Community Detection in Large Scale Big Data Networks

Pravin Chopade[1], Justin Zhan[2]
[1,2]Department of Computer Science
[1,2]North Carolina A&T State University, Greensboro, NC-27411, USA
pvchopad@ncat.edu[1] , zzhan@ncat.edu[2]

Abstract

Many complex systems in the real world can be modeled as graphs or networks. One of the most relevant features of graphs representing real systems is community structure, or clustering. In this paper we propose a method to extract the community structure of large scale/Big data networks. Such communities appear to be connected with unique spectral property of the graph Laplacian of the adjacency matrix. In this paper, we develop and explore communities from structural parameters of the complex network such as node degree distribution, algebraic connectivity, clustering coefficient i.e. using node attributes and edge structure. New modified algorithm statistically models the interaction between the network structure and the node attributes which leads to more accurate community detection as well as helps for identifying robustness of the network structure. We also show that any community must contain a dense Erdos-Renyi (ER) subgraph. We carried out comparisons of the Chung and Lu (CL) and Block Two-Level Erdos-Renyi (BTER) models with two real-world data sets. Results demonstrate that it accurately captures the observable properties of many real-world networks.

Modularity is a property of a network and a specific proposed division of that network into communities. In this paper we also propose modularity optimization based on a greedy agglomerative method. Our proposed modified algorithm is linearly scalable for efficient identification of communities in huge directed/undirected networks. The proposed algorithm shows great performance and scalability on benchmark networks in simulations and successfully recovers communities in real network applications.

Keywords: Community, Large Scale Network, Big Data, Modularity, Optimization.

1. Introduction

Social, technological and information systems can often be described in terms of complex networks that have a topology of interconnected nodes combining organization and randomness [1] [2]. Community is formed by individuals such that those within a group interact with each other more frequently than with those outside the group. Community detection is discovering groups in a network where individuals' group memberships are not explicitly given. Identifying network communities is one of the most important tasks when analyzing complex networks. Most of these networks possess a certain community structure that has substantial importance in building an understanding regarding the dynamics of the network.

This research was partially supported by the following grants: NSF No. 1137443, NSF No. 1247663, NSF No. 1238767, The United States Department of Defense, DoD No. W911NF-13-0130, DoD No. W911NF-14-1-0119, and the Data Science Fellowship Award by the National Consortium for Data Science.

The basic premise behind the network community detection is that communities have distinct connectivity patterns and thus one aims to detect them based (only) on the network connectivity structure. Thus means that one can use self-identified groups as "ground-truth" communities. If the goal of network community detection is to use the network connectivity structure to extract sets of nodes that have a common functional or social role, then our premise is that we can use these same sets of nodes with a common role or property (i.e., self-identified groups) a ground-truth communities.

2. Community Detection in a Large Scale Network

The aim of community detection in graphs is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology. Community detection is important for other reasons, too. Identifying modules and their boundaries allows for a classification of vertices, according to their structural position in the modules. So, vertices with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group; vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities [3].

Communities are of interest for a number of reasons. They have intrinsic interest because they may correspond to functional units within a networked system [4]. Different community structures and models are discussed below.

2.1 Different Community Structures

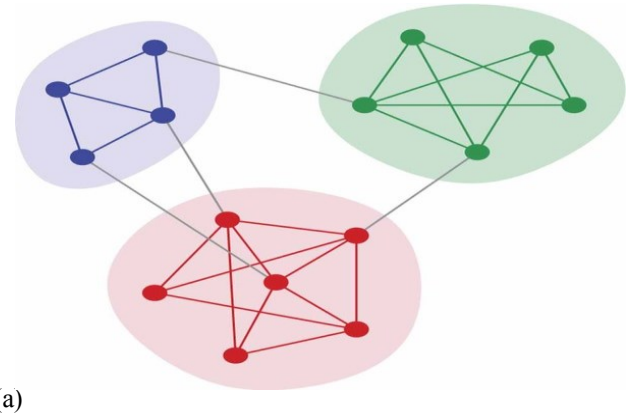
The best-studied form of large-scale structure in networks is modular or community structure [3] [5]. A community, in this context, is a dense sub network within a larger network, such as a close-knit group of friends in a social network or a group of interlinked web pages on the World Wide Web as shown in figure 1(a). In many networks it is found that the properties of individual communities can be quite different. Consider; for example, figure 1(b), which shows a network of collaborations among a group of scientists at a research institute. The network divides into distinct communities as indicated by the colors of the nodes [4].

