ATLANTIS PRESS
Advances in Computer Science Research, volume 44
3rd International Conference on Wireless Communication and Sensor Network (WCSN 2016)

A Survey of Clustering Algorithms in Community Detection of Complex Networks

Le Bo
Department of Computer Science, National University of Defense Technology
Changsha, China
yebo792653@126.com

Yang Shi
Department of Computer Science, National University of Defense Technology
Changsha, China
Shiyang7@qq.com

Hao Fang
Department of Computer Science, National University of Defense Technology
Changsha, China
306477486@qq.com

Mei Wen
Department of Computer Science, National University of Defense Technology
Changsha, China
Wenmei8086@163.com

***Abstract*—Complex network analysis begins in the 1930s, and the community structure in complex networks is a major characteristic that has been widely concerned. This paper introduces the basic and core ideas of several typical clustering algorithms in community detecting, and analyzes the advantages and disadvantages of each one. And it also reviews the background, the significance and the performance of these algorithms. Some algorithms may perform well in some specific areas but not in community detecting, while the newest algorithms still need to be tested in the future.**

Keywords-Complex Networks; Clustering; Community Detection

I. INTRODUCTION

Networks, which is called graph in mathematics, was studied by Euler in 1736 on the Konigsberg's seven bridges problem [1]. However, the research about graph developed slowly until 1936, in which the first book about Graph Theory was published.

In 1960s, two Hungarian mathematicians Erdos and Renyi set up a random graph theory [2], which was regarded as a systematic study of complex networks theory in mathematics. In the next 40 years, people had been using random graph theory as the basic theory of complex networks research. However, most of the real networks are not completely random. In 1998, Watts and Strogatz [3] revealed the "small world" character of complex networks. Subsequently, in 1999, Dr. Barabasi and Albert [4] revealed the feature "scale-free" and from then on a new era started in the study of complex networks. Fig. 1 shows a kind of scientific collaboration networks (SCN).

With further study, more and more properties of complex networks were found out. One of the most important is Girvan and Newman's research [5]which pointed out that the clustering is ubiquitous in complex networks and each cluster could be regarded as a community. The community detection problem has been studied a lot after that and a large number of algorithms have been generated.

A simple principle of judging the quality of the detection on the community is to determine the edges in the community as much as possible. Another complex but common method is the modularity proposed by Newman [6].

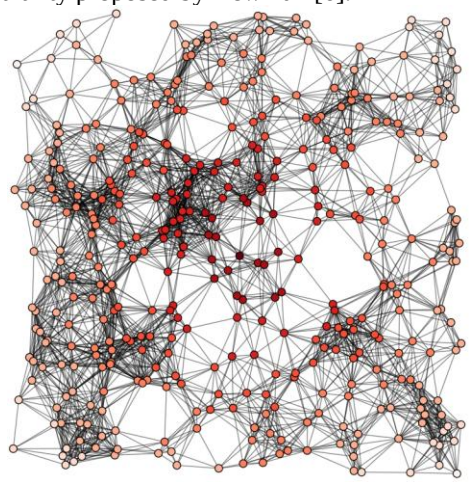

Figure 1. Scientific collaboration networks

The basic idea is that the summation of the dissimilarity between the values of all the sub networks is called modularity of this complex network. The formula is expressed as follows:

$$\mathrm{Q} = \frac{1}{2\mathrm{m}}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) = \sum_{c=1}^{n_c}[\frac{l_c}{m} - (\frac{d_c}{2m})^2]$$

where A_{ij} is the adjacency matrix, k_i is the degree of point i , m is the number of edges, $\mathrm{k}_i * \mathrm{k}_j/2\mathrm{m}$ is the expectation between the edges of point i and point j. Further, the modular degree can be changed into the form of the right side of the equation, where n_c is the total number of associations, l_c is the number of edges in the community, and d_c is the sum of the degree of points in the community.

For different clustering, the states and requirements must be figured out before the selection of algorithm such as

- The clustering is exclusive or overlapping.
- Based on hierarchy or partitioning.
- The number of cluster is fixed or unlimited.
- Based on distance or distribution model.

