International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Benchmarks for Overlapping Community Detection Algorithms

Dr. M. Nagaratna[1], S. Swarajya Lakshmi[2]

[1]JNTU College of Engineering, JNTU, Hyderabad, India

[2]JNTU College of Engineering, JNTU, Hyderabad

Abstract: ***Recent studies on networks obtained from different domains such as social networking sites, internet / web-pages, protein-protein interaction networks etc., have shown that they share many common properties. One of the common properties of all these real networks is that they follow a community structure where the set of nodes with common interest are grouped together. The nodes within a group interact with each other very frequently and have relatively less interaction with the nodes of the other groups. The other essential property observed about these communities is that they have an overlapping nature, where some nodes can be members of more than one community. Several overlapping community detection algorithms like GCE and CFinder were proposed to detect the overlapping nature of communities of a given network. This requires performance comparison of these algorithms on a set of networks with known community structure or in other words a set of benchmark networks. There are different types of benchmark networks. In this work we studied the existing commonly used benchmark, namely, the LFR benchmark. The actual problem of this LFR benchmark is that it can create a network in which every node belongs to at least one community i.e, membership is at least one. But in real networks such as social or biological or computer networks, some nodes may not belong to any community(i.e., zero membership) but still have sparse interactions with the other nodes that are either members of any community or nodes like them(these are actually called the homeless nodes or orphan nodes). We have extended the presently existing LFR benchmark with orphan nodes. In our experiments we have observed that in some cases extended LFR benchmark (benchmark with orphan nodes or ELFR) is 5% better than the normal LFR benchmark.***

Keywords: community, overlapping, LFR, GCE, CFinder

1. Introduction

Community is an entity, where set of nodes with common features are grouped together. Facebook groups in social networks and protein complexes in protein-protein interactions (PPI) in biological networks are some of the examples of communities. People who are planning some kind of orders in e-commerce websites, such as Flipkart or Amazon.com, also form a community. It will be an easy task to suggest some product for an e-commerce website. If its users are organized in the form of groups or communities based on their common interests. Community detection algorithms like Greedy Clique Expansion (GCE) and CFinder were proposed earlier. It's a tough question to decide which algorithm is better among the available set of community detection algorithms. Different benchmarks were also proposed in the recent past to estimate the performance of community detection algorithms. LFR is a Synthetic benchmark, where some artificial networks are created along with some known properties. The disadvantage of LFR benchmark is that it keeps every node inside atleast one community. But in the real world Facebook groups there can be some people who may not members of any groups or community but still can maintain interactions with the people who are in groups. In order to address this issue we extended LFR benchmark with orphan nodes (nodes with zero membership) so that the performance evaluation is more accurate. In our evaluations, we have observed that extended LFR benchmark is performing more accurately compared with normal LFR benchmark.

2. Types of Benchmarks

There are two kinds of benchmarks

- **Real world benchmarks**, where ground truth communities are embedded into the data sets. Real world benchmarks are nothing but empirical benchmarks, e.g, Zachary's Karate Club. It is worth mentioning that they are mostly "hand-curated" and small.
- **Synthetic benchmarks**, where a particular model creates an artificial dataset, e.g, LFR benchmark. Synthetic benchmark performs well since communities are planted by the user in the network and we know them.

3. LFR Benchmark

This benchmark creates a connected, undirected, and unweighted graph (Fig 1). It makes use of power law distribution for each nodes degree and size of the community. According to power law of degree distribution, there is more number of nodes with less degree and probability of having degree k decreases as the degree of a node increases. Let μ be the fraction of edges that links to the nodes in other communities. Then, (1-μ) fractions of nodes lie within the community. d_{min} and d_{max} are the minimum degree and maximum degrees of the network, respectively in such a way that the average degree is d. C_{min} and C_{max} are the minimal and maximal sizes of the community, and they are chosen in such a way that $C_{min} > d_{min}$ and $C_{max} > d_{max}$. Thus, these constraints ensure that all nodes belong to at least one community using the above condition. Initially, all the nodes are orphan (i.e., they are not a member to any community).