Most real networks typically contain parts in which the nodes are more highly connected to each other than to the rest of the network. Most real networks are characterized by well-

 ISBN: 978-1-62561-000-3

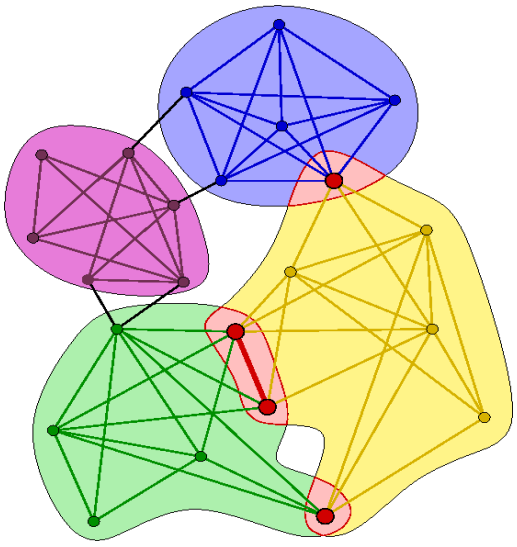
defined statistics of overlapping and nested communities. Such a statement can be demonstrated by the numerous communities each of us belongs to, including those related to our scientific activities or personal life (school, hobby, family) and so on [6].

The basic observation on which our community definition relies is that a typical community consists of several complete (fully connected) sub graphs that tend to share many of their nodes. All these can be explored systematically and can result in a large number of overlapping communities as shown in figure 1c for illustration [6].



(a)

(b)



(c)

Figure 1. Different community structure examples (a) Non-overlapping communities, the nodes of this network are divided into three groups, with most connections falling within groups and only a few between groups. (b) A network of collaborations among scientists at a research institute. (c) Communities with overlapping or non-overlapping communities [4] [6].

2.2 The CL and BTER Community Model

Community analysis can be reveal important patterns, decomposing large collections of interactions into more meaningful components.

Consider a graph G with N vertices and degrees $d_1, d_2, \ldots . d_N$.

Let

$$m = \frac{1}{2}\sum_{i=1}^{N} d_i \qquad (1)$$

denote the number of edges. A sub graph S has high modularity if S contains many more internal edges than predicted by a null model, which says vertices i and j are connected with probability $\frac{d_i d_j}{2m}$. Seshadhri et al. [7] refer to the null model as the CL model, based on its formalization by Chung and Lu [8] [9]; see also Aiello et al. [10] and the edge-configuration model of Newman et al. [11].

Considering idea of a graph comprising Erdos-Renyi (ER) communities, Seshadhri et al. [7] proposed the Block Two-Level Erdos-Renyi model (BTER). The advantages of the BTER model are that it has community structure in the form of dense ER subgraphs and that it matches well with real-world graphs. The first phase (or level) of BTER builds a collection of ER blocks a way that the specified degree distribution is respected. The second phase of BTER interconnects the blocks. The BTER model allows one to construct a graph with any degree distribution. Real-world degree distributions might be idealized as power laws, but it is by no means a completely accurate description [12] [13]. When the degree distribution is heavy tailed, then the BTER graph naturally has scale-free ER subgraphs. The internal connectivity of the ER graphs is specified by the user and can be tuned to match observed data.

