Estudo comparativo de métricas de ranking em Redes Sociais

Samuel O. Silva[1], **Bruno O. Goulart**[2], **Maria Júlia M. Schettini**[3], **Carolina Ribeiro Xavier** [4], **João Gabriel Rocha Silva**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT)
Rodovia MT-473, s/n, 78250-000, Pontes e Lacerda - MT - Brasil

[2]Universidade do Estado de Mato Grosso (UNEMAT)
Av. Tancredo Neves, 1095 - Cavalhada II, Cáceres - MT - Brasil

[3]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Rua José Peres, 558 - Centro, Leopoldina - MG - Brasil

[4]Universidade Federal de São João del Rei (UFSJ)
Rodovia BR-494, Vila São Paulo - CEP: 36301-360 - São João del-Rei/MG - Brasil

joaogabriel.comp@gmail.com

Abstract. *The use of modeling and application of complex networks in several areas of knowledge have become an important tool for understanding different phenomena; among them some related to the structures and dissemination of information on social medias. In this sense, the use of a network's vertex ranking can be applied in the detection of influential nodes and possible foci of information diffusion. However, calculating the position of the vertices in some of these rankings may require a high computational cost. This paper presents a comparative study between six ranking metrics applied in different social medias. This comparison is made using the rank correlation coefficients. In addition, a study is presented on the computational time spent by each ranking. Results show that the Grau ranking metric has a greater correlation with other metrics and has low computational cost in its execution, making it an efficient indication in detecting influential nodes when there is a short term for the development of this activity.*

Resumo. *A utilização da modelagem e aplicação de redes complexas em diversas áreas do conhecimento tem se tornado uma importante ferramenta para compreensão de diferentes fenômenos, dentre eles, relacionados às estruturas e disseminação de informações em redes sociais. Neste sentido, a utilização de rankings de vértices de uma rede podem ser aplicados na detecção de nós influentes e possíveis focos de difusão de informação. Entretanto, o cálculo de posição dos vértices em alguns desses rankings podem exigir um alto custo computacional. Neste trabalho é apresentado um estudo comparativo entre seis métricas de rankings aplicados em diversas redes sociais. Esta comparação é realizada a partir da utilização de coeficientes estatísticos de correlação de postos. Além disso, é apresentado um estudo sobre o tempo computacional gasto por cada ranking. Resultados apontam que a métrica de ranking Grau apresenta uma maior correlação para com as outras métricas e possui baixo custo computacional em sua execução, tornando esta uma indicação eficiente*

na detecção de nós influentes quando dispõe-se de curto prazo para o desenvolvimento desta atividade.

1. Introdução

O conceito de uma rede pode representar diferentes sistemas do mundo real, para isso é necessário apenas a definição de elementos que simbolizem os vértices e as arestas deste sistema [Rubinov and Sporns 2010]. Normalmente, os indivíduos ou seres de estudo no sistema são modelados como os vértices - também conhecido como nós - e as conexões entre estes seres como as arestas. Assim, baseado na teoria de grafos surge o conceito de redes complexas, que podem ser definidos como grandes grafos.

Devido a sua fácil e eficiente aplicação encontram-se trabalhos utilizando os conceitos de redes complexas em diferentes áreas do conhecimento, como por exemplo: na biologia [Donner et al. 2019] e física estatística [Cimini et al. 2019].

Outra área do conhecimento onde o estudo das redes complexas podem ser aplicados são em pesquisas que envolvem redes sociais e suas dinâmicas. Neste sentido, torna-se possível o estudo de propagação de informação, usuários mais influentes e estudos de comunidades e grupos de usuários [Liben-Nowell and Kleinberg 2007].

No aspecto de detecção de usuários mais importantes/influentes de uma rede social, tem-se o conceito de ranking. Este pode ser definido como a classificação dos vértices da rede a partir de um determinado critério. Conhecer o vértice mais importante de uma rede social permite por exemplo, conhecer o usuário com maior probabilidade de propagar uma *fake news* bem como qual o usuário tem uma maior probabilidade de influenciar outros usuários [Mo and Deng 2019].

