A Força dos Relacionamentos pode Medir a Qualidade de Comunidades?

Mariana O. Silva, Michele A. Brandão, Mirella M. Moro

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{mariana.santos,micheleabrandao,mirella}@dcc.ufmg.br

Abstract. *In social networks, community detection provides valuable data about relationships between individuals. There are various metrics to validate the quality of communities, but there is no consensus on the performance of these metrics. In this paper, we evaluate whether strength metrics can also be used to measure the quality of algorithms that detect communities. The results are positive to confirm such hypothesis.*

Resumo. *Em redes sociais, a detecção de comunidades fornece valiosos dados sobre as relações entre indivíduos. Há diversas métricas para validar a qualidade de comunidades, mas não há um consenso sobre o desempenho dessas métricas. Neste artigo, avaliamos se métricas para força dos relacionamentos podem também ser usadas para medir a qualidade de algoritmos que detectam comunidades. Os resultados são positivos para confirmar tal hipótese.*

1. Introdução

Estudos para a análise de interações entre pessoas ou organizações, bem como a detecção de padrões nessas interações, permitem prever o comportamento de uma rede e analisar diferentes aspectos da mesma. No contexto acadêmico, uma rede social de co-autoria pode ser representada por colaborações científicas que possuem padrões e características relacionados às interações dos indivíduos envolvidos. Tais indivíduos são representados por pesquisadores e os vínculos relacionais, as colaborações entre eles.

Em tais redes, existem grupos de pessoas que possuem relacionamentos mais fortes e compartilham interesses semelhantes [Brandão and Moro 2017] e são chamados de comunidades (agrupamentos ou *clusters*). A detecção de tais comunidades tornou-se um problema fundamental [Yang et al. 2016], podendo ajudar a obter valiosos dados sobre a existência de grupos que colaboram mais densamente, a identificação de relacionamentos mais intensos entre determinados autores ou ainda autores com maior grau de colaboração [Procópio et al. 2011]. Uma vez que este problema não possui uma solução exata, muitas heurísticas foram propostas para encontrar *clusters*. No entanto, não há garantia formal de que os resultados obtidos através deles sejam os melhores possíveis [Almeida et al. 2012].

Um dos aspectos mais importantes do processo de identificação de comunidades é a avaliação da qualidade dos algoritmos que as detectam. É fundamental não só para medir a eficiência dos algoritmos de agrupamento, mas também para prover uma visão sobre a dinâmica de relacionamentos em uma determinada rede. Muitas métricas para validar a qualidade de comunidades já foram propostas na literatura, mas não há um consenso sobre como elas se comparam e quão bons são seus desempenhos [Almeida et al. 2012].

Nesse contexto, este trabalho realiza uma análise comparativa entre métricas para a força dos relacionamentos em redes de co-autoria e métricas usadas para validar comunidades. Especificamente, utilizamos tais métricas para avaliar a qualidade de comunidades detectadas por três algoritmos comumente aplicados em grafos não direcionados [Brandão and Moro 2017]: *Louvain Method* (LM) [Blondel et al. 2008], *Clique Percolation Method* (CPM) [Palla et al. 2005] e *Markov Cluster Algorithm* (MCL) [Van Dongen 2000].

Após apresentar os trabalhos relacionados (Seção 2), descrevemos a metodologia para realização deste trabalho (Seção 3). Em seguida, apresentamos os resultados obtidos (Seção 4). Finalmente, discutimos as principais conclusões (Seção 5).

2. Trabalhos Relacionados

O processo de agrupamento (*clustering*) tem sido aplicado em diversos campos, incluindo Engenharias, Ciência da Computação, Ciências Médicas e Economia [Xu and Wunsch 2005]. Na área de ciência da computação, em especial, essas técnicas de agrupamento são utilizadas para detectar comunidades em redes sociais. Este processo é um problema fundamental, uma vez que permite a análise das interações e relacionamentos entre os participantes das redes. Especialmente em redes sociais acadêmicas, a detecção de comunidades auxilia na descoberta de padrões que podem aumentar a produtividade dos pesquisadores bem como entender a formação de grupos.

