2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Structure and Attributes Community Detection Benchmark and a Novel Selection Method

Haithum Elhadi
Illinois Institute of Technology
Computer Science Department
Chicago, IL 60616
helhadi1@iit.edu

Gady Agam
Illinois Institute of Technology
Computer Science Department
Chicago, IL 60616
agam@iit.edu

***Abstract*—In recent years due to the rise of social, biological, and other rich content graphs, several new graph clustering methods using structure and node's attributes have been introduced. In this paper, we proposed an effective benchmark to evaluate these new methods. Our benchmark is an attributes extension to a widely used structure only benchmark. We also developed a new clustering method, termed Selection method, that uses the graph structure ambiguity to switch between structure and attribute clustering methods. Using the new benchmark and Normalized Mutual Information (NMI) metric, we evaluated the Selection method against five clustering methods: three structure and attribute methods, one structure only method and one attribute only method. We showed that the Selection method outperformed that state-of-art structure and attribute methods.**

Keywords: Community detection, graph clustering benchmark, structure and attributes clustering

I. Introduction

Complex network community detection (i.e., graph clustering) is a long studied problem in machine learning and graph theory. The clustering goal evolved beyond the structure of the graph (vertices and edges) to the consideration of the vertices' attributes. The motivation is driven by the popularity of social network with rich content associated with the vertices. Structure and attributes clustering is based on three key well established assumptions in the research: 1) There are underlined clusters in the graph. 2) Members of the clusters have strong ties between them and weak ties to other clusters. 3) Member of the clusters exhibit attributes similarity between them compared to members of other clusters. The tendency to exhibit attributes similarity within a cluster is known as homophily or assortative mixing [1]. The homophily behavior can be observed in many complex network, such as social network, citation network, and others [2].

Several new clustering methods that use both structure and attributes of graphs are introduced in recent years [3]–[7]. Some of these methods are SA-Cluster, Entropy Based, SAC, and BAGC. SA-Cluster converts the attributes to edges and uses random walk along the structure and attribute edges to determine the clusters. Entropy Based and SAC methods use modified similarity objective functions, while the BAGC introduces a new Bayesian model approach.

Benchmarking the performance of community detection methods has been identified as critical step for improving these methods. Testing the performance of structure and attributes methods is conducted using real networks and several quality measure (i.e, modularity, entropy, density). Unfortunately due to the lack of ground truth in the used real network, it is not possible to objectively compare the methods performance. To overcome the lack of ground truth in most of the real networks, several computer generated model graphs have been proposed [8]–[12]. Most of the existing benchmarks focus on the structure aspect of the graphs. Some of these benchmarks are LFR, and Girvan and Newman.

In this research, we selected the widely used LFR benchmark graphs as foundation for the structure and attribute benchmark. The LFR benchmark is an improvement over Girvan and Newman benchmark [11]. LFR uses power law distributions for both vertex degree and community size, and a mixing parameter which is a more realistic representation of real life network than Girvan-Newman benchmark. Mixing parameter is the ratio between the node external degree (edges to nodes outside the node's cluster) and the node degree. Both of the two benchmarks are a realization of the planted l-partition model by Condon and Karp [13].

The main goal of this research is to provide objective benchmark to analyze and assess the performance of structure and attributes clustering methods. Further more, we propose a new structure and attribute clustering method that is flexible and adaptable to different type of complex networks. Finally, the Selection method is evaluated against five clustering methods: three structure and attribute methods BAGC [6], Entropy Based [4], and SA-Cluster [3]. One structure only method Louvain [14]. One attribute only method Kmeans [15].

II. Proposed Structure and Attribute Benchmark

In this section we detail the proposed LFR-EA benchmark. It assumes the assignment attributes domain labels and attributes noise is based on uniform random distribution. The construction LFR-EA dataset proceeds as follow:

1) The structure only datasets (nodes, edges and clusters) are generated as in LFR benchmark.
2) The creation of the attributes dataset is controlled by the following inputs: i) number of attributes (nattr) ii) size of domain values for each attribute (dom_i) where i is

ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN

the attribute index. iii) Assignment influence parameter (ainf), which specify the random selection with replacing (ainf = 0) or without replacing (ainf =1).

3) All the nodes in a cluster are assumed to share the same attribute domain values.
4) If ainf is set to 1 and domain size is less than number of clusters, list of available domain values is constructed by repeating domain value until their number equal to the number of clusters.
5) For each cluster, all of the nodes in a cluster is assigned a random domain value.
6) Lastly, nodes in the cluster are selected to host the noise. The noise is a random domain value that are different that the cluster domain value. The noise level can be set differently for each attribute.
7) Steps 3 through 5 are repeated for each attribute.

