Community Detection: Comparison of State of the Art Algorithms

Josiane Mothe
IRIT, UMR5505 CNRS
& ESPE, Univ. de Toulouse
Toulouse, France
e-mail: josiane.mothe@irit.fr

Karen Mkhitaryan
Institute for Informatics and
Automation Problems, National
Academy of Sciences of Armenia
Yerevan, Armenia
e-mail:
karenmkhitaryan@gmail.com

Mariam Haroutunian
Institute for Informatics and
Automation Problems, National
Academy of Sciences of Armenia
Yerevan, Armenia
e-mail: armar@ipia.sci.am

ABSTRACT

Real world complex networks may contain hidden structures called communities or groups. They are composed of nodes being tightly connected within those groups and weakly connected between them. Detecting communities has numerous applications in different sciences such as biology, social network analysis, economics and computer science. Since there is no universally accepted definition of community, it is a complicated task to distinguish community detection algorithms as each of them use a different approach, resulting in different outcomes. Thus large number of articles are devoted to investigating community detection algorithms, implementation on both real world and artificial data sets and development of evaluation measures. In this article several state of the art algorithms and evaluation measures are studied which are used in clustering and community detection literature. The main focus of this article is to survey recent work and evaluate them using artificially generated networks.

Keywords

Network Science, Community Detection, Stochastic Block Model

1. INTRODUCTION

In recent years network science gained more attention with the advent of modern computational machines enabling to challenge more complex problems and rapid increase in the amount of data. Real world networks are usually represented as undirected, directed or weighted graphs, composed of nodes and edges, where edges serve as connections between the nodes. While random graphs imply an homogeneous degree distribution, that is to say the probability of having an edge between every two nodes is the same, real world networks have inhomogeneous structure resulting in groupings of nodes being tightly connected to each other and weakly connected with nodes from other groups.

This property of real world complex networks is known to be a community structure, where communities or clusters are defined as groups of nodes having higher intra-connectivity (inside the groups) and weak inter-connectivity (between groups) [1]. The aim of community detection is to identify the groups with high concentrations using the information encoded in the graph topology.

Revealing communities in networks proved to have countless applications in protein-to-protein interactions from biology, social network analysis, recommendation systems from on-line product purchasing, machine learning problems, etc. Although there is no universally accepted definition of community, various structural definitions and scoring functions exist [2] to quantitatively assess how community-like are the groups of nodes(e.g., conductance, modularity, triangle participation ratio) that we will describe later.

Modern networks may have up to millions or billions of nodes and edges which obscures the process of community detection due to computational issues, however there are well designed algorithms with low complexity [1][3] to overcome these obstacles and give promising outcomes.

The results of the algorithms differ from network to network and it is mandatory to test and compare them on many networks to make acceptable conclusions. However the number of large real world networks is limited and benchmark models such as Lancichinetti Fortunato Radicchi (LFR) benchmark [4] or the Stochastic Block Model (SBM) [5] generating networks with community structure resembling real world networks are used to overcome the limitations.

Large number of articles are devoted to comparing community detection algorithms using LFR benchmark [3][4] so we will only use SBM for our investigation. After a community detection algorithm is implemented and graph is partitioned into communities, another research problem is to analyze how "good" or "bad" are the detected communities.

This can be done by comparing the estimated community structure with reference structure or ground truth using external measures or by assessing the quality of detected communities internally. The aim of this paper is to provide the reader with an overview of the methods that have been developed for community detection. The paper is organized as follows: In Section 2 we describe the benchmark models, community detection algorithms and evaluation measures and show some experimental results in Section 3.

2. COMMUNITY DETECTION ALGORITHMS AND EVALUATION MEASURES

Various community detection algorithms have been developed which differ in terms of complexity and network types they target (e.g., undirected, directed, weighted, etc.). The process of community detection is rather simple in terms of sequential processes being implemented which are network selection, implementation of an algorithm and evaluation of the final results.

2.1 Real World vs Simulated Network

In recent years with the rapid increase of data collection various large networks became available. However even large networks are available, the number of networks with pre-known community structure or ground truth is limited. This limitation is overpassed by generating unlimited networks similar to real networks using the LFR benchmark and the SBM.