Identificar comunidades em redes sociais de co-autoria é geralmente uma tarefa difícil, sendo necessário utilizar ferramentas para detectar e entender o comportamento das mesmas. De acordo com Mishra et al. [2007], modularidade, mutualidade, acessibilidade e cadeias de Markov, são exemplos de estratégias para detectar comunidades em redes sociais. Diante da variedade de técnicas, aplicamos três comumente utilizadas em grafos não direcionados, o *Louvain Method* que é baseado em modularidade [Blondel et al. 2008], *Clique Percolation Method* que considera o conceito de mutualidade [Palla et al. 2005] e *Markov Cluster Algorithm* que utiliza cadeias de Markov [Van Dongen 2000].

Para validar a qualidade das comunidades detectadas pelos algoritmos de agrupamento escolhidos, utilizamos as métricas *neighborhood_overlap* e *co-authorship frequency*. No entanto, de acordo com Brandão e Moro [2017], considerar apenas métricas para a força dos relacionamentos de uma rede não é suficiente para definir a qualidade de algoritmos de clusterização. Portanto, aplicamos também métricas que são comumente utilizadas para avaliar a qualidade de comunidades. Nosso estudo investiga se métricas para a força dos relacionamentos podem medir a qualidade das comunidades de forma adequada.

3. Metodologia

A metodologia deste trabalho possui cinco passos: inicialmente, criamos as redes sociais de co-autoria para analisar a detecção de comunidades; em seguida, aplicamos duas métricas para a força dos relacionamentos dos pesquisadores; então, utilizamos três algoritmos de agrupamento para detectar comunidades; finalmente, validamos a qualidade dos grupos através de três índices e analisamos os resultados das diferentes métricas.

Redes Sociais de Co-autoria. Utilizamos um conjunto de dados coletado da DBLP[1] com aproximadamente 15 milhões de registros para construir as redes sociais. Estes

[1]DBLP *Digital Bibliography & Library Project*): `http://dblp.uni-trier.de/`

Tabela 1: Descrição das redes sociais criadas.

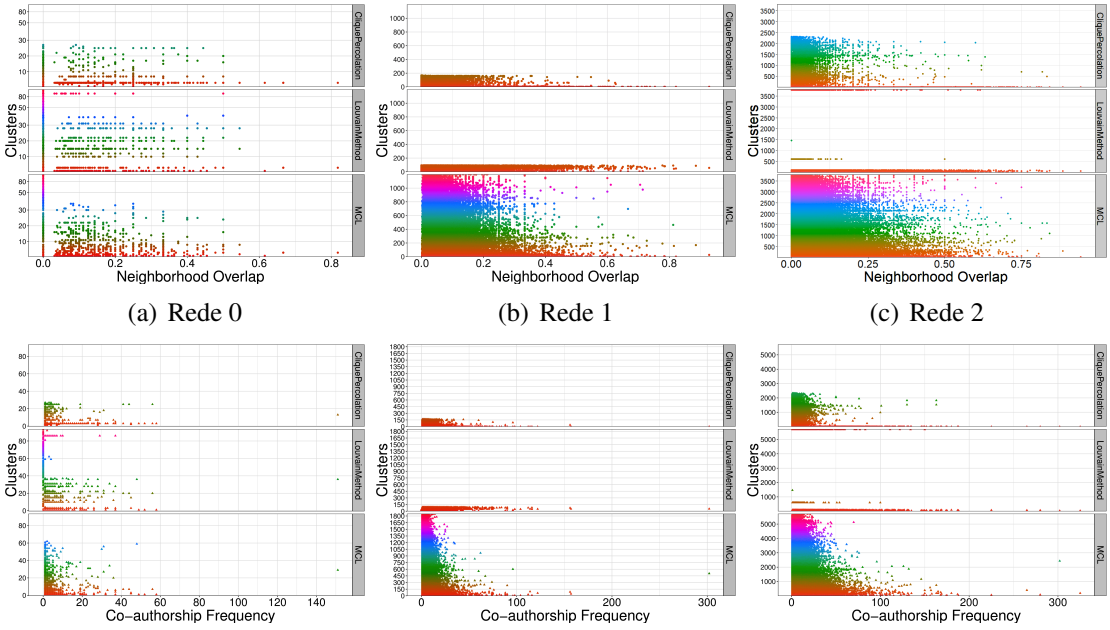
Rede	# autores	# pares (# dist)	MedPubA	Modularidade	Coeficiente de clusterização médio
0	394	3898 (738)	9,89	0,686	0,377
1	68.397	249.352 (110.357)	3,65	0,86	0,672
2	314.444	1.297.929 (540.571)	4,13	0,691	0,524

