carmo	182	1,3973E-04
Eneas Salati	187	1,3733E-04
Jacob Palis Junior	220	1,2473E-04
Ailton de Souza Gomes	229	1,2208E-04

Engenharias		
Nome	Rank	PRI
Neri dos Santos	42	2,6677E-04
Ricardo Miranda Barcia	47	2,4914E-04
Leonardo Ensslin	80	1,9973E-04
Giulio Massarani	95	1,8782E-04
Luiz Bevilacqua	109	1,7844E-04
Rosalvo Tiago Ruffino	125	1,6560E-04
Francisco José Kliemann Neto	143	1,5262E-04
Marcius Fantozzi Giorgetti	149	1,5077E-04
Ivo Barbi	153	1,4925E-04
Evaristo Chalbaud Biscaia Junior	175	1,4114E-04

Linguística, Letras e Artes		
Nome	Rank	PRI
Cidmar Teodoro Pais	6	4,9876E-04
Maria Lucia Santaella Braga	17	3,6748E-04
Maria Antonieta Alba Celani	18	3,5178E-04
Regina Zilberman	45	2,5781E-04
Anthony Julius Naro	141	1,5422E-04
Mary Aizawa Kato	150	1,5044E-04
Cleonice Seroa da Motta Berardinelli	165	1,4558E-04
Maria Aparecida de Campos Brando Santilli	166	1,4545E-04
Leila Barbara	199	1,3315E-04
Eni de Lourdes Puccinelli Orlandi	219	1,2492E-04

Ciências Humanas		
Nome	Rank	PRI
Dermeval Saviani	2	7,0162E-04
Geraldina Porto Witter	9	4,7095E-04
Carlos Guilherme Santos Serôa da Mota	21	3,3564E-04
Marilena de Souza Chaui	24	3,3007E-04
Gabriel Cohn	50	2,4543E-04
Carmen Sylvia de Alvarenga Junqueira	61	2,2439E-04
Silvia Tatiana Maurer Lane	65	2,2241E-04
Jose Sebastiao Witter	66	2,2021E-04
Eunice Ribeiro Durham	73	2,0746E-04
João Baptista Borges Pereira	76	2,0348E-04

Ciências Sociais Aplicadas		
Nome	Rank	PRI
Lucrecia D'alessio Ferrara	8	4,8881E-04
José Manoel de Arruda Alvim Netto	15	3,8734E-04
Sérgio de Iudícibus	19	3,4866E-04
Nelson Nery Junior	44	2,6476E-04
Fatima Cristina Trindade Bacellar	48	2,4847E-04
Jose Alfredo de Oliveira Baracho	69	2,1455E-04
Armando Catelli	78	2,0258E-04
Dalmo de Abreu Dallari	87	1,9251E-04
Tercio Sampaio Ferraz Junior	91	1,8953E-04
Jose Marques de Melo	97	1,8537E-04

considerável, uma vez que essa ocorre apenas entre os 80 maiores PageRanks invertidos. Isso pode ser explicado pela grande quantidade de orientações que cada um dos pesquisadores presentes nas listas tem, em oposição aos números menores e mais recorrentes de orientações de outros pesquisadores.

Esse fenômeno ocorre também em cada uma das áreas descritas. Apesar de apresentarem valores mais homogêneos, não sendo a variação dentro delas tão grande, é notável a diferença entre o primeiro e o último pesquisador de cada uma. Por exemplo, para áreas como Ciências Agrárias, Ciências da Saúde, Ciências Exatas e da Terra e Linguística, Letras e Artes a diferença entre os ranks dos pesquisadores chega a ser de mais de 200 posições.

Além da variação total, existe uma visível diferença entre os valores máximos de cada uma das áreas. Esses valores podem ser associados aos pesquisadores mais influentes que atuam nelas. A diferença, que chega a ser comparavelmente grande àquela do primeiro e do último valor de todas as listas, mostra a disparidade entre o nível de desenvolvimento em termos de formação de alunos nas distintas áreas. Com a análise dos ranks de cada área, podemos ver que áreas como Ciências Biológicas, Humanas e Sociais Aplicadas estão melhor colocadas que as outras, podendo isso ser uma evidência de uma maior produção de pesquisadores e também da idade destas áreas.

Por fim, vemos que muitos dos pesquisadores presentes na seleção já são renomados e reconhecidos por seus trabalhos e contribuições para o desenvolvimento de suas áreas e da Ciência Brasileira. Esse fato, apesar de qualitativo (e ainda a ser melhor explorado), sinaliza que o processo de identificação utilizado pode, de fato, indicar os pesquisadores mais relevantes do meio científico através da análise de sua genealogia acadêmica.

8.2 PageRank invertido local

Com os resultados de PageRank invertido global calculados e a verificação de que o algoritmo é, de fato, capaz de identificar pesquisadores influentes em um grafo, partiu-se para a próxima etapa do projeto descrita no módulo 5.2 (Extração de sub-grafos).

Para cada vértice de cada grafo, diversos sub-grafos foram gerados considerando de duas a cinco gerações a partir do vértice de interesse. Desse modo, podemos comparar os resultados

e estudar como o PageRank local se comporta para diferentes valores de k. A decisão de contar as gerações a partir do vértice do qual o sub-grafo foi gerado (e não considerar sua ascendência) foi tomada com base na noção de que o pesquisador deve ser tratado como o vértice de maior relevância local e, consequentemente, é interessante que acumule mais PageRank que os outros.

Note que, apesar dos sub-grafos com um maior número de gerações aparentarem ser muito grandes para uma análise local, existem grafos com um número de gerações muito maior, como o grafo da *Academic Family Tree* que possui 23. Sendo assim, a quantidade de gerações utilizadas ainda é comparativamente pequena. Entretanto, para grafos menores (como o do Lattes, que tem apenas 8 gerações) estes números podem ser de fato excessivos e sub-grafos menores são desejáveis.

O algoritmo descrito em Algoritmo 2 foi utilizado para calcular o PageRank de cada um dos sub-grafos e, ao final do cálculo, apenas o PageRank do pesquisador a partir do qual o sub-grafo foi gerado é armazenado. A partir destes primeiros resultados notou-se que, apesar de calculado corretamente, valores altos de PageRank foram atribuídos para vértices com poucos descendentes[8] e valores baixos para aqueles com muitos. Isso ocorre pois quanto menor a quantidade de vértices, menor será a distribuição do PageRank para cada um deles e, dessa forma, maior será o valor médio obtido por cada vértice. Um exemplo desse fenômeno são os sub-grafos de pesquisadores sem descendentes que, por conterem apenas um vértice, acabam concedendo-o um PageRank igual a 1.

Assim, pesquisadores com uma grande relevância local, apesar de terem PageRanks altos se comparados com outros vértices de seus sub-grafos, acabam sendo penalizados pela quantidade de descendente que possuem. Tendo isso em vista, seria necessário selecionar os pesquisadores com os menores PageRanks locais dentre todos os outros. Para facilitar essa seleção, o PageRank local de todos os vértices foi normalizado de acordo com o método descrito na Seção 4.2.