LFR benchmark generates networks with pre- known community structure where degree and community size distributions are heterogeneous and power-law. Mixing parameter μ is used to control the fraction of nodes a node shares inside and outside the community [4].

SBM is a generative model for random graphs, generating networks with community structure with predefined number of vertices, community sizes and probability matrix of intra and inter community edges [5]. These two approaches can be used to generate unlimited benchmark models resembling real world networks to implement and compare the algorithms and evaluate the results.

2.2 Algorithms

Solving community detection problems on modern real world networks can sometimes be much complicated due to computational complexity as networks may have large numbers of nodes and edges where exact detection can be NP-hard problem. Even nowadays distinguishing between algorithms and characterizing which algorithm works best on particular network is a hard task. In such cases heuristics or approximation algorithms are used to approximately optimize some objective function to detect almost "real" communities. Despite these barriers, plenty of successful algorithms exist in the literature, including those that were initially developed for cluster analysis. These algorithms are mainly classified into the following categories: modularity-based algorithms, spectral algorithms, algorithms based on random walks, label propagation and information-theoretical measures [1] that we develop in the next sub-sections.

Algorithms Based on Modularity Optimization

Modularity and other community scoring functions are characterizing how community-like are the groups of nodes in the network. Algorithms based on modularity optimization such as Newman's greedy algorithm [6] and its updated version by Clauset et. al (Fast Greedy) [7] join vertices which result in highest increase in modularity. After iterative process when modularity cannot be maximized any more, the network is partitioned into communities. Another popular modularity optimization method is Louvain algorithm which initially finds small communities by optimizing modularity locally and then aggregating nodes belonging to the same community and creating a network whose nodes represent the communities. This process is iterated until maximum modularity is reached and a hierarchy of communities are produced [8].

Algorithm based on eigenvectors of modularity matrix

This algorithm by Newman (Leading Eigenvector) [9] uses eigenspectrum of modularity matrix. Initially this algorithm initially creates the modularity matrix and finds eigenvector of the largest eigenvalue. Finally it labels nodes in corresponding communities knowing the sign of the elements in the eigenvector.

Random Walks

In general communities in networks have more intra connectivity then inter connectivity. Thus it is expected to have more edges inside those groups than between them. When implementing a short random walk, the probability that both the starting and ending points will be in the same group rather than in different groups, is higher. Algorithms based on random walks like Walktrap [10] use this idea to detect communities in networks.

Infomap

Infomap is an information theoretic method used to reveal community structure in the networks. At the beginning every node is assigned to its own community. Then nodes are moved to neighboring communities that results in the largest decrease of the map equation. After an iterative process when no move results in decrease of the map equation, network splits into communities [11].

Label Propagation

Unlike other community detection algorithms, label propagation does not optimize any given objective function and it does not require to have a priori information about the network structure. Initially every node has its own label and during an iterative process nodes gain the label which is frequent in their neighborhood. When every node has the label that the maximum number of its neighbors have, algorithm stops, resulting in densely connected groups. Among discussed algorithms label propagation is preferred due to its near linear time complexity [12].

2.3 Comparative Evaluation of Algorithms

After a community detection algorithm is implemented and the network is partitioned into communities, it is of paramount importance to interpret the results i.e. to know which algorithm performed well and detected meaningful communities.

Algorithms can be compared by their performance which is the time taken to partition the network and by qualitatively assessing how "good" are the derived communities. Measures used to assess the quality of detected communities are divided into two main categories:

- *Internal:* Evaluating communities internally by using community scoring functions.
- *External:* Comparison of communities derived by the algorithm with reference structure or ground truth.

Internal Measures

Internal measures are used to quantitatively assess how community-like is the given set of nodes in the network. As the global definition of community is based on the idea that it has high connectivity within a group and weak connectivity with other groups, scoring functions are based on this intuition. Here we will point out conductance, triangle participation ratio and modularity with the reason that conductance and triangle participation ratio give optimal results when identifying ground truth communities [2] and modularity which is the most widespread evaluation criteria used in the literature.

