josif 2022
14ª JORNADA CIENTÍFICA E TECNOLÓGICA
11º SIMPÓSIO DE PÓS-GRADUAÇÃO
IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

ISSN: 2319-0124

# AVALIAÇÃO DE MÉTRICAS DE QUALIDADE DE COMUNIDADES OBTIDAS A PARTIR DE REDES COMPLEXAS

**Amanda C. E. SOUZA[1]; Diego SAQUI[2]**

## RESUMO

Uma rede complexa é formada por um grafo composto por nós interligados com arestas que pode ser usada para representar perspectivas do mundo real e solucionar problemas. Dentro do estudo de redes complexas temos a análise de comunidades que consiste na divisão das redes em grupos de nós que são semelhantes entre si, mas diferentes do restante da rede fornecendo informações importantes sobre a organização e comportamento dessa rede. Este trabalho consiste na aplicação das métricas de qualidade de comunidades obtidas a partir de redes complexas chamadas de *Community Score* e *Community Fitness* de custo computacional menor e a avaliação da correlação entre tais métricas com a Modularidade, com alto custo computacional. A vantagem da identificação dessa correlação possibilita a adequação ou criação de novos métodos e algoritmos de detecção de comunidades de custo computacional menor. Foram executados diferentes algoritmos de detecção de comunidades em diferentes redes complexas, depois foram aplicadas métricas de qualidade nas diferentes comunidades obtidas e então a correlação entre os valores foi analisada.

**Palavras-chave:**
Grafos; Redes Complexas; Métricas de Avaliação; Comunidades.

## 1. INTRODUÇÃO

A Internet, redes biológicas, redes de comunicação e transporte, redes sociais são exemplos de redes complexas (CORDASCO; GARGANO, 2010). Cada uma dessas redes consiste em um conjunto de nós ou vértices representando, por exemplo, computadores na Internet ou pessoas em uma rede social, conectados entre si por links ou arestas, representando conexões de dados entre computadores, amizades entre pessoas, etc (CLAUSET; NEWMAN; MOORE, 2004).

Dado uma rede complexa de $V$ vértices e $E$ arestas, pode-se representá-la por um grafo $G=(V, E)$ e por meio dessa estrutura realizar diversos estudos.. Uma propriedade interessante dessas redes, é a estrutura da comunidade, ou seja, a divisão das redes em grupos de nós que são semelhantes entre si, mas diferentes do resto da rede (CORDASCO; GARGANO, 2010). Logo, um dos desafios do estudo dessas redes é calcular e classificar a compatibilidade de seus vértices e arestas buscando a detecção de possíveis comunidades, bem como avaliar sua qualidade.

Os métodos disponíveis para a detecção de comunidades se atentam às conexões da rede e não somente a distância ou semelhança de cada elemento, bem como elaboram as comunidades sem definir uma contagem pré definida de agrupamentos (FORTUNATO, 2010). Entre os principais

[1] Bolsista FAPEMIG, IFSULDEMINAS – *Campus* Muzambinho. E-mail: amandaeleuterio2@gmail.com.
[2] Orientador, IFSULDEMINAS – *Campus* Muzambinho. E-mail: diego.saqui@muz.ifsuldeminas.edu.br.

algoritmos para identificar comunidades são conhecidos o Girvan Newman (GN), o Clauset-Newman-Moore (CNM) e o Label Propagation (LP)[3]. O GN detecta comunidades removendo a aresta de maior centralidade em cada etapa do grafo inicial. Posteriormente, o resultado é retratado por um dendrograma onde pode-se definir a quantidade de comunidades desejadas. O CNM usa a maximização da modularidade gulosa para encontrar as comunidades, inicialmente os nós são estabelecidos em sua respectiva comunidade e posteriormente, os conjuntos que levam a maior modularidade serão separados em duplas de forma repetitiva, até que nenhum aumento adicional na modularidade seja possível. Por fim, o LP usa uma estratégia responsável por propagar rótulos por toda a rede formando comunidades com base nesse processo de propagação. A cada iteração de propagação, cada nó atualiza seu rótulo para aquele ao qual o número máximo de seus vizinhos pertencem e são considerados da mesma comunidade.