Para essa nova métrica os valores de PageRank obtido são maiores, estando todos (com

[8]Como discutido na Seção 6.1, os sub-grafos extraídos pelo Algoritmo 1 podem ter menos que k gerações e, consequentemente, existem sub-grafos com pouquíssimos vértices. Optamos por essa abordagem para não excluir pesquisadores muito jovens da avaliação, tornando-a, mais apropriada para avaliação de pesquisadores de diferentes idades acadêmicas.

exceção dos pesquisadores com PageRank 0) na casa de 10^{-1}. Isso ocorre pois agora o tamanho dos sub-grafos é pequeno (de, no máximo 10^3 pesquisadores), o que acaba produzindo PageRanks com valores mais altos.

Com os novos valores de PageRank, gráficos de dispersão de PageRank global por local e histogramas de PageRank local foram gerados. Os gráficos originados de ambos os grafos para $k = 2$ e $k = 5$ estão expostos nas Figuras 7 e 8.

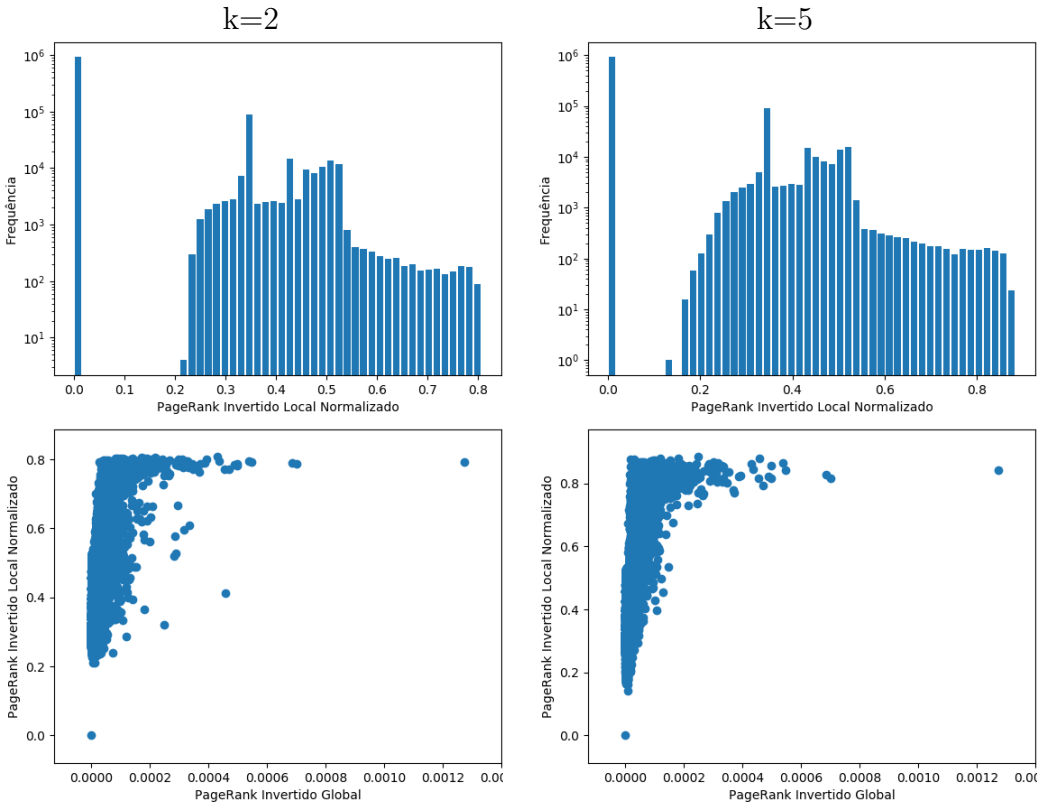


Figura 7: *Histogramas e gráficos de dispersão de PageRanks locais calculados para o grafo de genealogia extraído da Plataforma Lattes.*

É importante ressaltar que os maiores sub-grafos (por quantidade de vértices) não estão, em nenhum dos casos expostos, relacionados aos maiores PageRanks locais. Como comprovado pelo coeficiente de correlação obtido para as duas métricas, que se encontra na faixa de $0,1$ a $0,5$ para todos os resultados, o tamanho do sub-grafo não dita o valor do PageRank de um vértice. Isso também evidencia a importância da morfologia do sub-grafo para o cálculo do PageRank e a sua diferença para outras métricas.

Como é possível observar pelos gráficos, o PageRank local se comporta de maneira muito semelhante para ambos os grafos testados. Tanto no grafo obtido pela Plataforma Lattes, quanto naquele da *Academic Family Tree*, os valores de PageRank local estão bem distribuídos

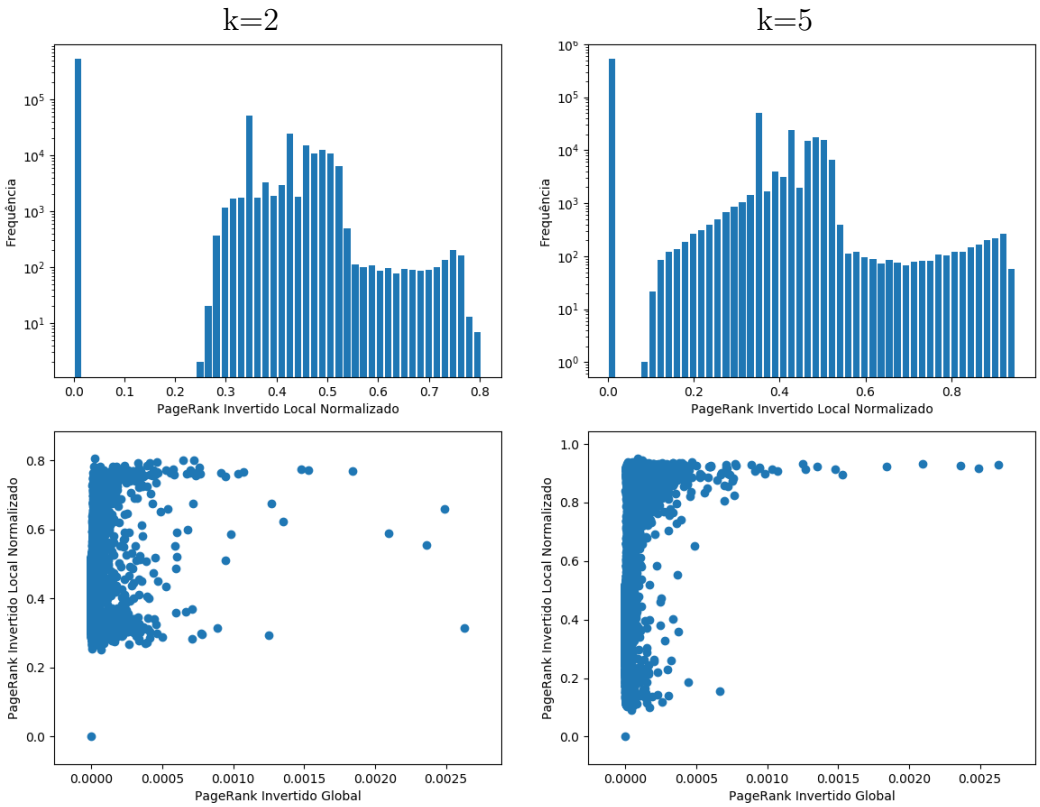


Figura 8: *Histogramas e gráficos de dispersão de PageRanks locais calculados para o grafo de genealogia extraído da Plataforma Academic Family Tree.*

em uma faixa $0,4$ a $0,8$. Diferente do PageRank global, o PageRank local não se comporta de acordo com uma lei da potência apesar de ainda existirem valores com alta incidência, os quais estão ligados aos pesquisadores com poucos ou nenhum descendente. Isso comprova, portanto, que a métrica é mais justa e proporcional, dado que os valores estão mais espalhados e que os pesquisadores não estão todos aglomerados em uma pequena faixa de valores baixos enquanto existem poucos com valores muito altos.