In each iteration, an orphan node will be assigned to a randomly picked community. And, this process continues

while satisfying the above mentioned degree and community size constraints. The stopping condition is when there are no more orphan nodes left to assign. It is worth mentioning that this benchmark can also act like the Girvan-Newman benchmark [1] when number of nodes is 128, four communities each with 32 nodes, average degree 16, and having no overlapping nodes.

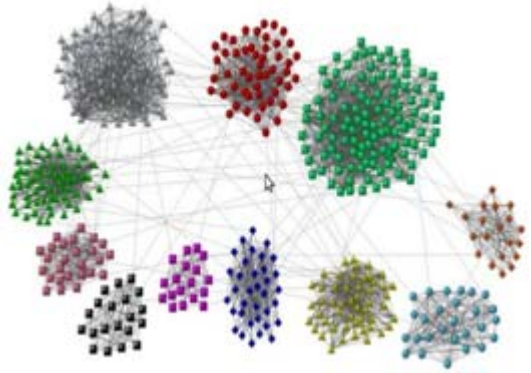
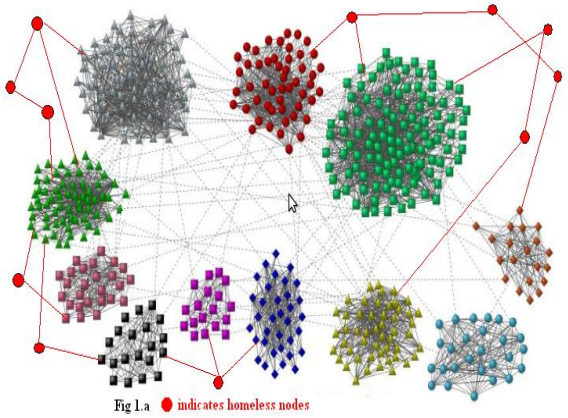


Figure 1: The LFR benchmark graph with 500 nodes [5]



4. ELFR Benchmark

ELFR stands for Extended LFR benchmark. According to LFR, every node must belong to at least one community. But, in real networks, like social networks (say, Facebook), there can be some people who are not members of any group but still can have relationship(friendships) with the people who are in different groups. Take the case of a celebrity, as an example as such. Another case may be, when a person newly opens an account in a social network and made some relationships but does not join any group, as yet. Here, in this project, we have extended the LFR benchmark to accept the varying memberships of overlapping nodes along with orphan nodes (Fig 1.a), which are nothing but the homeless nodes (nodes with zero membership).

5. Experiments Performed

We performed an extensive and rigorous analysis of the unweighted, undirected, and connected LFR benchmark. We used some common overlapping community detection algorithms.

5.1 Algorithms Used

The following overlapping community detection algorithms have been tested for assessing their performance.

1. CFinder [8]: It is the implementation mechanism for CPM (Clique Percolation Method), where all cliques of size k are identified in a given network and adjacent cliques are combined as one community.
2. Greedy Clique Expansion (GCE) [6]: It begins by identifying cliques as seeds and they greedily expand by optimizing a local fitness function.

5.2 Test with varying memberships

The varying membership implementation of the author's program takes a membership file. In addition to that, the program modified by me takes the percentage of overlapping nodes also, with different values and then using this new membership file we generate a benchmark graph.

5.3 Test with varying and zero memberships

The program is further extended to generate orphan nodes in the network. Then these orphan nodes are added to the original benchmark graph as orphans.

5.4 Parameters used for test

For our test, we took μ value from 0.1 to 0.4 for both benchmarks with orphan and without orphan.

The test was done for small graph where number of nodes (N) is 1000 and big graph, where number of nodes is 5000. Also, we set the community size as - for small community, the minimum community size (C_{min})is 10, and maximum (C_{max})is 50; for big graph, the minimum community size (C_{min}) is 20 and maximum (C_{max}) is 100. In both the cases average degree(k) is 20 while maximum degree (d_{max}) is taken as 50. Parameters used for GCE is -minimum Clique Size K = 4, overlap To Discard Eta = 0.6, fitness Exponent Alpha = 1.0, Clique Coverage Heuristic threshold Phi = .75. In case of CFinder (unweighted and undirected), the default values of the parameters are used.

