Estimating the similarity of community detection methods based on cluster size distribution

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca

Estimating the similarity of community detection methods based on cluster size distribution

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca

IMT Atlantique - Lab-STICC CNRS
UMR 6285 F-29238 Brest, France
{vinh.dao, cecile.bothorel, philippe.lenca}@imt-atlantique.fr,

Abstract. Detecting community structure discloses tremendous information about complex networks and unlock promising applied perspectives. Accordingly, a numerous number of community detection methods have been proposed in the last two decades with many rewarding discoveries. Notwithstanding, it is still very challenging to determine a suitable method in order to get more insights into the mesoscopic structure of a network given an expected quality, especially on large scale networks. Many recent efforts have also been devoted to investigating various qualities of community structure associated with detection methods, but the answer to this question is still very far from being straightforward. In this paper, we propose a novel approach to estimate the similarity between community detection methods using the size density distributions of communities that they detect. We verify our solution on a very large corpus of networks consisting in more than a hundred networks of five different categories and deliver pairwise similarities of 16 state-of-the-art and well-known methods. Interestingly, our result shows that there is a very clear distinction between the partitioning strategies of different community detection methods. This distinction plays an important role in assisting network analysts to identify their rule-of-thumb solutions.

Keywords: community detection, similarity metric, community size, comparative analysis

1 Introduction

Community detection discloses interesting information about the heterogeneous structure of complex networks and opens promising perspective in many theoretical as well as applied domains [8,23,24]. Although showing a high similarity with traditional unsupervised data clustering, community detection techniques have just been becoming prosperous in the last two decades remarked by the invention of modularity [14] and the availability of a large volume of networks from small scale to very large scale thanks to the development of Internet and notably social platforms. Since then, a numerous number of detection techniques with various approaches have been proposed [3,5] to solve this network decomposition

problem. Even though communities are widely assumed to be sub-graphs where nodes are more densely connected relatively to the rest of the network, there is no commonly accepted standard process to evaluate the accuracy of detection methods. Indeed, the notion of goodness varies according to contextual objective and also the assumption about the underlying network model. By consequence, there is normally a confusion when one needs to find the most suitable method among available ones that is presumed to satisfy some specific requirements in outcome quality.

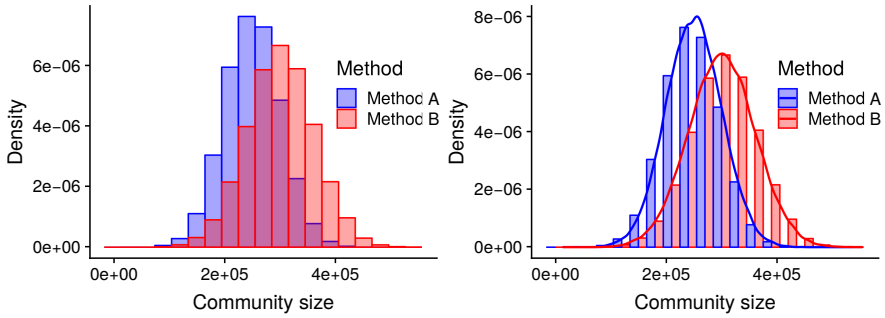
Meanwhile, the stated issue leaves behind rooms for developing theoretical and empirical techniques for comparing community detection algorithms. Actually, new methods are usually introduced in accompany with quality evaluation based on many variants of Mutual Information [27] or modularity. These two approaches work well to validate the functionality of proposed methods in ad-hoc networks but are not directly interpretable in a comparative evaluation of clustering quality. Actually, the former ones do not provide structural information of detected communities and the latter ones are dependent on hypotheses about null models. In other words, equivalent scores do not directly ensure an equivalence of partition quality.

In this paper, we are interested in estimating pairwise similarity of community detection methods based on expected community size. These estimates also reveal information about the closeness in terms of number of communities - a very important and intuitive characteristic of clustering algorithms - which is considered as an essential perspective in community detection literature [5] and recently addressed by many in-depth researches [7,19]. Specifically, we conduct an empirical experiment to inspect a large number of state-of-the-art and widely used community detection methods and estimate their similarity using size distribution of communities that they discover on a large dataset of networks across several domains. The result of our analysis implicates that community detection methods can be classified in three well discernible groups exhibiting three essential strategies of node partition. These strategies produce a great impact on the outcome of community detection methods, making them very distinctive. We will show that this taxonomy exposes very useful information for proposing appropriate methods according to expected analysis strategy.