II. Classification of Clustering Algorithms

A. Based on Hierarchy(Agglomerative)

The hierarchical method calculates the distance among samples first. Based on the distance, it merges the nearest points to the same class every time. Then, calculate the distance between two classes, and merge the nearest classes into a larger class. The combination will not stop until the synthesis of all the classes. To calculate the distance between the classes, there are several methods such as the shortest distance method, the longest distance method, the middle distance method, the class average method, etc.

1) Newman Algorithm: It is a kind of aggregation algorithm based on the greedy algorithm, which can be used to analyze the complex network with 1 million nodes.

a) Basic idea: choose the two communities ceaselessly which influence the modularity most. Q represents the change of modularity. The time complexity is O(m(m+n)). We can get a tree structure of the decomposition from the community. Choosing breakpoints in different positions can generate different community structure.

b) Optimization: merging multiple communities at each iteration, normalize ΔQ, and eliminate the impact of community size. Clause, Newman and Moore reduced the time complexity to $O(nlog^2n)$ [7]

c) Advantages: It reduces the time complexity and does not need to specify the number of communities in advance.

d) Disadvantages: the merging process is irreversible and the algorithm tends to be more sensitive to some single points. Due to the lack of global objective function like K-means, there is no local minimum problem or the difficulty to choose the initial point. The operation of the merging is often the final, once the merging of two clusters are not revoked. Of course, the cost of storage is expensive.

2) BRICH Algorithm[8]*:* It is mainly used to process a large amount of numeral data. Firstly, the object set is partitioned by the tree structure, and then other clustering methods are applied to optimize the clustering.

3) ROCK Algorithm[9]*:* Mainly used for the categorical data type with a random sampling technology.

B. Based on Hierarchy(Divisive)

The basic idea is to find out the edges that most likely to be located in the edge of the community. Removing these edges will naturally produce a different community. The representative algorithm is proposed by Girvan and Newman [10].

1) GN Algorithm: First of all, the edge betweenness refers to the number of the shortest path through the edge. The GN algorithm is a typical divisive hierarchical clustering algorithm. The basic idea is to remove the edge with the greatest betweenness continuously.

- Calculate each betweenness of edges in the network
- Remove the edge with the greatest betweenness
- Recalculate the remaining betwenness
- Repeat 2 and 3 until each node is a single community

2) CURE Algorithm[11]: In this algorithm, each data point is considered as a cluster, and then merge the nearest cluster until the number of clusters is required. CURE algorithm extract points with a fixed number and better distribution as representative points rather than the traditional representation, such as the center, radius or points. These points are multiplied by a proper contraction factor, which makes them closer to the center of the cluster. CURE algorithm uses random sampling and segmentation method to improve the space and time efficiency, and the algorithm used in the heap and K-d tree structure to promote the efficiency of the algorithm.

3) Chameleon Algorithm[12]: It configures a K-nearest neighbor graph G_k by the data set, then a graph partitioning algorithm is applied to map G_k into a large number of subgraphs. Each subgraph represents an initial sub cluster, finally an agglomerative hierarchical clustering algorithm is used to generate composite anti sub clusters and find the real result of the cluster. The clustering effect of Chameleon is considered to be very powerful, and it is better than BIRCH, but the computational complexity is $O\ (n^2)$, higher than BIRCH.

4) Pros&Cons:

a) Advantages:

- Clustering granularity can be controlled flexibly.
- Doesn't need specify the number of clustering in advance
- The hierarchical relationship among clusters can be easily detected
- Suitable for the data set with arbitrary shape and attribute of arbitrary type, relatively high scalability in general.

b) Disadvantages:

- Relatively high in time complexity in general
- Singular value make a big dissimilarity
- No backtracking
- The number of clusters needed to be preset

5) Comparision: The agglomerative and divisive algorithm are corresponding, one is bottom-up and the other is top-down; one merge points together and another remove the edge to divide points. The pros and cons of divisive clustering are similar to the agglomerative clustering.

C. Based on Partition

The basic idea of partitioning clustering algorithms is to regard the center of data points as the center of the corresponding cluster. Besides K-means [13] and K-medoids [14], the typical clustering algorithms based on partition include PAM [15], CLARA [16], CLARANS [17].

