Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods

Haithum Elhadi Gady Agam

Computer Science Department
Illinois Institute of Technology
Chicago, IL 60616
{helhadi1,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 compare our novel clustering method, termed Selection method, against seven clustering methods: three structure and attribute methods, one structure only method, one attribute only method, and two ensemble methods. The Selection method uses the graph structure ambiguity to switch between structure and attribute clustering methods. We shows that the Selection method out performed the state-of-art structure and attribute methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms Experimentation

Keywords

Social Network Analysis, Community Detection, Structure and Attributes Clustering

1. INTRODUCTION

Complex network community detection (i.e., graph clustering) is a long studied problem in machine learning and graph theory. The original goal of community detection is to group the graph vertices (i.e., nodes) into components (i.e., clusters) that contain dense connections (i.e., edges) within those components, and small number of connections to other components. The goal has been extended in recent years to go beyond the structure of the graph (vertices and edges) to the consideration of the vertices' attributes. The motivation of considering the vertices attributes is driven by the popularity of social network with rich content associated with the vertices. The new goal is to cluster the graph by using both its structure and attributes.

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 [16]. The homophily behavior can be observed in many complex network, such as social network, citation network, and others [15].

Several new clustering methods that use both structure and attributes of graphs are introduced in recent years [23, 4, 5, 22, 2]. Some of these methods are SA-Cluster, Entropy Based, SAC, and BAGC. SA-Cluster convert the attributes to edges and use 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 introduced 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 network, several computer generated model graphs have been proposed [10, 14, 9, 6, 20]. Most of existing benchmarks focus on the structure aspect of the graphs. Some of these benchmark are LFR, and Girvan and Newman.

In this research, LFR benchmark graphs is selected as foundation for our structure and attribute benchmark. The LFR benchmark is an improvement over the benchmark proposed by Girvan and Newman [6], which is a graph of 128 vertices divided into four equal communities of 32 each. LFR use power law distributions for both vertex degree and community size, a mixing parameter which is a more realistic rep-

The 7th SNA-KDD Workshop '13 (SNA-KDD'13), August 11, 2013, Chicago, United States.

resentation 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 realization of the planted l-partition model by Condon and Karp [3].

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.

Our key contributions in this paper are as follow:

- We evaluate the performance of seven clustering methods on a new benchmark using NMI heat maps.
- Describes an attributes extension to widely used LFR benchmark [12, 18, 11, 19]. The new benchmark is called LFR-EA. It provides the capability to evaluate all of three types of clustering algorithms: structure only, attribute only, and structure and attributes methods.
- A novel structure and attribute clustering method, termed Selection method, based on estimating the clustered graph mixing parameters.

The rest of the paper is organized as follow: Section 2 describes the related clustering methods. Section 3 details the benchmark. Section 4 details the Selection method, while section 5 contains the results. Section 6 provides conclusion to this paper.

2. RELATED WORK

Clustering large complex network is crucial to be able to understand and analyze these networks. Most of the clustering methods are focused on structure aspect of the complex network or the attribute aspect. In this section we discuss two examples of these methods (Louvain and Kmeans).

Later in this section, we summarize new type of clustering methods that bring the two aspects of clustering (structure and attributes) together. All of the the methods in this section are contrasted in section 5 against Selection method, which is detailed in section 4.

Louvain Method

Louvain [1] is a structure only method. The aim of this method is the optimization of modularity [17] using a hierarchical approach. The first pass partitions the original network into smaller communities. This helps in maximizing modularity in regard to vertice local movement. The step turns the first generation of clusters into super vertices with lesser-weighted graphs. This procedure is then repeated until modularity reaches a maximum point. The strengths of this method are fastness and appropriateness to analyze large graphs. Its weakness is its bias to the intrinsic limits of modularity maximization [13]

Kmeans Method

Kmeans [7] is an attribute only method. This is considered the most famous clustering algorithm. It comprises a simple procedure of re-estimation where data points are randomly assigned to K number of clusters. The centroid of each cluster is calculated, and then every data point is allocated to the cluster with the closest centroid to that point. These steps are interchanged until no change to data points assignment to clusters, and a stopping criterion is reached.

