



HAL
open science

Community structure: A comparative evaluation of community detection methods

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca

► **To cite this version:**

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca. Community structure: A comparative evaluation of community detection methods. *Network Science*, 2020, 8 (1), pp.1-41. 10.1017/nws.2019.59 . hal-01976587

HAL Id: hal-01976587

<https://hal.science/hal-01976587>

Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Community structure: A comparative evaluation of community detection methods

VINH-LOC DAO, CECILE BOTHOREL and PHILIPPE LENCA

IMT Atlantique,

Lab-STICC CNRS UMR 6285

F-29238, Brest, France

(*e-mail*: vinh.dao, cecile.bothorel, philippe.lenca@imt-atlantique.fr)

Abstract

Discovering community structure in complex networks is a mature field since a tremendous number of community detection methods have been introduced in the literature. Nevertheless, it is still very challenging for practitioners to determine which method would be suitable to get insights into the structural information of the networks they study. Many recent efforts have been devoted to investigating various quality scores of the community structure, but the problem of distinguishing between different types of communities is still open. In this paper, we propose a comparative, extensive and empirical study to investigate what types of communities many state-of-the-art and well-known community detection methods are producing. Specifically, we provide comprehensive analyses on computation time, community size distribution, a comparative evaluation of methods according to their optimisation schemes as well as a comparison of their partitioning strategy through validation metrics. We process our analyses on a very large corpus of hundreds of networks from five different network categories and propose ways to classify community detection methods, helping a potential user to navigate the complex landscape of community detection.

Keywords: community detection, community structure, comparative analysis, empirical analysis, computation time, community size, structural quality function, validation metric, decision-making assistance for practitioners

Contents

1	Introduction	2
2	Community detection methods and dataset	4
2.1	Community detection methods	4
2.2	Experimental dataset	8
3	Preliminary analysis of community detection methods	9
3.1	Computation time performance	9
3.2	Analysis on community size distribution	14
3.3	Similarity based on community size distribution	25
4	Goodness profiling of community detection methods	29
4.1	Fitness functions	29
4.2	Co-performance index	32

2

V.L. Dao, C. Bothorel and P. Lenca

5 Partitioning strategy comparison	35
5.1 Validation metrics	36
5.2 Empirical results	39
6 Related work	40
7 Conclusion	42
References	43

1 Introduction

In network science, *community detection*, sometimes called *graph clustering*¹ is one of fundamental challenges to discover the structure of networks in mesoscopic level. However, it is an ill-defined problem² such that there exists no universal definition or closed form formula of what kind of objects one should be looking for (Fortunato & Hric, 2016), and consequently there is ambiguity on what should be used as a golden standard to assess the quality of a community structure and the performance of a detection algorithm.

The most frequently found definition of community in the network science literature is derived from the mechanism of connection preference. It implies that *a community is a group of nodes (a subgraph) in a graph where there must be many edges (denser) connecting them together than edges connecting the community with the rest of the graph* (Radicchi *et al.*, 2004), (Fortunato, 2010). Newman defines a community as a *"group of vertices with a higher-than-average density of edges connecting them"* (Newman, 2006). Depending on the context, a community may be called a *cluster*, a *module*, a *class* or a *modular group*. This definition is the most basic that sets the fundamental requirement for most of its derivative definitions. Many different variations of community could be found in (Wasserman, 1994), for instance *LS-set*, which is a set of nodes in a network such that each of its proper subsets has more ties to its complement within the set than outside; or *k-core*, which is a subgraph in which each node is adjacent to at least a minimum number k of the other nodes in the subgraph. However, in recent developments of community detection algorithms, there is no consensus of the quantity of edges in reality that could be considered as *"many"*, communities are just algorithmically defined, i.e. they are final products of the algorithm without any precise a priori definition (Fortunato, 2010).

