GA-LP: Um algoritmo genético baseado no *Label Propagation* para detecção de comunidades em redes direcionadas

Rodrigo Francisquini,Valério Rosset e Mariá Cristina Vasconcelos Nascimento
Universidade Federal de São Paulo
Av. Cesare M. Giulio Lattes, 1201 - Eug. Melo - São José Dos Campos
francisquini.r@gmail.com, vrosset@unifesp.br, mcv.nascimento@unifesp.br

RESUMO

Neste trabalho, o aluno de Iniciação Científica, Rodrigo Franciquini, sob orientação dos coautores, desenvolveu um algoritmo genético para detecção de comunidades em redes direcionadas, otimizando a medida de modularidade para redes direcionadas. Após uma análise de estruturas de rotulação de vértices sugeridas, identificaram-se os melhores operadores genéticos para a solução do problema. Como a maioria dos métodos existentes utiliza informações globais sobre a rede, esses exigem um custo computacional muito alto para redes de larga escala. Além disso, detectar comunidades considerando o significado de relações assimétricas entre os vértices da rede é conhecidamente um grande desafio. Após revisão de trabalhos correlatos e estudo de codificação dos indivíduos, definiu-se um algoritmo baseado no *Label Propagation*, adaptado para redes direcionadas. A estratégia alcançou resultados excelentes, sendo competitiva com os principais algoritmos da literatura e alcançando os melhores resultados nas redes de larga escala testadas. Esse trabalho foi publicado em periódico internacional.

PALAVRAS CHAVE. Detecção de Comunidades em Redes Direcionadas; Maximização da Modularidade.

Área Principal: TAG - Teoria e Algoritmos em Grafos, MH - Metaheurísticas.

ABSTRACT

In this project, Rodrigo Franciquini, advised by the coauthors, developed a genetic algorithm for the community detection problem in directed networks, by optimizing the modularity designed for directed networks. After an analysis of suggested vertex labeling structures, the best genetic operators were identified for the solution of the problem. Most of the existing methods uses global information of the network, requiring a very high computational cost for large scale networks. In addition, detecting communities considering the significance of asymmetric relationships between network vertices is known to be a big challenge. After a review of related works and a study about the individuals decoding, we defined an algorithm based on Label Propagation, adapted for directed networks. The proposed strategy showed outstanding results, being competitive with the main algorithms of the literature and achieving the best results in the large scale networks tested. This study was published in an international journal.

KEYWORDS. Clustering in directed networks, Modularity maximization.

Main Area: TAG - Theory and Algorithms in Graphs, MH - Metaheuristics.

1. Introdução

Na literatura, um número considerável de algoritmos foram desenvolvidos para detectar comunidades (também conhecidas como *clusters*) em redes. Nesse caso, esses algoritmos têm por objetivo encontrar comunidades com vértices que possuam alta conectividade entre si, em comparação com vértices de outros grupos. Diversos algoritmos são baseados na otimização de medidas de avaliação de partições como a modularidade e a *map equation* [Newman e Girvan, 2004; Rosvall e Bergstrom, 2008; Malliaros e Vazirgiannis, 2013]. Brandes et al. [2008] mostraram que o problema de maximização da modularidade é NP-completo e como as aplicações envolvem grafos de, no mínimo, centenas de vértices e arestas, a forma mais eficiente de resolvê-lo é por meio de heurísticas.

Encontrar uma medida que avalie efetivamente a qualidade das comunidades é um grande desafio. No caso das redes direcionadas, o desafio é ainda maior. Redes direcionadas são caracterizadas por relações assimétricas entre os vértices, e essa característica deve ser considerada na análise dessa classe de redes. Nesse contexto, surge a necessidade de se desenvolver algoritmos especialmente para redes direcionadas, e algumas das métricas de avaliação, como a modularidade em redes direcionadas [Leicht e Newman, 2008] devem ser mais exploradas.

