Redes Sociais Científicas: análise topológica da influência dos pesquisadores

Vitor Horta[1], **Victor Ströele**[1], **Fernanda Campos**[1], **José Maria N. David**[1], **Regina Braga**[1]

[1]Programa de Pós-graduação em Ciência da Computação
Universidade Federal Juiz de Fora (UFJF) – MG – Brazil

{vitor.horta, victor.stroele}@ice.ufjf.br,

{fernanda.campos,jose.david, regina.braga}@iufjf.edu.br

Abstract. *Communities in social networks are composed by people with common interests who influence or are influenced by themselves. In this work, complex network analysis concepts are applied to verify the influence level among researchers, analyzing the structure of the scientific social network and its communities. We propose a bidirectional graph-based model to analyze the influence between researchers, and two algorithms to analyze the network structure, to identify scientific communities and to locate multidisciplinary researchers. To evaluate the model and the algorithms, a large scientific database is used in a use case. The results point to the solution viability and effectiveness.*

Resumo. *Comunidades em redes sociais são compostas por pessoas com interesses comuns, que influenciam ou são influenciadas por elas mesmas. Neste trabalho são aplicados conceitos de análise de redes complexas para verificar o nível de influência entre os pesquisadores, analisando a estrutura da rede social científica e suas comunidades. São propostos um modelo baseado em grafo bidirecional para analisar a influência entre os pesquisadores e algoritmos para analisar a estrutura da rede, identificar comunidades científicas e localizar pesquisadores multidisciplinares. Para avaliação do modelo e dos algoritmos é utilizada uma base de dados científicos de grande porte em um caso de uso. Os resultados apontam para a viabilidade e eficácia da solução.*

1. Introdução

Uma rede complexa representa um conjunto de objetos que se conectam de maneira não-trivial. Para se entender como os objetos se relacionam e se agrupam é necessária uma análise cuidadosa sobre a estrutura e características dessas redes [Wasserman and Faust 1994].

Neste trabalho serão utilizados os conceitos e técnicas de redes complexas para analisar como ocorrem as interações entre pesquisadores em uma rede social científica, com o objetivo de identificar (i) pesquisadores influenciadores, (ii) comunidades científicas e (iii) pesquisadores que participam de mais de uma comunidade, ou seja, pesquisadores que influenciam ou são influenciados por diferentes grupos de pesquisadores. Este trabalho avança as pesquisas do NEnC (Grupo de Pesquisa em Engenharia de Conhecimento) em análise de redes sociais científicas [Ströele et al. 2013] [Almeida et al. 2016].

A rede social científica modelada neste trabalho é representada por um grafo bidirecional cujos vértices representam os pesquisadores e as arestas correspondem às relações científicas entre eles. Um grafo direcionado é utilizado para que seja possível definir o grau de influência entre os pesquisadores, permitindo a identificação de pesquisadores influenciadores e influenciados em suas áreas de pesquisa.

A principal contribuição deste trabalho é a proposta e desenvolvimento de um algoritmo de agrupamento que considera as características de pesquisas multidisciplinares de alguns pesquisadores, permitindo que eles pertençam a mais de um grupo. Com essa abordagem é possível identificar pessoas que possuem mais de uma área de atuação, e que participam de duas ou mais comunidades científicas. Outra característica importante do algoritmo proposto é o sub-agrupamento, cujo objetivo é identificar subgrupos de comunidades de pesquisa.

Além da Introdução, este trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 descreve o modelo proposto para a realização deste estudo, na Seção 4 é feita a análise da topologia da rede social científica, na Seção 5 é descrito o algoritmo de agrupamento desenvolvido, um estudo de caso é realizado na Seção 6 e, finalmente, na Seção 7 são apresentadas as considerações finais.

2. Trabalhos Relacionados

[Luo et al. 2011] utiliza métodos hierárquicos divisivos para detecção de comunidades. Tais métodos consistem, essencialmente, em retirar as arestas com maior índice de centralidade (*edge betweenness*), e identificam comunidades em um dendograma. Essa abordagem é frequentemente utilizada, porém a alta complexidade computacional do método não favorece o seu uso em redes sociais de grande porte.

O DENGRAPH [Falkowski et al. 2007] foi desenvolvido utilizando a estratégia de agrupamento por densidade através de uma implementação do DBSCAN [Ester et al. 1996] adaptada para grafos não-direcionados. Como resultado alcança um melhor desempenho em performance e maior capacidade de detectar ruídos em redes com maior volume de dados.

Como o DBSCAN em sua implementação original não prevê que o conjunto de dados pode possuir múltiplas granularidades na densidade, [Gialampoukidis et al. 2016] desenvolveram o DBSCAN*-Martingale. Este algoritmo adapta os parâmetros de densidade de forma iterativa com a finalidade de descobrir grupos com diferentes níveis de similaridade de seus respectivos membros.