3. Elements of Large Scale Big Data Networks For Community Detection

3.1 Node Degree Distribution

The degree of a vertex is the number of edges connecting to it, and it is one of the simplest and most important properties of a single vertex. The degree distribution, usually denoted by $P(k)$, is the probability that a vertex chosen uniformly at random has degree k, or equivalently, the fraction of vertices in the network with degree k. The average degree k of a network with N vertices and m edges is always $\bar{k} = \frac{2m}{N}$. It is a measure of the edge density and the amount of redundant edges in the network. In a tree, which has no redundant edges,

$\bar{k} = 2 - \frac{2}{N}$, which is approximately 2 for large networks. In many real networks it has been found that the degree distribution follows a power-law, i.e. $P(k) \sim k^{-\alpha}$, where α is the scaling coefficient, it is typically between 1 and 3 [14]. Networks with a power-law are often called scale-free, since there is no characteristic scale in the degree distribution. Most vertices have low degrees, while a small but not negligible number of vertices, so-called "hubs", have very high degrees.

3.2 Algebraic Connectivity

The algebraic connectivity of a graph G is the second-smallest eigenvalue of the Laplacian matrix of G. This eigenvalue is greater than 0 if and only if G is a connected graph. This is a corollary to the fact that the multiplicity of the eigenvalue 0 is the number of connected components in the graph. The magnitude of this value reflects how well connected the overall graph is, and has been used in analyzing the synchronizability of networks. Furthermore, the value of the algebraic connectivity is bounded above by the traditional (vertex) connectivity of the graph [15] [16].

The smallest non-zero eigenvalue of L is called the spectral gap. If the algebraic connectivity $\lambda_2(L)$ is close to zero, the network is close to being disconnected. Otherwise, if $\varepsilon\lambda_2(L)$ tends to be 1, where ε is a normalized constant related with network size, the network tends to be fully connected. This eigenvalue is related to several important graph invariants, and it has been extensively investigated. Most of the results are consequences of the well-known Courant-Fischer principle [17] which states that:

$$\lambda_2(G) = \min_{\substack{x \perp 1 \\ x \neq 0}} \frac{x^T L(G) x}{x^T x} \tag{2}$$

Where $0 = (0, 0, \ldots\ldots, 0)^T$, and $1 = (1, 1, \ldots\ldots, 1)^T$ is an eigenvector of $\lambda_1 = 0$. T is the transpose.

Using Courant-Fischer principle, Fiedler [18] obtained expression for λ_2. If square matrix L is Laplacian matrix of graph G with all zero row sums has an eigenvalue (λ) '0' and a corresponding eigenvector e, where e is vector $(1,1,\ldots.1)^T$. If L is positive-semidefinite then the second smallest eigenvalue is equal to

$$\lambda_2 = \min_{x \in W} x^T . L . x \tag{3}$$

The algebraic connectivity plays a decisive role in determining the synchronizability of a given network [19]. The two graphs shown in figures 2 and 3 have the same degree sequence, but the graph in figure 2 seems weakly connected.

Intuitively we expect the network in figure 3 to be more synchronizable. The Graph in figure 2 has $\lambda_2 = 0.1531$ and in figure 3 has $\lambda_2 = 0.9249$.

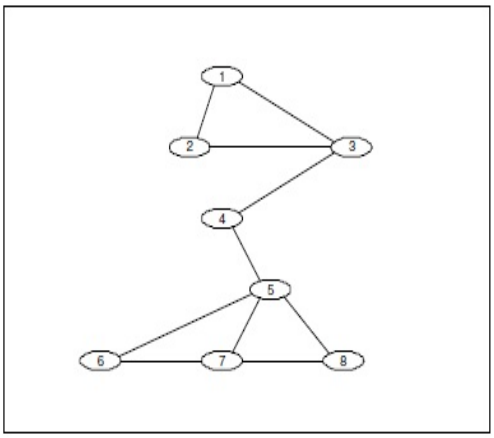


Figure 2. Graph with algebraic connectivity 0.1531.