BAGC Method

BAGC is a composite structure and attribute method. It is a model-based approach for attributed graph clustering where a Bayesian probabilistic model was developed for attributed graphs and the clustering problem was formulated as a probabilistic inference problem. The cluster label of each vertex is depicted as a hidden variable. The model implements the intra-cluster similarity by maintaining the dependence of the attribute values and edge connections of a vertex on its cluster label. The algorithm proposed by this method, is considered to be an efficient and approximate approach to solve the inference problem. Moreover, the probabilistic model defines a joint probability distribution that covers all possible clustering and all possible attributed graphs [22]

Entropy Based Method

Entropy Based is a composite structure and attribute method. The algorithm proposed in this method works by maximizing the modularity by changing the composition of the communities locally. The steps to create a graph of communities include modularity optimization followed by community aggregation. Entropy optimization is included as an intermediate step between optimization and aggregation. This is done to minimize semantic disorder of the nodes by moving nodes among the clusters found during modularity optimization. These steps are iterated until the modularity is not improving any further. [4]

SA-Cluster Method

SA-Cluster is a composite structure and attribute method. This method proposes the usage of graph augmentation to define the attribute similarity by edge connectivity. The method uses the neighborhood random walk model on the attribute-augmented graph to compute a unified distance between vertices. This is based on the paths consisting of both structure and attribute edges. This results in the natural combination of the structural closeness and attribute similarity. [23]

HGPA Method

HyperGraph Partitioning Algorithm (HGPA) [21] is an ensemble method we use to combine Louvain and Kmeans clusters labels. This algorithm is a direct approach where cluster ensemble problem is posed as a partitioning problem of a hypergraph by cutting a minimal number of hyperedges. It approximates the maximum mutual information objective with minimum cut objective constrains.

CSPA Method

Cluster-based Similarity Partitioning Algorithm (CSPA) [21] is an ensemble method we use to combine Louvain and Kmeans clusters labels. In this algorithm, binary similarity matrix is used to signify relationship between objects in the same

cluster in order to establish a pairwise similarity measure that yield a combined clustering. CSPA is considered an efficient, simple and obvious heuristic to solve the cluster ensemble problem. However, its computational and storage complexity are both quadratic in number of nodes, while HGPA is almost linear.

3. STRUCTURE AND ATTRIBUTE BENCHMARK

In this section we details the LFR-EA benchmark, which assumes that the assignment of both attributes domain labels and attributes noise to a cluster is based on uniform random distribution. The construction of the attributes data set of our benchmark proceeds through the following steps:

1. The structure only data sets (nodes, edges and clusters) are generated as in LFR benchmark [14], which assumes that degree and the community size follow power laws distributions.

2. The creation of the attributes data set is controlled by the following inputs: i) number of attributes (nattr) ii) size of domain values for each attribute (dom_i) where i is the attribute index. iii) Assignment influence parameter (ainf), which specify the random selection with replacing (ainf = 0) or random selection without replacing (ainf =1).

3. All the nodes in a cluster are assumed to share the same attribute domain values.

4. The size of domain values dom_i is compared to the number of clusters in the case ainf is set to 1. If domain size is less than number of clusters, we construct the list of available domain values 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 be able to evaluate clustering methods on all of different setting of structure mixing and attribute 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):

$$CNMI = \frac{\sum^{\mu}\sum^{\nu} NMI}{S} \quad (1)$$

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).

4. NOVEL SELECTION METHOD

Most of the structure and attributes methods use modified objective functions to combine the two aspects on the complex network. In this section we detail a new approach that switch between structure and attribute based on the ambiguity of the network structure.

The level of information in attributed graph can be grouped in four cases as shown in Figure 1. These four groups are: 1) clear structure and clear attributes 2) clear structure and ambiguous attributes 3) ambiguous structure and clear attributes 4) ambiguous structure and ambiguous attributes.

