Um Estudo sobre a Evolução Temporal de Comunidades Científicas

Bruno Leite Alves[1], **Orientador: Alberto H. F. Laender**[1], **Coorientador: Fabrício Benevenuto**[1]

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

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

Abstract. *abstract*

Resumo. *resumo*

1. Introdução

1.1. Motivação

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. Por um lado, influência postula que indivíduos mudam para se tornarem mais similares a seus amigos na comunidade. Por outro lado, homofilia postula que indivíduos criam conexões sociais dentro de uma comunidade justamente porque já são similares. Esforços recentes têm mostrado evidências quantitativas que ambas as forças [Cha et al. 2010, Backstrom et al. 2006], teorias [Rogers 1962, Watts and Dodds 2007] e modelos existentes [Kempe et al. 2003, Kempe et al. 2005] 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.

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. Intuitivamente, quando pesquisadores importantes e com grande influência em suas áreas decidem se juntar ou deixar uma comunidade científica, levam com eles recursos, experiência e até mesmo estudantes, e possivelmente influenciam outros membros

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

a fazerem o mesmo. 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. A primeira é a sociológica, vindo da necessidade de compreender como partes da sociedade evoluem, bem como responder a perguntas de longa data relacionadas com à interação entre os diferentes tipos de participante. Em contrapartida, sob uma perspectiva tecnológica, compreender estes aspectos é crítico para muitas aplicações relacionadas a predição de *links* [Getoor and Diehl 2005]. 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.

1.2. Contribuições

A seguir listamos as principais contribuições deste trabalho:

- Definição de uma métrica, chamada *CoScore*, que captura tanto a prolificidade quanto o envolvimento de um pesquisador em uma comunidade científica. Desta forma, nossa métrica é capaz de quantificar a importância de um determinado pesquisador em uma dada comunidade científica [Alves et al. 2013].
- Definição do conceito de núcleo de uma comunidade a partir da métrica proposta [Alves et al. 2013].
- 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 [Alves et al. 2013].
- Visualização das comunidades estudadas, em que é possível observar os componentes da rede e como os membros dos núcleos se organizam nestes componentes.
- Página Web que permite a visualização de tais comunidades: http://hidra.lbd.dcc.ufmg.br/graphs/

2. Comunidades

2.1. Comunidades Científicas

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.

[2]http://dblp.uni-trier.de/

[3]http://www.acm.org/sigs

[4]http://www.acm.org

A fim de construir um conjunto de comunidades científicas, coletamos dados da DBLP[5] [Ley 2009], uma biblioteca digital que contém mais de 2,1 milhões de publicações de 1,2 milhões de autores e que provê informações bibliográficas dos principais anais de conferências e periódicos da área de Ciência da Computação. 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. Assim, definimos uma comunidade científica como formada por pesquisadores interligados entre si por serem coautores de um algum artigo dessas conferências, fazendo com que elas atuem como comunidades onde coautorias são formadas. No total, 22 comunidades científicas foram construídas, sendo assim, removemos conferências sem dados suficientes para uma análise temporal, bem como conferências cujo histórico completo não está registrado na DBLP.

2.2. 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 e (ii) a frequência do envolvimento daquele pesquisador com a comunidade em um certo período de tempo.

Em primeiro lugar, a fim de capturar a prolificidade de um pesquisador, usamos o índice h [Hirsch 2005], uma métrica largamente adotada para esse propósito. Essa métrica consiste de um índice que tenta medir tanto a produtividade quanto o impacto dos trabalhos publicados de um dado pesquisador. Ela baseia-se no conjunto de artigos mais citados de um pesquisador e no número de citações desse pesquisador com pelo menos *h* citações. Mais especificamente, um pesquisador tem um índice h i se publicou i artigos que receberam pelo menos i citações. Assim, por exemplo, se um pesquisador possui 10 artigos com pelo menos 10 citações, seu índice h final é 10.

Em segundo lugar, como uma tentativa de capturar a importância de um pesquisador em uma comunidade específica em um certo período de tempo, 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, $\mathit{CoScore}_{p,c,t}$, 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.

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

[5]http://dblp.uni-trier.de/

Como podemos ver, a primeira parte da equação captura a importância de um pesquisador para a comunidade científica como um todo em um determinado período de tempo, independentemente de qualquer área de pesquisa específica, e a segunda parte pesa essa importância baseada na atividade do pesquisador em uma certa comunidade. 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.