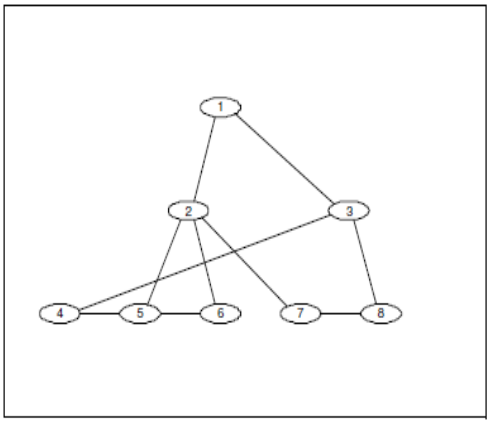


Figure 3. Graph with algebraic connectivity 0.9249.

3.3 Clustering Coefficient

The clustering coefficient, CC_i, of a vertex i is the ratio between the actual number of edges that exist between the vertex and its neighbors and the maximum number of possible edges between these neighbors. The CC of the network is defined as:

$$CC = \frac{1}{n} \sum_{i \in V} CC_i = \frac{1}{n} \sum_{i \in V} \frac{M_i}{k_i (k_i - 1)/2}, \tag{4}$$

where CC_i is the local clustering coefficient, M_i is the number of edges that exist between the neighbors of vertex i, and k_i is the number of neighbors for vertex i. The denominator $k_i(k_i - 1)/2$ is the maximum possible number of edges that can exist between the neighbors of vertex.

3.4 Modularity Optimization

The problem of detecting and characterizing different community structure has attracted considerable recent attention. One of the most sensitive detection methods is

optimization of the quality function known as "modularity" over the possible divisions of a network [20].

Let consider a network, represented by means of a graph $G=(V,E)$, let e_{ij} be the fraction of edges in the network that connect vertices in group i to those in group j, and let $a_i=\sum_j e_{ij}$ then modularity Q

$$Q=\sum_i^c (e_{ii}-a_i^2) \quad (5)$$

is the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure. If a particular division gives no more within-community edges than would be expected by random chance we will get $Q=0$. Values other than 0 indicate deviations from randomness, and in practice values greater than about 0.3 appear to indicate significant community structure.

By assumption, high values of modularity indicate good partitions. So, the partition corresponding to its maximum value on a given graph should be the best, or at least a very good one. This is the main motivation for modularity maximization, by far the most popular class of methods to detect communities in graphs. An exhaustive optimization of Q is impossible, due to the huge number of ways in which it is possible to partition a graph, even when the latter is small. Besides, the true maximum is out of reach, as it has been recently proved that modularity optimization is an NP-complete problem, so it is probably impossible to find the solution in a time growing polynomially with the size of the graph. However, there are currently several algorithms able to find fairly good approximations of the modularity maximum in a reasonable time [3].

The problem of community detection requires the partition of a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. Precise formulations of this optimization problem are known to be computationally intractable. Several algorithms have therefore been proposed to find reasonably good partitions in a reasonably fast way [21] [22]. Fast algorithm known as 'Fast Newman algorithm' or 'Girvan-Newman (GN)' introduced by Newman [23] is based on the idea of modularity. The Girvan–Newman algorithm detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Given any network, the GN community structure algorithm always produces some division of the vertices into communities, regardless of whether the network has any natural such division. One of the proposed algorithm is by Greedy sketch method for modularity optimization [24]. It is an agglomerative hierarchical clustering method, where groups of vertices are successively joined to form larger communities such that modularity increases after the merging. Greedy optimization method attempts to optimize the "modularity" of a partition of the network. The optimization is performed in two steps. First, the method looks for "small" communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. This algorithm is given below

Algorithm: Greedy algorithm sketch for modularity optimization.