Pela análise dos histogramas podemos ver, ainda, que existem algumas regiões (de $0,0$ a $0,2$) que o número de pesquisadores é muito baixo se comparado com outros intervalos, assim como existe um pico para valores entre $0,3$ e $0,5$ que decresce para valores mais altos. Para os casos onde existe uma pequena ocorrência de pesquisadores, isso é explicado pelo grande salto que ocorre entre o valor de PageRank que vértices sem nenhum descendente assumem $(0,0)$, e o valor que o próximo sub-grafo com menor PageRank (com apenas um descendente ocorrem). Já os casos onde existem picos estão relacionados à grafos de alta ocorrência e morfologicamente semelhantes, como no caso de pesquisadores com um ou dois descendentes.

Além disso, a análise desses resultados mostrou que não existe uma forte correlação entre

o PageRank invertido global e local. Para todos os valores de k testados em ambos os grafos o coeficiente de correlação se manteve numa faixa entre $0,3$ e $0,4$. Esses valores decrescem para a vizinhança de $0,15$ quando apenas os 100 pesquisadores com maiores PageRanks globais são considerados. Este fato está representado nos gráficos de dispersão apresentados, nos quais existe uma grande aglomeração de pontos no canto superior esquerdo e pouca relação de linearidade. Isso implica que existem pesquisadores com PageRank invertido global baixo que se destacam no PageRank invertido local, mostrando sua capacidade de identificar pesquisadores jovens.

Pesquisadores que se destacam com o PageRank invertido global, em sua maioria, também se destacam como PageRank invertido local, como vemos pela grande área sem ocorrências no canto inferior direito do gráfico de dispersão, representando pesquisadores com grande PageRank global, mas pouco PageRank local. Esse resultado mostra que pesquisadores relevantes de modo global geralmente têm grande destaque local, o que é intuitivo para a maior parte dos casos.

Existem, porém, pesquisadores que são prejudicados pela análise local, como vemos pela análise para k=2 no grafo da Academic Family Tree. Esses pesquisadores são aqueles que têm poucos alunos próximos, mas por serem muito relevantes passam uma parcela de seu PageRank para seu orientador. Esses são, portanto, exatamente os casos que queremos identificar e tornar mais justos com a análise local. Para um valor de k mais alto esse fenômeno já não é evidenciado, dado que esses pesquisadores começam a ganhar relevância de gerações maiores.

Antes da obtenção dos resultados, havíamos previsto que o PageRank de pesquisadores mais antigos se aproximaria do PageRank global conforme mais gerações sejam incluídas. Essa afirmação tinha como base o fato de que com a adição de novos vértices os sub-grafos se assemelharão cada vez mais aos grafos completos, tanto em tamanho quanto em morfologia. Entretanto, não houve uma variação considerável no coeficiente de correlação entre os números de gerações considerados. Existe pouca variação entre os valores de k testados para a grande maioria dos pesquisadores que não se encontram no topo do ranking. Para todos os valores testados, o PageRank local continua evidenciando de forma semelhante os mesmos pesquisadores, também pelo fato da grande maioria não ter mais do que duas ou três gerações

de descendentes e, portanto, não se beneficiar de um número maior de gerações consideradas no cálculo.

Por outro lado, notou-se que conforme aumentamos o número de gerações dos sub-grafos, pesquisadores mais antigos começaram a assumir PageRanks locais mais altos. Deste modo, vemos que um valor de k muito alto (próximo ao número máximo de gerações de um grafo) não proporciona todos os resultados desejados, dado que continua a evidenciar pesquisadores que são relevantes apenas por sua senioriedade. Considera-se preferencialmente utilizar valores de k igual a 2 ou 3, representando uma influência direta um pesquisador apenas sobre seus netos ou bisnetos.

Conclui-se, portanto, que existe uma diferença entre a relevância global e local de um pesquisador, e que a análise por meio de sub-grafos pode vir a identificar pesquisadores que, por serem mais jovens, não se destacam na análise global.

Pesquisadores com maiores PageRanks locais

De maneira semelhante ao que foi feito na Seção 8.1 e com base nos resultados obtidos pelo processamento do grafo da Plataforma Lattes, os pesquisadores com maiores PageRanks locais para $k = 2$ e $k = 5$ e de cada área foram selecionados para uma análise qualitativa. Os dados utilizados nesta análise estão apresentados nas Tabelas 6 a 9.

Diferente dos valores encontrados para os maiores PageRanks globais, os PageRanks locais se encontram muito mais agrupados, estando todos em torno de $0,8$, e sendo necessário até quatro casas decimais para diferenciar alguns deles. Entretanto, a diferença entre os maiores e menores PageRanks ainda é perceptível, especialmente para $k = 5$.

Apesar de o PageRank global e local estarem pouco correlacionados (Seção 8.2), a análise dos maiores PageRanks locais mostrou que existem pesquisadores relevantes globalmente que também se destacam de maneira local, dado que tanto para $k = 2$ quanto $k = 5$ os maiores pesquisadores estão entre os 1000 primeiros ranks de PageRank global (dentre os mais de um milhão do grafo). Isso é esperado, uma vez que um pesquisador com grande relevância local também tem uma grande relevância global caso seja suficientemente sênior. Alguns dos pesquisadores que mais acumularam PageRank são renomados por sua atuação na academia.