2.3. Definição dos Limiares

Nossa estratégia para definir os dois limiares necessários para definir o núcleo das comunidades consiste em variar cada um deles e quantificar como eles impactam nas mudanças dos membros desse núcleo. Para medir essas mudanças, calculamos a métrica *resemblance*, conforme definida por [Viswanath et al. 2009], que mede a fração dos membros do núcleo no tempo t_0 que permanecem no núcleo no tempo t_1. Para cada comunidade, variamos o tamanho da janela de 1 a 5 anos e o tamanho do núcleo de 10% a 60% do total dos respectivos pesquisadores e após realizarmos os cálculos de métricas como o coeficiente angular, definimos o tamanho da janela temporal como sendo de 3 anos e o tamanho do núcleo da comunidade de 10%.

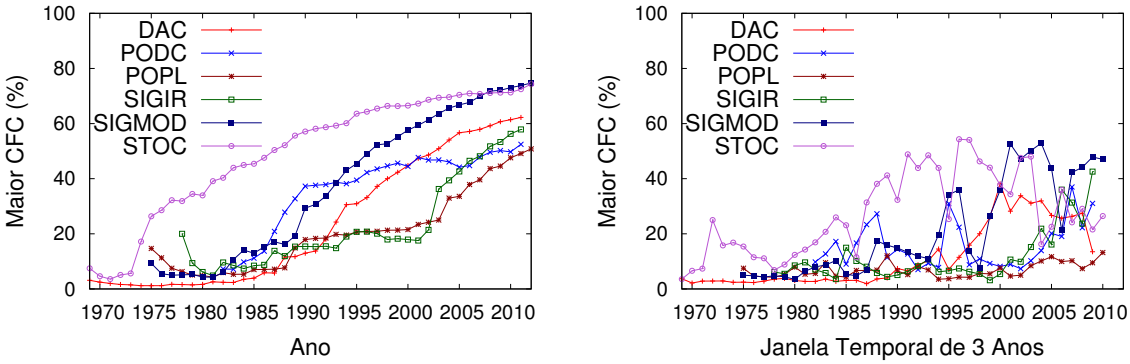
3. Análise das Comunidades

3.1. 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. Essa perspectiva nos permite observar a estrutura final de uma comunidade em função do tempo. A segunda perspectiva consiste em analisar instâncias construídas com base em nodos e arestas criados em uma janela de tempo predefinida. Esta análise nos permite investigar as variações da rede com potencial para impactar a sua estrutura final.

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. Com o tempo, alguns pesquisadores (e.g., estudantes) deixam suas instituições e começam a colaborar com outros grupos de pesquisa. Além disso, como a comunidade evolui, líderes de grupos de pesquisa tendem a colaborar com outros colegas da mesma comunidade. Assim, com o tempo, pesquisadores de diferentes grupos tendem a colaborar e aumentar o tamanho do maior CFC.

De forma geral, observamos que as comunidades científicas têm características de evolução semelhantes e que essas propriedades são dinâmicas, mudando ao longo do tempo. Mais importante ainda, nossas observações sugerem que um pequeno grupo de



(a) Maior CFC final

(b) Maior CFC por janela

Figura 1. Maior CFC das comunidades científicas

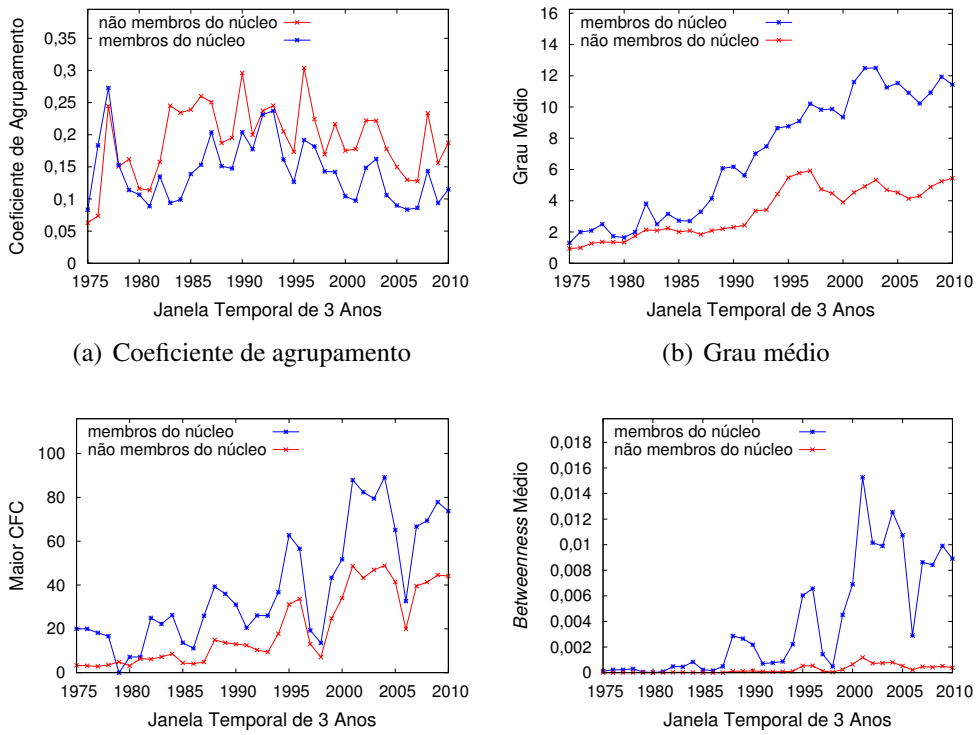
pesquisadores que fazem parte do núcleo são responsáveis por criar caminhos entre grupos de pesquisa menores e mais conectados. A fim de investigar melhor os pesquisadores que fazem parte do núcleo, na próxima seção comparamos membros e não membros dos núcleos das comunidades.

