Prestígio em grafos de genealogia acadêmica: Uma proposta baseada em PageRank

Arthur V. Kamienski[1], **Rafael J. P. Damaceno**[1], **Jesús P. Mena-Chalco**[1]

[1]Centro de Matemática, Computação e Cognição - Universidade Federal do ABC (UFABC)
Av. dos Estados 5001, Santo André, SP 09210-580, Brasil

Abstract. *Identifying prestigious researchers in their fields of research is an arduous task. The greatest difficulty is the evaluation of the prestige of each researcher on the local structure that surrounds them. In this work we present a new adaptation of the PageRank algorithm on an academic genealogical graph, called Normalized Local Inverted PageRank, in means of identifying the most prestigious researchers in terms of human resources production constrained by the number of academic generations. As a case study, we use data gathered from more than a million of researchers.*

Resumo. *Identificar pesquisadores com grande prestígio nas suas áreas acadêmicas é uma tarefa árdua. A maior dificuldade é a avaliação do prestígio de cada pesquisador na estrutura que o envolve. Neste trabalho apresentamos uma nova adaptação do algoritmo PageRank para grafos de genealogia, denominado PageRank Invertido Local Normalizado, de modo a identificar pesquisadores de maior prestígio em termos de formação de recursos humanos limitados pelo número de gerações de acadêmicos. Como estudo de caso usamos a genealogia obtida a partir de mais de um milhão de acadêmicos.*

1. Introdução

Um pesquisador pode ser caracterizado por indicadores que vão além de sua produção científica. Métodos clássicos para identificar seu prestígio, como a análise de suas premiações e avaliação por meio de pares, são comumente qualitativos e de difícil implementação, considerando que carregam consigo o juízo de quem o aplica. Assim, são suscetíveis a fatores pessoais que podem comprometer o resultado [Isaac and Michael 1971].

Ao impacto (ou prestígio) de um pesquisador também podem ser associadas medidas derivadas de sua produtividade bibliométrica, tais como o número de artigos publicados e de citações recebidas [Garner et al. 2018, Tol 2013]. Mais recentemente, outro aspecto inerente à produção científica despertou interesse da comunidade: a formação de recursos humanos. Trata-se de uma alternativa, de base quantitativa, aos métodos existentes para tornar o processo de avaliação mais assertivo e que considera a propagação do conhecimento científico a partir dos relacionamentos de orientação entre professor e aluno. Tendo em vista a relevância de se produzir avaliações sob múltiplos olhares, neste trabalho apresentamos uma nova adaptação do algoritmo PageRank para análise de importância para o contexto de redes de genealogia acadêmica[1], avaliando dessa forma, o prestígio ou importância de pesquisadores.

[1]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.

2. Trabalhos correlatos

Desde o seu surgimento como ferramenta para avaliar a importância de páginas web, o PageRank despertou interesse da comunidade para a avaliação em redes de citação. No trabalho de [Ma et al. 2008], esse algoritmo é comparado com técnicas de contagem de citações em mais de 230 mil artigos científicos da área de Bioquímica e Biologia Molecular. [Fiala and Tutoky 2017] utilizaram cerca de dois milhões de artigos científicos da área de Ciência da Computação para comparar o PageRank a medidas baseadas em citação, com o intuito de destacar os cientistas que receberam os Prêmios Turing e Codd.

Já a partir do olhar da genealogia acadêmica, busca-se extrair conhecimento por meio de métricas que caracterizem as relações de orientação entre professor e aluno. No trabalho de [David and 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. 2018] apresentaram medidas topológicas específicas para a genealogia acadêmica.

Diferentemente desses estudos, neste trabalho apresentamos uma nova adequação do algoritmo PageRank para atribuir a pesquisadores um indicador vinculado às suas características na rede de relações de orientação que compõem. Busca-se explorar as diferentes configurações do algoritmo para identificar pessoas proeminentes, o que pode prover informações complementares às bibliométricas já existentes.

3. Adaptação do algoritmo PageRank para grafos de genealogia

Um grafo de genealogia acadêmica $G = (V, E)$ é um conjunto de vértices (V) e arestas (E) em que cada vértice representa um pesquisador, e cada aresta uma relação de orientação acadêmica. Como a estrutura do grafo se assemelha a uma floresta na qual existem poucos vértices que agem como raízes, o algoritmo de PageRank obtido para estes grafos resultará em valores altos para orientados e baixos para orientadores.