1. Divide in as many clusters as there are nodes
2. Measure modularity variation Q for each candidate partition where a pair of clusters are merged
3. Select the network with the highest Q
4. Go back to step 2

4. Application to Large Scale Big Data Networks

For above discussed elements of community detection, we developed new MATLAB algorithm (See Appendix I) along with Greedy sketch method, and tested it for real world large scale/Big data networks. We consider comparisons of the BTER model with two real-world data sets i.e. the Football dataset a large scale directed network example describes the 22 soccer teams which participated in the World Championship in Paris, 1998 and US Air flights, 1997 network which represents undirected weighted graph. Properties of these data sets are shown in Table 1. Table 1 shows detected communities with respect to network structure and size. From values of modularity we can identify community structure.

Table 1. Experimental evaluation of large scale network.

Dataset	Football	US Air flights, 1997
Type of network	Directed weighted graph	Undirected weighted graph
N	35	332
m	118	2126
C	6	44
λ_2	0.3767	0.2854
CC	0.3390	0.6252
Q	0.2068	0.3190

N : Number of Nodes, m : Number of Edges,
C : Number of Communities, λ_2 : Algebraic Connectivity,
CC : Clustering Coefficient, Q : Modularity.

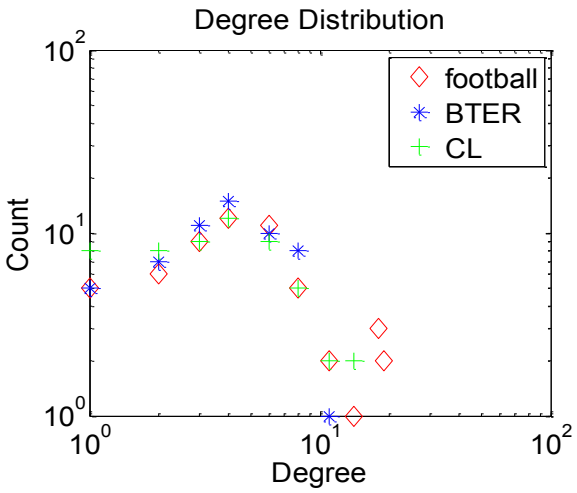


Figure 4 (a)

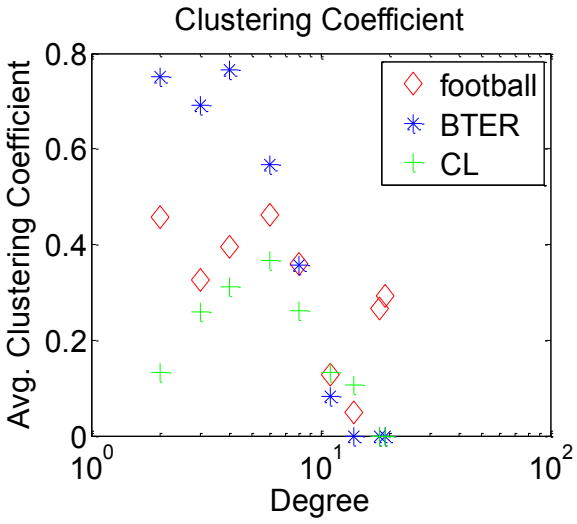


Figure 4 (b)

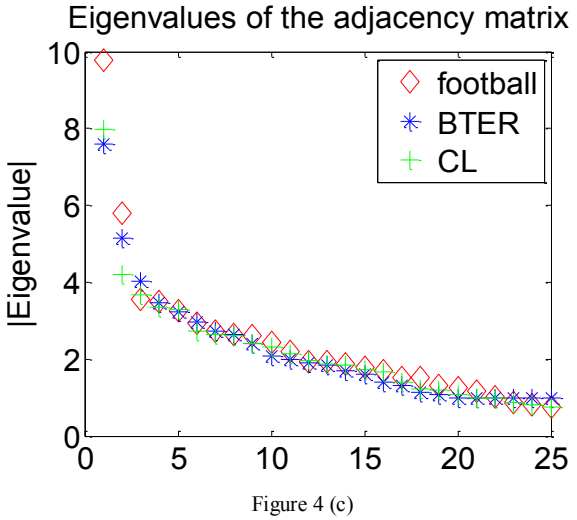


Figure 4 (c)

Figure 4: Properties of Football network, compared with the BTER and CL models (a) Degree distribution (b) Clustering Coefficient and (c) Eigenvalues of the adjacency matrix.