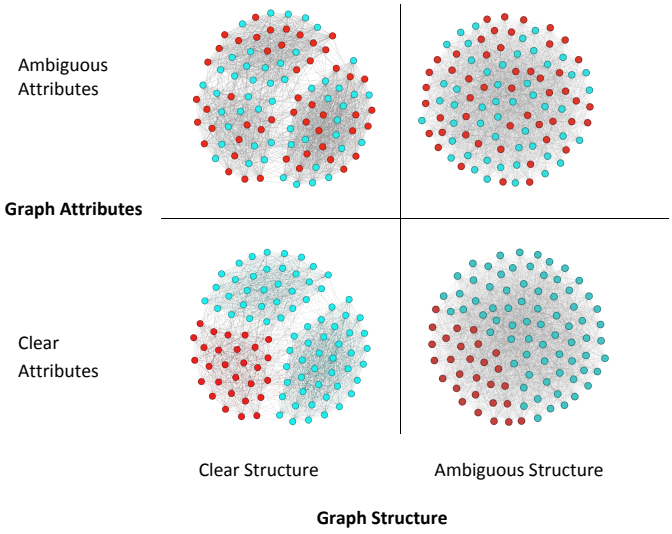


Figure 1: Graph Structure and Attributes Content

The structure only methods have proved to be scalable [1] and more resilient in recovering the correct underlined clusters (high NMI) even in the presence of many of edges to the outside clusters (mixing parameter up to 0.5). However, the structure methods suddenly loose the ability to recover the correct clusters when the cluster nodes have more edges to nodes in other clusters than to nodes in the same cluster (ambiguous structure). While this behavior of failing to recover the correct clusters is dependent on the structure method and the type of the graph,but typically occurs when graph mixing parameter is 0.6 or grater. In contrast, the attribute only methods (e.g., Kmeans) are sensitive to noise within the cluster because they try to make use of all data point.

The goal of the Selection method is to be able to achieve the following:

1. Rely on structure methods when the graph has clear structure content.
2. Detect the boundary between clear and ambiguous graph structure content.
3. Use the attributes only methods as an extension to structure only method when the graph has ambiguous structure content.

4. Allow the exchange of structure only and attribute only methods as needed to suite the different type of graphs.

It is critical to detect the boundary between clear and ambiguous structure setting. Mixing parameter is one way to accomplish that. The mixing parameter is used by the Selection method to detect the boundary between clear and ambiguous graph structure content. It can be 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

The Selection method is formulated in Equation (3) :

$$C_{sm} = \begin{cases} C_S & \text{if } \mu_s < \mu_{limit} \\ C_A & \text{otherwise} \end{cases} \tag{3}$$

where: C_{sm} is graph partition based on Selection method, C_S is partition based on structure method, C_A is partition based on attribute method, μ_s is the estimated mixing parameter for C_S, and μ_{limit} is the boundary between clear and ambiguous graph structure content.

The Selection method algorithm is listed in Algorithm 1.

Algorithm 1 Selection Method

Input:

Structure method, Attribute method, $G(V, E, A)$, μ_{limit}

Output: Node community assignment C_{sm}

Phase 1:

Run structure only method to obtain C_S

Calculate the graph mixing parameter μ_S

Phase 2:

$if(\mu_{est} < \mu_{limit})then$

return C_S

$else$

Run attribute only method to obtain C_A

return C_A

end

5. EXPERIMENTAL RESULTS

The are two ways to evaluate the performance of clustering methods: computer generated datasets and real network datasets. The computer generated datasets allow the creation of ground truth to assess the clustering methods ability in recovering them. The weakness of computer generated datasets is its limitation in representing a real network behavior. Testing with real dataset solve this issue, however, the presence of ground truth is lacking in most cases. Therefore, we always need to test using both methods. The following section details the tow types of datasets that are used in this paper.

