Evaluation of Community Mining Algorithms in the Presence of Attributes

Reihaneh Rabbany and Osmar R. Zaïane

Department of Computing Science,
University of Alberta
{rabbanyk,zaiane}@ualberta.ca

Abstract. Grouping data points is one of the fundamental tasks in data mining, commonly known as clustering. In the case of interrelated data, when data is represented in the form of nodes and their relationships, the grouping is referred to as *community*. A community is often defined based on the connectivity of nodes rather than their attributes or features. The variety of definitions and methods and its subjective nature, makes the evaluation of community mining methods non-trivial. In this paper we point out the critical issues in the common evaluation practices, and discuss the alternatives. In particular, we focus on the common practice of using attributes as the ground-truth communities in large real networks. We suggest to treat these attributes as another source of information, and to use them to refine the communities and tune parameters.

Keywords: Network Clusters, Community Mining, Networks with Attributes, Community Evaluation, Community Validation

1 Introduction and Related Works

One fundamental property of real networks is that they tend to organize according to an underlying modular structure [9]. Clustering networks (a.k.a community mining) has direct application such as module identification in biological networks; for example clusters in protein-protein interaction networks outline protein complexes and parts of pathways [47]. Clustering networks is also an intermediate step for further analyses of networks such as link and attribute prediction which are the basis of targeted advertising and recommendation systems; for example clusters of hyperlinks between web pages in the WWW outline pages with closely related topics, and are used to refine the search results [2].

A cluster in a network a.k.a community is loosely defined as groups of nodes that have relatively more links between themselves than to the rest of the network. This definition is interpreted in the literature in many different ways, e.g. a group of nodes that: have structural similarity [48], are connected with cliques [33], within them a random walk is likely to trap [34], follow the same leader node [36], coding based on them gives efficient compression of the graph [43], are separated from the rest by minimum cut, or conductance [22], the number of links between them is more than chance [29, 1].

Fortunato [8] shows that the different community mining algorithms discover communities from different perspective and may outperform others in specific classes of networks. Therefore, an important research direction is to evaluate and compare the results of different community mining algorithms. An intuitive practice is to validate the results partly by a human expert [24]. However, the community mining problem is NP-complete; the human expert validation is limited, and is based on narrow intuition rather than on an exhaustive examination of the relations in the given network, specially for large real networks.

There is a congruence relation between defining communities and evaluating community mining results. In fact, the well-known Q-modularity by Newman and Girvan [28] which is commonly used as an objective function for community detection, was originally proposed for quantifying the goodness of the community structure, and is still used for evaluating the algorithms [4, 41]. More generally, the *internal evaluation* practice verifies whether a clustering structure produced by an algorithm matches the underlying structure of the data, using only information inherent in the data [13]. The main problem with this type of evaluation is the assumption it makes about what are good communities, and hence is not appropriate to validate results of algorithms built upon different assumptions. In our earlier works in Rabbany et al. [35, 38], we presented an extensive set of general objectives for evaluation of network clustering algorithms, mostly adapted from clustering background such as Variance Ratio Criterion, Silhouette Width Criterion, Dunn index, etc. Our experiments revealed that the ranking of these measures depends on the experiment settings, and there is not one to rule them all. This is not surprising as an evaluation criterion encompasses the same non-triviality as of the community mining task itself.

Another common evaluation practice is the *external evaluation*, which involves measuring the agreement between the discovered communities and the ground-truth structure in benchmark datasets [17, 32, 12, 3, 7]. There are few and typically small real world benchmarks with known communities available. Therefore the external evaluation is usually performed on synthetic benchmarks or on large networks with explicit or predefined communities. In the following, we discuss the issues and considerations with these types of evaluation.