Para obter um resultado onde os orientadores sejam exaltados, o sentido de fluxo de PageRank deve ser invertido. Com o PageRank Invertido (PRI), pudemos encontrar uma ordenação de pesquisadores na qual indivíduos são recompensados por realizarem orientações ao invés de serem orientados. A fórmula para o cálculo do PRI é dada por

$$PRI(v) = \frac{1-d}{n} + d \times \sum_{v \in V} \frac{PRI(v)}{S(v)} \tag{1}$$

em que v representa um pesquisador no conjunto V de pesquisadores, d é um número real entre 0 e 1, n é o número de vértices, e $S(v)$ é o grau de entrada do vértice v.

Para exemplificar as diferenças entre Pageranks consideramos um grafo composto de 11 vértices (Fig. 1). O eixo horizontal representa a hierarquia na orientação, com a orientação emergindo do orientador e incidindo no orientado. Os valores de PageRank mais altos estão representados em cor mais escura. Para o PageRank original (Fig. 1a), os vértices que mais se destacaram são aqueles encontrados na folhas da árvore. Já os resultados do PRI (Fig. 1b) exaltam vértices com muitos descendentes.

Ainda que o PRI (Eq. 1) permita destacar pesquisadores mais antigos na hierarquia, existem distorções se compararmos um pesquisador mais antigo com um mais

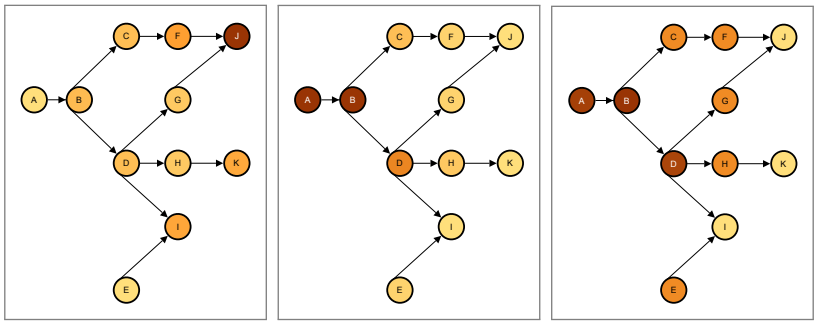

Figura 1. Diferenças entre as medidas de: (a) PageRank, (b) PRI e (c) PRILN (k=2).

jovem, visto que os descendentes do primeiro são em maior número. Pesquisadores para os quais o número de descendentes é grande apenas para gerações distantes podem não ter grande influência na formação destas últimas gerações. Assim, restringimos o grafo utilizado para o cálculo do PageRank, ao gerar um sub-grafo contendo apenas o pesquisador em questão (i.e., raiz do sub-grafo) e um número k de suas gerações descendentes.

O valor do PageRank Invertido Local (PRIL) de um pesquisador corresponde ao seu PRI dentro do sub-grafo extraído a partir dele. Essa adaptação permite que analisemos um pesquisador para um grupo de orientações arbitrariamente próximo, desconsiderando a influência de vértices que estão fora do sub-grafo.

Por outro lado, como o valor da soma de todos os PageRanks de um grafo é igual a 1, quanto maior o número de vértices menor o PageRank atribuído para cada vértice. Portanto, tomando dois pesquisadores com sub-grafos de tamanhos muito distintos, aquele com maior sub-grafo terá um valor de PageRank menor que o outro.

Para atingir uma métrica mais justa, aplicamos uma normalização aos valores de PRIL obtidos para subgrafos de tamanhos distintos, mantendo o ranking de sub-grafos de tamanho igual. Primeiro, inverte-se o valor do PRIL de um vértice escolhido como o representante para um tamanho de sub-grafo. O valor do PageRank se tornará, dessa forma, maior para sub-grafos maiores. O ranking dos demais vértices é calculado com base no valor obtido para o representante do tamanho de seu sub-grafo. A fórmula resultante para o cálculo do PageRank Invertido Local Normalizado (PRILN) é

$$PRILN(v) = \frac{PRIL(v)}{PRIL(w)} \times (1 - PRIL(w)) = \frac{PRIL(v)}{PRIL(w)} - PRIL(v) \qquad (2)$$

em que v e w tem o mesmo tamanho de sub-grafo e w é o representante deste conjunto de vértices. Neste trabalho consideramos como representante o vértice com maior valor de PRIL para cada tamanho de sub-grafo.

A Figura 1c apresenta um exemplo do cálculo desta nova adaptação que coloca em evidência o vértice B como o de maior prestígio do grafo. A métrica é mais justa que o PRI, onde A é o vértice de maior prestígio apenas por ser uma raiz da árvore (ver Fig. 1b). Finalmente, note que o PRILN se mostra melhor distribuído pelo grafo, não estando centralizado nos vértices mais altos da árvore.