2 Estimating the similarity of community detection methods

We present a novel approach to determine the similarity of community detection methods using the size distribution of communities that they discover. Certainly, this is only one among interesting quality aspects that differentiate one method from the others. Nonetheless, it allows to get more insight into the difference in terms of partitioning strategy.

Specifically, a very naive but efficient approach to evaluate the similarity of two methods is to inquire into the *"closeness"* of the two corresponding community size distributions. As such, two methods could be supposed to be similar



(a) Overlapped sizes (b) Interlaced sizes

Fig. 1. The size distributions of communities detected by two different methods.

if their corresponding density distributions expose a large intersection area as shown in Fig. 1(a). From this notice, we can define our new similarity function as follows.

First, we denote two 2-tuples $(\mathcal{A}, n^a)$ and $(\mathcal{B}, n^b)$ being the multisets representing all communities detected on a set of networks $\mathcal{G} = \{G\}$ by method A and method B respectively, where $\mathcal{A} = \{x_1^a, x_2^a, ..., x_r^a\}$ and $\mathcal{B} = \{x_1^b, x_2^b, ..., x_s^b\}$ being the ascending ordered sets of sizes of communities: $1 \leq x_1^a < x_2^a < ... < x_r^a$ and $1 \leq x_1^b < x_2^b < ... < x_s^b$. The multiplicity functions $n^a : \mathcal{A} \to \mathbb{N}_{\geq 1}$ and $n^b : \mathcal{B} \to \mathbb{N}_{\geq 1}$ measure the number of communities of sizes x_i^a and x_i^b respectively. Let $N^a = \sum_{i=1}^{r} n^a(x_i^a)$ and $N^b = \sum_{i=1}^{s} n^b(x_i^b)$ being the total number of communities of all sizes detected by each method, we define a similarity function describing the closeness of A and B on $\mathcal{G}$ as:

$$S_{\mathcal{G}}(A, B) = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{s} \min \left\{ \frac{n^a(x_i^a)}{N^a}, \frac{n^b(x_j^b)}{N^b} \right\} \delta(x_i^a, x_j^b), \tag{1}$$

where $\delta(x_i^a, x_j^b) = 1$ if $x_i^a = x_j^b$ and 0 otherwise. Equation (1) is simply the common fraction of same-size communities detected on $\mathcal{G}$ by both A and B: $0 \leq S_{\mathcal{G}}(A, B) \leq 1$. This definition seems to be intuitive but does not work well in practice. As illustrated in Fig. 1(b), when the sizes interlace each other, a low score will be produced although the similarity in this case is as much as that of the case of Fig. 1(a). Choosing an appropriate binning interval would mitigate the problem. This solution is, however very inflexible, sensible to the characteristic of data as well as to the functionality of the methods in use. A straightforward alternative can be envisioned by using a kernel density estimator to uncover the probability density functions as shown by the solid lines in Fig. 1(b). In this way, we approximate the common fraction of same-size communities of Equation (1) by the overlapping area of two corresponding continuous distributions. The premise behind this estimation is that two similar methods must not compulsorily produce a large portion of exactly same-size communities, but a large

portion of comparable-size ones. Hence, we consider the following estimator to take in local information of community size x_0:

$$\widehat{f}(x_0) = \frac{1}{hn} \sum_i K\left(\frac{x_i - x_0}{h}\right), \tag{2}$$

where h is the bandwidth controlling the neighborhood interval around x_0 and K is the kernel function controlling the weight given to the observations $\{x_i\}$ chosen as Gaussian in our analysis. Using this estimator, we rewrite the similarity function defined in Equation (1) as follows:

$$S_{\mathcal{G}}(A, B) = \int \min\{\widehat{f}^{(a)}(x), \widehat{f}^{(b)}(x)\}dx, \tag{3}$$

where

$$\widehat{f}^{(u)}(x) = \frac{1}{hN^u} \sum_i^{N^u} \left[n^u(x_i^u) K\left(\frac{x_i^u - x}{h}\right)\right], \tag{4}$$

with $u \in \{a, b\}$. In the estimations of this paper, the bandwidth h is selected based on the normal reference rule [26] to minimize the mean integrated squared error. The only exception is the cases illustrated Fig. 3 where a higher value has been chosen to get a higher smoothing quality for a better illustration.