1) K-means:

a) Basic idea: K-means divides n objects into k clusters,

and assures that there's a striking similarity in the cluster and few affinity among different clusters. First of all, choose k cluster centroids randomly as:

$$\mu_1\mu_2\mu_3 \in R^n$$

calculate $c^{(i)}$ *of each rest cluster:*

$$c^{(i)} := \arg min_j \left|\left|x^{(i)} - u_j\right|\right|^2$$

Recalculate the cluster centroids:

$$u_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\}x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

This process will be repeated constantly until the convergence and the centroid does not change significantly.

K is the number of cluster centroids specified in advance, $c^{(i)}$ represents the nearest cluster to cluster centroids in rest clusters and its value is between 1 and k. u_j represents the conjecture at the center of a sample of the same cluster. When the result is dense and the dissimilarity between the clusters is obvious, K-means performs better. It is relatively scalable and efficient when processing large-scale data sets. The time complexity is $O(nkt)$. Usually $k << n$ and $t << n$ (t represents iterations) which means the algorithm often ends with a local optimal result.

b) Advantages:

- Easy to understand and implement
- Low time complexity

c) Disadvantages: the initial value of k is very sensitive to noise and outliers and only applicable for the numerical datatype and convex data. Among these shortcomes, the sensitivity of the value of k is the most important. Users must determine the number of k in advance. The choice of k is generally based on the previous experiences and results of different experiments, the value of k is not available for reference on different data sets

d) Optimization: Each of optimization is corresponding to the disadvantages and there is no necessity to determine K in advance: such as K-means++, intelligent K-means, genetic K-means; K-medoids, K-medians, K-modes and kernel K-means.

2) K-medoids:

a) Basic idea: K-medoids is the cure of sensitivity to noise and outliers in K-means. It proposes a new way to select cluster centroids(which is named medoid here). After each iteration, the medoid is selected from the sample points of the cluster, and the selection depends on that whether the new medoid can improve the clustering quality and make the cluster more compact, and the absolute error of new medoid tends to reduce constantly before stabilization.

b) Comparsion: The difference between k-means and k-medoids is similar to the difference between the mean and the median of a data sample: the former's range can be any value in a continuous space, and the latter can only be selected at the point where the sample is given. One of the most direct reason is that the K-means is too strict with the data, because it uses the Euclidean distance to describe the dissimilarity between data points, which can be directly calculated by the center point. However, most data set cannot meet such requirements. For example, height can be very natural to deal with in this method, but categorical type cannot. Therefore, the Euclidean distance of the original objective function J is changed to an arbitrary measure dissimilarity function ν:

$$\tilde{J} = \sum_{n=1}^{N}\sum_{k=1}^{K} \Upsilon_{nk}\, \nu(x_n, \mu_k)$$

In addition, since the medoid is selected in the existing data points, it is unlikely to be affected by error, make it more robust.

3) Some other algorithm: PAM is a typical k-medoids algorithm. CLARA can handle larger data sets than PAM and its effectiveness depends on the size of the samples. However, when the center of a sample is not the best center point, CLARA cannot generate the global best results. CLARANS is proposed on the basis of the CLARA algorithm. Being different with CLARA, CLARANS is not in any given time limited to any sample, but in the search for each step with a certain random selection of a sample.

D. Based on Density

The hierarchical clustering algorithm and partition clustering algorithm are only suitable for convex cluster. In order to make up for this deficiency, the density algorithm was proposed to find clusters of arbitrary shapes. Density-based algorithm is not sensitive to noise and is capable to filter the low density area and find the dense sample point.

1) DBSCAN: DBSCAN [18] is a traditional algorithm which is sensitive to the main two parameters (ε and $minPts$).

a) Basic idea: DBSCAN requires that the number of objects (points or other spatial objects) in a cluster and not smaller than a given value, which is a density threshold. DBSCAN is not sensitive to the order of the samples in the database, which means the input sequence of patterns has little impact on the final results. However, it depends on the order how the boundary clusters are detected.

b) Advantages: Comparing with K-means, DBSCAN doesn't require the number of clusters in advance. The clustering is fast and can effectively deal with the noise points and find the spatial clustering of arbitrary shape.

c) Disadvantages: DBSCAN cannot be a good reflection of high dimensional data or the density of the data set and its consumption of memory and I/O is considerable when data scale is huge. The point is extremely sparse in high dimensional data, and the density is difficult to define. When the density of spatial clustering is not uniform, and the dissimilarity of the distance between clusters is very quite obvious. Therefore, some points may be mistaken for outliers or boundary points.