No âmbito de medidas de centralidade (influência) de vértices, encontram-se disponíveis na literatura, métricas que calculam *scores* de nós com o intuito de classificá-los [Silva et al. 2015]. Estes cálculos possuem diferentes estratégias e de acordo com seus conceitos, podem exigir alto tempo computacional em sua execução. Dentre outras, destacam-se: Grau [Albert and Barabási 2002], o PageRank [Page et al. 1999], Hub e Authority [Kleinberg 1999b, Kleinberg 1999a], Closeness [Stephenson and Zelen 1989] e o Betweenness [Freeman 1977].

Neste projeto é apresentado um estudo comparativo entre as métricas citadas, no que se refere a correlação de ranking (o quão próximo uma métrica classifica os nó de uma rede em relação a outra métrica) e, o estudo do tempo computacional na execução de cada uma delas. Para o estudo das correlações, foi aplicado os coeficientes estatísticos de correlação propostos por [Spearman 1910] e [Ghent 1963].

2. Aspectos Teóricos

2.1. Redes Complexas e redes sociais

Os estudos das redes complexas foram iniciados em meados de 1930, quando sociólogos utilizavam essas redes com a finalidade de estudar o comportamento da sociedade e a relação entre os indivíduos [Metz et al. 2007]. Uma rede complexa é, por muitas das vezes, analisada como a abstração matemática de um Grafo, a qual é usualmente caracterizada como um conjunto de elementos (nós) conectados entre si através de uma ligação (aresta) [Gabardo 2015].

As redes complexas podem ser classificadas como direcionadas e não direcionadas. As redes direcionadas ocorrem quando as arestas possuem uma direção ou caminho para serem percorridas entre os vértices. Já as não direcionadas dizem respeito às redes cujas arestas não possuem uma direção, sendo tratada como uma aresta bilateral [Gabardo 2015].

Outro conceito interessante é o de assortatividade (Assort.) , o qual tem a finalidade de medir e identificar a tendência de um vértice se conectar ao outros vértices semelhantes a ele [Barbosa et al.]. Esta medida segue o domínio [-1,1] onde quanto mais próximo de 1 a rede tem mais tendência dos vértices se conectarem a nós com características similares, -1 o contrário, e mais próximo de 0 uma rede sem tendências.

O conceito de rede social é caracterizado pelo relacionamento de algum grupo de pessoas entre si, como por exemplo, uma rede de amigos, pessoas com algum tipo de parentesco, pessoas vinculadas profissionalmente, ou então indivíduos ligados a uma rede social virtual [Machado and Tijiboy 2005]. Um exemplo de relação social é a disseminação de um vírus, a qual só ocorre quando os nós (indivíduos) tem algum contato social (arestas) com outro indivíduo

A ideia de rede social virtual se originou do desenvolvimento e expansão de ferramentas tecnológicas, aquelas criadas com intuito de expandir e melhorar a comunicação e o relacionamento entre pessoas. Devido essa ferramenta, é possível analisar de diferentes maneiras a sociedade como um todo, sendo uma dessas formas o estudo do vínculo e relação que um indivíduo tem com o outro. [Wasserman et al. 1994].

Dessa forma, a rede social virtual pode ser representada facilmente como uma rede complexa, sendo os usuários os nós dessa rede e as relações entre eles as arestas, assim como uma conta no Instagram representa um vértice e os seguidores suas arestas direcionadas.

2.2. Métricas de Ranking

Métricas de ranking definem a centralidade de um vértice na rede. Afim de então, definir qual é o nó mais influente para, por exemplo, em redes sociais, em focos de difusão de *fake news*.

A métrica **Grau** é definida como o número de conexões que cada nó possui, ou seja, com quantos outros vértices ele está conectado [Gabardo 2015]. Já a métrica **PageRank** leva em consideração a qualidade das arestas que o nó se relaciona, quanto mais arestas com nós influentes, maior será seu score [Page et al. 1999]. Esta métrica qualifica os nós de acordo com seu grau e a colocação dos vértices adjacentes a ele.