Tabela 6: *Listas dos 30 pesquisadores com maior valor de PageRank invertido local com k=2.*

Nome	Rank PR	Rank PRIL	PRIL k=2
Friedrich Gustav Brieger	13	1	8,0717E-01
Bernard Pottier	68	2	8,0554E-01
Miguel Reale	119	3	8,0447E-01
José da Costa Boucinhas	104	4	8,0419E-01
Louis Michel	173	5	8,0331E-01
Aniela Ginsberg	90	6	8,0285E-01
Arnold Rothe	67	7	8,0279E-01
Raul Valentin da Silva	303	8	8,0224E-01
Virender Kumar Handa	304	9	8,0224E-01
Cândido Procópio Ferreira de Camargo	362	10	8,0223E-01
Fritz Joachim Von Rintelen	477	11	8,0200E-01
Louis Richard Anderson	425	12	8,0189E-01
Paul Hugon	389	13	8,0158E-01
Warwick Estevam Kerr	14	14	8,0144E-01
Lenita Correa Camargo	35	15	8,0130E-01
Karl Heinz Schwab	269	16	8,0078E-01
Orlando Magalhães Carvalho	74	17	8,0067E-01
José Theophylo do Amaral Gurgel	85	18	8,0044E-01
Joseph Alan Roper	84	19	8,0033E-01
Antonio Rubbio Muller	412	20	8,0030E-01
Florestan Fernandes	53	21	8,0023E-01
João Paulo de Almeida Magalhães	70	22	7,9958E-01
Paul Arboussebastide	279	23	7,9934E-01
Domingos Gallo	31	24	7,9927E-01
David Mayburylewis	194	25	7,9911E-01
Wilson da Silva Sasso	63	26	7,9846E-01
Gillesgaston Granger	502	27	7,9846E-01
Justo Moretti Filho	261	28	7,9824E-01
Guido Ranzani	94	29	7,9808E-01
Jose Martiniano de Azevedo Neto	127	30	7,9726E-01

Tabela 7: *Listas dos 30 pesquisadores com maior valor de PageRank invertido local com k=5.*

Nome	Rank PR	Rank PRIL	PRIL k=5
Fatima Cristina Trindade Bacellar	48	1	8,8396E-01
Andre Dreyfus	10	2	8,8077E-01
Maurice Byé	105	3	8,7990E-01
Tese de Catedrático	256	4	8,7873E-01
Widinei Alves Fernandes	996	5	8,7714E-01
Samuel Devons E Aubrey Jaffe	950	6	8,7564E-01
Frederico Gustavo Brieger	374	7	8,7335E-01
Barbara Mcclintock	375	8	8,7335E-01
Erastus H Lee	145	9	8,7313E-01
Lourival Gomes Machado	341	10	8,7264E-01
Crenildo Campelo	450	11	8,7110E-01
Gerald A Fleischer	451	12	8,7110E-01
James Greene	299	13	8,7055E-01
Ph D David A Thompson	912	14	8,6927E-01
Carl E Taylor	764	15	8,6904E-01
Italo Bettarello	410	16	8,6900E-01
J F Eastham	625	17	8,6898E-01
Gerhard Salinger	698	18	8,6835E-01
Leslie John Kastner	846	19	8,6791E-01
Jorge Pereira Lima	33	20	8,6715E-01
Paula Beiguelman	215	21	8,6714E-01
Jose Maria Nogueira da Costa	970	22	8,6713E-01
Reginald H Painter	971	23	8,6650E-01
Jack Peter Green	27	24	8,6627E-01
Cornelius Bernard Van Niel	172	25	8,6598E-01
Giovanni Gazzinelli	231	26	8,6563E-01
Maurice Chaumont	368	27	8,6551E-01
Sanjit Kumar Mitra	832	28	8,6488E-01
Marcello Iacomini	661	29	8,6438E-01
Arthur C Smith	964	30	8,6425E-01

Entretanto, quase nenhum dos pesquisadores encontrados nas listas de pesquisadores com maiores PageRanks locais estão presentes na lista global, indicando que melhores pesquisadores de maneira global são, de fato, prestigiados pela sua senioridade e não pela sua relevância local. Isso também ocorre entre as duas listas de PageRanks locais, revelando que a a inclusão de novas gerações produz um efeito similar de exaltação de pesquisadores mais antigos.

Tabela 8: *Listas de pesquisadores com maior valor de PageRank local para $k = 2$ separados em suas respectivas áreas.*

Ciências Agrarias		
Nome	Rank	PRI
Antonio Fernando Lordelo Olitta	154	7,8328E-01
José Fernando Coelho da Silva	173	7,8167E-01
Cyro Paulino da Costa	174	7,8164E-01
Octavio Nakano	196	7,7892E-01
Otto Jesu Crocomo	211	7,7720E-01
Salassier Bernardo	236	7,7502E-01
Norberto Mario Rodriguez	239	7,7456E-01
Horacio Santiago Rostagno	282	7,6954E-01
Paulo Afonso Ferreira	284	7,6931E-01
Tuneo Sediyama	303	7,6700E-01

Ciências da Saúde		
Nome	Rank	PRI
Guilherme Rodrigues da Silva	119	7,8722E-01
György Miklós Böhm	120	7,8715E-01
José Mondelli	132	7,8498E-01
Ruy Laurenti	137	7,8451E-01
José Eduardo Dutra de Oliveira	152	7,8341E-01
Vilma de Carvalho	160	7,8252E-01
Emilia Luigia Saporiti Angerami	169	7,8174E-01
Oslei Paes de Almeida	203	7,7835E-01
Eduardo Moacyr Krieger	242	7,7419E-01
Décio Rodrigues Martins	243	7,7405E-01

Ciências Biológicas		
Nome	Rank	PRI
Warwick Estevam Kerr	15	8,0144E-01
Crodowaldo Pavan	42	7,9598E-01
Almiro Blumenschein	62	7,9328E-01
Isaias Raw	84	7,9089E-01
Giovanni Gazzinelli	108	7,8876E-01
Leopoldo Magno Coutinho	111	7,8844E-01
Elisaldo Luiz de Araujo Carlini	118	7,8730E-01
Jesus Santiago Moure	128	7,8568E-01
Armando Freitas da Rocha	129	7,8545E-01
Leopoldo de Meis	150	7,8372E-01

Ciências Exatas e da Terra		
Nome	Rank	PRI
Eneas Salati	92	7,8972E-01
Sergio Mascarenhas Oliveira	101	7,8893E-01
Nei Fernandes de Oliveira Junior	116	7,8797E-01
Eduardo Fausto de Almeida Neves	127	7,8576E-01
Klaus Reichardt	183	7,8004E-01
Osvaldo Antonio Serra	195	7,7893E-01
Manfredo Perdigao do Carmo	201	7,7851E-01
Blanka Wladislaw	202	7,7836E-01
Otto Richard Gottlieb	210	7,7721E-01
Carlos José Pereira de Lucena	219	7,7698E-01

Engenharias		
Nome	Rank	PRI
Luiz Bevilacqua	56	7,9405E-01
Rosalvo Tiago Ruffino	63	7,9326E-01
Manuel de Jesus Mendes	126	7,8606E-01
Isaias de Carvalho Macedo	141	7,8415E-01
Giulio Massarani	155	7,8324E-01
Marcius Fantozzi Giorgetti	158	7,8282E-01
Pérsio de Souza Santos	159	7,8276E-01
Leonardo Ensslin	164	7,8204E-01
Ivanildo Hespanhol	165	7,8200E-01
Carlos Ignacio Zamitti Mammana	189	7,7965E-01