Using Equations (3) and (4) to estimate the similarity between pairs of detection methods on a large dataset will help us discovering different behaviors of community detection methods. Since the accuracy of the estimator depends on the networks of the dataset that we analyze, the result will obviously relativized. However, a large and representative corpus would help to reduce the dependency impact.

3 Community detection methods

A community is roughly described as a group of nodes in a graph where there must be many edges connecting them together than edges connecting the community with the rest of the graph [5]. However, in practice, this concept is mathematically or algorithmically formulated in different ways engendering various discovery approaches. In this paper, we select a representative set of state-of-the-art and widely studied detection methods whose approaches spread out over the most commonly used in the literature. These methods are summarized in Table 1 with corresponding information. Their approaches could be briefly summarized as follows:

– **Edge removal**: In this approach, inter-community edges in a network are gradually removed in order to disconnect densely connected groups. The problem of community detection is translated to identifying candidates for inter-community edges based on their topological positions. Popular techniques include using edge betweenness centrality (GN in Table 1) or edge clustering coefficient, which could be based on triangular (RCCLP-3) or quadrangular (RCCLP-4) patterns.

Table 1. Community detection methods and associated implementations involved in our analyses grouped by different methodological approach. The label column denotes the corresponding abbreviations used in our paper.

Approach	Method	Label	Source
Edge removal	Girvan-Newman [14]	GN	igraph[1]
	Radicchi *et al.* [16]	RCCLP-3 (for $g = 3$)	Authors[2]
	Radicchi *et al.* [16]	RCCLP-4 (for $g = 4$)	Authors[2]
Modularity optimization	Clauset *et al.* [2]	CNM	igraph[1]
	Blondel *et al.* [1]	Louvain	Authors[3]
	Newman [13]	SN	igraph[1]
Dynamic process	Pons *et al.* [15]	Walktrap	igraph[1]
	Rosvall *et al.* (2007) [22]	Infomod	Authors[4]
	Rosvall *et al.* (2009) [21]	Infomap	Authors[5]
Statistical inference	Lancichinetti *et al.* [10]	Oslom	Authors[6]
	Riolo *et al.* [19]	(DC)SBM	Authors[7]
Other methods	Reichardt *et al.* [18]	RB	igraph[1]
	Raghavan *et al.* [17]	LPA	igraph[1]
	Xie-Szymanski [28]	SLPA	Authors[8]
	Demeo *et al.* [12]	Conclude	Authors[9]

[1] http://igraph.org/ (Available in R, Python and C/C++)
[2] http://homes.sice.indiana.edu/filiradi/resources.html
[3] https://sourceforge.net/projects/louvain/
[4] http://www.tp.umu.se/~rosvall/code.html
[5] http://www.mapequation.org/
[6] http://www.oslom.org/
[7] http://www-personal.umich.edu/~mejn/
[8] https://sites.google.com/site/communitydetectionslpa/
[9] http://www.emilio.ferrara.name/code/conclude/

- **Modularity optimization**: Methods in this approach use a common objective function called *modularity* [14], but have different optimization strategies. The modularity function measures the quality of a partition by calculating the difference between the fraction of intra-community edges of the partition with the expected of such fraction in the associated partition whose edges are redistributed randomly following a null model.
- **Dynamic process**: Methods in this group do not use directly topological information in order to deduce densely connected subgraphs. Instead, they exploit stochastic information from various dynamic models regulated by network structure in order to deduce community structure.
- **Statistical inference**: This approach takes into consideration the statistical significance of community structure based on different theoretical network models. Such methods usually optimize likelihood functions to find the best configuration fitting hypothetical assumptions using different searching strategies.
- **Other methods**: Some approaches define implicitly or explicitly requirements about community structure or mix different traditional approaches to

Table 2. A summary of network used in this analysis where *Size* is the number of networks in each category, *Nodes* and *Edges* indicates the average number of nodes and edges of networks respectively. This dataset is collected from [9,11,20].