Hub classifica um vértice que dado a sua concentração de arestas, juntamente com as a classificação de autoridade do vértices da rede [Kleinberg 1999b, Kleinberg 1999a] relacionados a ele. **Authority** é calculada a partir da soma da quantidade de hubs com o nó que está conectado [Kleinberg 1999b]. **Closeness** calcula a média dos caminhos com menor distância daquele vértice para os outros vértices da rede [Stephenson and Zelen 1989]. **Betweenness** expressa a quantidade de vezes que um vértice faz o papel de ponte entre o menor caminho entre dois vértices [Freeman 1977].

2.3. Coeficientes de Correlação

Em estatística, coeficientes de correlação são medidas que mensuram a correlação entre variáveis e suas representações. Nesta técnica é possível identificar a relação e variabilidade entre conjuntos de informações.

Um coeficiente de correlação vastamente utilizado e aplicado foi apresentado em [Spearman 1910] e é descrito pela (Equação 1), onde, d_i descreve a diferença de posição entre elementos relacionados (variáveis de correlação) e n simboliza o número de pares de variáveis contidas nos conjuntos correlacionados.

$$\rho = 1 - \frac{6 \sum d_i^2}{(n^3 - n)} \tag{1}$$

Outro coeficiente de correlação proposto na literatura é o coeficiente de Kendall-Tau. Este, descrito em [Ghent 1963] utiliza em seu cálculo a quantidade de pares de variáveis em posições iguais e diferentes. O cálculo deste coeficiente é descrito pela Equação 2.

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)} \tag{2}$$

onde n_c expressa a quantidade de pares concordantes - variáveis que ocupam a mesma colocação; e n_d expressa a quantidade de pares discordantes - variáveis que ocupam posições distintas nos postos ocupados. Por fim, n representa a soma dos números de pares iguais (n_c) e número de pares distintos (n_d). Em ambos os coeficientes de correlação os valores calculados estão contidos no intervalo [-1,1] onde 1 é a maior correlação possível e -1 a menor.

3. Materiais e Métodos

Com objetivo de conhecer qual métrica de ranking tem maior correlação às outras, visto este fato poder contribuir para indicações de uma métrica que sintetiza características de diversas formas de ranking, uma série de testes foram realizados. Dando o início a fase de experimentação com intuito de aumentar a precisão dos resultados, oito bases de dados de redes sociais foram selecionadas, suas características são apresentadas na Tabela 1.

Nome	Vértices	Arestas	Assort.	Sobre
Epinions	75879	508837	-0.041	Relações de confiança
Wiki-Vote	8298	103689	-0.083	Votações dos usuários
Pretty G. Privacy	10680	24316	0.255	Relações sobre criptografia
Google+	23628	39242	-0.028	Relações de amizade
Slashdot0902	82168	948464	-0.051	Relações de confiança de notícias
Slashdot0811	77360	905468	-0.048	Relações de confiança de notícias
Email-EuAll	265214	420045	-0.210	Rede de e-mails de universitários
Twitter lists	23370	33101	-0.089	Relações de amizade

Tabela 1. Características das Bases utilizadas na pesquisa

As bases de dados selecionadas foram extraídas de sites de grupos de pesquisa das Universidades de Stanford e de Koblenz por serem bases utilizadas em outros trabalhos científicos e disponíveis de livre acesso na internet [Kunegis 2013], [Leskovec and Krevl 2016].

De modo subsequente, seis métricas de ranking foram selecionadas para análise devido às suas diferentes formas de cálculos de centralidade. Estas são: Grau, PageRank, Hub, Authority, Closeness e Betweenness. Para cada uma das oito bases de dados selecionadas os vértices foram ranqueados de acordo com cada uma das métricas.