2) OPTICS [19]*:* This algorithm overcomes the biggest shortcoming in DBSCAN—the sensitivity to parameters.

a) Basic idea: OPTICS algorithm generates an ordered list of objects, and each object has two properties-the core distance and reachable distance. Using this list, we can get the clustering of any radius smaller than the ε.In other words, the clustering results from this list can be obtained based on anyεand $minPts$.

b) Optimization: There are many types of optimizations for DBSCAN, such as DENCLUE [20] and DBCLASD [21]. They all improve the DBSCAN for the different density distribution of the data.

3) CFSFDP[22]*:* This algorithm is based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. And it forms the basis of a clustering procedure in which the number of clusters arises intuitively. Outliers are automatically spotted and excluded from the analysis, and clusters are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. This algorithm can achieve a very good clustering result for all kinds of data scale.

E. Based on Spectral Graph Theory

Spectral clustering is a clustering method based on graph theory. The weighted undirected graph is divided into two or more of the optimal sub graphs. Making sure that the interior of the sub graph is as similar as possible, while the distance between the subgraphs are as far as possible. The above clustering is capable of identifying any shape of the sample space and converging to the global optimal solution.

The computational bottleneck of the spectral algorithm is the characteristic value of the matrix, and the essence of the spectral algorithm is matrix decomposition. The basic idea of matrix decomposition is to map the point from one space to another, and to cluster in the new space using the traditional clustering method

1) SM[23]*:* The core idea of SM which is usually used for image segmentation and it minimizes the normalized cut by heuristic method, based on the eigenvector.

2) NJW[24]*: N*JW carries out the clustering analysis in the feature space constructed by the eigenvectors corresponding to the k largest eigenvalues of the Laplacian matrix. Laplacian matrix transforms the discrete clustering into continuous feature vectors, and the smallest series of feature vectors correspond to the graph optimal series partition method.

F. Based on Distribution

The basic idea is that the data, generated from the same distribution, belongs to the same cluster if there exists several distributions in the original data. The typical algorithms are DBCLASD and GMM [25]. DBCLASD has been discussed in the density-based algorithm.

The basic idea of GMM is that GMM consists of several Gaussian distributions from which the original data is generated and the data, obeying the same independent Gaussian distribution, is considered to belong to the same cluster. In fact, GMM is similar to K-means. In short, in K-means, each data points will be assigned into one of the cluster, but in GMM, the possibility of these data points being assigned into each cluster is figured out, so it's called soft assignment.

III. COMPREHENSIVE PERFORMANCE EVALUATION

The detailed and comprehensive comparisons of all the discussed clustering algorithms are listed in Table 1.

TABLE 1. COMPREHENSIVE PERFORMANCE EVALUATION

Category	Algorithm	Complexity	Scalability	For large scale data	For high dimensional data
Based on Hierarchy	Cure	$O(n^2 * \log n)$	High	Yes	Yes
	BRICH	$O(n)$	High	Yes	No
	ROCK	$O(n^2 * \log n)$	Middle	No	Yes
	Chameleon	$O(n^2)$	High	No	No
	Newman	$O((m+n)n)$	Middle	Yes	Yes
	GN	$O(n * m^2)$	Middle	Yes	Yes
Based on Partition	K-means	$O(knt)$	Middle	Yes	No
	K-medoids	$O(k(n-k)^2)$	Low	No	No
	PAM	$O(k^3 * n^2)$	Low	No	No
	CLARA	$O(ks^2 + k(n-k))$	High	Yes	No
	CLARANS	$O(n^2)$	Middle	Yes	No
Based on Density	DBSCAN	$O(n * \log n)$	Middle	Yes	No
	Optics	$O(n * \log n)$	Middle	Yes	No
	CFSFDP	$O(n * \log n)$	Middle	Yes	Yes
	DENCLUE	$O(n * \log n)$	Middle	Yes	Yes
	DBCLASD	$O(n * \log n)$	Middle	Yes	Yes
Based on Spectral Graph Theory	SM	High	Middle	No	Yes
	NJW	High	Middle	No	Yes
based on distribution	GMM	$O(n^2 * kt)$	Middle	No	No