Category	Size	Nodes	Edges	Notable networks
Biological	7	1860	10763	Yeast, brain, protein-protein interactions
Communication	9	39595	195032	Email, forums, message exchanges
Information	25	38358	159812	Amazon, DBLP, citation & education webs
Social	37	6888	49666	Facebook, Youtube, Google plus networks
Technological	19	18431	48494	Internet, AS Caida, Gnutella P2P networks
Miscellaneous	11	4298	49033	Ecology, power-grid, synthetic networks
Total size *	108	1.99M	9.08M	

*The total number of networks, nodes and edges in the whole dataset respectively.

take the advantages of each one. In many cases, they can be classified into one theoretical family or another. To simplify the theoretical taxonomy, we present them in a common group.

To maintain the controllability of the experiment and to ensure the reproducibility of the analysis, all of the above presented methods are studied with the default parameters determined by the authors. The following section will be dedicated to an introduction of the network dataset that will be used in our experiment.

4 Network dataset

Our experiment requires a large number of networks in order to reduce the impact of the irregularity which could be presented in a small set of ad-hoc networks. Hence, presuming that networks in different domains possess various structural particularities [4], we collect networks spanning of a variety of categories which are widely studied in the research community. To ensure that distribution of community size is allowed to be spread in a wide range, it is very essential to gather networks from small scale to large scale. We cover networks from around 30 vertices up to 320000 vertices and a million edges. Table 2 encapsulates the principle information of networks involved in our experiment.

Fig. 2 exhibits the distribution of network size in each category. As we can see, the marginal distributions on top imply that inside each category, networks also span in a relatively wide range of size with some slightly differences from one category to another. Additionally, the networks in this dataset are quite sparse. The majority of networks has an average degree of approximately from 10 to 20 connections. Also, the number of edges increase linearly by the number of nodes with equivalent rates among categories as can be deduced from the gradients of the linear estimates.

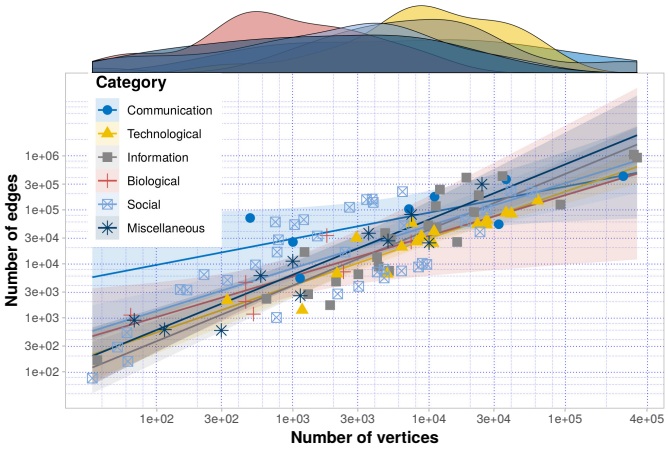


Fig. 2. Structural information of network employed in the experiment. The solid lines represent estimated relations between number of nodes and number of edges in each network category using a linear regression model. Accordingly, the translucent color backgrounds represent the corresponding 95% confidence intervals.

5 Experimental results

The experimental procedure is simple. We gradually employ each method presented in Section 3 to discover community structures on the whole set of networks summarized in Table 2. When the whole set of communities is identified, we proceed to measure the volumes of communities detected by each method to identify the elements of the corresponding 2-tuples. Finally, we use the similarity function defined by Equation (3) to estimate the closeness between each pair of methods.

Due to the huge number of necessary experiments, only processes having a reasonable theoretical estimated time and memory consumption are maintained (less than a few days and require at most 30 to 40 GBytes of memory). The outcome distributions are illustrated in Fig. 3.

As we can see, there is a clear difference in the densities of community size, showing that these methods have various partitioning strategies. Knowing that methods belonging to the same theoretical group (as shown in Table 1) are placed next to each other, we can notice some agreements between the theoretical families with practical outcomes as follows:

> **Edge removal**: GN and RCCLP-3 have very similar distributions where a large number of communities are very small. This is due to the fact that in some highly local centralized networks having star-like structures, they have a tendency to remove edges connecting hub and peripheral nodes, hence create singletons (single node community). This phenomenon is less distinguishable on RCCLP-4 since there are much less quadrangular than triangular connections in networks.