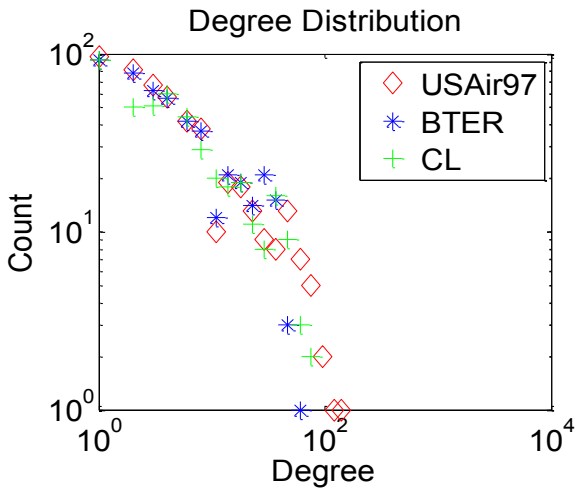


Figure 5 (a)

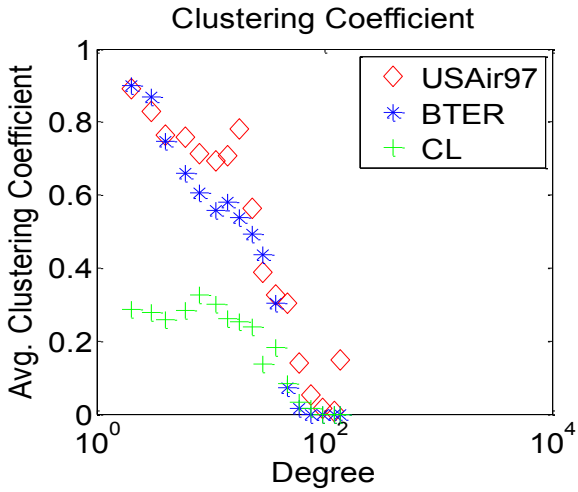


Figure 5 (b)

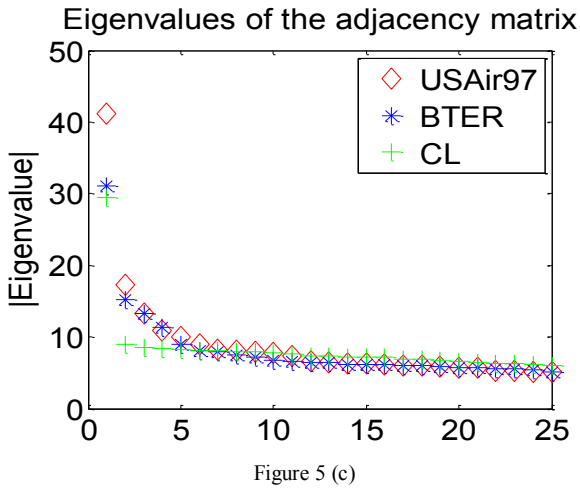


Figure 5 (c)

Figure 5: Properties of US Air Flight 1997 network, compared with the BTER and CL models (a) Degree distribution (b) Clustering Coefficient and (c) Eigenvalues of the adjacency matrix.

From Q value of US Air flight 1997 network it indicates significant community structure. We can determine robustness of Football and US Air flight 1997 network from its algebraic connectivity values.