A modularidade usada no CNM, devido a promoção de bons resultados, é bastante comum em diversos algoritmos, porém, como sua aplicação requer um custo computacional elevado pode ser inviável para redes de grande porte (BABAK, et al. 2013). Neste contexto, objetiva-se avaliar se outras métricas de custo computacional menor, chamadas de Community Score (CS) e Community Fitness (CF), apresentam boa correlação com a modularidade. Caso isso aconteça, algoritmos tão bons quanto os que utilizam a modularidade podem ser projetados com tais métricas.

## 2. MATERIAL E MÉTODOS

Inicialmente realizou-se uma análise da literatura, onde selecionou-se o artigo de Babak, et al. (2013) com o intuito de replicar duas métricas de qualidade de comunidades: o CS (eq. 1), que mede a densidade das comunidades obtidas e o CF (eq. 2) que minimiza os links externos. Para o desenvolvimento das métricas foi considerado um grafo não direcionado *G=(V, E)* que conecta dois vértices *V* através de uma aresta *E*. *G* pode ser representado por uma matriz de adjacência *A,* onde se houver uma aresta de *vi* a *vj* então *Aij=1* caso contrário *Aij=0*.

$$cs = \sum_{i=1}^{k} score(C_i) \quad (1)$$

Para encontrarmos a $CS$, devemos calcular a média de potência de $C$ de ordem $r$, denotada por $M(C) = \frac{\sum i \epsilon C (\mu i)^r}{|C|}$, sendo $C$ um subgrafo de $G$ $(C \subset G)$, $\mu i = \frac{1}{|C|} k_i^{in}(C)$ a fração de arestas que conectam o nó $i$ aos outros nós em $C$, onde $|C|$ é a cardinalidade de $C$ e $k_i^{in}(C) = \sum j \epsilon C A_{ij}$ o número de arestas que conectam $i$ aos outros nós em $C$. Igualando-se a $0 \leq \mu i \leq 1$ o expoente $r$

[3] https://networkx.org/documentation/stable/reference/algorithms/community.html. Acesso em Julho/2022

aumenta os pesos dos nós com muitas conexões com outros nós pertencentes ao mesmo módulo e diminui o peso daqueles com poucas conexões dentro de *C*. Em seguida calculamos o volume $vc$ de uma comunidade $C$ a partir da equação $vc = \sum i, j \epsilon C\, A_{ij}$ no qual é definido como o número de arestas conectando vértices dentro de $C$, ou seja o número 1 na matriz de adjacência $A$ corresponde a $C$. Com isso, chegamos a função interna para calcular a $CS$, chamada $score(C) = M(C) \times vc$.

$$P(C) = \sum_{i \epsilon S} \frac{k_i^{in}(C)}{(k_i^{in}(C) + k_i^{out}(C))^{\alpha}} \quad (2)$$

Para a $CF$ utilizamos a $P(C)$, no qual $k_i^{in}(C) = \sum j \epsilon C A_{ij}$ e $k_i^{out}(C) = \sum j \notin C A_{ij}$ corresponde aos graus internos e externos dos nós pertencentes à comunidade *C*, e o parâmetro de valor real e positivo α controla o tamanho das comunidades (BABAK, et al., 2013).

| | Graphs | community_score | community_fitness_pc | modularity |
|---|---|---|---|---|
| 0 | girvan_newman + karate_club_graph | 31.666310 | 17.765686 | NaN |
| 1 | girvan_newman + complete_graph | 9.000000 | 3.000000 | NaN |
| 2 | girvan_newman + circular_ladder_graph | 9.444444 | 4.666667 | NaN |
| 3 | girvan_newman + dorogovtsev_goltsev_mendes_graph | 19.528345 | 19.375000 | NaN |
| 4 | girvan_newman + ladder_graph | 10.888889 | 5.333333 | NaN |
| 5 | girvan_newman + multilayered_graph | 75.643414 | 12.142857 | NaN |
| 6 | label_propagation_communities + karate_club_graph | 31.111111 | 2.000000 | 0.325115 |
| 7 | label_propagation_communities + complete_graph | 16.000000 | 5.000000 | 0.000000 |
| 8 | label_propagation_communities + circular_ladde... | 9.000000 | 2.666667 | 0.240000 |
| 9 | label_propagation_communities + dorogovtsev_go... | 26.632868 | 1.500000 | 0.367627 |
| 10 | label_propagation_communities + ladder_graph | 12.000000 | 3.333333 | 0.414062 |
| 11 | label_propagation_communities + multilayered_g... | 75.643414 | 12.142857 | 0.416896 |