Conductance

Conductance is the fraction of total edges that goes outside the community and is defined as:

$$Conductance = \frac{O_c}{2I_c + O_c}$$

where O_c and I_c are the number of edges pointing outside from community c and the number of edges in c respectively.

Using conductance as a community goodness metric Leskovec et.al showed that best possible communities get less community like when they grow in size [13]. In their

other study while experimenting on 230 large real world networks, conductance and triangle participation ratio gave best results in identifying ground truth communities [2].

Triangle participation ratio
Triangle participation ratio is the fraction of nodes that belong to a triangle and is defined as:

$$TriangleParticipationRatio = \frac{T_c}{N_c}$$

where T_c is the number of vertices that form a triangle in c and N_c is the number of nodes in c.

Modularity
Modularity is the difference of fraction of the edges that fall within communities and expected number of edges in a random graph

$$Modularity = \frac{1}{2M}\sum_{xy}(A_{xy} - \frac{d_x d_y}{2M})\delta(c_x, c_y).$$

Experiments on both real and artificial networks show that modularity suffers from resolution limit merging small groups in case of low resolution and splitting large groups in case of high resolution i.e. missing important structures in the network [14] and often it is not possible to eliminate both biases simultaneously.

External Measures

Normalized Mutual Information
Mutual Information (MI) is an information-theoretic measure that quantifies the mutual dependence between two random variables. In other terms MI measures how much information can be obtained about one random variable through another.
Normalized Mutual Information (NMI) between two random variables X and Y is defined as the ratio of mutual information $I(X,Y)$ and the average of entropies of X and Y

$$NMI(X,Y) = \frac{2I(X,Y)}{H(X)+H(Y)},$$

where $H(X)$ and $H(Y)$ are the entropies of random variables X and Y respectively.
Considering X and Y as two different partitions,
$NMI(X,Y)$ shows the similarity of the two partitions.

Adjusted Rand Index
Adjusted Rand Index is a similarity measure of two different partitions of a network like NMI. Given a set of n elements $S = (d_1, d_2, ..., d_n)$ and two partitions of S, X and Y respectively, where X and Y partition S into different subsets. Adjusted rand index is defined as:

$$AdjustedRandIndex = \frac{SS + DD}{SS + SD + DS + DD}$$

where
SS is the number of pairs of elements in S that are in the same subset in X and in the same subset in Y.
DD is the number of pairs of elements in S that are in different subsets in X and in different subsets in Y.
SD is the number of pairs of elements in S that are in the same subset in X and in different subsets in Y.
DS is the number of pairs of elements in S that are in different subsets in X and in the same subset in Y.

Purity
Purity is also used to compare two partitions.
Consider $X = (x_1, x_2, ..., x_p)$ and $Y = (y_1, y_2, ..., y_q)$ to be two random variables representing different partitions of the network, where x_p and y_q are parts of these partitions. Denote N_{x_p} and N_{y_q} number of nodes in x_p and y_q parts respectively, N_{x_p,y_q} number of nodes in $x_p \cap y_q$ and N number of nodes in the network.
The purity of partition X related to partition Y is defined as

$$Purity(X,Y) = \frac{1}{N}\sum_{p}\max_{q} N_{x_p,y_q}.$$

According to Orman and Labatut, these three common evaluation measures ignore the network topology [15]. Based on this idea Labatut introduced modified versions, which enabled to include the topological importance of the nodes. The idea is based on assigning a weight to each node by combination of the degree and community embeddedness. Tests on artificial networks assume that modified NMI was able to assess the correspondence with reference structure in terms of community memberships and topological properties [16]. Another novel approach was proposed by Rossetti et. al and Zhang. Rossetti et. al used community precision and community recall, where community precision quantifies the level of label homophily between community and ground truth while community recall quantifies the correspondence between a community and ground truth. Unlike NMI, this method works fast in large networks [17].
Zhang proposed a *relative normalized mutual information* (rNMI) measure which considers statistical significance of NMI by comparing it with expected NMI of random partitions. Zhang claims that regular NMI is affected by errors when the network size is finite and rNMI overcomes this barrier [18].
In this paper we use modularity to assess the quality of detected communities by algorithms. We will also measure effectiveness considering the processing time of the algorithms in various configurations.