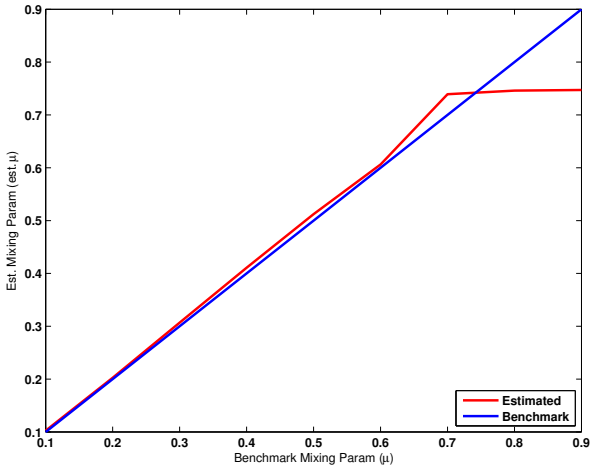


Figure 3: Estimated mixing parameter

5.1 Data Sets

LFR-EA Dataset

The parameters to generate LFR-EA dataset is shown in Table 1. The table shows the structure and attributes parameters that allow the testing of all the eight methods described in Sections 2 and 4. The values of node degree and community size parameters are chosen to reflect a real mobile social network. A subset of monthly call detailed records (CDR) of a real mobile network (300K nodes) are analyzed. Only two attributes were selected to accommodate a limitation in one method code. The attribute assignment influence parameter (ainf), is set to random selection without replacing, to coverage each domain values across the difference clusters.

DBLP84K Dataset

This dataset is known as DBLP co-author network. It contains 84,170 nodes for scholars in 15 research fields: database, data mining, information retrieval, artificial intelligence, machine learning, computer vision, networking, multimedia, computer systems, simulation, theory, architecture, natural language processing, human-computer interaction, and programming language. Each scholar is associated with two attributes: prolific and primary topic. Prolific attribute has domain size of three: highly prolific for the scholars with $\geq$ 20 publications; prolific for the scholars with $\geq$ 10 and $<$ 20 publications; and low prolific for the scholars with $<$ 10 publications. The domain of the attribute primary topic consists of 100 research topics extracted by a topic model from a collection of paper titles [8].

5.2 Performance Results

Figure 3 show the plot of estimated structure mixing parameter for Louvain method against actual mixing parameter value that is used to generate the test graph by LFR-EA benchmark. Each point corresponds to 100 graph samples.

Table 1: LFR-EA Benchmark Settings

Structure Parameters	Attributes Parameters
Number of nodes (N) = 1000	Number of Attributes (nattr) = 2
Avg. node degree (k) = 25	Attribute's domain cluster assignments (ainf) = 1
Max node degree (maxk) = 40	
Mixing parameter (μ) = 0.1, 0.2, ..., 0.9	Attribute # 1:
Exponent for the degree ($\tau 1$) = 2	Attribute domain size (dom1) = 3
Exponent for the community size ($\tau 2$) = 1	Attribute noise ($\nu 1$) = 0.0, 0.1, ..., 0.9
Minimum for the community sizes (minc) = 60	
Maximum for the community sizes (maxc) = 100	Attribute # 2:
Number of overlapping nodes (on) = 0	Attribute domain size (dom2) = 15
Number of memberships of the overlapping nodes (om) = 0	Attribute noise ($\nu 2$) = 0.0, 0.1, ..., 0.9

The results show that the estimated Louvain method mixing parameter is very close to the benchmark value until it reaches 0.6 value, which is the same point where Louvain method NMI significantly drops as shown in LFR benchmark [14].