[Li et al. 2016] observaram a necessidade de se considerar a direção nas relações dos membros em redes sociais e propuseram um método que utiliza a estratégia de agrupamento por densidade para detecção de comunidades em grafos direcionados. Entretanto, não consideram a possibilidade de um elemento pertencer a dois ou mais grupos distintos.

Como diferencial deste trabalho pode-se destacar o desenvolvimento de um algoritmo de agrupamento e subagrupamento, visto que o processo iterativo proposto por outros trabalhos é muito custoso tornando-se inviável para grandes bases de dados. Além disso, o algoritmo proposto considera diferentes níveis de influência entre pares de pesquisadores (grafo bidirecional), e fornece também um parâmetro que pode ser ajustado para aprimorar a decisão de se incluir membros com menor influência ou que já estejam

alocados em outras comunidades, localizando pesquisadores multidisciplinares.

3. Modelagem da Rede Social Científica

No modelo da rede científica proposto neste trabalho os nós do grafo representam os pesquisadores e as arestas possuem pesos que representam o nível de influência entre eles. Neste modelo, a influência que um pesquisador exerce sobre o outro não necessariamente é igual à influência que ele recebe desse mesmo pesquisador. Por isso, o grafo social que representa esse modelo é um grafo bidirecional no qual os pesos (ida e volta) da relação entre dois pesquisadores são analisados de forma diferenciada. Dessa forma dois vértices P_A e P_B estão conectados sempre por duas arestas direcionadas, que por sua vez são definidas através da relação de coautoria [Lopes et al. 2010].

$$IP_{AB} = \frac{\parallel P_A \bigcap P_B \parallel}{\parallel P_A \parallel} \tag{1}$$

4. Análise Topológica

Para analisar a estrutura da DBLP, composta por 1.306.546 vértices e 9.915.146 arestas, foi calculada a sua distribuição de graus (equação (2)), propriedade que caracteriza a topologia de uma rede complexa, e é obtida calculando-se quantos nós possuem determinado grau para todos os valores de graus existentes na rede.

$$f(k) = \frac{\textit{Número de vértices com grau } k}{\textit{Número total de vértices}} \tag{2}$$

Os gráficos da Figura 1 apresentam os resultados obtidos, onde é possível observar que há um grande número de pesquisadores com poucas relações (grau baixo) e poucos pesquisadores com muitas relações (grau alto). Esta característica é típica das redes livres de escala, nas quais poucos elementos centralizam a maioria das relações da rede, obedecendo uma lei de potência.

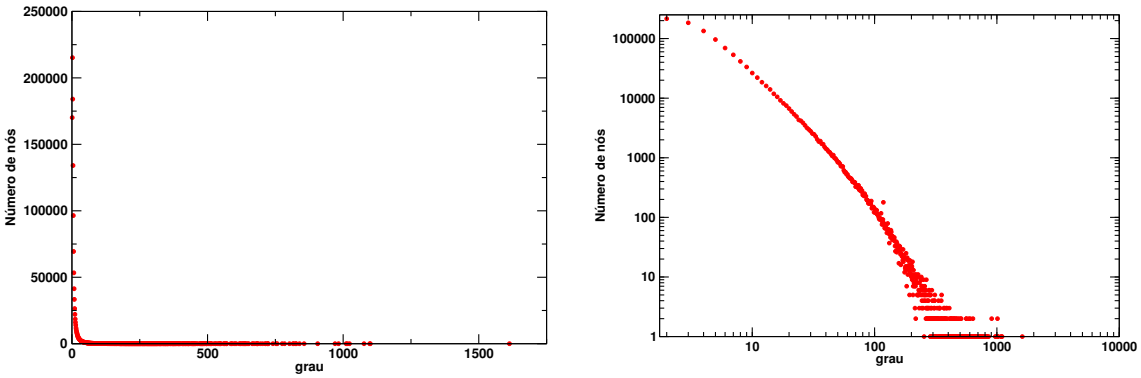

Figura 1. Distribuição de grau (esquerda) e gráfico log-log (direita)

Na análise de distribuição de grau foi detectado que existem poucos nós com grau elevado. Por outro lado, na análise da distribuição das influências entre os pesquisadores, observou-se que uma grande quantidade de nós possui a influência máxima. Para que essas duas análises sejam verdadeiras de forma simultânea, é necessário que a rede da DBLP

não seja totalmente conectada, sendo formada por componentes conexas independentes que contêm poucos nós de muita influência.