We compare BTER with the real data as well as the corresponding CL model. From plots of figures 4 (a), (b) and (c) and figures 5 (a), (b) and (c), we see the comparison of the degree distributions, clustering coefficient and eigenvalues of the adjacency matrix. The degree distribution for Football network has a slight "kink" mid-way and does not conform to any standard degree distribution such as lognormal or power law. Nonetheless, both BTER and CL are able to match it. The degree distribution for US Air Flight 1997 network is fairly close to a power law, and matched well by both BTER and CL. Observe the close match of the clustering coefficients of the US Air flight 1997 network and BTER, in contrast to CL. Additionally, the eigenvalues of the BTER adjacency matrix are close to those of the Football and US Air flight 1997 network.

The difference between BTER and CL is highlighted when we instead consider the clustering coefficient, shown in the plots of figure 4 (b) and figure 5 (b). As noted previously, CL cannot have a high clustering coefficient and a heavy tail, and this is evident in these examples. BTER, on the other hand, has a close match with the observed clustering coefficients. The dense ER graphs ensure that all nodes have high clustering coefficient.

5. Conclusions

The problem of discovering the community structure in large networks has been widely investigated during last years. The main drawback of the existing techniques is that they do not consider complete network or topology information. In this paper we consider elements of community detection from node attributes and edge structure point of view. We observed that node attributes and edge structure plays important role into community detection. Some of the advantage is the improved accuracy in community detection. Another advantage is that the node attributes provide cues for interpreting detected communities.

We prove that any community must contain a dense ER subgraph. Therefore, any graph model that captures community structure must contain dense sub-structures in the form of dense ER graphs. This observation leads naturally to the BTER model, which explicitly builds communities of varying sizes and simultaneously generates a heavy tail. Our experimental results show that BTER has proper-ties that are remarkably similar to real-world data sets. We contend that this makes BTER an appropriate model to use for testing algorithms and architectures designed for interaction graphs.

Our approach is able to discover the community structure in, possibly large scale big data networks. Our experimental evaluation, carried out over real-world networks, proves the efficiency and the robustness of the proposed strategy. To investigate the robustness of performance under an unreliable network structure, in our future work we will explore the problem of detecting communities from partially observed networks where some fraction of edges are missing while the node attributes are fully available.

Acknowledgment

Our thanks to The United States Department of Defense (DoD), National Science Foundation (NSF) and National Consortium for Data Science (NCDS) for their support and finance for this project. This research was partially supported by the following grants: NSF No. 1137443, NSF No. 1247663, NSF No. 1238767, The United States Department of Defense, DoD No. W911NF-13-0130, DoD No. W911NF-14-1-0119, and the Data Science Fellowship Award by the National Consortium for Data Science.

References

[1] R. Albert and A-L. Barab´asi, 2002 *Rev. Mod. Phys.* 74 4797.

[2] M. Newman, A-L. Barab´asi and D. Watts, "The Structure and Dynamics of Networks", (Princeton University Press, Princeton, 2006).

[3] S. Fortunato, "Community detection in graphs", *Physics Reports,* vol. 486, no. 3-5, pp. 75–174, 2010.

[4] M. Newman, "Communities, modules and large-scale structure in networks", *Nature Physics,* Vol 8 January 2012, pp. 25-31. DOI:10.1038/NPHYS2162

[5] M. Girvan and M. Newman, "Community structure in social and biological networks', *Proc. Natl Acad. Sci.* USA 99, 7821-7826 (2002).

[6] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", arXiv:physics/0506133v1 [physics.soc-ph] 15 Jun 2005.

[7] C. Seshadhri, T. Kolda and A. Pinar, "Community structure and scale-free collections of Erdos and Renyi graphs", arXiv: 1112.3644v1 [cs.SI] 15 Dec 2011.

[8] F. Chung, L. Lu, "The average distances in random graphs with given expected degrees", *Proceedings of the National Academy of Sciences*, 99, 15879 (2002).