Linguística, Letras e Artes		
Nome	Rank	PRI
Anthony Julius Naro	136	7,8457E-01
Maria Aparecida de Campos Brando Santilli	143	7,8402E-01
Aryon Dall'igna Rodrigues	153	7,8336E-01
Maria Antonieta Alba Celani	157	7,8298E-01
Clylton José Galamba Fernandes	162	7,8229E-01
Cidmar Teodoro Pais	171	7,8173E-01
Maria do Carmo Peixoto Pandolfo	209	7,7769E-01
José Cavalcante de Souza	213	7,7711E-01
Leila Barbara	288	7,6907E-01
Gilberto Mendonça Teles	302	7,6707E-01

Ciências Humanas		
Nome	Rank	PRI
Jose de Souza Martins	82	7,9105E-01
Eunice Ribeiro Durham	85	7,9085E-01
Jorge Lobo Miglioli	86	7,9073E-01
Jose Sebastiao Witter	97	7,8931E-01
Carlos Guilherme Santos Serôa da Mota	110	7,8854E-01
Dermeval Saviani	112	7,8835E-01
Roberto Cardoso de Oliveira	114	7,8820E-01
Fernando Antonio Novais	142	7,8403E-01
Wanderley Guilherme dos Santos	145	7,8393E-01
Marilena de Souza Chaui	146	7,8389E-01

Ciências Sociais Aplicadas		
Nome	Rank	PRI
Elza Salvatori Berquó	78	7,9165E-01
José Manoel de Arruda Alvim Netto	91	7,9004E-01
Luis Alberto Warat	105	7,8890E-01
Luiz Fernando Coelho	131	7,8503E-01
Maria da Conceicao de Almeida Tavares	135	7,8461E-01
Lucrecia D'alessio Ferrara	144	7,8397E-01
Nicolau Reinhard	207	7,7790E-01
Geraldo Luciano Toledo	228	7,7564E-01
Rodolfo Hoffmann	246	7,7362E-01
Carlos Roberto Jamil Cury	248	7,7330E-01

Com a análise dos maiores pesquisadores de cada área para $k = 2$, percebemos que os valores estão muito mais agrupados, sendo a diferença dentro de cada área muito menor que para o PageRank global e a maior diferença de ranks de 170. Além disso, a diferença entre as áreas é muito menos perceptível, indicando que áreas mais antigas não têm tanta vantagem quanto as outras.

Tabela 9: *Listas de pesquisadores com maior valor de PageRank local para $k = 5$ separados em suas respectivas áreas.*

Ciências Agrarias		
Nome	Rank	PRI
Jose Maria Nogueira da Costa	23	8,6713E-01
Carlos Holger Wenzel Flechtmann	181	8,4297E-01
Carlos Jorge Rossetto	361	8,2251E-01
Otto Jesu Crocomo	390	8,1863E-01
Antonio Fernando Lordelo Olitta	522	8,0558E-01
Salassier Bernardo	555	8,0137E-01
José Fernando Coelho da Silva	574	7,9865E-01
Cyro Paulino da Costa	575	7,9861E-01
Octavio Nakano	615	7,9407E-01
Julio Marcos Filho	617	7,9372E-01

Ciências da Saúde		
Nome	Rank	PRI
Edna Paciência Vietta	146	8,4887E-01
Vicente Amato Neto	221	8,3874E-01
Antonio Spina França Netto	227	8,3813E-01
György Miklós Böhm	251	8,3551E-01
Guilherme Rodrigues da Silva	275	8,3282E-01
José Alberto de Souza Freitas	298	8,2967E-01
Carmen Fontes de Souza Teixeira	307	8,2887E-01
Ruy Laurenti	344	8,2475E-01
Antonio Ruffino Netto	355	8,2343E-01
Vilma de Carvalho	381	8,1996E-01

Ciências Biológicas		
Nome	Rank	PRI
Marcello Iacomini	30	8,6438E-01
Crodowaldo Pavan	37	8,6360E-01
Pedro Henrique Saldanha	53	8,6042E-01
Almiro Blumenschein	223	8,3846E-01
Carlos Augusto de Figueiredo Monteiro	295	8,3018E-01
Isaias Raw	315	8,2797E-01
Claudio Gilberto Froehlich	326	8,2656E-01
Warwick Estevam Kerr	345	8,2436E-01
Luiz Rodolpho Raja Gabaglia Travassos	350	8,2381E-01
Wilma Pereira Bastos Ramos	362	8,2235E-01

Ciências Exatas e da Terra		
Nome	Rank	PRI
Widinei Alves Fernandes	5	8,7714E-01
Sergio Mascarenhas Oliveira	166	8,4453E-01
Amaury de Souza	175	8,4393E-01
Mauricio Matos Peixoto	184	8,4291E-01
Carlos José Pereira de Lucena	249	8,3567E-01
Vadlamudi Brahmananda Rao	284	8,3176E-01
José Moacyr Vianna Coutinho	287	8,3153E-01
Nei Fernandes de Oliveira Junior	292	8,3048E-01
Erasmo Madureira Ferreira	318	8,2725E-01
Eneas Salati	324	8,2680E-01

Engenharias		
Nome	Rank	PRI
Affonso Carlos Seabra da Silva Telles	153	8,4728E-01
Marcos Alves de Magalhães	180	8,4309E-01
Amilton Martins dos Santos	194	8,4149E-01
Luiz Bevilacqua	197	8,4117E-01
Rosalvo Tiago Ruffino	198	8,4111E-01
Jose Abel Royo dos Santos	205	8,4023E-01
Leonardo Ensslin	226	8,3833E-01
Fernanda Margarida Barbosa Coutinho	291	8,3091E-01
Yaro Burian Junior	293	8,3036E-01
Carlos Augusto Guimaraes Perlingeiro	303	8,2924E-01

Linguística, Letras e Artes		
Nome	Rank	PRI
Leyla Beatriz Perrone Moisés	217	8,3930E-01
Maria Antonieta Alba Celani	255	8,3526E-01
Leonor Lopes Favero	309	8,2857E-01
Cidmar Teodoro Pais	400	8,1739E-01
Maria Aparecida de Campos Brando Santilli	428	8,1497E-01
Aryon Dall'igna Rodrigues	453	8,1304E-01
Leila Barbara	477	8,1049E-01
Maria do Carmo Peixoto Pandolfo	535	8,0393E-01
Clylton José Galamba Fernandes	540	8,0314E-01
Anthony Julius Naro	559	8,0089E-01

Ciências Humanas		
Nome	Rank	PRI
Eunice Ribeiro Durham	136	8,5068E-01
Jorge Lobo Miglioli	141	8,4946E-01
Lygia Maria Sigaud	159	8,4618E-01
Moacir Gracindo Soares Palmeira	211	8,3974E-01
Roberto Cardoso de Oliveira	225	8,3835E-01
Jose Sebastiao Witter	256	8,3526E-01
Jose Augusto Guilhon Albuquerque	266	8,3395E-01
Jose Sergio Leite Lopes	335	8,2564E-01
Francisco Correa Weffort	351	8,2378E-01
Leonidas Hegenberg	356	8,2306E-01

Ciências Sociais Aplicadas		
Nome	Rank	PRI
Fatima Cristina Trindade Bacellar	1	8,8396E-01
Luiz Alberto Machado	114	8,5332E-01
Ivan Guérios Curi	187	8,4260E-01
Joao Luiz Maurity Saboia	250	8,3554E-01
Dalmo de Abreu Dallari	260	8,3514E-01
Elza Salvatori Berquó	272	8,3315E-01
Maria da Conceicao de Almeida Tavares	311	8,2834E-01
Lucrecia D'alessio Ferrara	358	8,2285E-01
José Manoel de Arruda Alvim Netto	360	8,2252E-01
Thereza Celina Diniz de Arruda Alvim	383	8,1989E-01

Essa afirmação não pode ser feita quando analisamos os pesquisadores com maiores PageRanks locais de cada área para $k = 5$. A diferença dentro de cada área teve uma grande aumento, assim como a diferença entre as áreas. Essa diferença se expressa tanto pela variação dos valores de PageRank, quanto para a variação de ranks. Isso mostra a vantagem obtida por pesquisadores e áreas mais antigas quando um número maior de gerações é considerada.