Este trabalho de Iniciação Científica (IC) apresenta uma estratégia concebida para detectar comunidades em redes direcionadas, distribuídas e de larga escala. Batizado de GA-LP, o algoritmo é baseado no *Label Propagation* (LP) [Raghavan et al., 2007], adaptado para redes direcionadas e refinado com operadores genéticos. Experimentos computacionais foram realizados com o objetivo de comparar os resultados do GA-LP com algoritmos da literatura, como Infomap [Rosvall e Bergstrom, 2008], LP e *Order Statistics Local Optimization Method* (OSLOM) [Lancichinetti et al., 2011]. Os resultados indicam que o GA-LP é robusto e os seus resultados superaram os de estratégias da literatura. Este estudo originou uma publicação em periódico internacional indexado, referenciado em [Francisquini et al., 2017].

2. Trabalhos Relacionados

Esta seção apresenta brevemente as medidas para avaliação de comunidades em redes direcionadas e algoritmos relacionados ao desenvolvido durante a IC.

Dentre as medidas mais populares para definir comunidades por meio de sua otimização, destacam-se a modularidade, o *map equation* [Rosvall e Bergstrom, 2008] e o *Order Statistics Local Optimization Method* (OSLOM). As versões direcionadas da modularidade [Leicht e Newman, 2008] são apontadas como medidas importantes para definir a função objetivo (de aptidão) de heurísticas desenvolvidas para detectar comunidades em redes direcionadas [Malliaros e Vazirgiannis, 2013]. Essa medida é apresentada na Seção 3 deste artigo e avalia a qualidade de uma partição de uma rede. Já a medida *map equation* é baseada no fluxo de uma dada rede e, no algoritmo proposto pelos autores, ela é usada para captar os padrões de fluxo dentro da rede para detectar comunidades. Algoritmos que otimizam essa medida, como o Infomap, têm apresentado resultados interessantes, em particular, em redes direcionadas. Recentemente proposto, o *Order Statistics Local Optimization Method* (OSLOM) é uma função objetivo apresentada por [Lancichinetti et al., 2011] que expressa a significância estatística das comunidades. Uma estratégia que otimiza essa medida foi proposta em [Lancichinetti et al., 2011].

Segundo Malliaros e Vazirgiannis [2013], poucos estudos e algoritmos podem ser encontrados na literatura especialmente desenvolvidos para o problema de maximização da modularidade em redes direcionadas. Santos et al. [2016], recentemente, otimizaram a modularidade para grafos direcionados por meio de uma estratégia consensual, chamada de *ConClus*, em que os resultados obtidos foram competitivos com os melhores resultados

encontrados na literatura. Porém, *ConClus* requer um elevado tempo computacional para obter as partições finais.

2.1. AGs para o Problema de Detecção de Comunidades em Redes

Algoritmos genéticos (AGs) são estratégias baseadas em conceitos de seleção natural, cruzamento entre indivíduos e operadores de mutação para investigar soluções de problemas em geral e, em particular, problemas de otimização combinatória [Goldberg, 1989]. O uso de AGs para o problema de detecção de comunidades em redes tem produzido resultados interessantes, como brevemente relatado nesta seção.

Shang et al. [2013] e Mu et al. [2015] propõem AGs que objetivam maximizar a modularidade para detectar comunidades em redes. Ambos os estudos consideram o problema de detecção de comunidades em redes não direcionadas e descrevem os indivíduos em um vetor n-dimensional, sendo n o número de vértices do grafo, que armazena em cada posição o rótulo da comunidade à qual aquele vértice pertence. De maneira geral, a maioria dos estudos encontrados na literatura empregam a modularidade como função de aptidão e possuem a mesma representação de cromossomos, que, em algumas situações, não é pertinente em aplicações de larga escala. A ligeira diferença entre esses algoritmos reside nas operações de cruzamento, quando combinam-se pares de soluções (indivíduos). Para atingirem bons resultados, a maioria desses algoritmos são hibridizados com estratégias de busca local, o que implica em um alto custo computacional, também, em se tratando de redes de larga escala. Por conseguinte, esses estudos, em sua maioria, abordam somente redes pequenas, como mostram os experimentos indicados pelos próprios autores. Dessa forma, este trabalho de IC diferencia-se de trabalhos da literatura por propor um algoritmo genético capaz de lidar com redes de larga escala que independe de estratégia de busca local e que define os indivíduos de maneira a melhor representar soluções para redes de larga escala.