**Figura 1.** *Dataframe* com as métricas CS, CF e Modularidade para diferentes grafos e métodos.

Foram implementadas as funções CS e CF nos algoritmos de detecção de comunidades GN e LP e em diferentes *datasets* usando o Colaboratory do Google que permite a escrita e execução de códigos em Python e, posteriormente, os resultados encontrados foram inseridos em um *dataframe* da biblioteca Pandas como mostrado na Fig. 1. A função modularity disponibilizada na biblioteca networkx é calculada utilizando o LP, portanto, na coluna modularity possui valores NaN pois não é possível calcular a modularidade com GN.

## 3. RESULTADOS E DISCUSSÕES

Para avaliar as funções CS e CF e evidenciar qual delas melhor se correlaciona com a Modularidade, foi utilizado a função .corr() da biblioteca Pandas que calcula a correlação de colunas em pares, excluindo valores NA/nulos e retorna a matriz de correlação que foi estruturada

em um mapa de calor representado na Fig. 2. Na correlação tem-se os limites de -1 e 1 que indicam relações mais fortes entre as duas variáveis. Observando a correlação entre *Community_Score* (CS), *community_fitness_pc* (CF) e Modularidade concluiu-se que a função que mais se aproximou com a Modularidade mesmo sendo uma correlação baixa é a CF.

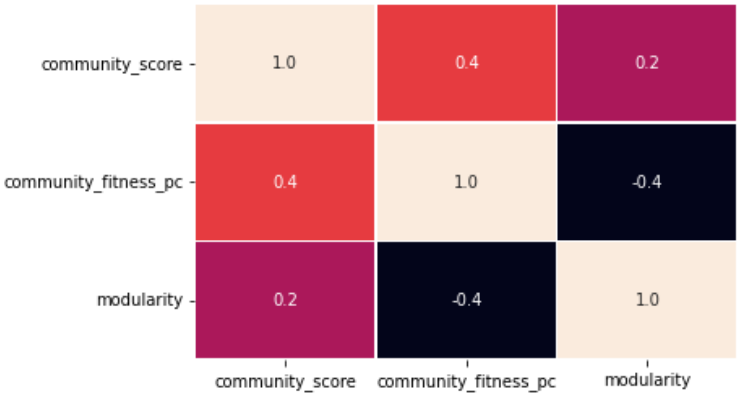


**Figura 2.** Mapa de Calor da Matriz de Correlações.

## 4. CONCLUSÕES

Este projeto possibilitou a avaliação de métricas de qualidade de comunidades de redes complexas e comparação com a modularidade. Algoritmos projetados para encontrar comunidades utilizando modularidade possuem alto custo computacional, portanto, métricas com custo computacional baixo e boa correlação com a modularidade favorecem a criação de bons algoritmos de detecção de comunidades que podem ser aplicados a redes complexas maiores. Neste estudo observou-se que a métrica CF foi a que mais se aproximou da modularidade mesmo obtendo baixa correlação, em novas pesquisas outras métricas serão estudadas tentando obter melhor correlação.

## AGRADECIMENTOS

Agradecemos à FAPEMIG, ao LabSoft e ao IFSULDEMINAS- Campus Muzambinho pela oportunidade e estrutura concedidas para realização dessa pesquisa.

## REFERÊNCIAS

BABAK, A., et al. **Multiobjective enhanced firefly algorithm for community detection in complex networks**. Knowledge-Based Systems, 2013, 46: 1-11.

CLAUSET, Aaron; NEWMAN, Mark EJ; MOORE, Cristopher. **Finding community structure in very large networks**. Physical review E, 2004, 70.6: 066111.

CORDASCO, Gennaro; GARGANO, Luisa. **Community detection via semi-synchronous label propagation algorithms.** In: 2010 IEEE international workshop on: business applications of social network analysis (BASNA). IEEE, 2010. p. 1-8.

FORTUNATO, S. **Community detection in graphs**. Physics reports, Elsevier, v. 486, n. 3, p. 75Ű174, 2010. Citado 5 vezes nas páginas 11, 12, 16, 17 e 26.