To evaluate clustering methods on all of different setting of structure and attributes noise, a modified NMI measure called CNMI is introduced. CNMI allow the integration of clustering performance across structure and attribute noise. CNMI is defined in Equation (1):

$$CNMI = \frac{\sum^{\mu}\sum^{\nu} NMI}{S} \tag{1}$$

where: μ is mixing parameter (0.1 to 0.9), and ν is attributes noise (0 to 0.9), and S is number of samples (normalization factor).

III. Proposed Structure and Attribute Method

The level of information in attributed graph can be grouped into four cases: 1) clear structure and clear attributes 2) clear structure and ambiguous attributes 3) ambiguous structure and clear attributes 4) ambiguous structure and ambiguous attributes. In this section a novel clustering approach is proposed termed Selection method, Algorithm 1. It uses the mixing parameter to detect the boundary between clear and ambiguous network structure. Mixing parameter obtained by Equation (2)

$$\mu = \frac{\sum_{i=1}^{N} \frac{d_{ei}}{d_i}}{N} \tag{2}$$

where: d_{ei} is node external degree, and d_i is node degree, and N is number of nodes

IV. Experimental Results

The performance of clustering methods is evaluated using LFR-EA computer generated datasets, and real network DBLP84K datasets. A real mobile network call detailed records (CDR) of 300K nodes is analyzed to obtain a realistic parameters for LFR-EA dataset node degree and community size. LFR-EA dataset have 1000 nodes, min node degree is 25 and max is 40, min community size is 60 and max is 100. Only two attributes were selected to accommodate a limitation in one method code. The DBLP84K contains 84,170 scholars in 15 research fields. Each scholar is associated with two attributes, prolific and primary topic. Prolific attribute has domain size of three. Domain of primary topic consists of 100 research topics extracted by a topic model from a collection of paper titles.

Algorithm 1 Selection Method

```
Input: Structure method, Attribute method, $G(V, E, A)$, and $\mu_{limit}$
Output: Node community assignment $C_{sm}$
Phase 1:
    Run structure only method to obtain $C_S$
    Calculate the graph mixing parameter $\mu_S$
Phase 2:
    $if(\mu_S < \mu_{limit})then$
        return $C_S$
    $else$
        Run attribute only method to obtain $C_A$
        return $C_A$
end
```