dados produziram uma rede social extremamente grande, tornando a análise de rede muito trabalhosa. Diante deste problema, utilizamos a técnica de amostragem não probabilística conhecida como *snowball sampling* [Goodman 1961]. A partir de uma rede origem formada pelos bolsistas vigentes (Abril de 2017) de produtividade do CNPq[2] da área de Ciência da Computação, foram feitas mais duas coletas para aumentar a amostra e criar três redes reais de tamanhos diferentes. Ao final, são três redes criadas a partir da DBLP: (0) formada apenas pelos bolsistas vigentes do CNPq (que fazem parte da DBLP); (1) formada pela rede 0 e seus vizinhos; e (2) formada pela rede 1 e seus vizinhos. Dessa forma, foi possível analisar como o tamanho das redes influencia na formação e validação de comunidades. A Tabela 1 apresenta os números de pesquisadores e de publicações, número médio de publicações por autor, número de pares de co-autores (e distintos), modularidade da rede e o coeficiente de clusterização médio de cada rede.

Configuração da Análise. Para a análise comparativa, as redes sociais de co-autoria foram modeladas como um grafo ponderado não dirigido $G^w = (V, E^w)$, onde V é o conjunto de nós e E^w o conjunto de arestas ponderadas. Os nós representam todos os pesquisadores na rede, enquanto as arestas representam a co-autoria em publicações. Além disso, o peso da aresta representa o número absoluto de publicações entre eles, chamado de *co-authorship frequency* (frequência de co-autoria). Para analisar as comunidades, aplicamos os três algoritmos de agrupamento: *Louvain Method* (LM), *Clique Percolation Method* (CPM) e *Markov Cluster Algorithm* (MCL). O método de Louvain é um dos algoritmos de *clustering* mais utilizados e baseia-se na otimização da modularidade de uma partição de rede. O segundo método é o Clique Percolation que é capaz de detectar comunidades com sobreposição, ou seja, os nós podem pertencer a mais de uma comunidade. Finalmente, o Markov Cluster detecta *clusters* alternando dois processos de Markov: expansão e inflação.

Para medir a força dos relacionamentos, utilizamos as métricas *co-authorship frequency* e *neighborhood_overlap* (sobreposição de vizinhança). Dados dois pesquisadores v_i e v_j, *neighborhood_overlap* é definida pela equação: $\frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j) - \{v_i, v_j\}|}$, onde $N(v_i)$ representa os co-autores do pesquisador v_i, e $N(v_j)$ os co-autores de v_j. Essa propriedade topológica calcula a colaboração entre dois nós em relação ao seus vizinhos. Seguindo [Brandão and Moro 2015], consideramos que um relacionamento é fraco quando o valor da métrica *neighborhood_overlap* está no intervalo $[0; 0.2]$ e forte, caso contrário. Da mesma forma, um laço é fraco quando a frequência de co-autoria está no intervalo $[1; 5]$ e forte, caso contrário. Para verificar se métricas para a força dos relacionamentos podem ser usadas para avaliar a qualidade do agrupamento, comparamos seus resultados com três métricas de validação interna: *C-index* (C), *Dunn Index* (Dunn) e *Davies Bouldin Index* (DB). Tais métricas utilizam noções de similaridade intra-comunidade, em contraste com as noções de separação inter-comunidades [Zaki and Meira Jr 2014]. Pela definição de

[2]Bolsistas de produtividade do CNPq: `http://cnpq.br/bolsistas-vigentes/`



(a) Rede 0 (b) Rede 1 (c) Rede 2

(d) Rede 0 (e) Rede 1 (f) Rede 2

Figura 1: Resultados dos algoritmos de agrupamento (cada cor representa um *cluster*.

clusters [Blondel et al. 2008; Palla et al. 2005; Van Dongen 2000], os vínculos intra-comunidade devem ser fortes e os vínculos inter-comunidades devem ser fracos. Logo, uma comunidade deve ter a maioria dos pares de pesquisadores classificados como fortes e a maioria dos laços que conectam diferentes grupos devem ser fracos.