3. Algoritmo Proposto: GA-LP

De complexidade quase-linear, o LP [Raghavan et al., 2007] é capaz de detectar comunidades em redes não direcionadas por meio de decisões consensuais entre os vértices da vizinhança e se caracteriza por não otimizar uma função objetivo. O LP é capaz de detectar eficientemente comunidades em redes de larga escala. Entretanto, da forma como foi proposto, o LP não considera a relação assimétrica entre nós, e, portanto, para detectar comunidades em dígrafos, é necessário transformar o dígrafo em grafo não direcionado para que o LP possa ser utilizado sem qualquer adaptação [Malliaros e Vazirgiannis, 2013]. Uma transformação possível é ignorar a direção dos arcos e considerar a rede como não direcionada. No entanto, essa transformação, que não leva em consideração as propriedades de uma rede direcionada, gera comunidades que não garantem as características necessárias, como o fluxo dentro das comunidades da rede. Apesar da perda de informação, essa transformação é amplamente utilizada [Malliaros e Vazirgiannis, 2013].

A função de aptidão do algoritmo proposto é a modularidade para grafos direcionados, cuja formulação, dada uma partição π, é apresentada na Equação (1).

$$q(\pi) = \frac{1}{m} \sum_{\forall \mathcal{C} \in \pi} \sum_{i,j \in \mathcal{C}} (a_{ij} - \frac{d_i^+ d_j^-}{m}) \tag{1}$$

Nessa equação, $\mathcal{C}$ é uma comunidade de π, m é o número de arcos do dígrafo e a_{ij} é o número de arcos com vértice i como origem e j como destino. Nas próximas seções discutem-se os detalhes da algoritmo proposto, nomeado GA-LP. Um pseudocódigo do algoritmo é apresentado no Algoritmo 3.

3.1. Cromossomos e População Inicial

Para evitar um consumo demasiado de memória, a representação dos indivíduos é armazenada nos vértices da rede. Assim, para um determinado indivíduo, o rótulo de seu vértice indica a comunidade à qual ele pertence, um valor inteiro que varia de 1 a n, em que n é o número de vértices do dígrafo cujas comunidades devem ser detectadas. A população inicial do GA-LP é um conjunto de soluções obtidas por meio de uma versão simplificada do algoritmo LP. Ao contrário do LP, que é executado até atingir um consenso entre os vértices vizinhos, para garantir a diversidade da população inicial no GA-LP, execuções independentes com uma única iteração do LP produzem cada indivíduo da população inicial. Além disso, uma adaptação nessa versão foi necessária porque a versão original do LP não considera a direção dos arcos. O pseudocódigo dessa estratégia está no Algoritmo 1.

Algoritmo 1: DIRECTED LP

Data: Um dígrafo conexo G
Result: A partição resultante $\mathcal{C}$
$\mathcal{C} \leftarrow$*Propagar o rótulo id* (G);
Marcar todos os vértices;
FASE DE REFINAMENTO$(G, \mathcal{C})$;

O algoritmo, de complexidade O(n), apresenta uma grande diversidade nas partições resultantes que formarão a população inicial. Para ajuste dos rótulos, é aplicado um algoritmo de refinamento de rótulos.