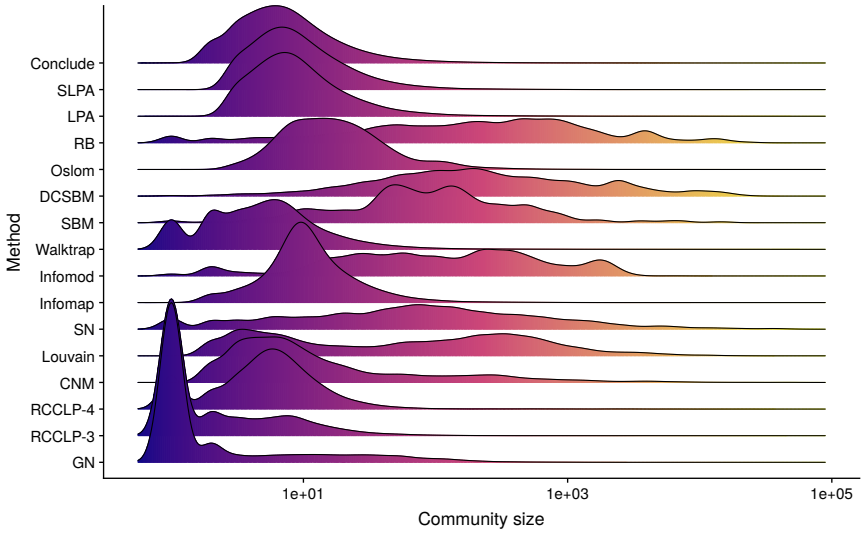


Fig. 3. The distribution of community size contained in the partitions detected on the networks of the dataset. The distribution are smoothed using a Gaussian kernel estimator. The illustrative gradient color is for the ease of view purpose.

Modularity optimization: Modularity is known to suffer from resolution limit phenomenon [6], which often aggregates small communities in large scale networks. We can see from Fig. 3 that Louvain and SN found very large communities as predicted. In the meanwhile, there are also a comparable number of small communities found on small graphs. The behavior is a little bit different on CNM method.
Dynamic process: Methods in this family show very discernible distributions. Although all based on dynamic processes, they have different assumptions about community structure and searching mechanisms. Therefore, the closeness in the concept does not lead to a similarity in practical results.
Statistical inference: the Bayesian SBM and DCSBM uses Monte Carlo sampling process which is very time demanding in order to sweep the solution space. This makes the method unfeasible if the maximum number of clusters is not limited. Indeed, in the default version, the maximum number of communities is limited at 25 making (DC)SBM methods find very large communities in general. On the other hand, Oslom method use a different discovery mechanism and identify globally smaller communities.
Other methods: In this group, LPA, SPLA (both based on label propagation) and Conclude display nearly identical distributions. RB methods, being based on a very close concept with modularity (with a tuning parameter), exhibits a similarity with modularity optimization based methods.

Quantitatively, applying the estimator presented in Equation (4) to compute pairwise similarities between the methods leads us to the results demonstrated in