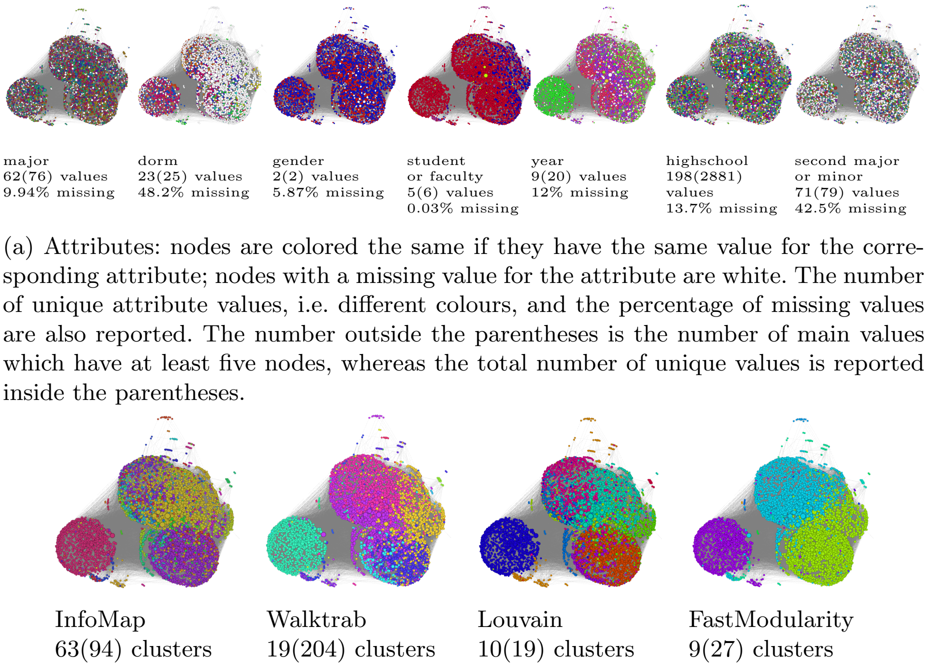
The external evaluation is not applicable in real-world networks, as the ground-truth is not available. However we assume that the performance of an algorithm on the synthetic benchmarks, is a predictor of its performance on real networks. For this assumption to hold, we need realistic synthetic benchmarks, with tunable parameters for different domains; since it has been shown that the characteristics of clusters in networks are remarkably similar between networks from the same domain [19, 30]. However, the current common generators used for synthesizing benchmarks, such as the LFR benchmarks [18], are domain-independent and also overlook some characteristics of the real networks [27, 31]. Consequently, there are recent studies which try to improve the synthetic benchmark generators, including our recent works in [20].

Alternative to generating benchmarks for the community detection task, large real world benchmarks are often used where the ground-truth communities are defined based on some explicit properties of the nodes such as user memberships in social network. Notably Yang and Leskovec [49] adapt this approach to compare different community detection algorithms based on their performance on large real world benchmarks; where characteristics such as social groups are considered "reliable and robust notion of ground-truth communities". For example, in a collaboration network of authors obtained from DBLP, venues are considered as the ground-truth communities, or in the Amazon product copurchasing network, product categories are considered as the ground-truth. A similar analysis is performed in Yang et al. [51], including a comparison between the result on large real social networks and the LFR benchmarks, arguing that the former is better indicator of the performance of the algorithms. However as Lee and Cunningham [21] elaborate, this ground-truth data is imperfect and incomplete and should be rather considered as metadata or labeled attributes correlated with the underlying communities.

In this paper we first investigate the correlations between attributes and community structure using our network specific agreement/external indexes proposed recently in [40]. Then we present the concept of *community guidance by attributes*, where we adapt our previously proposed TopLeaders[37] community detection method, to find the right number of communities in the given network, based on the available attributes information.

