Um Estudo sobre a Evolução Temporal de Comunidades Científicas

Bruno Leite Alves,
Alberto H. F. Laender (Orientador), Fabrício Benevenuto (Coorientador)

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{bruno.leite,fabricio,laender}@dcc.ufmg.br

Resumo. *Diversos esforços têm sido realizados para compreensão e modelagem de aspectos dinâmicos de comunidades científicas. Apesar do grande interesse, pouco se sabe sobre o papel que diferentes membros têm na formação da estrutura topológica da rede dessas comunidades. Neste trabalho, investigamos o papel que os membros do núcleo de comunidades científicas têm na formação e evolução de sua rede de colaboração. Para isso, definimos o núcleo de uma comunidade com base em uma métrica, denominada CoScore, derivada do índice h que captura tanto a prolificidade quanto o envolvimento dos pesquisadores na comunidade. Nossos resultados subsidiam uma série de observações importantes relacionadas à formação e aos padrões de evolução das comunidades. Notamos que variações no conjunto de membros que compõem o núcleo das comunidades tendem a ser fortemente correlacionadas com variações nas propriedades de rede. Mostramos ainda que nossas observações são importantes para caracterizar o papel dos principais membros na formação e na estrutura das comunidades.*

1. Introdução

Desde os seus primórdios, a sociedade tem se organizado em comunidades, ou seja, grupos de indivíduos com interesses em comum[1]. Particularmente, a proliferação de novas tecnologias de comunicação baseadas na Internet tem facilitado a rápida formação e crescimento de comunidades *online* [Kleinberg 2008]. Comunidades possuem uma grande quantidade de características e servem a vários propósitos. Elas podem ser desde pequenos grupos hermeticamente comprometidos com temas específicos, como comunidades científicas de determinadas áreas, até mesmo grupos de milhões de usuários ligados por um interesse em comum, tais como comunidades relacionadas a esporte ou fãs de uma celebridade.

Geralmente, indivíduos que são socialmente conectados em comunidades tendem a compartilhar interesses comuns e outras similaridades. Embora existam muitos fatores que possam determinar a formação de uma comunidade e o seu crescimento, existem duas forças que explicam a formação de comunidades: influência e homofilia. Esforços recentes têm mostrado evidências quantitativas que ambas as forças dependem da identificação de um grupo influente de indivíduos com o poder de afetar não somente a estrutura topológica de uma comunidade, mas também interferir na difusão e no fluxo de informação da comunidade.

[1]http://www.merriam-webster.com/dictionary/community

Neste trabalho, apresentamos uma perspectiva diferente e um estudo complementar desse problema. Aqui, nos concentramos em estudar os papéis que indivíduos influentes em comunidades científicas desempenham na evolução das propriedades de tais comunidades. Para esse estudo, usamos dados da DBLP[2] para construir comunidades científicas, representadas pelas principais conferências dos SIGs[3] (*Special Interest Groups*) da ACM[4] (*Association for Computing Machinery*). Então, propomos uma estratégia para definir o núcleo de uma dada comunidade científica, juntamente com seus líderes em um dado período de tempo. Finalmente, investigamos como os aspectos do núcleo impactam a estrutura topológica da comunidade.

O estudo do núcleo de comunidades científicas pode ser visto de duas perspectivas diferentes: sociológico e tecnológico. A primeira é necessidade de compreender como partes da sociedade evoluem, enquanto a segunda compreende estes aspectos críticos para muitas aplicações, como predição de links. Tal estudo, entretanto, tem sido difícil devido à caracterização de alguns conceitos, como conexões humanas e uma definição apropriada de liderança, sendo difícil até mesmo de se reproduzir em grande escala dentro de um laboratório de pesquisa.

A seguir listamos as principais contribuições deste trabalho [Alves et al. 2013]:

- Definição de uma métrica, chamada *CoScore*, que captura tanto a prolificidade quanto o envolvimento de um pesquisador em uma comunidade científica.
- Definição do conceito de núcleo de uma comunidade a partir da métrica proposta.
- Caracterização de mais de vinte comunidades científicas e uma discussão de como a métrica *CoScore* afeta as propriedades topológicas das redes ao longo do tempo.
- Visualização das comunidades estudadas[5], em que é possível observar os componentes da rede e como os membros dos núcleos se organizam nestes componentes.

2. Comunidades

Dada uma rede social, uma comunidade pode ser compreendida como um grupo denso de nodos dessa rede que possuem mais arestas interligando-os entre si, do que arestas interligando-os ao restante da rede. Existem múltiplas definições e estratégias para identificar comunidades e elas variam de acordo com o contexto [Kleinberg 2008, Leskovec et al. 2010]. No nosso contexto, uma comunidade científica pode ser definida em termos de uma grande e consolidada conferência científica capaz de reunir pesquisadores que trabalham em uma mesma área de pesquisa ao longo de vários anos.