Por fim, inicialmente havíamos previsto que conforme o número de gerações (k) aumentasse, os valores de PageRank local se assemelhariam cada vez mais com o PageRank global. Como pode-se ver, entretanto, os melhores pesquisadores para k igual a dois estão mais bem colocados no ranking global que os melhores pesquisadores para k igual a cinco. Esse fenômeno ainda deve ser explorado com cautela para que suas causas/relações sejam descobertas.

8.3 Comparação com Bolsistas de Produtividade em pesquisa do CNPq

Mesmo que o PageRank invertido local tenha oferecido resultados promissores, ainda é necessário que esses sejam validados. A análise qualitativa como foi feita para o PageRank global é dificultada pois nem todos os pesquisadores de destaque local são reconhecidos em seus ramos, consequência também de suas carreiras acadêmicas mais novas.

Para realizar uma análise quantitativa dos resultados decidiu-se utilizar a lista de pesquisadores com bolsas de produtividade em pesquisa oferecidas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). A lista conta com o nome de aproximadamente 14 mil pesquisadores de todas as áreas do conhecimento que recebem bolsas de produtividade de nível 2, 1(A-D) e SR (sênior). Todos os bolsistas analisados pertencem ao grafo de pesquisadores da Plataforma Lattes e, portanto, puderam ser localizados pelos seus respectivos códigos identificadores (ID Lattes).

Acreditamos que, por ser uma avaliação humana (partindo do CNPq), a lista de pesquisadores representa um conjunto (apesar de pequeno) dos pesquisadores mais influentes na comunidade e que, portanto, devem ser identificados pelas nossas métricas.

Assim, em uma primeira análise, identificou-se a posição de todos os bolsistas no ranking de todos os PageRanks. A Figura 9 mostra os a quantidade acumulada de bolsista para cada parcela do rank (sendo 0 o rank 1 e 1,0 o maior rank) para cada uma das métricas.

A figura mostra que para o PageRank global (em azul) a quantidade de bolsistas segue uma linha quase diagonal, indicando que conforme o rank aumenta, o número de bolsistas também aumenta de forma proporcional. Assim, o PageRank global não exalta os bolsistas, de modo que apenas uma pequena quantidade está realmente nas primeiras posições do rank.

Para os PageRanks locais (em laranja), no entanto, vemos que uma grande quantidade de bolsistas se encontra nas primeiras posições do rank, estando quase todos eles (aproximadamente 12 mil) em posições menores que 60% do total, e mais da metade deles nos primeiros 20%.

Além disso, a linha verde mostra a quantidade acumulativa de bolsistas pela sua posição

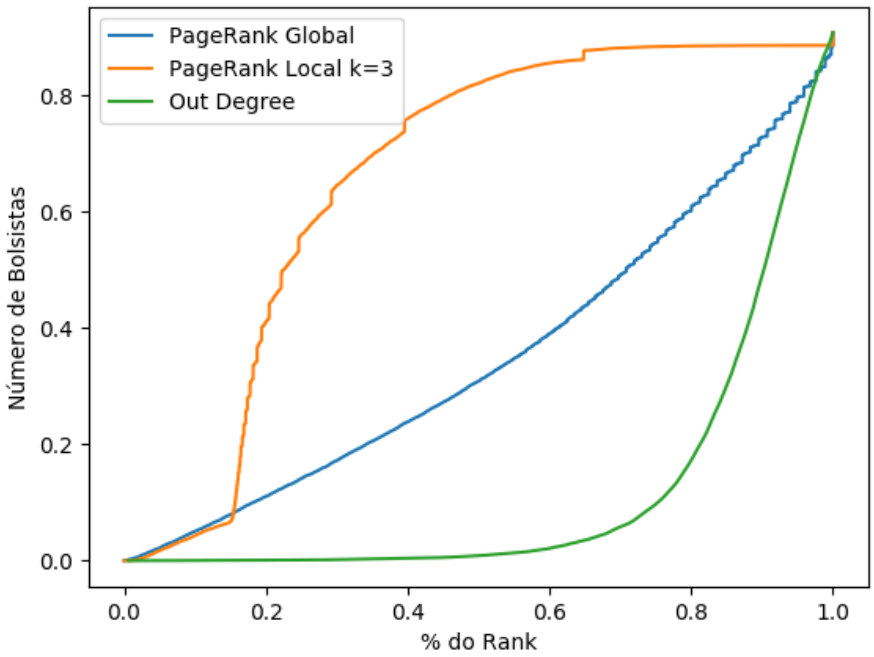


Figura 9: *Gráfico da quantidade acumulada de bolsistas por parcela do rank para as métricas de PageRanks invertido (em azul) e locais para k=3 (em laranja), e grau de saída (em verde).*

no rank de grau de saída (indicando o número de alunos orientados). A partir dessa linha, vemos que os pesquisadores são prejudicados, dado que a maioria se encontra nos últimos 80% do rank. Isso mostra que essa métrica é inviável para a exaltação de pesquisadores relevantes, e também explora o fato de que bolsistas de produtividade costumam ter poucos alunos.

Conclui-se, desse modo, que para a identificação de pesquisadores relevantes o PageRank local é de fato mais apropriado, de forma que exalta pesquisadores já renomados pela comunidade. Esta evidência será melhor estudada nos projetos que serão considerados como continuação deste projeto de Iniciação Científica.

Existe um fenômeno não explorado para ranks entre 0% e 20% do PageRank local, nos quais o número de pesquisadores cresce de forma linear (semelhante ao PageRank global), antes de sofrer um aumento muito grande. Também nota-se que as curvas de PageRank local não são exatamente suaves, havendo certos degraus ao longo de seu comprimento. Acredita-se que esses fatos estão relacionados à quantidade de PageRanks semelhantes para grafos morfologicamente similares, como já discutido na Seção 8.2. Considera-se analisar com maior profundidade essa característica em estudos futuros.