2 Correlation of Communities and Attributes

Traud et al. [45] show that a set of node attributes can act as the primary organizing principle of the communities; e.g. House affiliation in their study of Facebook friendship network of five US universities. In computing the correlation between attributes and relations, Traud et al. [45] use the basic clustering agreement indices for communities comparison. They observe that the correlation significantly depends on this agreement index and differs significantly even between those indices that have been known to be linear transformation of each other. Here we perform similar experiments, but in the context of evaluating community mining algorithms. In more details, we compare the agreements of the results from four different community mining algorithms, with each attribute in the dataset; see Figure 1 for a visualized example. First, the community mining algorithms are applied on the dataset, which are InfoMap [42], WalkTrap [34], Louvain [1], and FastModularity [29]. Then the correlations between the resulted communities from these algorithms and the attributes are measured using clustering agreement indices. More specifically, we measure the agreement assuming the unique attribute values are grouped together and formed a clustering. For example for the attribute 'year', all nodes that have value '2008' are in the same group or cluster. Figure 2 shows the agreements of the community mining algorithms with each attribute averaged over all the networks in the Facebook 100 dataset. The agreements, between two groupings/clusterings of the dataset, are measured with eight different agreement indices: Jaccard Index, F-measure, Variation of Information(VI), Normalized Mutual Information(NMI), Rand Index(RI), Ad-



(a) Attributes: nodes are colored the same if they have the same value for the corresponding attribute; nodes with a missing value for the attribute are white. The number of unique attribute values, i.e. different colours, and the percentage of missing values are also reported. The number outside the parentheses is the number of main values which have at least five nodes, whereas the total number of unique values is reported inside the parentheses.

(b) Communities: nodes are colored the same if they belong to the same community in the results of corresponding community mining algorithms. The number of clusters, i.e. colours, with at least five members is reported, whereas the total number of clusters in the result is given inside the parentheses.

Fig. 1: Visualization of correlations between attributes and communities for the American75 dataset from *Facebook 100 dataset*[46]. This network has 6386 nodes and 217662 edges (friendships which are unweighted, undirected). Visualization is done with Gephi, and an automatic layout is used which positions nodes only based on their connections.

justed Rand Index(ARI), and two structure based extensions of ARI tailored for comparing network clusters: with overlap function as the sum of weighted degrees($\mathcal{ARI}_{x^2}^{\Sigma d}$), and the number of common edges($\mathcal{ARI}_{x^2}^{\xi}$) [38].

Unlike the previous study, we observe very similar rankings with different agreement indexes. The most agreements are observed with the attribute ‘year’, followed not so closely by ‘dormitory’. We can however see that the ranking across different attributes is not the same, whereas Walktrap is the winner according to the ‘year’, and Infomap performs the best if we consider the agreement with the ‘dormitory’. Therefore, although we observe a correlation between the attributes and the communities, it is not wise to compare the general performance of community mining algorithms based on their agreements with a selected attribute as the ground-truth. Instead one should treat attributes as another source of information. In the next section, for example, we use this information to fine tune