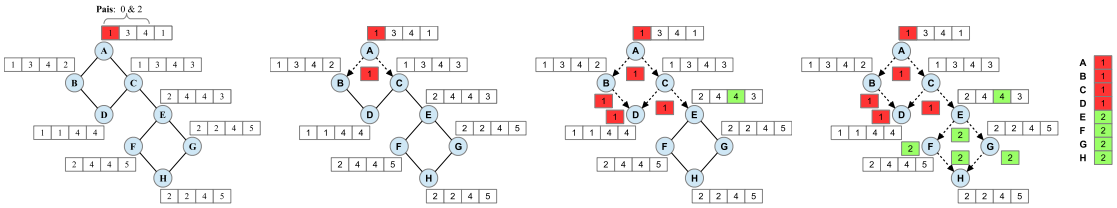
3.2. Algoritmo de Seleção e Operadores Genéticos

GA-LP usa a seleção por roleta que é um método de seleção proporcional a função objetivo, muito empregada em estratégias genéticos. O operador de cruzamento aqui proposto segue uma estratégia local para propagar as comunidades dos pais escolhidos. A sistemática da distribuição dos rótulos entre os vértices está baseada na busca em profundidade (DFS) em que, a partir do primeiro vértice de G, sorteia-se um dos pais para a propagação da comunidade indicada no filho correspondente. A propagação da primeira comunidade ocorre até saturar o rótulo sorteado, ou seja, até todos os vértices do pai que tiverem rótulo sorteado terem sido propagados no filho. Em seguida, dado o primeiro vértice sem rótulo, realiza-se o sorteio de um dos pais para definir quem são os vértices que receberão o segundo rótulo, representados pela comunidade indicada no vértice sem rótulo do pai correspondente. Para ajuste dos rótulos dos filhos, é aplicado um algoritmo de refinamento de rótulos.O Algoritmo 2 apresenta essa estratégia.

Algoritmo 2: PROPAGAR-DFS

Data: Um dígrafo G, um rótulo l, k e o vértice inicial i, o pai replicado $\mathcal{C}$
Result: Um dígrafo conectado G com os vértices marcados e o vértice i
Marcar o vértice i com o valor de k;
forall *vizinhos* n_i *do vértice* i **do**
  if n_i *não foi visitado* **then**
    PROPAGAR-DFS$(G,l,k,n_i, \mathcal{C})$;
    Marcar n_i como visitado;
    if *o rótulo de* n_i *em* $\mathcal{C}$ *é* l *e* n_i *não está marcado* **then**
      Atribuir a n_i o rótulo l;
    end
  end
end

A Figura 1 apresenta um exemplo de cruzamento entre pais de uma população exemplo.



(a) Escolha dos pais. (b) Propagação do rótulo. (c) Escolha aleatória. (d) Propagação.

Figura 1: Uma ilustração do processo de *crossover* dos indivíduos localizados nas posições 0 e 2 da matriz da geração atual. Fonte: Francisquini et al. [2017].

Nesse exemplo, há uma população de 4 indivíduos, cada um representado em uma posição do vetor indicado no vértice. Selecionam-se os cromossomos pais nas posições 0 e 2 dos vetores e a propagação inicia no vértice A. O processo de recombinação seleciona aleatoriamente o pai 0 e propaga o rótulo 1 de acordo com os membros com esse rótulo nesse pai. Então, além de A, os vértices B, C e D recebem o rótulo 1. Em seguida, o vértice E é selecionado aleatoriamente entre os vértices não marcados (E, F, G e H), bem como o pai 2. Dessa forma, a segunda comunidade do filho, representando o rótulo 4 no pai, 2, é propagado aos vértices que são membros dessa comunidade no pai e não foram marcados ainda. Em seguida, o algoritmo para, pois E, F, G e H possuem rótulo 4 no pai 2 e nenhum vértice não marcado permanece.