3.2. 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.

Podemos fazer observações importantes a partir dessas análises. Primeiro, podemos observar que o grau médio dos membros dos núcleos é consideravelmente maior em comparação com o dos não membros, que tendem a estabelecer mais e mais conexões em função do tempo. No entanto, o coeficiente de agrupamento dos membros do núcleo tendem a ser ligeiramente menor quando comparados com o dos não membros, o que significa que eles podem atuar como *hubs*, conectando diferentes grupos com pequenas interseções. Ao analisar a fração de membros do núcleo que fazem parte do maior CFC, podemos notar que é muito maior do que a fração de não membros, sugerindo que eles podem estar conectando componentes menores. Confirmamos essas observações, analisando a centralidade desses grupos de pesquisadores através da métrica *betweenness*. Podemos notar que o *betweenness* médio do núcleo da comunidade é muito maior, o que significa que um maior número de caminhos mais curtos incluem esses nodos.

Na próxima seção investigamos como aspectos dos membros dos núcleos podem impactar a estrutura geral das comunidades.



(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

3.3. 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. Podemos especular inúmeros fatores que são capazes de explicar essas pequenas variações, incluindo a expansão ou redução do número de artigos publicados, a ascensão e queda de temas com capacidade de atrair pesquisadores ou a perda de membros importantes do núcleo, membros envolvidos na organização de outras conferências, etc. No entanto, 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.

3.4. 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

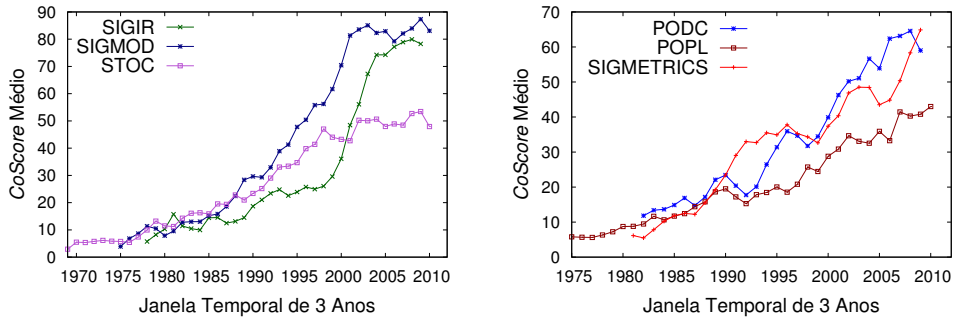
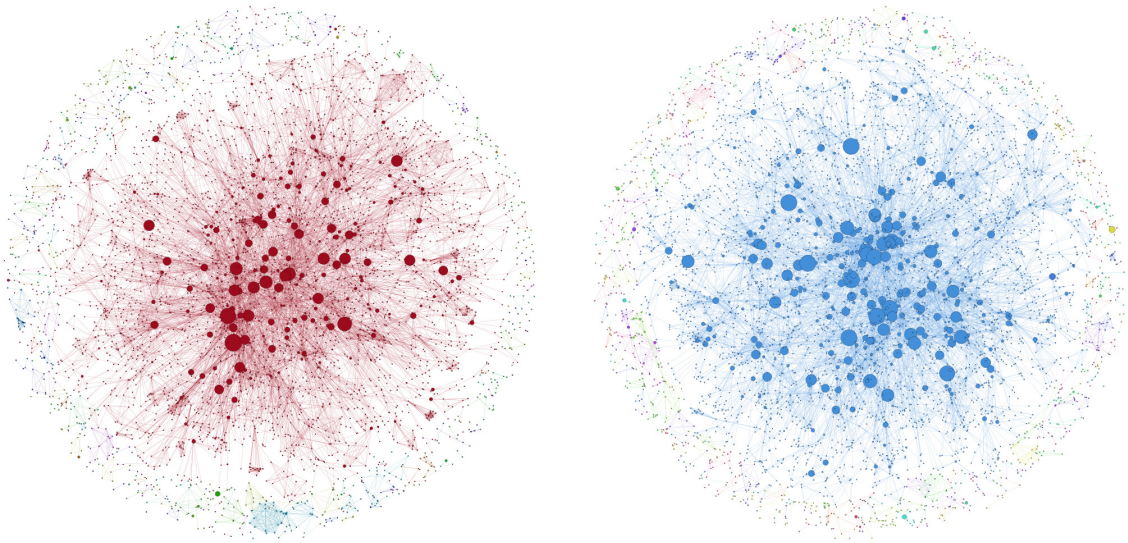