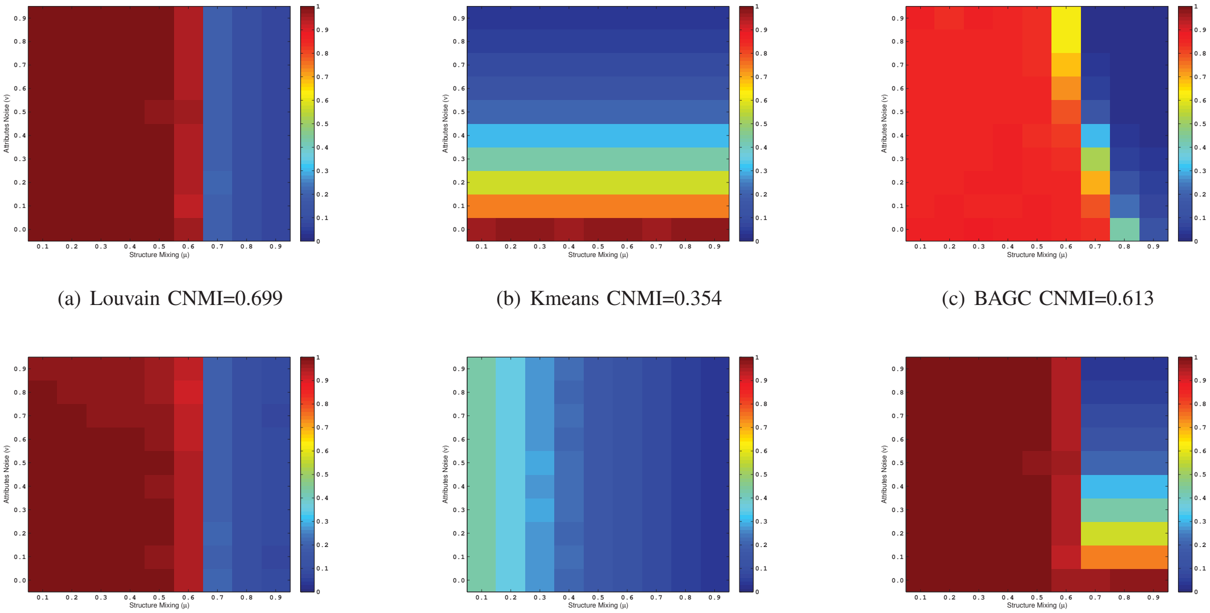
Performance Results

A Heatmap for each method is shown in Figure 1: the color range is based on a mean NMI. Each NMI mean value corresponds to 100 graph samples. The x-axis is the structure mixing parameter (μ) and y-axis is the attributes Noise (ν). (a) Louvain NMI results are constant in the y-axis because the method is structure only, and the results along the x-axis show stable high NMI results until the boundary of ambiguous structure of 0.6 mixing parameter is reached. (b) Kmeans NMI is constant in the x-axis because it is an attribute only method. Its NMI in the y-axis is sensitive to attribute noise. (c) In our setting, BAGC performed very well and shows sensitivity towards both structure and attribute. (d) In our setting, EntropyBased results are identical to structure only method. (e) SA-Cluster perform the worse in our setting, and was not able to recover the correct clusters. (f) Selection method shows ability to detect the boundary between clear and ambiguous graph structure, and combined Louvain and Kmeans results. CNMI for each method is also shown under each heatmap. The Selection method outperformed all of the other tested method.

TABLE I
Methods Modularity on DBLP84K Dataset

Methods	Modularity
Louvain	**0.62**
Kmeans	0.22
BAGC	0.53
SA-Cluster	0.15
Selection	**0.62**

Table I shows the modularity result on DBLP84K Dataset. Modularity measure was used instead of NMI or CNMI because the DBLP84K dataset lacks the ground truth. The Selection method reflects the structure only (Louvain) results



(a) Louvain CNMI=0.699

(b) Kmeans CNMI=0.354

(c) BAGC CNMI=0.613

(d) EntropyBased CNMI=0.696

(e) SA-Cluster CNMI=0.193

(f) Selection CNMI=**0.776**

Fig. 1. NMI Heatmaps of the evaluated methods on LFR-EA dataset

because the estimated mixing parameter value of 0.345 is less than the 0.6 limit value.

V. Conclusion

An effective structure and attribute benchmark and an evaluation Heatmap with CNMI measure that provide ability to show the strengths and weakens under four cases of graph information contents are introduced. Further more, a simple Selection method is introduced and outperformed the state-of-art BAGC method on computer generated and real network datasets. The Selection method allows a flexible selection of structure only and attribute only methods. The Selection method requires selection of input mixing parameter, which is dependent on the chosen structure only method and the type of the graph. The input mixing parameter can be obtained by testing the desired structure only method on LFR benchmark.

Acknowledgment

The authors would like to thank Yiping Ke, Hong Cheng, and Juan David Cruz Gomez, for providing the code of their clustering methods.

References

[1] M. E. J. Newman, “Assortative Mixing in Networks,” *Physical Review Letters*, vol. 89, no. 20, pp. 208 701+, Oct. 2002.

[2] J. Moody, “Race, School Integration, and Friendship Segregation in America,” *American Journal of Sociology*, vol. 107, no. 3, pp. 679–716, 2001.

[3] Y. Zhou, H. Cheng, and J. X. Yu, “Graph clustering based on structural/attribute similarities,” *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 718–729, Aug. 2009.

[4] J. Cruz, C. Bothorel, and F. Poulet, “Entropy based community detection in augmented social networks,” in *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, 2011, pp. 163–168.

[5] T. A. Dang and E. Viennet, “Community detection based on structural and attribute similarities,” in *International Conference on Digital Society (ICDS)*, Jan. 2012, pp. 7–14.

[6] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, “A model-based approach to attributed graph clustering,” in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’12. New York, NY, USA: ACM, 2012, pp. 505–516.

[7] D. Combe, C. Largeron, E. Egyed-Zsigmond, and M. Gery, “Getting clusters from structure data and attribute data,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, 2012, pp. 710–712.

[8] M. Kim and J. Leskovec, “Multiplicative attribute graph model of real-world networks,” *Internet Mathematics*, vol. 8, no. 1-2, pp. 113–160, 2012.

[9] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 78, no. 4, 2008.

[10] M. Kim and J. Leskovec, “Modeling social networks with node attributes using the multiplicative attribute graph model,” in *UAI*, 2011, pp. 400–409.

[11] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[12] D. A. Rachkovskij and E. M. Kussul, “Datagen: a generator of datasets for evaluation of classification algorithms,” *Pattern Recogn. Lett.*, vol. 19, no. 7, pp. 537–544, May 1998.

[13] A. Condon and R. M. Karp, “Algorithms for graph partitioning on the planted partition model,” *Random Structures and Algorithms*, vol. 18, no. 2, pp. 116–140, 2001.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10 008+, Jul. 2008.

[15] J. A. Hartigan and M. A. Wong, “A K-means clustering algorithm,” *Applied Statistics*, vol. 28, pp. 100–108, 1979.