4. Resultados

Esta seção compara os resultados das métricas para a força dos relacionamentos com as métricas de validação interna considerando as comunidades detectadas pelos LM, CPM e MCL. A Figura 1 apresenta a comparação das comunidades detectadas por cada algoritmo em relação às duas métricas de força dos relacionamentos, *neighborhood_overlap* e *co-authorship frequency*. Nota-se que o CPM encontra um número menor de comunidades em relação aos demais métodos, na maioria dos casos. Por outro lado, o MCL detecta o maior número de grupos. Ademais, é importante enfatizar que algumas relações fortes foram excluídas no método CPM, pelo fato dessas arestas pertencerem a um 2-clique. Isso ocorre porque na aplicação do método, foi escolhido k=3. Essa limitação pode ter restringido o número de comunidades identificados pelo algoritmo, fazendo com que ele tenha sido o método com menor número de comunidades formadas.

Analisando as principais vantagens de cada algoritmo, observamos que o LM não foi capaz de detectar pequenas comunidades em grandes redes. Por outro lado, o CPM permite a sobreposição de comunidades, sendo uma situação bem rotineira em redes sociais reais (pesquisadores publicam com pesquisadores de outras comunidades). No entanto, pela necessidade da escolha do parâmetro *k*, o método CPM não teve desempenho melhor que o MCL, visto que o MCL não detectou relacionamentos entre pesquisadores de diferentes comunidades nas três redes. Ou seja, uma vez que o objetivo de um algoritmo de agrupamento é maximizar o número de arestas intra-comunidade e minimizar as inter-comunidade, este foi o algoritmo que forneceu o melhor resultado.

Tabela 2: Resultados das métricas para a força dos relacionamentos.

	Neighborhood Overlap			Co-authorship Frequency		
Rede	**1º**	**2º**	**3º**	**1º**	**2º**	**3º**
0	MCL	LM	CPM	MCL	LM	CPM
1	MCL	CPM	LM	MCL	CPM	LM
2	MCL	CPM	LM	MCL	CPM	LM

Tabela 3: Resultados dos índices de validação interna.

		Neighborhood Overlap			Co-authorship Frequency		
Índice	**Rede**	**1º**	**2º**	**3º**	**1º**	**2º**	**3º**
C	**0**	MCL	LM	CPM	MCL	LM	CPM
C	**1**	-	-	-	-	-	-
C	**2**	-	-	-	-	-	-
Dunn	**0**	LM	CPM	MCL	LM	CPM	MCL
Dunn	**1**	MCL/LM	CPM	-	MCL/LM	CPM	-
Dunn	**2**	-	-	-	-	-	-
DB	**0**	LM/MCL	CPM	-	LM/MCL	CPM	-
DB	**1**	CPM	MCL	LM	CPM	MCL	LM
DB	**2**	-	-	-	-	-	-

Tabela 4: Resultados dos diferentes tipos de métricas.

Rede	**NO**	**C-index**	**Dunn Index**	**DB Index**	**CF**	**C-index**	**Dunn Index**	**DB Index**
0	MCL	MCL	MCL	LM/MCL	MCL	MCL	LM	LM/MCL
1	MCL	-	MCL/LM	CPM	MCL	-	MCL/LM	CPM
2	MCL	-	-	-	MCL	-	-	-

Em relação ao tamanho das redes, podemos notar que ao aumentar o número de nós conectados, a frequência de co-autoria aumentou menos do que o esperado. Da *Rede 1* para a *Rede 2* (Figuras 1(e) e 1(f), respectivamente), não houve um crescimento significativo no nível de colaboração. Por outro lado, foi nítido o crescimento no valor da sobreposição de vizinhança, à medida que o tamanho das redes aumenta. Além disso, nas Figuras 1(a), 1(b) e 1(c), nota-se que há uma alta concentração de arestas somente até o valor 0.4 em todas as redes. Considerando a frequência de co-autoria, as Figuras 1(d), 1(e) e 1(f) mostram uma alta concentração de arestas com frequência de co-autoria menor do que 40 na *Rede 0* e menor do que 100 nas outras duas redes.