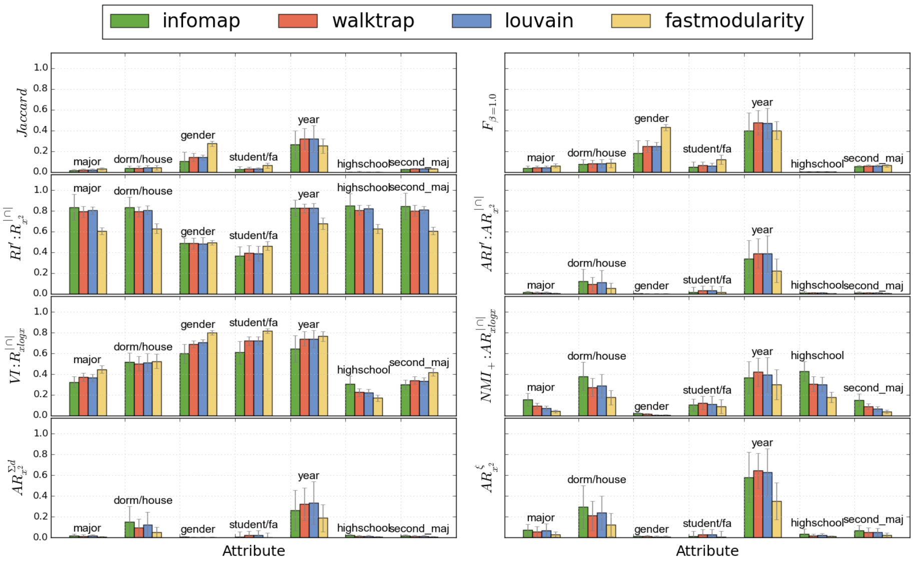


Fig. 2: The agreement of different community detection algorithms with each attribute, averaged over datasets from *Facebook 100 dataset*.

the parameters of a community mining algorithm, so that it results in a community structure which compiles most with our selected attribute. Before that we present a discussion on the effect of missing values on the agreement indices.

2.1 Missing Values and Agreement Indices

The definitions of original agreement indices assume the two clusterings are covering the same set of datapoints. Therefore to use these indices, nodes with missing values should be either removed, or grouped all as a single cluster. The implementations we use here are based on our generalized formula proposed recently in [40]. Unlike the original definitions, these formulae do not require the assumption that the clusterings cover the whole dataset. Hence they can be directly applied to the cases where we have un-clustered datapoints, which will be ignored. For the sake of comparison, in Figure 3 three bars are plotted per <attribute, community mining> pair, corresponding to how the missing values can be handled: (i) when nodes with missing values are removed from both groupings before computing the agreement, (ii) when all the nodes with missing attribute value are grouped into a single cluster, and (iii) when computing the agreements with lifting the covering assumption, using the formulations of [40]. This comparison is in particular important here, since we have many nodes with *missing values* for some of the attributes, such as ‘dormitory’ or ‘second major’; which can significantly increase the agreements if missing values are removed altogether, as seen in the Figure 3.