[9] F. Chung, L. Lu, "Connected components in random graphs with given degree sequences", *Annals of Combinatorics* 6, 125 (2002).

[10] W. Aiello, F. Chung, L. Lu, "A random graph model for power law graphs", *Experimental Mathematics* 10, 53, (2001).

[11] M. Newman, D. Watts, S. Strogatz, "Random graph models of social networks", *Proceedings of the National Academy of Sciences* 99, 2566 (2002).

[12] C. Clauset, R. Shalizi, M. Newman, "Power-law distributions in empirical data", *SIAM Review* 51, 661, (2009).

[13] Sala, H. Zheng, B. Zhao, S. Gaito, G. Rossi, "Brief announcement: revisiting the power-law degree distribution for social graph analysis", PODC '10: *Proceeding of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (ACM, 2010), pp. 400-401.

[14] K. Sun, "Complex Networks Theory: A New Method of Research in Power Grid", 2005 IEEE/PES Transmission and Distribution Conference & Exhibition: Asia and Pacific Dalian, China, pp. 1-6.

[15] R. Merris, "Laplacian graph eigenvectors", *Linear Algebra and its Applications,* Vol. 278, No. 1-3. (15 July 1998), pp. 221-236, doi:10.1016/S0024-3795(97)10080-5.

[16] L. Barriere, C. Huemer, D. Mitsche, D. Orden, "On the Fiedler value of large planar graphs", Electronic Notes in Discrete Mathematics, *Elsevier B.V.,* 38, 2011, pp. 111-116, www.elsevier.com/locate/endm, doi:10.1016/j.endm.2011.09.019.

[17] A. Torres and G. Anders, "Spectral Graph Theory and Network Dependability", *IEEE Fourth International Conference on Dependability of Computer Systems,* DepCos-RELCOMEX '09, 2009, pp. 356 - 363.

[18] M. Fiedler, "Algebraic connectivity of graphs", *Czechoslovak Math.* J.

23 (98) (1973) 298–305.

[19] F. Atay, T. Biyikoglu and J. Jost , "Synchronization of networks with prescribed degree distributions", Circuits and Systems I: Regular Papers, *IEEE Transactions*, vol.53, no.1, pp. 92- 98, Jan. 2006 doi:10.1109/TCSI.2005.854604 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1576889&isnumb r=33334

[20] M. Newman, "Modularity and community structure in networks", arXiv:physics/0602124v1 [physics.data-an] 17 Feb 2006.

[21] V. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", arXiv: 0803.0476v2 [physics.soc-ph], 25 Jul 2008.

[22] M. Newman, 2006 *Proc. Natl. Acad. Sci.* 103 8577.

[23] M. Newman, "Fast algorithm for detecting community structure in networks", arXiv:cond-mat/0309508v1 [cond-mat.stat-mech] 22 Sep 2003.

[24] M. Newman, 2006 *Phys. Rev.* E 74 036104.

Appendix I

Algorithm: New algorithm for large scale big data network analysis and Community detection

1. G the initial network
2. Compute Adjacency matrix (Adj) of G
3. Compute Degree Distribution (G)
   Compute Avg Degree Distribution
4. Compute
   Laplacian: L (Adj)
   Algebraic Connectivity: ac(Adj)
   Clustering Coefficient: CC(Adj)
   Avg. Clustering Coefficient
5. B= BTER Model
   CL= CL Model
6. Compare graphs (G ,B, CL, graph_name, true)
7. Compute Community C and Modularity Q
   Optimization with Fast_Newman Greedy algorithm
8. repeat
9. Put each node of G in its own community
10. while some nodes are moved do
11. for all node N of G do
12. place N in its neighboring community including its own which maximizes the modularity gain
13. end for
14. end while
15. if the new modularity is higher than the initial then
16. G = the network between communities of G
17. else
18. Terminate
19. end if
20. until