As Tabelas 2, 3 e 4 apresentam os resultados das métricas para a força dos relacionamentos, dos índices de validação interna e a comparação entre as diferentes métricas, respectivamente. Nas duas primeiras tabelas, os resultados foram ordenados de acordo com o desempenho dos métodos de agrupamento (primeiro, segundo e terceiro melhor algoritmo). Na Tabela 2, comparando os resultados das duas métricas, observa-se que os algoritmos apresentaram os mesmos desempenhos independente da métrica analisada. Além disso, o algoritmo que realizou o melhor agrupamento foi o MCL em todos os casos. Em redes maiores, o segundo melhor método foi o CPM, seguido pelo LM. Já na menor rede (*Rede 0*), o LM apresenta o segundo melhor resultado. Isso pode ser explicado pelo fato de que algoritmos de otimização de modularidade, geralmente, apresentam dificuldades em detectar pequenas comunidades em grandes redes.

Na Tabela 3, a ordem de desempenho dos algoritmos qualificados pelas medidas de validação interna é similar aos resultados apresentados na Tabela 2 . Já a Tabela 4 apresenta uma comparação entre as métricas para força dos relacionamentos e as métricas para validação interna. Analisando apenas a *Rede 0*, nota-se que na maioria dos casos as

medidas de validação interna indicam o mesmo método com melhor resultado: o algoritmo MCL. Finalmente, os três algoritmos detectam comunidades com relacionamentos fracos e fortes. No entanto, o MCL mostrou-se capaz de detectar grupos com relacionamentos mais fortes do que fracos. Similarmente, analisando os resultados das métricas de validação interna, o algoritmo MCL apresenta comunidades mais coesas na maioria dos casos.

5. Conclusão

Neste trabalho, aplicamos três algoritmos de agrupamento em três redes sociais de co-autoria criadas a partir de dados coletados da DBLP. Os resultados mostram que o MCL é o melhor algoritmo de agrupamento a ser aplicado em redes sociais de co-autoria quando comparado ao LM e CPM, pois a força dos relacionamentos *inter-cluster* tendem a ser mais fortes nessa técnica do que nas demais. Ademais, os resultados mostram que as métricas para a força de relacionamento, principalmente a *neighborhood_overlap*, podem também auxiliar na avaliação da qualidade de algoritmos que detectam comunidades. Pesquisas futuras incluem realizar a mesma comparação gerando redes sociais de co-autoria sintéticas. Além disso, planejamos aplicar mais métricas para força dos relacionamentos e avaliar mais algoritmos de detecção de comunidades.

Agradecimentos. Trabalho parcialmente financiado por CAPES, CNPq e FAPEMIG.

Referências

Almeida, H., Guedes Neto, D., Meira Jr., W., and Zaki, M. J. (2012). Towards a better quality metric for graph cluster evaluation. *JIDM*, 3(3):378.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008.

Brandão, M. A. and Moro, M. M. (2015). Analyzing the strength of co-authorship ties with neighborhood overlap. In *DEXA*, pages 527–542, Valencia, Espanha.

Brandão, M. A. and Moro, M. M. (2017). A comparative analysis of the strength of co-authorship ties in clusters. In *AMW*, Montevideo, Uruguai.

Goodman, L. A. (1961). Snowball sampling. *Ann. Math. Statist.*, 32(1):148–170.

Mishra et al., N. (2007). Clustering social networks. In *WAW*, pages 56–67, S.Diego, USA.

Palla et al., G. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Procópio, P., Laender, A. H., and Moro, M. M. (2011). Análise da rede de coautoria do simpósio brasileiro de bancos de dados. In *SBBD Short Papers*, Florianópolis, Brasil.

Van Dongen, S. M. (2000). *Graph clustering by flow simulation*. PhD thesis, University of Utrecht.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(30750).

Zaki, M. J. and Meira Jr, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.