IV. CONCLUSION

The main purpose of the paper is to introduce the basic and core idea of each commonly used clustering algorithm, specify the source, and analyze the advantages and disadvantages of each one. The mentioned clustering algorithms above, with high practical value and well studied, are discussed in detail so as to give readers a systematical and clear view of the important data analysis method in community detection.

Many algorithms are currently the only discussed theoritically, based certain assumptions, such as clustering can be separated, no prominent outlier data. However, the reality is usually very complicated and noisy. How to effectively eliminate the influence of noise and improve the ability of processing real data remains to be further to improved.

ACKNOWLEDGMENTS

It's my pleasure to thank my friends Wenxiang Yang and Ning Li for their help and discussions during preparation of the manuscript.

REFERENCES

[1] Shekhar S, Xiong H. Problem of Seven Bridges of Konigsberg [M]. Springer US, 2008.

[2] Panagiotoua K, Spöhela R. Explosive Percolation in Erdos-Rényi-Like Random Graph Processes [J]. Electronic Notes in Discrete Mathematics, 2011, 22(1):699–704.

[3] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. [J]. Nature, 1998, 393(6684):440-2.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. science, 286(5439):509–512, 1999

[5] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):7821-7826.

[6] M. E. J. Newman (2006).Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A. 103 (23): 8577–8582..

[7] Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. Phys Rev E, 69(2), 26113.

[8] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases [J]. Acm Sigmod Record, 1996, 25(2):103-114.

[9] Guha S, Rastogi R, Shim K. Rock: A robust clustering algorithm for categorical attributes ☆ [J]. Information Systems, 2000, 25(5):345-366.

[10] Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.

[11] Guha S, Rastogi R, Shim K. Cure: an efficient clustering algorithm for large databases ☆ [J]. Information Systems, 1998, 26(1): 35-58.

[12] Nguyen A T, Li B, Eliassen F. Chameleon: Adaptive Peer-toPeer Streaming with Network Coding [J]. Proceedings - IEEE INFOCOM, 2010:2088-2096.

[13] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm [J]. Applied Statistics, 1979, 28(1):100-108.

[14] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering [J]. Expert Systems with Applications, 2009, 36(2):3336-3341.

[15] Wu G D, Chen J, Hoffmann C, et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes [J]. Science, 2011, 334(6052):105-8.

[16] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment [J]. Journal of the Acm, 1999, 46(5):604-632.

[17] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment [J]. Journal of the Acm, 1999, 46(5):604-632.

[18] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// 2008:226--231.

[19] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[J]. ACM SIGMOD Record, 1999, 28(2):49-60.

[20] Hinneburg A, Keim D A. A General Approach to Clustering in Large Databases with Noise[J]. Knowledge & Information Systems, 2003, 5(4):387-415.

[21] Xu X, Ester M, Kriegel H P, et al. A distribution-based clustering algorithm for mining in large spatial databases[C]// International Conference on Data Engineering. IEEE, 1998:324-331.

[22] Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks. [J]. Science, 2014, 344(6191):1492-6.

[23] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment [J]. Journal of the Acm, 1999, 46(5):604-632.

[24] Ng A Y, Jordan M I, Weiss Y. On Spectral Clustering: Analysis and an algorithm [J]. Proceedings of Advances in Neural Information Processing Systems, 2002, 14:849--856.

[25] Rother C, Kolmogorov V, and Blake A. "Grab Cut": interactive foreground extraction using iterated graph cuts [J]. A cm Transactions on Graphics, 2004, 23(3): págs. 307-312.

[26] Bhaskar DasGupta and Devendra Desai. On the complexity of newmans community finding approach for biological and social networks. Journal of Computer and System Sciences, 79(1):50–67, 2013.