The following section illustrates the heatmaps for each method in Figure 2. The color range is based on mean NMI, each NMI mean value corresponds to 100 graph samples. The x-axis represents the structure mixing parameter (μ) and the y-axis represents the attributes noise (ν). Louvain NMI results in heatmap (a) is constant in the y-axis because the method is structure only, and the results along the x-axis show stable high NMI until the boundary of ambiguous structure of 0.6 mixing parameter is reached. In heatmap (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. In heatmap (c), BAGC performed very well and was sensitive to both structure and attribute. However, The BAGC method didn't use the clear attribute content to overcome the structure ambiguity effect. In our setting, Entropy Based method shown in heatmap (d) didn't use the attributes information and its result was identical to structure only method. SA-Cluster in heatmap (e) perform the worse in our setting, and was not able to recover the correct clusters. We use the ensembles methods in (f) and (g) to combine structure only (Louvain) and attribute only (Kmeanse). The results of the ensembles methods were affected by the low Kmeans NMI in the clear structure region of Louvain, which resulted in a low CNMI overall. Heatmap (h) shows that Selection method was able to detect the boundary between the clear and ambiguous graph structure and combined Louvain and Kmeans results.

CNMI for each method on the LFR-EA dataset is shown in Table 2. The results show that the Selection method outperformed all of the other tested methods.

The modularity result on DBLP84K Dataset is shown in Table 3. Modularity measure is used instead of NMI or CNMI because the DBLP84K dataset lacks the ground truth which is a requirement for NMI based measures. The Selection method reflects the structure only (Louvain) results because the estimated mixing parameter value of 0.345 is less than the 0.6 limit value.

Table 2: Methods Cumulative NMI Results on LFR-EA dataset

Method	Type	CNMI
Louvain	Structure	0.699
Kmeans	Attributes	0.354
BAGC	Composite	0.613
EntropyBased	Composite	0.696
SA-Cluster	Composite	0.193
Selection	Switching	**0.776**
HGPA	Ensemble	0.454
CSPA	Ensemble	0.482

Table 3: Methods Modularity on DBLP84K Dataset

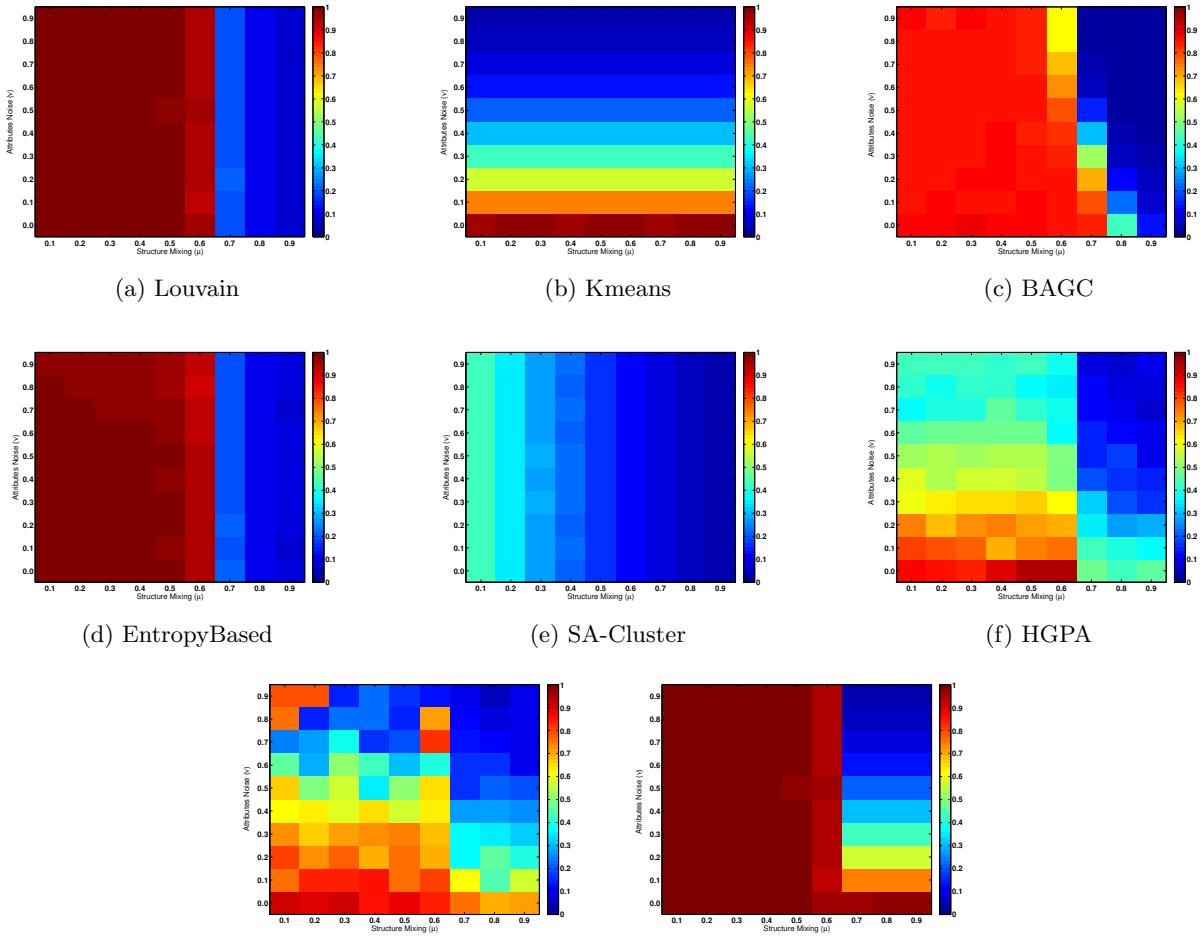
Method	Modularity
Louvain	**0.62**
Kmeans	0.22
BAGC	0.53
SA-Cluster	0.15
Selection	**0.62**