5.5 Evaluation Criteria

To evaluate these above modifications done, a similarity measure called Normalized Mutual Information (NMI) [4] from information theory was considered.

6. Result

6.1 Without Orphan nodes (LFR with varying memberships)

The following graphs (from Fig 2.a to Fig 2.d) are generated based on the NMI values of different overlapping community detection algorithms. We can see that GCE is performing well in finding the overlapping communities with varying memberships in both small and big graphs. We can also see that both the algorithms performance is degrading as the fraction of external degree (μ) is increasing.

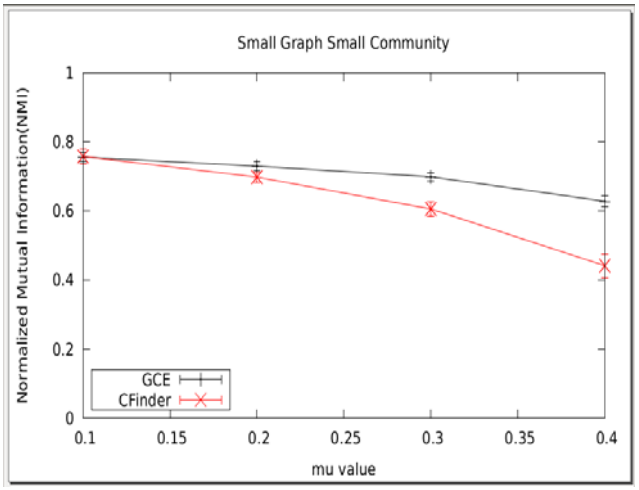


Figure 2 (a): Small Graph and Small Community

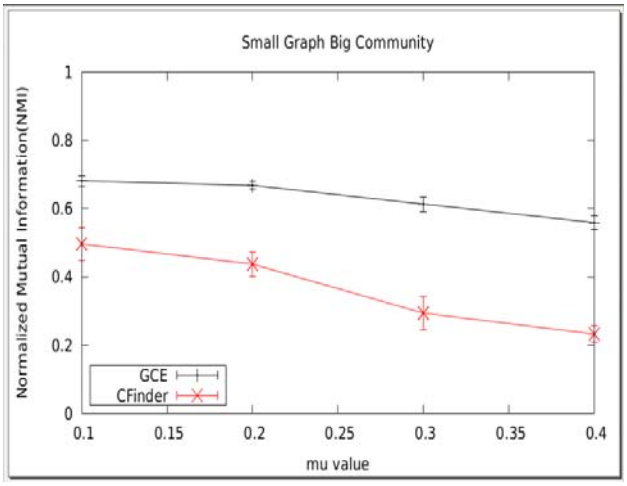


Figure 2 (b): Small Graph and Big Community

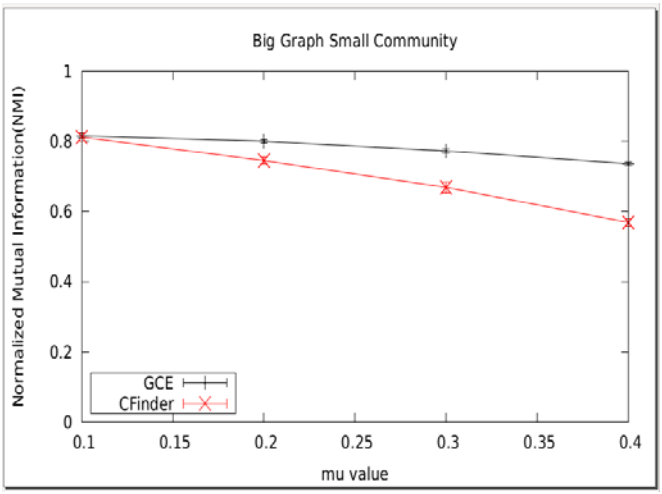


Figure 2 (c): Big Graph and Small Community

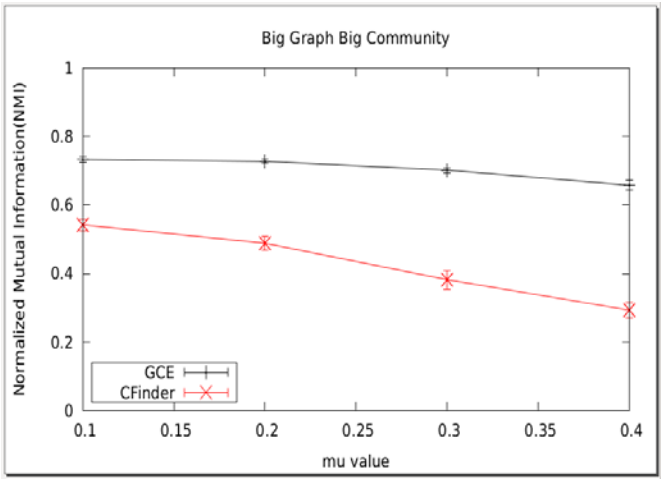


Figure 2 (d): Big Graph and Big Community

6.2 With orphan nodes (LFR with varying and zero memberships)

Since every overlapping community detection algorithm may not assign every node into a community, no particular algorithm is really suitable to test the performance of it on orphan nodes. Instead, the existing algorithms were tested in such a way that they cannot put these orphan nodes into any community and thereafter, NMI is tested. In this test the major observation is that GCE is again performing well in both small (Fig 3.a and Fig 3.b) and Big graphs (Fig 3.c and Fig 3.d). We also observe that CFinder is also competing well with GCE
algorithm in the case of smaller communities.