Algoritmo 3: GA-LP

Data: Um dígrafo G, tamanho da população, p
Result: A partição resultante π
Gere a população C com o Directed LP(G);
repeat
  for `i=0; i<p; i++` **do**
    Desmarcar todos os vértices de G;
    $\mathcal{C}_1, \mathcal{C}_2 \leftarrow$ Sorteio de dois pais distintos em (C);
    O rótulo k do filho recebe 1;
    repeat
      $\mathcal{C}_c \leftarrow$ Sorteio de um dos pais ($\mathcal{C}_1$ ou $\mathcal{C}_2$);
      $l \leftarrow$ o rótulo do vértice j em $\mathcal{C}_c$;
      Marcar todos os vértices como não visitados;
      PROPAGAR-DFS(G,l,k,j,$\mathcal{C}_c$);
      $k \leftarrow k + 1$;
    until *existir um vértice j desmarcado*;
    $nC[i] \leftarrow$ rótulos dos vértices marcados de G;
  end
  Substituir os 60% mais fracos de $\mathcal{C}$ pelos 60% mais fortes de nC;
until *número de gerações sem melhoria for menor que 5*;
Atribuir a π o mais forte de $\mathcal{C}$;

O processo de substituição segue o paradigma de elitismo no qual, a cada nova geração, até 60% dos mais fortes da nova geração substituem os 60% mais fracos da geração atual. A percentagem exata de substituição depende se alguns dos 60% indivíduos mais fortes possuem menos aptidão do que qualquer um dos 60% indivíduos mais fracos da geração atual. Se assim for, os indivíduos mais aptos são mantidos para a próxima geração.

4. Experimentos Computacionais

Dois experimentos computacionais atestam a qualidade do GA-LP em comparação com os melhores algoritmos da literatura: LP, Infomap e OSLOM. Um conjunto de redes

artificiais foi gerado utilizando o software introduzido em [Lancichinetti e Fortunato, 2009], conhecidas como redes LFR. Como o algoritmo GA-LP visa resolver o problema de detecção de comunidades em redes de larga escala, além de gerar grafos com o número de vértices mais habitual nesse tipo de experimento, entre 1000 e 5000 vértices, consideramos também redes com 10000 e 50000 vértices.

O conjunto de redes LFR criado possui 5 redes direcionadas para cada combinação de valor dos parâmetros sugeridos pelos autores. Os valores índice de mistura gerados, μ_t, que definem comunidades mais difíceis de detectar quando esse valor for próximo de 1, foram entre 0.1 e 0.8. Foram realizados experimentos para o refinamento dos parâmetros do algoritmo e testes com outras redes, como redes densas e redes com comunidades maiores. Porém, para que esse artigo não fique extenso, esses experimentos não serão abordados. O trabalho completo pode ser encontrado em [Francisquini et al., 2017].

5. Experimento I

Para avaliar os resultados obtidos por GA-LP com as redes LFR discutidas anteriormente, avaliaram-se as partições obtidas com as esperadas usando a medida *Normalized Mutual Information* (NMI). Quanto mais próximos os resultados são de 1, melhores eles são. As Figuras 2 e 3 mostram os resultados médios (NMI e Tempo) dos algoritmos desse experimento.

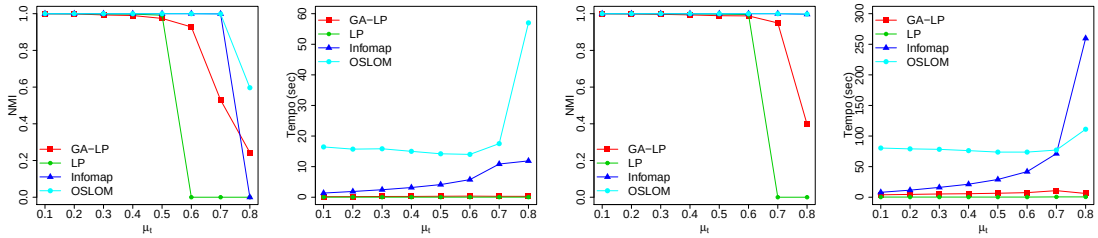


Figura 2: NMI e tempo médio para as redes com 1000 e 5000 vértices, respectivamente.