A fim de construir um conjunto de comunidades científicas, coletamos dados da biblioteca digital, DBLP[6]. Para o propósito do nosso trabalho, consideramos uma comunidade científica como um grafo em que os nodos representam pesquisadores e as arestas ligam os coautores de artigos de uma mesma comunidade. A fim de definir tais comunidades, focamos nas publicações das principais conferências (*flagship*) dos SIGs da ACM. No total, 22 comunidades científicas de diferentes áreas foram construídas e analisadas.

[2]http://dblp.uni-trier.de/
[3]http://www.acm.org/sigs
[4]http://www.acm.org
[5]http://hidra.lbd.dcc.ufmg.br/graphs/
[6]http://dblp.uni-trier.de/

3. Definição de Núcleo das Comunidades

Tentativas anteriores para identificar o núcleo de comunidades científicas são baseadas em abordagens algorítmicas que visam identificar conjuntos densos de nodos na rede [Seifi and Guillaume 2012]. Entretanto, como planejamos investigar o papel do núcleo na estrutura da rede, qualquer abordagem que faz uso da estrutura da rede para identificar tais nodos poderia nos levar a um conjunto de pesquisadores enviesado. Em vez disso, focamos no desenvolvimento de uma métrica que quantificasse o envolvimento de um pesquisador em uma comunidade científica durante um certo período de tempo. Intuitivamente, essa métrica deveria ser capaz de capturar, (i) a prolificidade de um pesquisador em diferentes comunidades, usamos o índice h, uma métrica largamente adotada para esse propósito e (ii) a frequência do envolvimento daquele pesquisador com a comunidade em um certo período de tempo, para isso, multiplicamos o valor do seu índice h ao final desse período pelo número de publicações desse pesquisador nessa comunidade no mesmo período.

Denominamos essa métrica de *Community Score* (*CoScore*) [Alves et al. 2013]. Mais formalmente, o *CoScore* de um pesquisador p em uma comunidade c durante um período de tempo t, $\textit{CoScore}_{p,c,t}$, é dado pelo seu índice h ($h_{p,t}$) ao final do período t multiplicado pelo seu número de publicações na comunidade c durante t (#publicações$_{p,c,t}$), como expresso pela Equação 1.

$$\textit{CoScore}_{p,c,t} = h_{p,t} \times \text{\#publicações}_{p,c,t} \tag{1}$$

A fim de computar o *CoScore* para os membros de uma comunidade, definimos o núcleo de uma comunidade em um certo período de tempo como sendo formado pelos pesquisadores com o melhor escore naquela comunidade em termos de seu *CoScore* em um dado período.

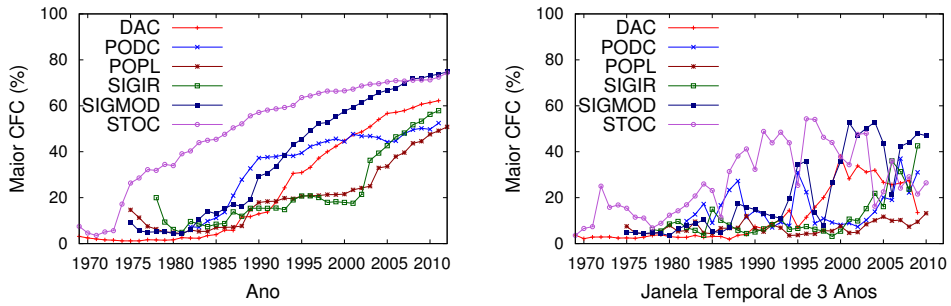
4. Definição dos Limiares

Utilizamos a métrica *resemblance*, conforme definida por [Viswanath et al. 2009], e coeficiente angular para determinar dois limiares importantes, o tamanho do núcleo da comunidade, sendo 10% da comunidade e o tamanho da janela temporal, 3 anos.

5. Análise de Evolução das Comunidades

A fim de estudar a evolução das principais propriedades estruturais das comunidades científicas, examinamos várias métricas de redes para cada uma das comunidades consideradas. Apresentamos aqui o tamanho do maior componente fracamente conectado (CFC). A Figura 1 mostra a métrica varia ao longo do tempo. Apresentamos essa métrica para um conjunto de seis comunidades científicas selecionadas entre aquelas que mais se estendem ao longo do tempo em nosso conjunto de dados. Nossas análises são realizadas sob duas perspectivas. A primeira consiste em analisar a evolução da rede ano a ano, acumulando nodos e arestas da instância final do grafo. A segunda consiste em analisar instâncias construídas com base em nodos e arestas criados em uma janela de tempo predefinida.