4. Resultados

A fim de testar o funcionamento do algoritmo em um caso real e obter resultados que podem ser validados, dados foram coletados da Plataforma Acácia[2]. O grafo de genealogia acadêmica disponibilizado por esta plataforma é formado por $1\,111\,544$ de pesquisadores (mestre e doutores) e $1\,208\,398$ de orientações formais concluídas.

Para cada vértice do grafo, diversos sub-grafos foram gerados considerando de uma a cinco gerações a partir do vértice de interesse. Os sub-grafos ainda podem ter menos que k gerações caso o pesquisador não tenha esse número de gerações abaixo dele. Optamos por essa abordagem para não excluir pesquisadores muito jovens da avaliação.

Pesquisadores com subgrafos triviais (tamanhos 1 e 2, para os quais só existe uma morfologia possível) apresentam valores comuns de PRILN e representam juntos 92% dos casos para o grafo utilizado. Por esses serem os pesquisadores com menor PRILN (i.e., menor prestígio), desconsideramos esses casos para as análises aqui descritas.

É importante ressaltar que os maiores sub-grafos (por quantidade de vértices) não estão relacionados aos maiores PageRanks locais, como comprovado pelo coeficiente de correlação (Pearson) obtido para as duas métricas, que se encontra na faixa de $0,6$ (para $k = 1$) a $0,2$ (para $k = 5$). Isso também evidencia a importância da morfologia do subgrafo para o cálculo do PageRank e a sua diferença para outras métricas.

Os valores de PRILN estão bem distribuídos em uma faixa $0,4$ a $0,8$. Diferentemente do PRI, o PRILN não se comporta de acordo com uma lei da potência, apesar de existirem valores com alta incidência, os quais estão ligados aos pesquisadores com poucos ou nenhum descendente. Além disso, a média encontrada para os valores de PRILN permanece ao redor de $0,45$ para todos os valores de k testados.

A análise mostrou que não existe uma forte correlação entre PRI e PRILN. Para todos os valores de k testados a correlação se manteve entre $0,2$ e $0,4$. Quando apenas os 100 pesquisadores com maiores PRILNs são considerados, esses valores decrescem para em torno de $0,3$ para $k = 1$ e $k = 2$, e permanecem em uma faixa de $0,0$ à $0,1$ para ks maiores. Isso implica na existência de pessoas com PRI baixo que se destacam no PRILN, mostrando uma provável capacidade de identificar pesquisadores jovens. No entanto, pesquisadores que se destacam com o PRI continuam se destacando com o PRILN. Dos 1000 pesquisadores com maior PRI, 25% também estão presentes nos 1000 maiores PRILNs. Quando são considerados os $10,000$ maiores, esse número sobe para 50%.

Ao compararmos o PRILN para diferentes valores de k, notamos grande variação da métrica para $k = 1$ e $k = 2$. A primeira, apesar de melhor distribuída no intervalo $[0,40; 0,55]$, não exalta os pesquisadores com PRILN mais alto como as outras, estando o terceiro e quarto quartis deslocados para valores mais altos. Já a segunda mostra uma distribuição mais centrada em valores médios de PRILN, e apresenta um quarto quartil que se espalha por um intervalo maior de valores, havendo *outliers* com PRILNs altos. A correlação encontrada entre as duas métricas foi de $0,79$.

As outras métricas apresentam pouca variação entre si. A partir de $k = 3$, a variação passa a ser mínima, havendo uma correlação maior que $0,95$ com ks maiores (desconsiderando subgrafos triviais). Entretanto, notou-se que conforme aumentamos o

[2] `http://plataforma-acacia.org`, último acesso em 12 de Maio de 2019.

Tabela 1. Lista dos 10 pesquisadores com os maiores valores para PRI e PRILN.

Ranking	PRI	PRILN ($k = 1$)	PRILN ($k = 2$)	PRILN ($k = 3$)
1	J Martins	FAP Fialho	EA Vieira	FG Brieger
2	D Saviani	PB Carvalho	AJ Naro	B Pottier
3	CM Bori	CAN Cosenza	A Catelli	M Reale
4	ED Franca	NFF Ebecken	CM Bori	JC Boucinhas
5	C Pavan	MLS Braga	ED Franca	L Michel
6	CT Pais	MH Diniz	J Martins	A Ginsberg
7	M Moisés	L Landau	AR Santanna	A Rothe
8	LD Ferrara	IM Beuren	MA Matos	VK Handa
9	GP Witter	GP Witter	D Saviani	RV Silva
10	A Dreyfus	ÁGR Lezana	O Nakano	CPF Camargo

número de gerações dos subgrafos, pesquisadores mais antigos começaram a assumir PageRanks locais mais altos. 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 de k usados, o PRILN evidencia de forma semelhante os mesmos pesquisadores. Deste modo, vemos que um valor de k muito alto (próximo ao número máximo de gerações do grafo) não proporciona todos os resultados desejados, dado que continua a evidenciar pesquisadores que são de prestígio apenas por sua senioriedade. Considera-se assim utilizar preferencialmente um valor de k de até 3.