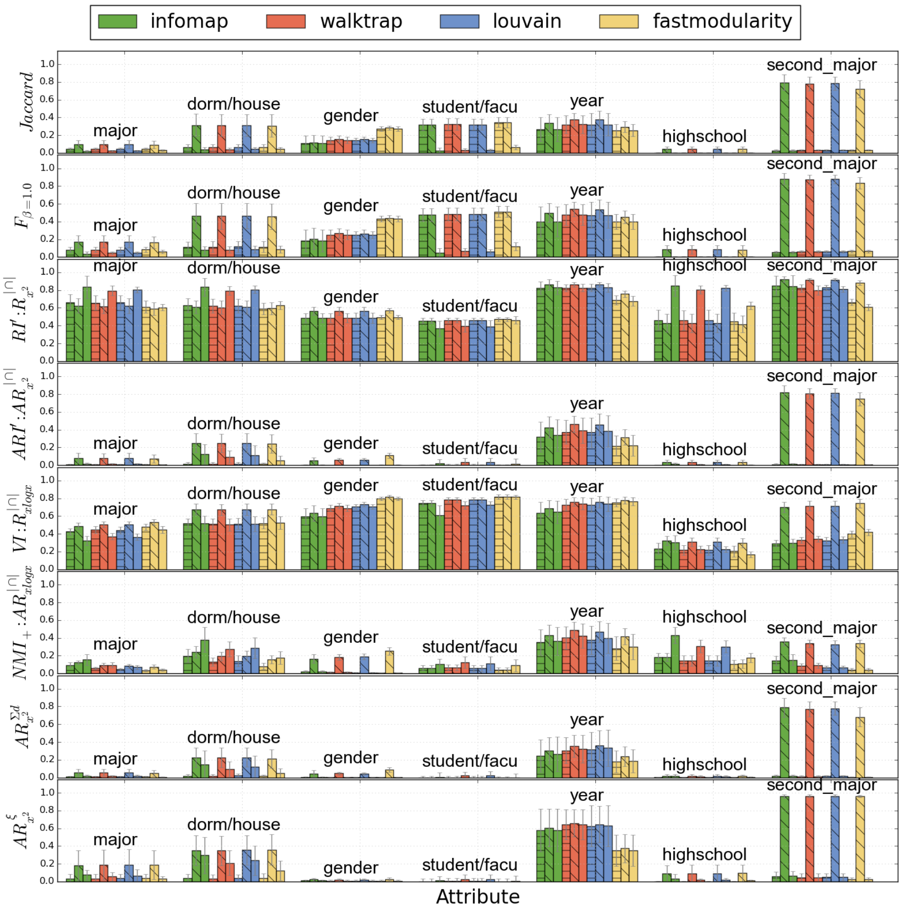


Fig. 3: The effect of missing values: bars with horizontal, diagonal, and solid fill correspond respectively to removing missing values, adding missing values as a single cluster, or lifting the covering assumption.

3 Community Guidance by Attributes

Many real world applications include information on both attributes of individual nodes as well as relations between the nodes, while there exists an interplay between these attributes and relations [5, 16, 23]. More precisely, the relations between nodes motivates them to develop similar attributes (influence), whereas the similarities between them motivates them to form relations (selection), a property referred to as homophily. This can also account for the correlation observed between community structure and attributes, i.e. self-identified user characteristics [45]; which has motivated defining ground-truth communities for real networks based on these explicit properties of nodes.

In the presence of attributes, a more plausible viewpoint is finding groups of nodes that are both internally well connected and having their members with homogeneous attributes. This grouping is referred to as structural attribute clustering by Zhou et al. [52] or cohesive patterns mining by Moser et al. [26]. Similar to community mining, several alternative approaches are proposed for this task [14, 26, 25, 50, 15, 11, 6]. Zhou et al. [52] propose clustering an attribute augmented network. The augmented network includes attribute nodes for each <attribute, value> and edges are added between original graph nodes to their corresponding <attribute, value> nodes[1]. The authors show that a straightforward distance function based on a linear combination of the structural and attribute similarities, fails to outperform a similar method that only considers structural or attribute similarities. In Mislove et al. [25], communities are found using a link based approach but are initialized using a clustering based on their attribute similarities. As another example in Cruz et al. [6], communities found by links are further divided into smaller sub-groups according to the attributes. In more details, the overlap of each community is computed with each cluster in the clustering of the same data according to the attributes. Then larger than average overlaps are cut from the main community to form smaller, more cohesive communities. All these works we have discussed so far further motivate combining attribute and link data, rather than validating one based on the other.

Here, we propose the concept of *community guidance by attributes*, where selected attribute is used to direct a community mining algorithm. More specifically, we guide our TopLeaders [37, 39] algorithm to find the right number of communities, based on the agreements of its result with the given attribute[2].

The number of communities, k for short, is the main parameter for the TopLeaders algorithm, similar to the k-means algorithm for data clustering. Figure 4 illustrates an example on the Amherst41 dataset, where the agreements of each attribute with the results of Topleaders are plotted as a function of k. For some of the attributes, such as 'student/faculty', we observe a clear peak around the true number of classes. We also plotted where other algorithms land. However, there has not been any parameter tuning for those algorithm, and hence they are indicated with a single point. The vertical lines show the true number of classes for the corresponding attribute, i.e. the distinct values[3].

Consequently, between the communities detected by the TopLeaders for different values of k, which only uses the links to discover communities, we select the one that has the most agreement with the given attribute. We used an exhaustive search to find the optimal k for each attribute, in the range of $[2, \sqrt{n}]$, where n is the total number of datapoints. A better optimization method is a future work. Figure 5 shows the agreements obtained through this approach,

[1] This graph representation has also been used in link recommendation, e.g. see [10].

[2] The concept is however general and can be applied to fine tune parameters of any community mining algorithm. Which is true for algorithms which are capable of providing different community structure perspectives, based on different values for the algorithm parameters.