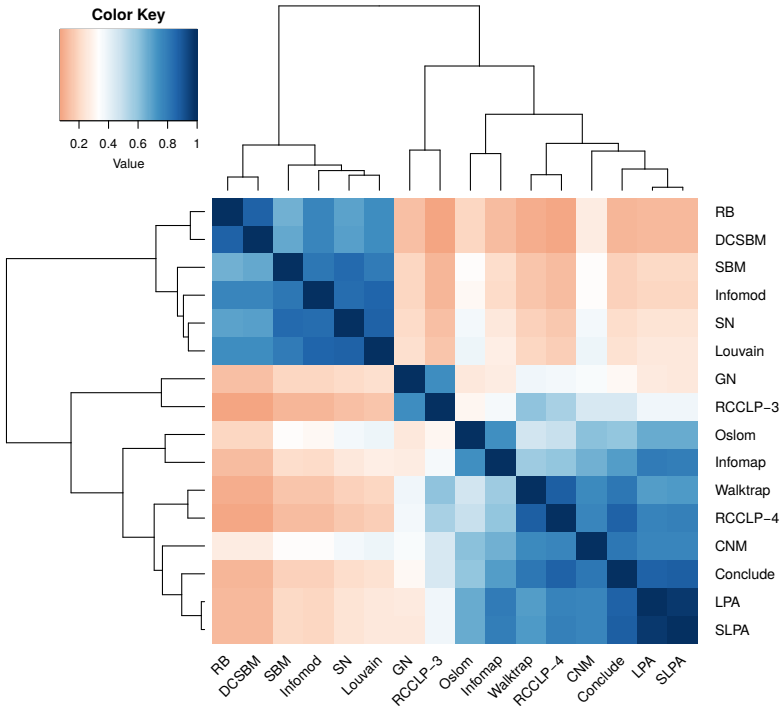


Fig. 4. The estimated proximity between detection methods. Similar methods share a large fraction of same-size communities. Methods are ordered using hierarchical clustering. The dendrogram proposes a hierarchical structure of the fitting closeness. Blue colors mean high similarity.

Fig. 4. As we can see, according to the community size criterion, these methods can be classified into different classes of partitioning strategy. The separations are very shaped showing that the distinction is very clear between groups. Therefore, we choose to characterize these methods by 3 (possibly 4) principle groups as follows:

Group 1: RB, DCSBM, SBM, Infomod, SN, Louvain: Methods in this group discover communities whose size vary in a very large spectrum, from very small to very large communities. The characterized community size distribution is quite flat, meaning all sizes are nearly equally considered on the dataset.
Group 2: GN and RCCLP-3: These two methods identify a huge number of very small communities including singletons regardless of network size. As a consequence, there are few variations in community volume.
Group 3: the others: These methods produce communities whose sizes approach bell-shaped distribution. The strategy can be translated as: not left not right, i.e. not too small and not too big communities.

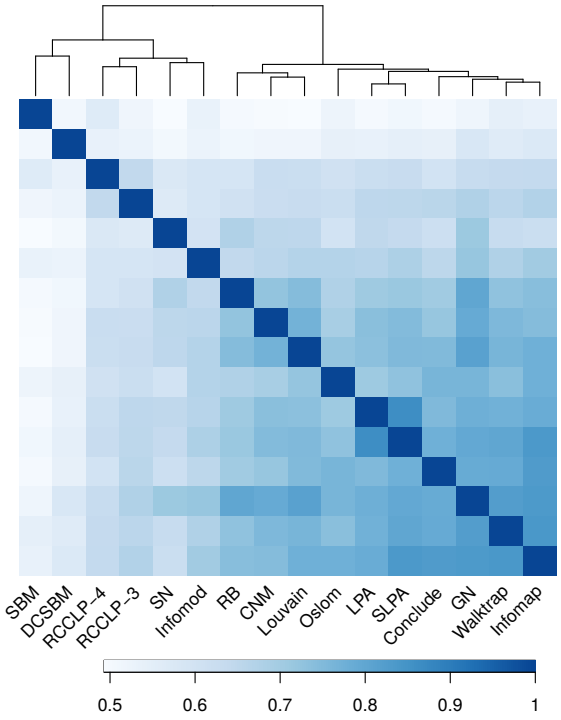


Fig. 5. The similarity of partitions average NMI

As we can see, the strategies that these methods partition networks are very discernible. In fact, some other alternatives have been used to parametrize the estimator such as regulating the bandwidth h by using cross validation or pilot estimation of derivatives [25]; adapting other kernel functions K. However, the only major difference is noticed at GN and RCCLP-3 methods, which are no more highly similar. This variation is explainable since both GN and RCCLP-3 create a large number of very small communities, there are high spurious peaks in the distributions making the estimation unfeasible. In the meanwhile, using the original function of Equation (1) helps us to validate the closeness between GN and RCCLP-3.

Also, it worth noting that we could not yet conclude whether two methods are similar in a wide sense if their distributions are close; on the contrary we are able to deduce that they are not similar if their distributions are too different.