A Figura 10 mostra os resultados obtidos para o número acumulativo de bolsistas por

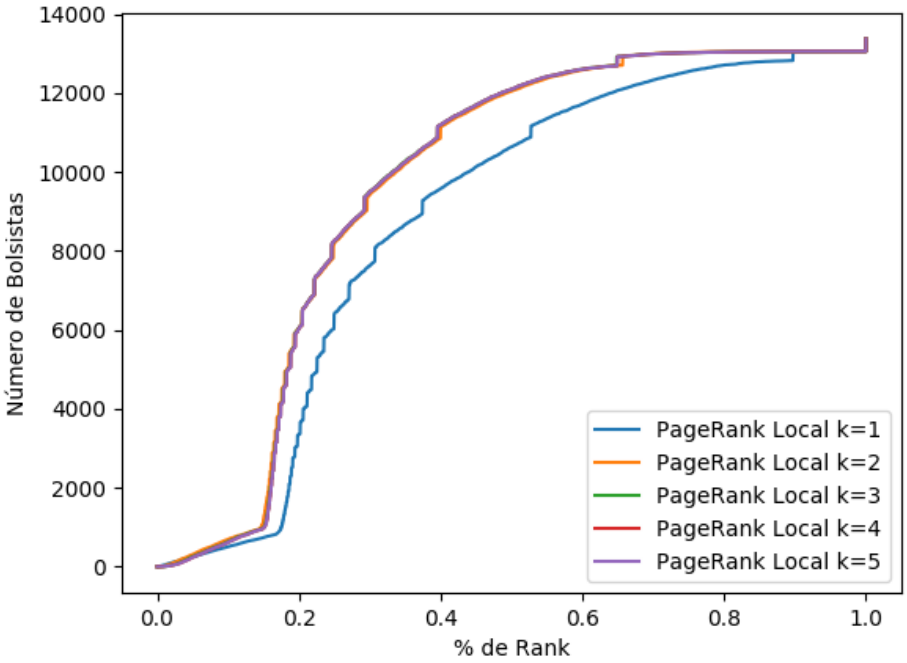


Figura 10: *Gráfico da quantidade acumulada de bolsistas por parcela do rank para as métricas de PageRanks locais de k=1 a k=5.*

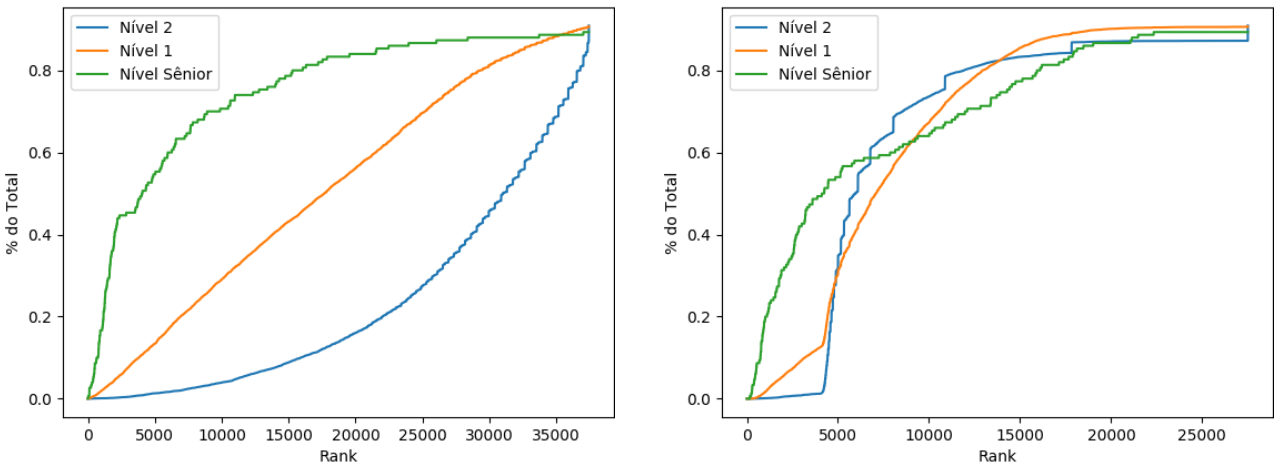
rank de PageRank Local para valores de k de 1 a 5. Vemos que os resultados para valores de k acima de 1 são muito parecidos, sendo quase indistinguíveis para valores a partir de 3. Isso pode ser uma indicação de que o valor ideal de k para ser utilizado é 3, dado que o rank dos pesquisadores parece convergir a partir desse ponto.

Comparação entre bolsistas de produtividade do CNPq de diferentes níveis

Além da análise feita com os bolsistas de todos os níveis sem distinção, também foi realizada uma análise comparando bolsistas dos diferentes níveis de bolsa existentes. Foram utilizados para a comparação os níveis 2, 1 (englobando bolsistas de 1A à 1D) e Sênior.

Para esta análise, acreditávamos que os bolsistas que mais seriam exaltados seriam aqueles de nível Sênior, dado que são pesquisadores com uma carreira acadêmica mais antiga e que já acumularam méritos durante sua trajetória. Também esperávamos que os bolsistas de nível 2 seriam os menos exaltados, sendo esses pesquisadores mais jovens e no início de sua carreira.

De fato, como vemos na Figura 11(a), a qual apresenta a quantidade acumulada de bolsistas (em porcentagem) pelo seu rank para o PageRank Global, os bolsistas Sênior se (repre-



(a) Global

(b) Local

Figura 11: *Gráfico da quantidade acumulada de bolsistas por parcela do rank para bolsistas de nível 2, 1 e Senior para o PageRank local com k=3 e PageRank global.*

sentados pela linha verde) foram os que mais se destacaram, estando quase 60% deles entre os 5000 primeiros ranks. Além disso, bolsistas de nível 2 (linha azul) são prejudicados, tendo uma maior ocorrência nos últimos ranks. Os bolsistas de nível 1 (em laranja) se mantiveram em uma reta diagonal, mostrando que estão bem distribuídos pelo rank.

Esse resultado é esperado para o PageRank Global, uma vez que esse, como discutido anteriormente, tende a exaltar pesquisadores que estão no topo do grafo de genealogia, em detrimento daqueles que estão mais próximos à base.

Já para o PageRank Local (Figura 11(b)), vemos que a situação não é mais a mesma: enquanto os pesquisadores de nível Sênior ainda se destacam entre os melhores ranks, o número de pesquisadores nestas posições é menor, sendo representado por um crescimento menor da linha que os representa. Por outro lado, ao utilizarmos essa métrica pesquisadores de nível 1 e 2 passaram a ser exaltados, mostrando que são relevantes de maneira local (por serem mais jovens), e atingindo uma porcentagem maior que pesquisadores Sênior para ranks entre 5000 e 15000.

Esses resultados comprovam que o PageRank Local pode identificar pesquisadores relevantes de forma apropriada, de modo que tanto pesquisadores mais sêniors quanto mais jovens são exaltados de foram similar. Além disso, vê-se que a métrica se assemelha à uma avaliação humana, identificando os pesquisadores que também foram renomados pela comunidade por

um método mais qualitativo. Temos aqui uma grande oportunidade de desenvolvimento de uma forma alternativa para avaliação e identificação de pesquisadores importantes em termos de formação de recursos humanos.