No objetivo de verificar a correlação entre os rankings, para cada base de dados, cada um de seus seis rankings foram comparados par a par utilizando duas medidas de correlação de postos: Spearman e Kendall-Tau. A fim de se obter um padrão entre os resultados das correlações, utilizou-se a média entre os valores encontrados pelos coeficientes entre as redes sociais utilizadas para cada par de combinações de métricas, experimento realizado tanto para Spearman quanto para Kendall-Tau.

Por fim, com o intuito de análise de tempo na execução de cada uma das métricas, foi calculada a média de tempo em cada métrica de ranking entre todas as bases de dados, para proporcionar também, análise de custo computacional.

Os experimentos foram desenvolvidos na linguagem de programação Python com a utilização das bibliotecas Python-igraph, Time, Spearman e Nltk e executados em uma máquina Ryzem 7 2700x, 1x8 GB Ram Ddr4 3000 mhz, Hd 1tb 7200rpm.

4. Resultados e Discussão

Os resultados aqui apresentados (Figuras 1, 2 e 3) foram obtidos através de gráficos de calor gerado pelos valores dos coeficientes de correlação combinando as métricas de ranking par a par. A diagonal principal do gráfico representa a maior correlação pois a métrica é comparada com ela própria, obtendo valor máximo de correlação: 1.

Além disso, estes gráficos são divididos em duas regiões a partir da diagonal principal, onde a região triangular inferior apresenta os valores de correlação utilizando o coeficiente de Spearman e a região triangular superior o de Kendall-Tau.

A Figura 1 apresenta os dados da rede social com a maior assortatividade (0.25) - rede *Pretty Good Privacy*. Sobre a Figura 1 é possível notar que, de um modo geral, Grau é a métrica com correlações mais altas quando comparado a outras métricas. Neste cenário específico, destaca-se também a alta correlação entre Authority e Closeness.

Na Figura 2 os cálculos foram aplicados à rede social com a menor assortatividade (-0.21) - rede *Email-EuAll*. Sobre a Figura 2 nota-se mais uma vez que o ranking da métrica Grau tem uma maior similaridade aos das demais métricas em ambos os coeficientes estudados. Neste cenário, a alta correlação entre PageRank e Authority também pode ser destacada.

A Figura 3 mostra a aplicação cálculo da média dos valores dos coeficientes entre todas as oito bases de dados selecionadas para o estudo. Objetiva-se ao analisar a Figura 3, obter uma visão geral das correlações para diferentes bases, ou seja, o encontro de um padrão. Nesta Figura, em consonância com as figuras anteriores, a métrica Grau é a que apresentou *rankings* mais relacionados com as demais métricas.

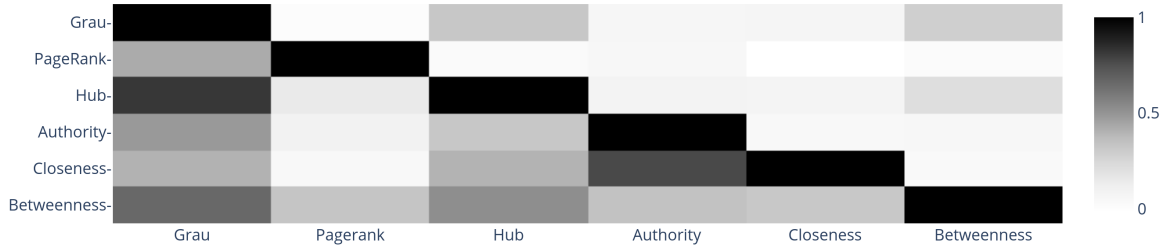

Figura 1. Estudo das correlações de ranking na rede Pretty Good Privacy

Figura 2. Estudo das correlações de ranking na rede Email-EuAll

Figura 3. Estudo das correlações de ranking na média das redes estudadas

É importante ressaltar, a diferença entre os valores encontrados nas correlações utilizando um coeficiente em detrimento ao outro. Os valores encontrados nas correlações pelo coeficiente de Kendall-Tau são mais baixos devido à sensibilidade de sua equação.