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. Para aprofundar essa ideia, apresentamos na Tabela 1 os 10 pesquisadores ranqueados com os maiores valores para PRI e PRILN ao serem consideradas vizinhanças correspondentes a k de 1 à 3.

Observa-se que a aplicação do PageRank segundo diferentes ks permitiu ranqueamentos distintos. Para $k = 3$, por exemplo, nenhum dos pesquisadores listados consta na avaliação por PRI, indicada na duas primeiras colunas da Tabela 1. Esse fenômeno também evidencia a capacidade do PRILN de destacar pesquisadores mais jovens. A média e mediana das idades acadêmicas[3] dos 10 pesquisadores mais importantes segundo o PRI são 57 e 53 respectivamente. Esses valores são reduzidos para 37 e $41,5$, respectivamente, ao serem analisados os 10 pesquisadores a partir do PRILN com $k = 1$, para $48,1$ e $49,5$ com $k = 2$, e para $51,9$ e 51 com $k = 3$.

A Figura 2 ilustra o comportamento dos valores de PRI e PRILN (para ks de 1 à 5) com relação à idade acadêmica dos $1\,111\,544$ pesquisadores. As curvas obtidas pelo PRILN apresentam um comportamento distinto daquela obtida pelo PRI. Cabe destacar o prestígio obtido por acadêmicos com 25 ou menos anos de idade acadêmica a partir do PRILN; acadêmicos esses que obtiveram importância quase nula utilizando-se o PRI para cálculo. Ainda, os pesquisadores com os melhores ranqueamentos situam-se em torno de 50 e 60 anos para $k = 2$, ao passo que para o PRI esse valor é de 75 anos.

[3] A idade acadêmica pode ser definida como a diferença entre o ano corrente (ou de obtenção dos dados) e o ano de titulação. Por exemplo, um pesquisador formado em 2008 teria idade acadêmica 9 ($2017 - 2008$).

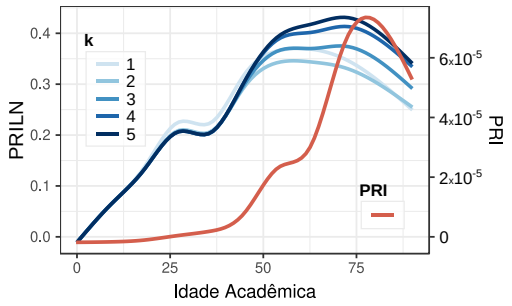

Figura 2. Comparação entre os valores de PageRanks.

5. Considerações finais

A medida PRILN aqui tratada permite obter insumos sobre o impacto local de um pesquisador na sua descendência, o qual pode ser um indicador de sua relevância no meio científico. Ao efetuar alterações ao algoritmo PageRank, considerando uma comunidade local influenciada por um pesquisador e normalizando seus valores, obtemos uma métrica que representa seu prestígio acadêmico.

Como desdobramentos futuros a fim de aprimorar a medida desenvolvida, outros fatores podem ser considerados em seu cálculo, como o grau acadêmico da orientação e seu tipo (orientação ou co-orientação). Similarmente, para contabilizar pelos recursos humanos na formação de um pesquisador, outras medidas de excelência acadêmica não baseadas em grafos (e.g., número e relevância de publicações, premiações) também podem auxiliar na avaliação e identificação dos pesquisadores de prestígio.

Referências

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

Fiala, D. and Tutoky, G. (2017). Pagerank-based prediction of award-winning researchers and the impact of citations. *Journal of Informetrics*, 11(4):1044–1068.

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

Garner, R. M., Hirsch, J. A., Albuquerque, F. C., and Fargen, K. M. (2018). Bibliometric indices: defining academic productivity and citation rates of researchers, departments and journals. *Journal of Neurointerventional Surgery*, 10(2):102–106.

Isaac, S. and Michael, W. B. (1971). *Handbook in research and evaluation - For Education and the Behavioral Sciences*. Robert R. Knapp, first edition.

Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810.

Rossi, L., Damaceno, R. J., Freire, I. L., Bechara, E. J., and Mena-Chalco, J. P. (2018). Topological metrics in academic genealogy graphs. *Journal of Informetrics*, 12(4):1042–1058.

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