[3] For attribute 'highschool', true k is 1075 and out of the plot's scale.

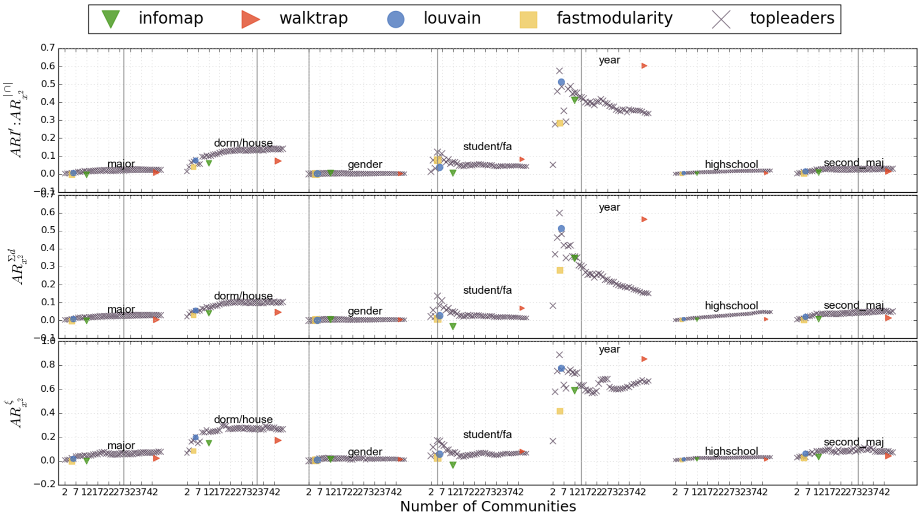


Fig. 4: Agreement of attributes with the results of algorithms plotted as a function of number of communities.

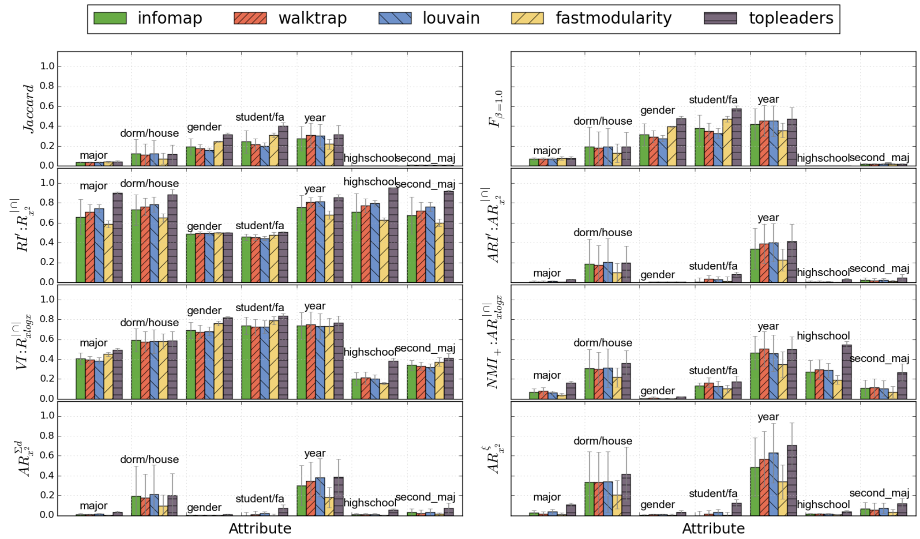


Fig. 5: TopLeaders performance when the number of communities are chosen according to the agreement of its results with the given attribute. This result is averaged over a subset of 5 datasets from the 100 Facebook networks, which are: Amherst41, Bowdoin47, Caltech36, Hamilton46, and Haverford76.

compared to the four commonly used community detection algorithms. We can see in Figure 5 that the communities found by this approach have comparable and in some cases better agreements with the attributes, compared to the methods which do not consider that extra information. This is more significant according to the structure based agreement measures, especially $\mathcal{ARI}^{\xi}_{x^2}$, which considers common edges as the cluster overlaps; and also for less trivial attributes which have a low agreement with the trivial communities, e.g. 'student/faculty', 'second_major', or 'highschool'. One should however note that this is not a comparison for the performance of these algorithms, since TopLeaders used the agreements with the attribute to find the k, which is not available to the other methods.