3. RESULTS

We used SBM to generate networks with community structure, where number of vertices, community sizes and edge probabilities in communities and between communities are known a priori. In our experiments, generated networks have 200 nodes and they are grouped into five equally sized communities. We compared six algorithms using modularity score for different $P_{out} \in [0,1]$ and $P_{in} = 1$ values, where P_{out} and P_{in} represent probability of edge between communities and in communities respectively.

Observing more than 300 random models and averaging the results we noticed that Louvain and leading eigenvector algorithms give best results identifying communities which have high modularity score compared with other methods (see Fig. 1).
Infomap and Label propagation reach to zero modularity sooner i.e. being unable to find "good" communities when P_{out} increases (See Fig. 1).
In the next stage of our experiments we compared these algorithms based on the time of detection, using P_{out} and the number of vertices in the network N.

Results displayed in Fig. 2 & 3 show that Louvain and label propagation algorithms remain relatively fast compared with infomap, fast greedy and walktrap, when the number of vertices in the network and probability of edge between communities increase.

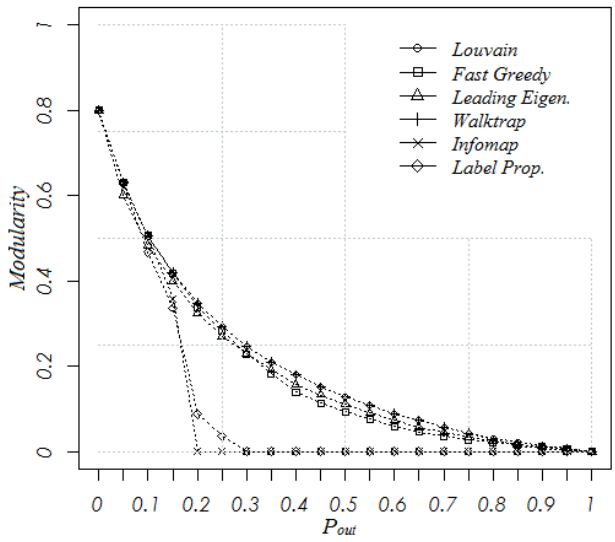

Figure 1: Probability of edge between communities (P_{out}) vs Modularity for $N = 200$ nodes and $P_{in} = 1$

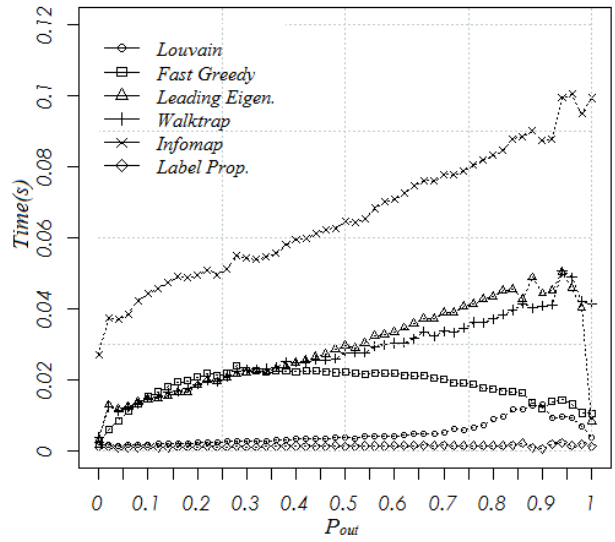

Figure 2: Probability of edge between communities P_{out} vs Time for $N = 200$ nodes and $P_{in} = 1$