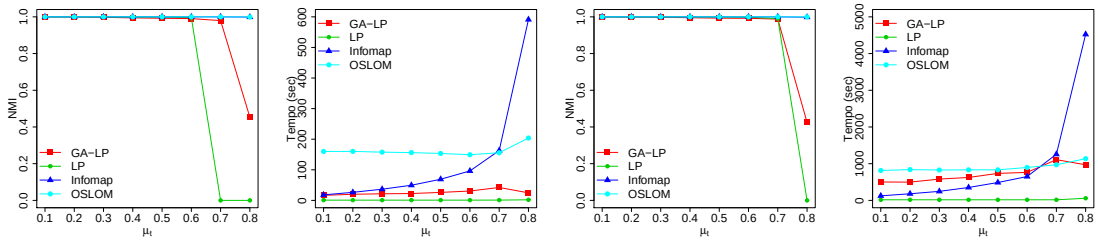


Figura 3: NMI e tempo médio para as redes com 10000 e 50000 vértices, respectivamente.

O objetivo de nossa estratégia foi alcançado com êxito. Ele se refere a estratégia local que refina os resultados do LP. Em todos os estudos de caso, GA-LP superou o LP. No entanto, considerando os algoritmos Infomap e OSLOM, a análise comparativa mostrou que o OSLOM foi o mais robusto e que alcançou os melhores resultados na maioria das redes testadas. O Infomap mostrou um desempenho muito bom, com valores elevados de NMI. Em média, observamos que GA-LP foi muito competitivo com estas duas estratégias, embora não tenha superado elas.

Por um lado, tanto OSLOM quanto Infomap apresentaram um alto tempo computacional. Por outro lado, GA-LP teve tempo médio próximo ao requirido pelo LP, ambos

muito pequenos se comparados aos outros algoritmos. Estes resultados corroboram que GA-LP é uma excelente estratégia para detectar comunidades em redes de larga escala. Em experimentos complementares, testamos GA-LP com redes ainda maiores. Ambos Infomap e OSLOM foram muito lentos na detecção das comunidades nessas redes, sendo muito difícil obter os resultados na máquina utilizada para executar os experimentos. GA-LP e LP apresentaram valores de NMI próximos aos obtidos com as redes de 50000 vértices.

Os resultados obtidos por GA-LP foram muito estáveis, superando as outras estratégias em média em todos as redes testadas. O tempo gasto por GA-LP e LP foram muito próximos e significativamente menor do que as outras estratégias. De acordo com esse experimento, GA-LP é uma estratégia altamente recomendada para detectar comunidades em redes de larga escala. A próxima seção apresenta um experimento realizado usando duas redes reais.

6. Experimento II

Esse experimento considera duas redes reais, uma não direcionada e outra direcionada. A rede não direcionada é a conhecida rede *afootball* [Girvan e Newman, 2002], de 115 vértices e 613 arestas. A partição esperada nessa rede corresponde ao grupo de 12 conferências da temporada de partidas de jogos da divisão I-A de futebol americano realizadas na temporada de 2000. A Figura 4 mostra a rede plotada usando o pacote de visualização *igraph*. O NMI da partição encontrada pelo GA-LP foi 0,91151. LP, Infomap e OSLOM obtiveram comunidades com valores de NMI iguais a $0,91085$, $0,92419$ e $0,91568$, respectivamente. Esse resultado mostra que o GA-LP ainda é eficiente detectando comunidades em redes não direcionadas.