Para verificar o comportamento descrito anteriormente, foi elaborado um algoritmo cuja função é identificar as componentes conexas da rede. Em vista do grande volume de dados da rede, o algoritmo disponibiliza um parâmetro *raio* que pode ser utilizado para viabilizar sua execução, impactando a performance do algoritmo sem influenciar o resultado final. Um valor elevado para o *raio* irá diminuir o tempo de execução do algoritmo, porém, o consumo de memória aumenta consideravelmente. Isso acontece porque, com o aumento do *raio*, aumenta-se também o número de vértices armazenados em memória, o que contribui para acelerar o processamento. Assim, este algoritmo permite que mesmo computadores com baixa capacidade de processamento encontrem todas as componentes conexas de uma rede social de grande porte.

5. NETSCAN: algoritmo de agrupamento

Após as análises da topologia da rede social científica DBLP, pode-se definí-la como sendo uma rede complexa livre de escala, que possui baixa conectividade global e alguns nós centralizadores, ou seja, alguns nós altamente conectados. Neste contexto surge o interesse em identificar quais são os grupos de pesquisa mais bem definidos, quais são os membros de cada grupo, bem como os pesquisadores com maior influência sobre seus respectivos grupos. Para responder essas questões, é proposto e desenvolvido um algoritmo de agrupamento por densidade, chamado NETSCAN, baseado no algoritmo DBSCAN [Ester et al. 1996] e que possa ser aplicado em redes de grande porte como a DBLP.

Uma das principais diferenças com relação ao DBSCAN é que o NETSCAN considera que a distância entre dois vértices depende da direção do relacionamento que está sendo analisada. Além disso, permite que um mesmo vértice seja incluído em mais de um grupo e, dessa forma, é possível identificar pesquisadores que contribuem em diferentes áreas e grupos de pesquisa distintos, mesmo que esses pesquisadores não sejam caracterizados como centralizadores.

Além dos parâmetros ε e *minPts*, originados do DBSCAN, o NETSCAN permite a definição de um *raio*, utilizado para buscar por elementos influenciados por um vértice em uma profundidade maior, ou seja, pode ser definido quantas camadas de vizinhos a partir de um *core* serão analisadas.

6. Estudo de Caso

Neste estudo de caso, o algoritmo NETSCAN foi executado na maior componente conexa da rede social científica DBLP composta por 1.100.504 pesquisadores e 9.598.808 relacionamentos. Ao todo foram identificadas 34.776 comunidades de pesquisa, utilizando os parâmetros: $\varepsilon = 1$, $minPts = 5$ e $raio = 1$. Esses parâmetros foram definidos após alguns experimentos realizados com uma parte da base de dados.

Dado o grande volume de dados utilizados neste estudo, alguns agrupamentos foram selecionados com o intuito de ilustrar o comportamento do algoritmo proposto e analisar os resultados obtidos. A Figura 2 mostra um dos grupos definidos pelo NETSCAN onde os nós centralizadores estão representados pelos vértices verdes e os *border points* pela cor cinza. Para uma melhor visualização do grafo, as medidas de influência representadas nas arestas foram multiplicadas por 1000.

Observa-se que este grupo possui dois vértices influenciadores e que apenas dois *border points* estão ligados nestes dois vértices simultaneamente. Isso indica que este grupo possui uma grande chance de ser separado em dois subgrupos caso o parâmetro ε seja escolhido de forma mais rígida, exigindo um maior nível de influência no grupo.