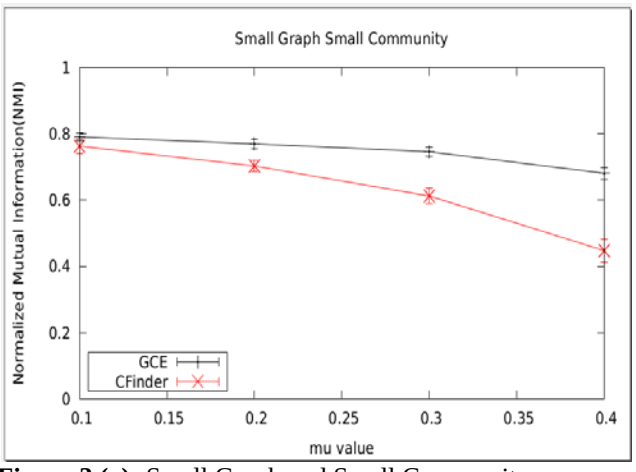


Figure 3 (a): Small Graph and Small Community

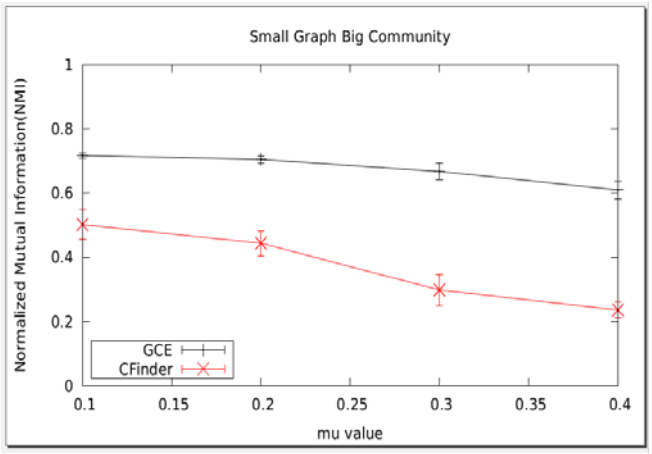


Figure 3 (b): Small Graph and Big Community

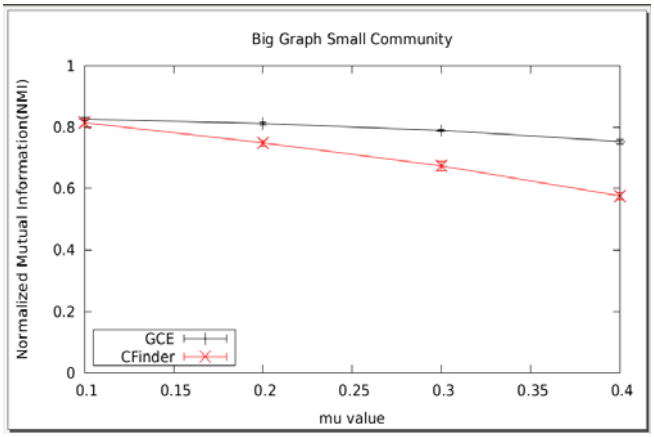


Figure 3 (c): Big Graph and Small Community

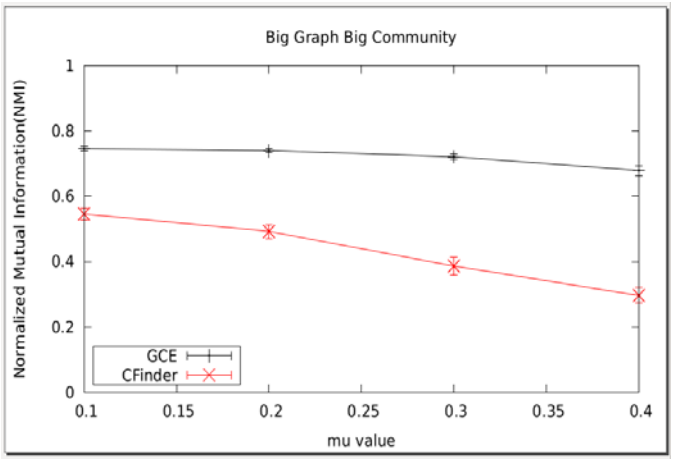


Figure 3 (d): Big Graph and Big Community

CASE	LFR		ELFR	
	GCE	CFinder	GCE	CFinder
Small Graph and Small	0.7033	0.6256	0.7569	0.6315
Small Graph and Big	0.6306	0.3651	0.6747	0.3699
Big Graph and Small	0.7815	0.6990	0.7958	0.7037
Big Graph and Big	0.7051	0.4262	0.7213	0.4301

6.3 Performance of Extended LFR benchmark against Normal LFR benchmark

Table: Comparing NMI values of GCE and CFinder algorithms of LFR and ELFR. For a small graph with smaller communities, we can orderly observe that extended LFR benchmark is 5% better than the normal LFR benchmark in GCE case and only 1% in case of CFinder. Same in the case with small graph with big communities by 3% and 1%. For a big graph with small communities their difference is not very high this just 1% in both the algorithms. In the last case, such as big graph with big communities also the difference is just 2% and 1%.

7. Conclusion & Future Work

We conclude that the tests were conducted on different sets of benchmark networks to compare the above overlapping community detection algorithms. In our observation, we found that GCE is performing well. We have also observed that extended LFR benchmark performance is better than normal LFR benchmark. However these set of benchmarks only takes care of topological properties of networks. Nevertheless, most of the real networks come with attributes. As the community structure must relate to these attributes, in future we would like to implement an LFR benchmark while taking care of topological properties as well as node attributes.

References

[1] Michelle Girvan and Mark EJ Newman. Community structure in so-cial and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821{7826, 2002.
[2] Peter D Gr•unwald, In Jae Myung, and Mark A Pitt. Advances in min-imum description length: Theory and applications. MIT press, 2005.
[3] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing com-munity detection algorithms on directed and weighted graphs with over-lapping communities. Physical Review E, 80(1):016118, 2009.
[4] Andrea Lancichinetti, Santo Fortunato, and Janos Kertesz. Detecting the overlapping and hierarchical community structure in complex net-works. New Journal of Physics, 11(3):033015, 2009.
[5] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Bench-mark graphs for testing community detection algorithms. Physical Re-view E, 78(4):046110, 2008.
[6] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detect-ing highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827, 2010
[7] Min Li, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. Mod-ifying the dpclus algorithm for identifying protein complexes based on new topological structures. BMC bioinformatics, 9(1):398, 2008.
[8] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043):814{818, 2005.
[9] Martin Rosvall and Carl T Bergstrom. Maps of information ow reveal community structure in complex networks. Technical report, 2007.
[10] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: the state of the art and comparative study. arXiv preprint arXiv:1110.5813, 2011.