Notamos a partir da Figura 1 que o maior CFC tende a aumentar largamente em função do tempo. Isto sugere que na fase inicial, as comunidades científicas são formadas por vários grupos de pesquisa pequenos e segregados. De forma geral, observamos que as comunidades científicas têm características de evolução semelhantes e que essas



(a) Maior CFC final (b) Maior CFC por janela

Figura 1. Maior CFC das comunidades científicas

propriedades são dinâmicas, mudando ao longo do tempo. Mais importante ainda, nossas observações sugerem que um pequeno grupo de pesquisadores que fazem parte do núcleo são responsáveis por criar caminhos entre grupos de pesquisa menores e mais conectados. A fim de investigar melhor os pesquisadores que fazem parte do núcleo, na próxima seção comparamos membros e não membros dos núcleos das comunidades.

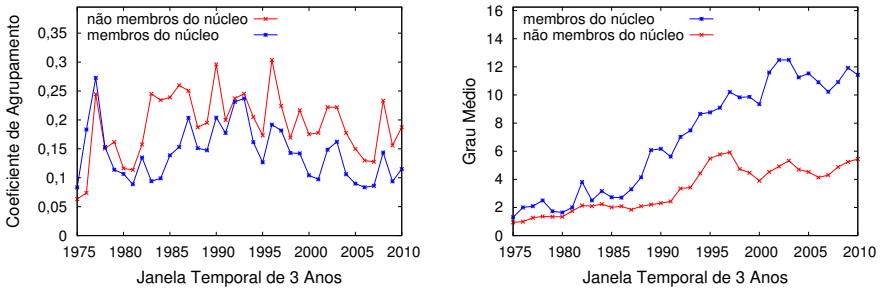
6. Caracterização dos Núcleos das Comunidades

Até que ponto as propriedades dos membros do núcleo diferem dos demais membros das comunidades? Para responder a essa pergunta, calculamos as propriedades de rede para os membros e não membros dos núcleos das comunidades. Consideramos a análise de janelas de tempo para compreender as variações que essas duas classes podem ter na medida global. A Figura 2 mostra o grau médio e o coeficiente de agrupamento médio calculados para membros e não membros do núcleo da comunidade SIGMOD. Além disso, também medimos a fração de membros do núcleo, bem como dos não membros que estão no maior CFC e calculamos o *betweenness* médio de cada um desses grupos de membros. Na próxima seção investigamos como aspectos dos membros dos núcleos podem impactar a estrutura geral das comunidades.

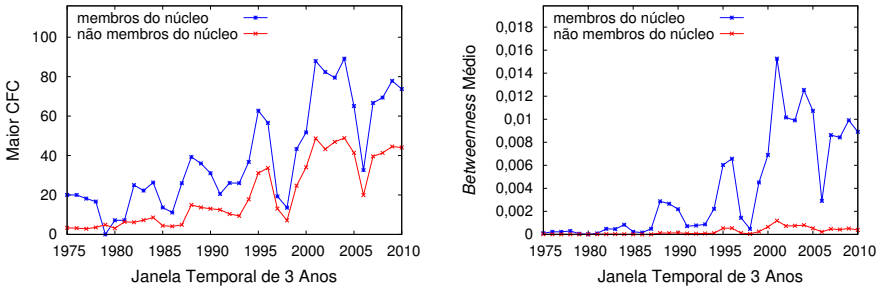
7. Impacto dos Membros dos Núcleos na Estrutura Topológica das Comunidades

Agora analisamos o quanto as variações no núcleo da comunidade afetam a estrutura da rede. Para isso, calculamos a média do *CoScore* dos membros de cada comunidade ao longo do tempo. Intuitivamente, essa medida captura a prolificidade global e o grau de participação dos membros do núcleo em uma comunidade científica. A Figura 3 mostra o *CoScore* médio para um conjunto específico de comunidades em função do tempo. Plotamos em duas figuras distintas para facilitar a visualização. Podemos observar em todas as comunidades que, apesar de pequenas quedas, de forma geral o valor aumenta ao longo do tempo sofrendo variações. Desconsiderando o que causou essas variações, queremos investigar se tais variações podem impactar diretamente a estrutura da rede.

Nossa abordagem para investigar esta questão consiste em calcular o coeficiente de correlação de Pearson entre a média do *CoScore* e as métricas de redes para cada comunidade. Obtivemos uma forte correlação possitiva para maioria das métricas mostrando que nossa métrica afeta diretamente a estrutura da rede.