6. CONCLUSION

In this research, the strengths and weakens of different community detection methods are evaluated under the spectrum of the four graph information contents cases. Further more the simple Selection method presented in this paper, outperform the rest of the tested methods 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 careful selection of mixing parameter threshold, which is dependent on the chosen structure only method and the type of the graph.

Acknowledgment

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



(a) Louvain (b) Kmeans (c) BAGC

(d) EntropyBased (e) SA-Cluster (f) HGPA

(g) CSPA (h) Selection

Figure 2: NMI Heatmaps of the evaluated methods on LFR-EA dataset

7. REFERENCES

[1] 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*, 2008(10):P10008+, July 2008.

[2] 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*, pages 710–712, 2012.

[3] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

[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*, pages 163–168, 2011.

[5] T. A. Dang and E. Viennet. Community detection based on structural and attribute similarities. In *International Conference on Digital Society (ICDS)*, pages 7–14, Jan. 2012. ISBN: 978-1-61208-176-2. Best paper award.

[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[7] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[9] M. Kim and J. Leskovec. Modeling social networks with node attributes using the multiplicative attribute graph model. In *UAI*, pages 400–409, 2011.

[10] M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

[11] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E (Statistical, Nonlinear, and Soft*

Matter Physics), 80(1):016118+, 2009.

[12] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117, Nov 2009.

[13] A. Lancichinetti and S. Fortunato. Limits of modularity maximization in community detection. *Phys. Rev. E*, 84:066122, Dec 2011.

[14] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 78(4), 2008.

[15] J. Moody. Race, School Integration, and Friendship Segregation in America. *American Journal of Sociology*, 107(3):679–716, 2001.

[16] M. E. J. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701+, Oct. 2002.

[17] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, Aug. 2003.

[18] G. K. Orman and V. Labatut. The Effect of Network Realism on Community Detection Algorithms. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 301–305. IEEE Computer Society, Aug. 2010.

[19] G. K. Orman, V. Labatut, and H. Cherifi. Qualitative comparison of community detection algorithms. In *DICTAP (2)*, pages 265–279, 2011.

[20] D. A. Rachkovskij and E. M. Kussul. Datagen: a generator of datasets for evaluation of classification algorithms. *Pattern Recogn. Lett.*, 19(7):537–544, May 1998.

[21] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, Mar. 2003.

[22] 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*, SIGMOD '12, pages 505–516, New York, NY, USA, 2012. ACM.

[23] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.*, 2(1):718–729, Aug. 2009.