Onde trata-se a diferença de postos de uma forma binária e não ponderada como o de Spearman. Dessa forma, uma pequena diferença nas posições de um vértice no coeficiente de Kendall-Tau, contribui para uma baixa da correlação. Esta observação pode ser confirmada nas Equações 1 e 2.

A Tabela 2 apresenta a média de tempo de execução (em segundos) de cada métrica entre todas as bases do estudo. Ao analisar a Tabela 2 nota-se que devido ao cálculo bastante criterioso e exigente a nível de comparações e custos computacionais, Closeness e Betweenness se mostram métricas mais demoradas em suas execuções quando comparadas as demais.

Métrica	**Tempo médio (s)**
Grau	0.098
PageRank	1.214
Hub	0.514
Authority	0.392
Closeness	1342.613
Betweenness	6646.016

Tabela 2. Média de tempo de execução de cada métrica estudada

É interessante observar que a métrica com o menor custo computacional (Grau) é a métrica que mais correlaciona com as demais, tornando esta, uma boa indicação para identificação de usuários em redes sociais em menor tempo.

5. Conclusão

Neste trabalho foi apresentado um estudo comparativo entre métricas de ranking aplicadas em um conjunto de diversas redes sociais virtuais. Estudar a correlação entre métricas de ranking pode ser relevante para uma situação em que seja necessário definir uma determinada métrica para detecção de nós influentes em diferentes perspectivas. Quando se consegue avaliar qual métrica possui índices de correlação mais próximos às outras, torna-se elementar decidir por aplicar uma métrica que a partir de características diferentes de cada ranking consegue manter uma correlação.

Sobre os resultados, reforça-se que a correlação entre grau e as outras métricas de ranking são as mais altas do estudo considerando a média entre a correlação de todas as bases de dados e as redes com maior e menor assortatividade. Além disso, o tempo de execução desta métrica é o menor entre as métricas estudadas. Reforçando a indicação deste ranking pela alta correlação e baixo custo computacional.

Referências

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97.

Barbosa, L. M., Attux, R., and Godoy, A. Uma análise de assortatividade e similaridade para artigos científicos.

Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., and Caldarelli, G. (2019). The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71.

Donner, R. V., Lindner, M., Tupikina, L., and Molkenthin, N. (2019). Characterizing flows by complex network methods. In *A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*, pages 197–226. Springer.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Gabardo, A. C. (2015). *Análise de redes sociais: uma visão computacional*. Novatec Editora.

Ghent, A. (1963). Kendall's "tau" coefficient as an index of similarity in comparisons of plant or animal communities. *The Canadian Entomologist*, 95(06):568–575.

Kleinberg, J. M. (1999a). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Kleinberg, J. M. (1999b). Hubs, authorities, and communities. *ACM Computing Surveys*, 31(4):5.

Kunegis, J. (2013). Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350.

Leskovec, J. and Krevl, A. (2016). Snap datasets: Stanford large network dataset collection (2014). *URL http://snap. stanford. edu/data*, page 49.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Machado, J. R. and Tijiboy, A. V. (2005). Redes sociais virtuais: um espaço para efetivação da aprendizagem cooperativa. *RENOTE-Revista Novas Tecnologias na Educação*, 3(1).

Metz, J., Calvo, R., Seno, E. R., Romero, R. A. F., Liang, Z., et al. (2007). Redes complexas: conceitos e aplicações.

Mo, H. and Deng, Y. (2019). Identifying node importance based on evidence theory in complex networks. *Physica A: Statistical Mechanics and its Applications*, 529:121538.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: bringing order to the web.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069.

Silva, J. G. R., Xavier, C. R., da Fonseca Vieira, V., and de Carvalho, I. A. (2015). Estudo comparativo de métricas de ranqueamento em redes complexas utilizando coeficientes de correlaçao. *Congresso Brasileiro de Inteligência Computacional*.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295.

Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37.

Wasserman, S., Faust, K., et al. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.