9 Cronograma de atividades

Inicialmente, este projeto de Iniciação Cientifica foi dividido em oito atividades (linhas 1-7, 10 da Tabela 10) a serem realizadas durante um ano.

O cronograma (Tabela 10) mostra as atividades previstas na proposta de iniciação científica e os períodos em que elas deveriam ser realizadas em vermelho (marcados por P). Em verde escuro estão indicados os meses nos quais as atividades foram de fato cumpridas desde o início do projeto. Apesar de termos seguido as atividades descritas na tabela, o decorrer do projeto se deu de maneira mais rápida que o planejado e, portanto, foi possível adiantar algumas das atividades.

O desenvolvimento do algoritmo do PageRank, em especial, ocorreu de maneira acelerada, o que permitiu o adiantamento do desenvolvimento de métricas auxiliares (atividade 4) e do algoritmo de extração de sub-grafos (atividade 5), e da análise de resultados (atividade 7).

Em virtude disso, houve um avanço considerável do cronograma, o que permitiu que novas atividades fossem incluídas para serem realizadas futuramente. Portanto, desenvolvemos um novo cronograma para ditar os próximos passos que serão realizados no restante do tempo.

As atividades 2 (estudo de leitura e representação de grafos) e 7 (análise de resultados) tiveram uma grande duração, dado que sempre buscando novas formas de representar os grafos analisados, e que neste projeto nos concentramos em analisar e comparar diferentes configurações do algoritmos PageRank (global e local), assim como com outras métricas.

Como pode-se ver, o cronograma foi satisfatoriamente cumprido e todos as atividades foram executadas como previsto, apesar de com algumas alterações em suas quantidades e tempo. Por fim, ainda pôde-se realizar novas atividades que não haviam sido previstas no cronograma inicial, o que adicionou novos horizontes à pesquisa.

Tabela 10: *Cronograma de atividades de trabalho. Os meses estão representados por números de 2 algarismos. As células marcadas em vermelho correspondem ao período planejado para a realização das atividades na proposta do projeto. Aquelas marcadas em verde escuro representam o momento em que as atividades foram de fato executadas.*

Atividade		Quadrimestre 1				Quadrimestre 2				Quadrimestre 3			
		01	02	03	04	05	06	07	08	09	10	11	12
1. Estudo teórico sobre o uso do PageRank	P												
	R												
2. Estudo sobre leitura e representação de grafos	P												
	R												
3. Desenvolvimento do algoritmo de cálculo de PageRank	P												
	R												
4. Desenvolvimento de métricas auxiliares	P												
	R												
5. Desenvolvimento do algoritmo de extração de sub-grafos	P												
	R												
6. Aplicação do PageRank aos sub-grafos	P												
	R												
7. Análise de resultados e comparação com outras métricas	P												
	R												
8. Desenvolvimento de variações do PageRank	P												
	R												
9. Desenvolvimento de métodos de análise e validação dos resultados	P												
	R												
10. Escrita de relatórios	P												
	R												

10 Considerações finais

Um pesquisador pode ser avaliado por diferentes fatores, dentre eles, pelo número e impacto de suas publicações, por sua relevância no cenário político-social no qual está inserido, pelo número de premiações e homenagens recebidas e, finalmente, pelas suas contribuições para o desenvolvimento da comunidade científica como um todo. A medida de PageRank, descrita neste projeto, permitiu obter o impacto local de um pesquisador na sua descendência, o qual pode ser um indicador de sua relevância no meio científico em que está inserido.

Ao efetuar alterações ao algoritmo, considerando apenas uma comunidade local influenciada diretamente por um pesquisador e normalizando os resultados, pôde-se obter uma métrica mais justa e condizente com outras medidas de relevância acadêmica.

A métrica obtida permitiu identificar corretamente pesquisadores antigos e jovens de forma similar, e sua efetividade pôde ser verificada ao ser utilizada para rankear bolsistas de produtividade em pesquisa do CNPq.

Definiu-se, assim, que a métrica ideal para identificar pesquisadores relevantes em grafos

de genealogia acadêmica é o PageRank Local Normalizado, quando são consideradas três gerações de descendência a partir de um pesquisador.

Ademais, a fim de aprimorar a métrica desenvolvida para que a relação de orientação seja melhor modelada, alguns fatores podem ser considerados em seu cálculo, como o grau acadêmico da orientação, seu tipo (orientação ou co-orientação) e o tempo de conclusão. Similarmente, para contabilizar pelos recursos humanos na formação de um pesquisador, outros indicadores de excelência acadêmica não baseados em grafos (e.g., número e relevância de publicações, premiações e bolsas de produtividade) podem também auxiliar na avaliação e identificação dos pesquisadores.

Acreditamos que o projeto permitiu gerar uma nova frente de análise quando comparado a outros já utilizados para a avaliação de pesquisadores, como o índice genealógico (Rossi *et al.* , 2017). Além de proporcionar uma forma de quantificar a relevância de um pesquisador, pode também ser usado de forma a complementar estudos mais aprofundados sobre avaliação de impactos em Ciência, Tecnologia e Inovação.

Referências Bibliográficas

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

Brin, S., & Page, L. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, **56**(18), 3825–3833.

Cronin, B., & Sugimoto, C. R. 2014. *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*. MIT Press.

Damaceno, R., Rossi, L., & Mena-Chalco, J. P. 2017. Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes. 10.

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

Elias, M. C., Floeter-Winter, L. M., & Mena-Chalco, J. P. 2016. The dynamics of Brazilian protozoology over the past century. *Memórias do Instituto Oswaldo Cruz*, **111**(1), 67–74.

Gargiulo, F., Caen, A., Lambiotte, R., & Carletti, T. 2016. The classical origin of modern mathematics. *EPJ Data Science*.

Gleich, D. F. 2015. PageRank beyond the Web. *SIAM Review*, **57**(3), 321–363.

Hey, T., Tansley, S., & Tolle, K. M. 2009. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. 2015. The Leiden Manifesto for research metrics. *Nature*, **520**(7548), 429.

Isaac, S., & Michael, W. B. 1971. Handbook in research and evaluation.

Langville, A. N., & Meyer, C. D. 2011. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

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

Page, L., Brin, S., Motwani, R., & Winograd, T. 1999. *The PageRank citation ranking: Bringing order to the web.* Tech. rept. Stanford InfoLab.

Patton, M. Q. 2005. *Qualitative research*. Wiley Online Library.

Rossi, L. 2015. *Caracterização de grafos de genealogia acadêmica por meio de métricas topológicas*. M.Phil. thesis, Universidade Federal do ABC, Brazil.

Rossi, L., & Mena-Chalco, J. P. 2014. Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. *Pages 1–12 of: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

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

Sugimoto, C. R. 2014. Academic Genealogy. *Pages 365–382 of:* Cronin, B, & Sugimoto, C R (eds), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, first edn. MIT Press.

Tenn, J. S. 2016. Introducing AstroGen: the Astronomy Genealogy Project. *arXiv preprint arXiv:1612.08908*.

Tol, R. S. J. 2013. Identifying excellent researchers: A new approach. *Journal of Informetrics*, **7**(4), 803–810.