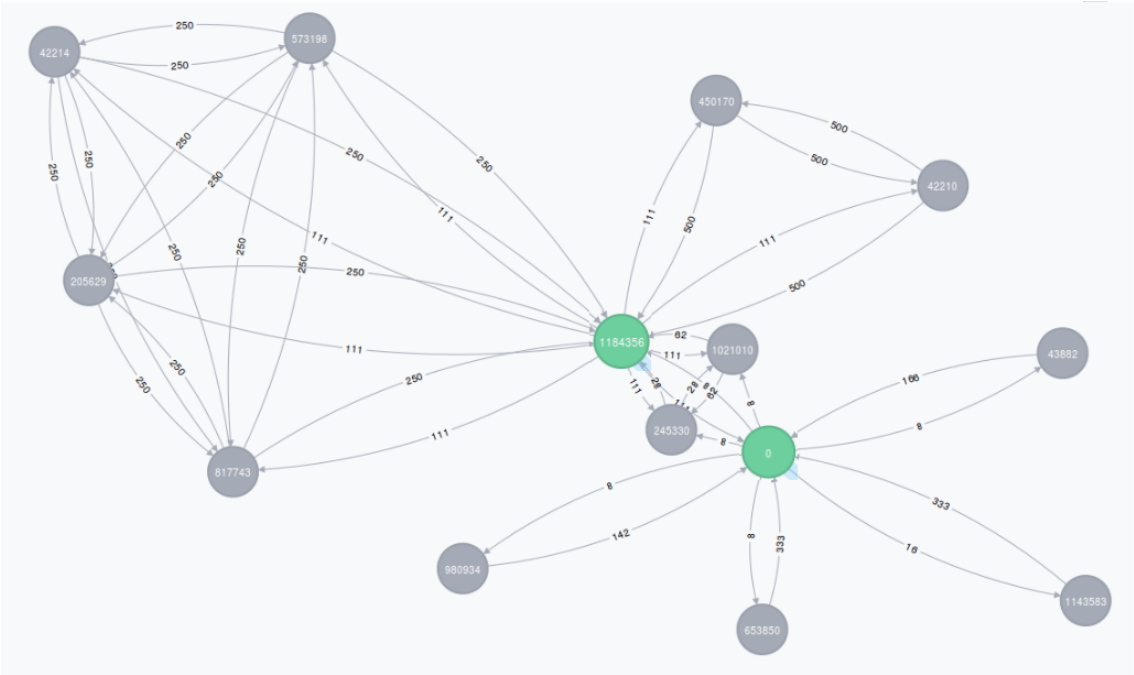

Figura 2. Grupo de pesquisadores identificado pelo NETSCAN

6.1. Análise dos Resultados

Através do NETSCAN foram identificadas comunidades científicas, pesquisadores influenciadores e pesquisadores que participam de mais de uma comunidade. Um dos desafios encontrados, ao analisar uma rede científica com grande volume de dados, foi a variação de densidade que dificulta a parametrização das entradas do algoritmo de agrupamento Para resolver este problema, sem a necessidade de se calcular os valores ideais de cada parâmetro repetidamente e em tempo de execução, foi proposto um método de subagrupamento que foi utilizado para refinar grupos formados por muitos pesquisadores.

Através de uma entrevista informal, com observação direta como fonte de coleta de dados, os resultados obtidos foram apresentados a seis pesquisadores para que eles os analisassem. Cinco deles reportaram que as comunidades nas quais eles estavam envolvidos estava condizente com a realidade dos seus grupos de pesquisa. Apenas um pesquisador reportou que, de fato, desenvolvia suas pesquisas com os pesquisadores que ele estava relacionado, mas considerou o grupo ao qual ele estava vinculado muito grande. Então, foi apresentado para ele os sub-agrupamentos do seu grupo original, o qual confirmou que o seu subgrupo refletia a realidade de seu grupo de pesquisa.

As análises realizadas apontam para a viabilidade tanto do modelo quanto dos algoritmos propostos. Através da avaliação realizada pode-se dizer que há indícios de que o resultado alcançado foi satisfatório e que os métodos apresentados podem ser evoluídos para alcançar uma coesão maior no agrupamento.

7. Considerações Finais

Neste trabalho foi modelada uma rede social científica (DBLP), onde os elementos representam pesquisadores e suas ligações como relacionamentos de coautoria. Para essa modelagem foi adotado um grafo bidirecional para possibilitar a representação de níveis de influência diferentes entre os pesquisadores.

Um dos objetivos da análise dessa rede social científica é compreender a sua estrutura e como os pesquisadores se influenciam no meio acadêmico. Para tal, foi desenvolvido um algoritmo de agrupamento por densidade, baseado no DBSCAN. Este possui adaptações para trabalhar em grafos com arestas direcionadas e com grande volume de dados. Os resultados obtidos no desenvolvimento deste trabalho apontam para a viabilidade da solução proposta.

Em trabalhos futuros uma evolução temporal das componentes conexas encontradas deve ser realizada para verificar se as comunidades menores tendem a serem incluídas na componente conexa dominante. Deseja-se também elaborar uma estratégia para a parametrização ótima no refinamento dos grupos. Por fim, avaliações adicionais devem ser realizadas, incluindo pessoas de diferentes regiões e países para que se possa explorar a solução em profundidade.

Referências

Almeida, R., Pereira, C. K., Campos, F., and Stroele, V. (2016). Recomendação de recursos educacionais para grupos: buscando soluções em redes sociais. In *Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)*.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press.

Falkowski, T., Barth, A., and Spiliopoulou, M. (2007). Dengraph: A density-based community detection algorithm. In *In Proc. of the 2007 IEEE / WIC / ACM International Conference on Web Intelligence,*, pages 112–115.

Gialampoukidis, I., Tsikrika, T., Vrochidis, S., and Kompatsiaris, I. (2016). Community detection in complex networks based on dbscan* and a martingale process. In *2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6.

Li, X., Tan, Y., Zhang, Z., and Tong, Q. (2016). Community detection in large social networks based on relationship density. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*. IEEE.

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. (2010). Cooperative Authorship Social Network. In *AMW*.

Luo, T., Zhong, C., Ying, X., and Fu, J. (2011). Detecting community structure based on edge betweenness. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE.

Ströele, V., Zimbrão, G., and Souza, J. M. (2013). Group and link analysis of multirelational scientific social networks. *Journal of Systems and Software*, 86:1819–1830.

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