(a) Coeficiente de agrupamento (b) Grau médio



(c) Maior CFC (d) *Betweenness* médio

Figura 2. Propriedades da comunidade SIGMOD para os membros e não membros do núcleo

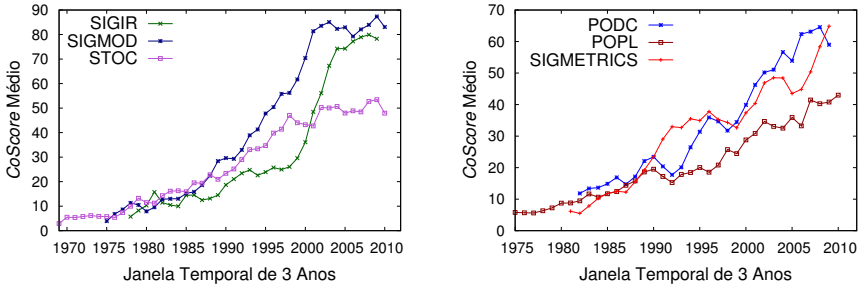


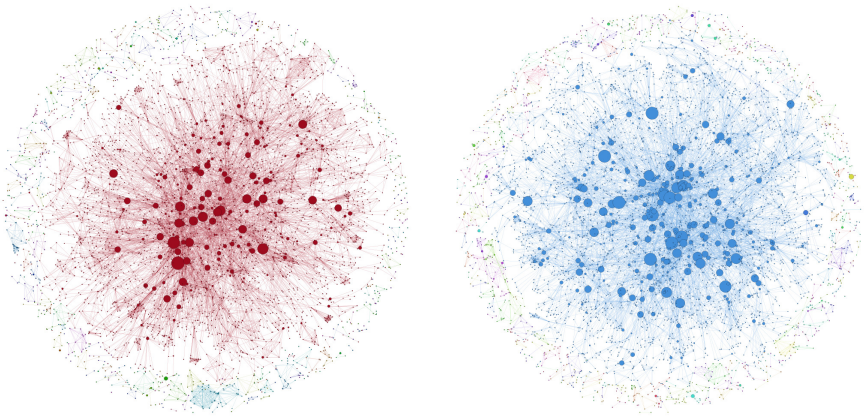
Figura 3. *CoScore* médio das comunidades científicas

8. Visualização das Comunidades

Em complemento a nossas análises, plotamos as comunidades científicas acumulando todos os seus nodos e arestas ao longo do tempo. A Figura 4 apresenta a plotagem das comunidades SIGMOD e CHI. Cada cor representa um componente conectado diferente e o tamanho dos nodos indica o número de vezes que o pesquisador apareceu no núcleo ao longo de todo o tempo de vida daquela comunidade

9. Conclusão e Trabalhos Futuros

Neste trabalho apresentamos uma investigação profunda dos papéis que os membros do núcleo das comunidades científicas desempenham na formação e evolução da estrutura da rede de coautoria. Definimos um núcleo para as comunidades baseado em uma nova métrica, *CoScore*, uma métrica derivada do índice h que captura tanto, a prolificidade, quanto o envolvimento de pesquisadores em uma comunidade. Observamos que as



(a) SIGMOD (b) CHI

Figura 4. Instância final das comunidades científicas

variações nos membros do núcleo da comunidade estão fortemente correlacionadas com as variações nas propriedades de rede. Nosso estudo também destaca a importância de estudar os membros do núcleo da comunidade e esperamos que nossas observações possam inspirar futuros modelos de formação da comunidade.

A partir dos resultados do nosso estudo algumas oportunidades de trabalhos futuros foram identificadas e são listadas a seguir:

- Aplicação do estudo em outros contextos.
- Utilização de outras métricas de prolificidade.
- Avaliação do *CoScore* em outros contextos.
- Geração de modelos de formação de comunidades.
- Aplicação da abordagem proposta para o estudo de *clusters*.
- Análise do impacto da migração de pesquisadores entre comunidades.

Referências

Alves, B. L., Benevenuto, F., and Laender, A. H. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proc. of WWW (Companion Volume)*, pages 649–656.

Kleinberg, J. (2008). The convergence of social and technological networks. *Commun. ACM*, 51(11):66–72.

Leskovec, J., Lang, K. J., and Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proc. of WWW*.

Seifi, M. and Guillaume, J.-L. (2012). Community cores in evolving networks. In *Proc. of MSND*.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in facebook. In *Proc. of WOSN*.