4 Conclusions

In this paper we discussed different evaluation approaches for community detection algorithms. In particular, we investigated the evaluation of communities on real-world networks with attributes, where there exist a correlation between the characteristics of individual nodes and their connections. We then proposed the concept of community guidance by attributes, where a community mining algorithm is guided to find a community structure which corresponds most to a given attribute. This is in particular useful in real world applications, since we often have access to both link and attribute information, and an idea of how communities will be used. For example, communities in protein-protein interaction networks are shown to be correlated with the functional categories of their members, which are used to predict the previously uncharacterized protein complexes [44]; in such case, one might be interested to select the community structure that corresponds most with the available functional categories.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
2. Chen, J., Zaiane, O., Goebel, R.: An unsupervised approach to cluster web search results based on word sense communities. In: Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on. vol. 1, pp. 725–729 (Dec 2008)
3. Chen, J., Zaïane, O.R., Goebel, R.: Detecting communities in social networks using max-min modularity. In: SIAM International Conference on Data Mining. pp. 978–989 (2009)
4. Clauset, A.: Finding local community structure in networks. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 72(2), 026132 (2005)
5. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 160–168. ACM (2008)

6. Cruz, J., Bothorel, C.: Information integration for detecting communities in attributed graphs. In: Computational Aspects of Social Networks (CASoN), 2013 Fifth International Conference on. pp. 62–67 (Aug 2013)
7. Danon, L., Guilera, A.D., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment (09), 09008 (2005)
8. Fortunato, S.: Community detection in graphs. Physics Reports 486(35), 75–174 (2010)
9. Fortunato, S., Castellano, C.: Community structure in graphs. In: Computational Complexity, pp. 490–512. Springer (2012)
10. Gong, N.Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E.C.R., Stefanov, E., Song, D., et al.: Jointly predicting links and inferring attributes using a social-attribute network (san). arXiv preprint arXiv:1112.3265 (2011)
11. Günnemann, S., Boden, B., Färber, I., Seidl, T.: Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors. In: Advances in Knowledge Discovery and Data Mining, pp. 261–275. Springer (2013)
12. Gustafsson, M., Hörnquist, M., Lombardi, A.: Comparison and validation of community structures in complex networks. Physica A Statistical Mechanics and its Applications 367, 559–576 (Jul 2006)
13. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17, 107–145 (2001)
14. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics 18(suppl 1), S145–S154 (2002)
15. Hu, B., Song, Z., Ester, M.: User features and social networks for topic modeling in online social media. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 202–209. IEEE (2012)
16. La Fond, T., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: Proceedings of the 19th international conference on World wide web. pp. 601–610. WWW '10, ACM, New York, NY, USA (2010)
17. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E 80(5), 056117 (2009)
18. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical Review E 78(4), 046110 (2008)
19. Lancichinetti, A., Kivelä, M., Saramäki, J., Fortunato, S.: Characterizing the community structure of complex networks. PloS one 5(8), e11976 (2010)
20. Largeron, C., Mougel, P.., Rabbany, R., Zaïane, O.R.: Generating attributed networks with communities. PloS one to appear (2015)
21. Lee, C., Cunningham, P.: Benchmarking community detection methods on social media data. arXiv preprint arXiv:1302.0739 (2013)
22. Leskovec, J., Lang, K.J., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th international conference on World wide web. pp. 631–640. ACM (2010)
23. Lewis, K., Gonzalez, M., Kaufman, J.: Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences 109(1), 68–72 (2012)
24. Luo, F., Wang, J.Z., Promislow, E.: Exploring local community structures in large networks. Web Intelligence and Agent Systems 6, 387–400 (2008)
25. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM

international conference on Web search and data mining. pp. 251–260. WSDM '10, ACM, New York, NY, USA (2010)

26. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining cohesive patterns from graphs with feature vectors. In: SDM. vol. 9, pp. 593–604 (2009)
27. Moussiades, L., Vakali, A.: Benchmark graphs for the evaluation of clustering algorithms. In: Proceedings of the Third IEEE International Conference on Research Challenges in Information Science. pp. 197–206. RCIS'09 (2009)
28. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004)
29. Newman, M.E.: Fast algorithm for detecting community structure in networks. Physical review E 69(6), 066133 (2004)
30. Onnela, J.P., Arbesman, S., González, M.C., Barabási, A.L., Christakis, N.A.: Geographic constraints on social network groups. PLoS one 6(4), e16939 (2011)
31. Orman, G.K., Labatut, V.: The effect of network realism on community detection algorithms. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining. pp. 301–305. ASONAM '10 (2010)
32. Orman, G.K., Labatut, V., Cherifi, H.: Qualitative comparison of community detection algorithms. In: International Conference on Digital Information and Communication Technology and Its Applications. vol. 167, pp. 265–279 (2011)
33. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
34. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005, pp. 284–293. Springer (2005)
35. Rabbany, R., Takaffoli, M., Fagnan, J., Zaiane, O., Campello, R.: Relative validity criteria for community mining algorithms. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 International Conference on (aug 2012)
36. Rabbany, R., Chen, J., Zaïane, O.R.: Top leaders community detection approach in information networks. In: Proceedings of the 4th Workshop on Social Network Mining and Analysis (2010)
37. Rabbany, R., Chen, J., Zaïane, O.R.: Top leaders community detection approach in information networks. In: SNA-KDD Workshop on Social Network Mining and Analysis (2010)
38. Rabbany, R., Takaffoli, M., Fagnan, J., Zaïane, O.R., Campello, R.: Relative validity criteria for community mining algorithms. Social Networks Analysis and Mining (SNAM) (2013)
39. Rabbany, R., Zaïane, O.R.: A diffusion of innovation-based closeness measure for network associations. In: IEEE International Conference on Data Mining Workshops. pp. 381–388 (2011)
40. Rabbany, R., Zaïane, O.R.: Generalization of clustering agreements and distances for overlapping clusters and network communities. CoRR abs/1412.2601 (2014)
41. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences 104(18), 7327–7331 (2007)
42. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4), 1118–1123 (2008)
43. Rosvall, M., Bergstrom, C.T.: Mapping change in large networks. PloS one 5(1), e8694 (2010)

44. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences 100(21), 12123–12128 (2003)
45. Traud, A.L., Kelsic, E.D., Mucha, P.J., Porter, M.A.: Comparing community structure to characteristics in online collegiate social networks. SIAM review 53(3), 526–543 (2011)
46. Traud, A.L., Mucha, P.J., Porter, M.A.: Social structure of facebook networks. Physica A: Statistical Mechanics and its Applications 391(16), 4165–4180 (2012)
47. Wagner, A., Fell, D.A.: The small world inside large metabolic networks. Proceedings of the Royal Society of London. Series B: Biological Sciences 268(1478), 1803–1810 (2001)
48. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.: Scan: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 824–833. ACM (2007)
49. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. p. 3. ACM (2012)
50. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 927–936. ACM (2009)
51. Yang, Y., Sun, Y., Pandit, S., Chawla, N.V., Han, J.: Perspective on measurement metrics for community detection algorithms. In: Mining Social Networks and Security Informatics, pp. 227–242. Springer (2013)
52. Zhou, Y., Cheng, H., Yu, J.X.: Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment 2(1), 718–729 (2009)