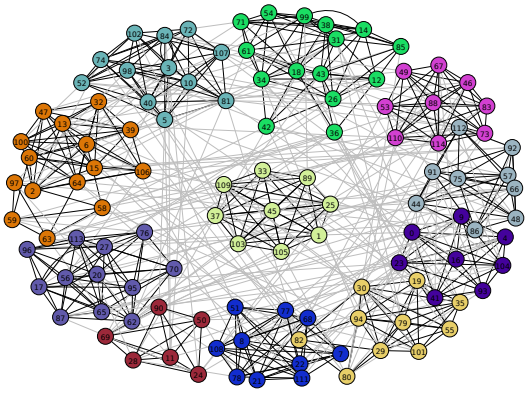


Figura 4: A rede não direcionada *afooball* particionada de acordo com as comunidades encontradas pelo GA-LP.

A segunda rede testada foi a rede *polBlogs*, uma representação das conexões entre 1490 blogs que discutem políticas americanas, com 18910 arcos. A partição esperada corresponde a classificação dos blogs em liberais e conservadores. O NMI das comunidades encontradas pelo GA-LP foi 0,6889. OSLOM, Infomap e LP obtiveram partições com NMI igual a $0,57208$, $0,43547$ e $0,38534$, respectivamente. Esse resultado demonstra que o GA-LP encontrou uma partição mais próxima da esperada.

7. Conclusões

Este projeto propôs um algoritmo genético local, chamado de GA-LP, para detectar comunidades em redes direcionadas de larga escala. Ele é baseado no algoritmo *Label Propagation* (LP), conhecido por ser uma estratégia para identificar comunidades em redes de larga escala. Porém, como no LP a propagação dos rótulos é baseada na frequência

entre os vizinhos, é muito difícil para o LP detectar comunidades em redes densas. Nessas redes, é muito comum encontrar uma única comunidade ou comunidades muito grandes, sendo mais provável a perda de informação decorrente da detecção de comunidades fracas. Mais robusto que o LP e com o mesmo custo computacional que o LP, GA-LP foi muito competitivo em comparação com os melhores algoritmos de detecção de comunidades em redes direcionadas.

Agradecimentos

À FAPESP (Proc.: 15/21660-4 e 15/18580-9) e ao CNPq (Proc.: 448614/2014-6).

Referências

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., e Wagner, D. (2008). On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172–188.

Francisquini, R., Rosset, V., e Nascimento, M. C. (2017). GA-LP: A genetic algorithm based on label propagation to detect communities in directed networks. *Expert Systems with Applications*, 74:127–138.

Girvan, M. e Newman, M. (2002). Community structure in social and biological networks. *National Academy of Sciences*, (12):7821–7826.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition. ISBN 0201157675.

Lancichinetti, A. e Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, p. 016118.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., Fortunato, S., et al. (2011). Finding statistically significant communities in networks. *PloS One*, 6(4):e18961.

Leicht, E. A. e Newman, M. E. J. (2008). Community structure in directed networks. *Physical Review Letter*, 100:118703.

Malliaros, F. D. e Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(95-142).

Mu, C.-H., Xie, J., Liu, Y., Chen, F., Liu, Y., e Jiao, L.-C. (2015). Memetic algorithm with simulated annealing strategy and tightness greedy optimization for community detection in networks. *Applied Soft Computing*, 34:485 – 501. ISSN 1568-4946.

Newman, M. E. J. e Girvan, M. (2004). Finding and evaluating community structure in networks. *American Physical Society*, 69(2):1–15.

Raghavan, U. N., Albert, R., e Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106.

Rosvall, M. e Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105:1118–1123.

Santos, C. P., Carvalho, D. M., e Nascimento, M. C. V. (2016). A consensus graph clustering algorithm for directed networks. *Expert Systems with Applications*, 54:121–135.

Shang, R., Bai, J., Jiao, L., e Jin, C. (2013). Community detection based on modularity and an improved genetic algorithm. *Physica A: Statistical Mechanics and its Applications*, 392(5):1215 – 1231. ISSN 0378-4371.