Figura 3. *CoScore* médio das comunidades científicas

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



(a) SIGMOD (b) CHI

Figura 4. Instância final das comunidades científicas

4. 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. Nosso esforço se baseia em estudos anteriores existentes, uma vez que se concentra no núcleo de comunidades, em vez de analisar os aspectos evolutivos de comunidades inteiras. Para isso, definimos um núcleo para as comunidades baseado em uma nova métrica, ou seja, *CoScore*, uma métrica derivada do índice h que captura tanto, a prolificidade, quanto o envolvimento de pesquisadores em uma comunidade. Nossas análises sugerem que membros do núcleo da comunidade atuam como pontes que conectam pequenos grupos de pesquisa. Além disso, observamos que os membros do núcleo

das comunidades tendem a aumentar o grau médio da rede e diminuir a assortatividade. Mais importante, observamos que as 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.** Nossas análises do núcleo das comunidades são aplicáveis a outros contextos, como jogos multijogador massivo, OSNs e repositórios de outras naturezas, como filmes e livros.
- **Utilização de outras métricas de prolificidade.** Existem outras métricas capazes de medir a prolificidade de um pesquisador além do índice h que poderiam também serem utilizadas no cálculo do *CoScore*, como o índice g [Egghe 2006].
- **Avaliação do *CoScore* em outros contextos.** Nossa métrica quantifica a importância dos membros das comunidades. Desta forma, também poderíamos utilizá-la em outros contextos, e.g., para predição de *links* e em sistemas de recomendação.
- **Geração de modelos de formação de comunidades.** O *CoScore* pode ser combinado a outras métricas para mapear o comportamento de como nodos e arestas surgem na rede, possibilitando a geração de modelos de formação de comunidades, conforme resultados prévios apresentados por [Leskovec et al. 2005, Leskovec et al. 2008].
- **Aplicação da abordagem proposta para o estudo de *clusters*.** Vários trabalhos na literatura utilizam abordagens algorítmicas para identificação de *clusters* e de nodos importantes na topologia da rede. No entanto, essas abordagens possuem um custo computacional considerável. Assim sendo, seria interessante aplicar a nossa abordagem em estudos semelhantes.
- **Análise do impacto da migração de pesquisadores entre comunidades.** A migração dos membros do núcleo de uma comunidade pode ser utilizada para prever o sucesso ou declínio dessa comunidade. Sendo assim, nossa métrica poderia ser aplicada em estudos que visem caracterizar a migração de membros de um comunidade ou até inspirar modelos capazes de predizer tais migrações.

Referências

Alves, B. L., Benevenuto, F., and Laender, A. H. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22nd International Conference on World Wide Web (Companion Volume)*, pages 649–656, Rio de Janeiro, Brazil.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, Philadelphia, PA, USA.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington DC, USA.

Egghe, L. (2006). An Improvement of the H-index: The G-index. *ISSI Newsletter*, pages 8–9.

Getoor, L. and Diehl, C. P. (2005). Link Mining: A Survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

Hirsch, J. E. (2005). An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the Spread of Influence through a Social Network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, Washington, D.C.

Kempe, D., Kleinberg, J., and Tardos, E. (2005). Influential Nodes in a Diffusion Model for Social Networks. In *Proceedings of the 32nd International Conference on Automata, Languages and Programming*, pages 1127–1138, Lisbon, Portugal.

Kleinberg, J. (2008). The Convergence of Social and Technological Networks. *Communications of the ACM*, 51(11):66–72.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic Evolution of Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470, Las Vegas, Nevada, USA.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187, Chicago, Illinois, USA.

Leskovec, J., Lang, K. J., and Mahoney, M. (2010). Empirical Comparison of Algorithms for Network Community Detection. In *Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, Raleigh, North Carolina, USA.

Ley, M. (2009). Dblp: Some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500.

Rogers, E. M. (1962). *Diffusion of Innovations*. Free Press.

Seifi, M. and Guillaume, J.-L. (2012). Community Cores in Evolving Networks. In *Proceedings of the 21st International Conference on World Wide Web (Companion Volume)*, pages 1173–1180, Lyon, France.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 37–42, Barcelona, Spain.

Watts, D. and Dodds, P. (2007). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458.