6 Discussion and Conclusion

Our experimental taxonomy discloses a new source of information that some traditional evaluation methods could not directly expose. For example, we demonstrate the similarity between partitions detected by these methods in Fig. 5 using

Normalized Mutual Information (NMI) metric [27]. As we can see, one would hardly identify and get insight into the differences between detection methods in an intuitive way. In the meanwhile, using only community size distribution to deduce the similarity of methods could lead to unexpected results. Specifically, two methods could produce exactly the same sequence of community size, but the way that nodes are distributed into the two partitions are totally different. In this case, combining the two evaluation paradigms discloses complement information that helps network practitioners to characterize and deduce appropriate community detection methods. For instance, taking SBM and DCSBM, the average NMI score is approximately 0.5, which means that the two partitions are quite different. However, Fig. 1 divulges that SBM and DCSBM produce communities having quite similar sizes. In the case of LPA and SLPA on the other hand, the two methods produce quite identical outcomes as there is a consistency in the average NMI and our similarity index, which are all nearly 1.

Finally, the variation of the similarity according to the changing of input dataset has not been investigated comprehensively to validate the consistency of our result. In fact, it seems that the distributions illustrated in Fig. 3 have a tendency to move to the left hand side if more small scale networks are involved and inversely, to the right hand side if large scale networks are more implicated. Also, the (in)consistency of each method to the variation of data would not be at the same magnitude, i.e. some methods could be more *"robust"* in controlling community size than others. A further investigation is deemed necessary and promising as perspective.

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), P10,008 (2008)
2. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6) (2004)
3. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining **4**(5), 512–546 (2011)
4. Dao, V.L., Bothorel, C., Lenca, P.: An empirical characterization of community structures in complex networks using a bivariate map of quality metrics. ArXiv e-prints (2018)
5. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3-5), 75–174 (2010)
6. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences **104**(1), 36–41 (2006). DOI 10.1073/pnas.0605965104
7. Ghasemian, A., Hosseinmardi, H., Clauset, A.: Evaluating Overfit and Underfit in Models of Network Community Structure. ArXiv e-prints (2018)
8. Hizanidis, J., Kouvaris, N.E., Zamora-López, G., Diaz-Guilera, A., Antonopoulos, C.G.: Chimera-like states in modular neural networks. Scientific Reports (19845) (2016)

9. Jerome, K.: The koblenz network collection. In: Proceedings Conference on World Wide Web Companion, pp. 1343–1350 (2013). URL http://konect.uni-koblenz.de
10. Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S.: Finding statistically significant communities in networks. PLoS ONE **6**(4), e18,961 (2011)
11. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection (2014). URL http://snap.stanford.edu/data
12. Meo, P.D., Ferrara, E., Fiumara, G., Provetti, A.: Mixing local and global information for community detection in large networks. Journal of Computer and System Sciences **80**(1), 72–87 (2014)
13. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006). DOI 10.1073/pnas.0601602103
14. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2) (2004)
15. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: p. Yolum, T. Güngör, F. Gürgen, C. Özturan (eds.) Computer and Information Sciences - ISCIS 2005, pp. 284–293. Springer Berlin Heidelberg (2005)
16. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences **101**(9), 2658–2663 (2004)
17. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3) (2007)
18. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Physical Review E **74**(1) (2006)
19. Riolo, M.A., Cantwell, G.T., Reinert, G., Newman, M.E.J.: Efficient method for estimating the number of communities in a network. Physical Review E **96**(3) (2017)
20. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015). URL http://networkrepository.com
21. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. European Physical Journal Special Topics **178**, 13–23 (2009). DOI 10.1140/epjst/e2010-01179-1
22. 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)
23. Schaub, M.T., Delvenne, J.C., Rosvall, M., Lambiotte, R.: The many facets of community detection in complex networks. Applied Network Science **2**(1) (2017)
24. Sekara, V., Stopczynski, A., Lehmann, S.: Fundamental structures of dynamic social networks. Proceedings of the National Academy of Sciences **113**(36), 9977–9982 (2016)
25. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society **53**(3), 683–690 (1991)
26. Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London New York (1986)
27. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
28. Xie, J., Szymanski, B.K.: Towards linear time overlapping community detection in social networks. In: Advances in Knowledge Discovery and Data Mining, pp. 25–36. Springer Berlin Heidelberg (2012)