[27] Daniel A Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning.SIAM Journal on Computing, 42(1):1–26, 2013.

[28] Steven H Strogatz. Exploring complex networks. Nature, 10(6825): 68–276, 2001.

[29] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world' networks. Nature, 93(6684): 40–442, 1998

[30] Yourim Yoon and Yong-Hyuk Kim. Vertex ordering, clustering, and their application to graph partitioning. Applied Mathematics & Information Sciences, 8(1), 2014.

[31] Y. Xiong and D. Yeung, "Mixtures of ARMA models for model-based time series clustering," in Proc. IEEE Int. Conf. Data Mining, 2002, pp. 717–720

[32] Borge-Holthoefer J., and A. Arenas, (2010). Semantic networks: structure and dynamics. Entropy 12, 1264–1302.

[33] Corominas Murtra B., S. Valverde, and R.V. Solé, (2007). Emergence of scale-free syntax networks. preprint arXiv:0709.4344

[34] Solé R.V., and L.F. Seoane, (2014). Ambiguity in language networks. The Linguistic Review 32(1), 5-35.

[35] Zhou S., G. Hu, Z. Zhang, and J. Guan, (2008). An empirical study of Chinese language networks. Physica A 387 3039–3047.

[36] Newman M.E.J.,(2012). Communities, modulesand large-scale structurein networks. Nature Phys. 8 25–31.

[37] Holovatch Yu., and V. Palchykov, (2007). Mykyta the Fox and networks of language. J. Phys. Stud. 11, 22–33.

[38] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. J. Mach. Learn. Res., 9:1981–2014, June 2008.

[39] Md Altaf-Ul-Amin, Yoko Shinbo, Kenji Mihara, Ken Kurokawa, and Shigehiko Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC bioinformatics, 7(1):207, 2006.

[40] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. Physical review E, 70(6):066111, 2004.

[41] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. CoRR, abs/1206.3552, 2012.

[42] Jie Chen and Yousef Saad. Dense subgraph extraction with application to community detection. Knowledge and Data Engineering, IEEE Transactions on, 24(7):1216–1230, 2012.

[43] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008.

[44] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005(09):P09008, 2005.

[45] Michel Crampes and Michel Plantié. A unified community detection, visualization and analysis method. CoRR, abs/1301.7006, 2013.

[46] Peng Jiang and Mona Singh. Spici: a fast clustering algorithm for large biological networks. Bioinformatics, 26(8):1105–1111, 2010

[47] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110(36):14534–14539, 2013.

[48] Santo Fortunato. Community detection in graphs. CoRR, abs/0906.0612, 2009.

[49] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. Physical review E, 80(5):056117, 2009.

[50] Jure Leskovec, Kevin J. Lang, and Michael W. Mahoney. Empirical comparison of algorithms for network community detection. CoRR, abs/1004.3539, 2010.

[51] Kathy Macropol and Ambuj K. Singh. Scalable discovery of best clusters on large graphs. PVLDB, 3(1):693–702, 2010.

[52] Fragkiskos D. Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. CoRR, abs/1308.0971, 2013.

[53] Michael Ovelgönne and Andreas Geyer-Schulz. An ensemble learning strategy for graph clustering. In Graph Partitioning and Graph Clustering, pages 187–206, 2012.

[54] Günce Keziban Orman, Vincent Labatut, and Hocine Cherifi. Comparative evaluation of community detection algorithms: a topological approach. Journal of Statistical Mechanics: Theory and Experiment, 2012(08):P08001, 2012.

[55] Mark EJ Newman. Finding community structure in networks using the eigen vectors of matrices. Physical review E, 74(3):036104, 2006.

[56] Farnaz Moradi, Tomas Olovsson, and Philippas Tsigas. An evaluation of community detection algorithms on large-scale email traffic. In SEA, pages 283–294, 2012.

[57] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti.Mixing local and global information for community detection in large networks.J. Comput. Syst. Sci., 80(1):72–87, 2014.

[58] Aaron F. McDaid, Derek Greene, and Neil J. Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. CoRR, abs/1110.2515, 2011.