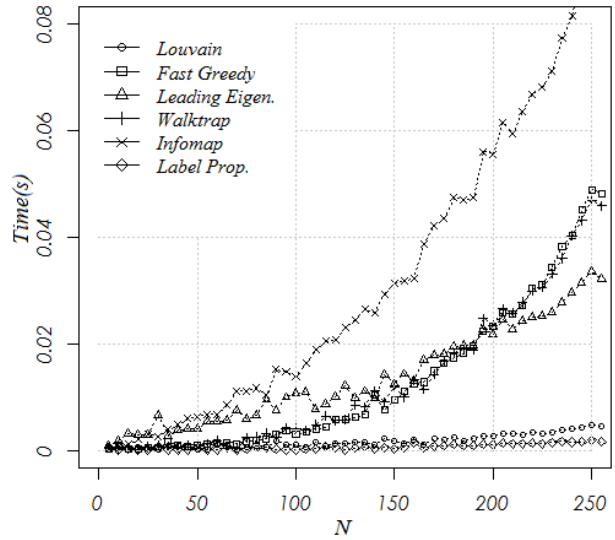

Figure 3: N vs Time for $P_{in} = 1$ nodes and $P_{out} = 0.3$

4. CONCLUSION AND FUTURE WORK

In this paper we surveyed six state of the art community detection algorithms.
Stochastic block model was used to generate random networks to compare the algorithms based on modularity score, detection time and network size.
In future we plan to include real world networks with ground truth communities, use more internal and external evaluation measures to assess both the quality of detected communities and correspondence with ground truth as well as more traditional algorithms such as Spinglass and Girvan-Newman algorithm.

REFERENCES

[1] Fortunato S.,"Community Detection in graphs",*Physics Reports 486*, pp. 75-174, 2010.

[2] Leskovec J., Yang J., "Defining and Evaluating Network Communities based on Ground-truth", *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, pp. 1-8, 2012.

[3] Lancichinetti A., Fortunato S., "Community detection algorithms: a comparative analysis", *Physical Review E 80, 056117*, 2009.

[4] Lancichinetti A., Fortunato S., Radicchi F., "Benchmark graphs for testing community detection algorithms", *Physical Review E78, 046110*, 2008.

[5] Abbe E., "Community detection and stochastic block models: recent developments", arXiv:1703.10146, 2017.

[6] Newman M., Girvan M., "Finding and evaluating community structure in networks", *Phys. Rev. E 69, 026113*, 2004.

[7] Clauset A., Newman M., Moore C., "Finding community structure in very large networks", *Phys. Rev. E 70, 066111*, 2004.

[8] Blondel V., Guillaume J., Lambiotte R., Lefebvre E., "Fast unfolding of communities in large networks", *J. Stat. Mech.,P10008*, 2008.

[9] Newman M., "Finding community structure in networks using the eigenvectors of matrices", *Phys. Rev. E 74, 036104*, 2006.

[10] Pons P., Latapy M., "Computing Communities in Large Networks Using Random Walks", *Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg*, pp. 284-293, 2005.

[11] Rosvall M., Bergstrom C., "Maps of random walks on complex networks reveal community structure", *PNAS, vol. 105, no 4* , pp. 1118-1123, 2008.

[12] Raghavan U., Albert R., Kumara S., "Near linear time algorithm to detect community structures in large-scale networks", *Physical Review E 76, 036106*, 2007.

[13] Leskovec J., Lang K., Dasgupta A., Mahoney M., "Statistical Properties of Community Structure in Large Social and Information Networks", *Proceedings of the 17th international conference on World Wide Web*, pp. 695-704, 2008.

[14] Lancichinetti A., Fortunato S., "Limits of modularity maximization in community detection", *Physical Review E84, 066122*, 2011.

[15] Orman G., Labatut V., Cherifi H., "Comparative Evaluation of Community Detection Algorithms: A Topological Approach", *Journal of Statistical Mechanics: Theory and Experiment, P08001*, 2012.

[16] Labatut V., "Generalized Measures for the Evaluation of Community Detection Methods", *International Journal of Social Network Analysis and Mining (SNAM),2(1)*, pp. 44-63, 2013.

[17] Rossetti G., Pappalardo L., Rinzivillo S., "A novel approach to evaluate community detection algorithms on ground truth", *Complex Networks VII. Studies in Computational Intelligence, vol 644. Springer, Cham*, 2016.

[18] Zhang P., "Evaluating accuracy of community detection using the relative normalized mutual information", *Journal of Statistical Mechanics: Theory and Experiment, Volume 2015*, 2015.