In practice, there are even more constraints, which are sometimes not explicitly expressed, than that appeared in the announcement of the problem. If one only look for a partition of graph that maximize the number of internal edges and minimize the number of external edges, then the graph itself can be considered as a big community and there is none external connection. Another solution is to let the node having the smallest degree of a graph into one community, and all other nodes into another community. This solution could also maximize the ratio between external and internal edges. However, these monotonous

¹ The concept of graph clustering might refer to two different meanings existing in the literature. The first one implies a categorization of many graphs into different sets within which graphs share a common similar feature. The second one relates to the problem of partitioning nodes of a graph into densely connected groups. Here we means graph clustering in the latter case.

² It means *"it does not have clear goals, solution paths or expected solution"* (Arifin *et al.*, 2017)

solutions seem not be a seductive one for most (if not to say all) analysts who consider using a community detection algorithm to detect communities. In fact, it is preferable to cluster a network into at least 2 relatively similar size communities or more³ (Newman, 2010). It means that somehow, the relative size of communities with respect to the network containing them is important without having explicitly been announced. Besides, there are many other criteria that could be mentioned such as community complete mutuality, reachability, vertex degree distribution and the comparison of internal versus external cohesion (Wasserman, 1994), (Fortunato, 2010). There exists a subtle compromise between adding new vertices as well as their edges into a community and conserving the common property that defines the group. In fact, different community detection methods usually have different ways to divide a network into multiple subsets of nodes and hence result in different community structures. There are many reasons that could lead to these contentions between detection methods:

- Different algorithms may have different notions of community meaning that the structure discovered may strongly depend on the assumptions made about expected community structure.
- When two algorithms define the same concept of community, it may also mathematically and algorithmically be formalized in different ways (the same objective but different objective functions) and hence lead us to different results.
- Even when two algorithms have exactly the same objective function, the algorithmic mechanism they employ to find communities also decides what they are going to find, especially in heuristic searching approaches.
- Initial configuration is also another important factor that affects the final result of an algorithm, many community detection methods are not deterministic.
- Each method may include a consideration between obtaining optimized results in its sense and providing a high-performance method (in terms of calculation time, memory consumption, etc.). This trade-off may be considered differently across the methods.
- Some algorithms are variable in function of input data and will prove more or less efficient on some kinds of inputs than on others.
- Variations due to implementation factors could also impact the final result of an algorithm.
- Finally, in some algorithms, there are tie-break situations where the algorithm have to chose randomly without any factor related their final objectives. It may also affect heavily the result that one would get if the tie-break problems have been resolved in a different way.

Due to many reasons stated above, choosing the community detection method that corresponds well to a particular scenario or to an expectation of quality is not straightforward. In this paper, we introduce some techniques and empirical analyses that help to get insights

³ Community detection is identified in the research community as the search for *natural groups* in networks without a given number of clusters. When the number and the size of clusters are specified, the problem is often referred as *graph partitioning* or *graph bisection* for a division into only two clusters.

into the differences between popular community detection methods according to many community structure quality aspects and algorithmic performance. The paper is organized as follows: Section 2 introduces some popular and state-of-the-art community detection methods that will be analyzed in this paper as well as the benchmarking dataset employed in our experiments. Then, Section 3 presents analyses on the most essential aspects of community detection performance including computation time, community size distribution as well as numbers of detected communities by each method. In Section 4, we address different structural quality aspects of community structure associated with the community detection methods introduced in Sections 2. It is followed by an comparative evaluation using many popular clustering validation metrics in Section 5, which are widely used in the context of community detection. Finally, we present some close results of related work that can be found in the literature in Section 6 and conclude our study by some discussions and recommendations in Section 7.

2 Community detection methods and dataset

2.1 Community detection methods

We present in this section some popular community detection methods that have been widely used and discussed in the literature. Note that in recent years, there are a large number of innovative methods which are proposed to solve either generic or specific cases. However, an empirical and exhaustive analysis of all methods would be impractical if not to say unrealizable. In the best of our knowledge, we try to include most important and representative methods among several approaches for the community detection task.

There are many possible theoretical taxonomies for community detection methods depending on the final objective of each categorization. For instance, one could classify methods according to differences in searching mechanisms, objective functions, assumptions about the structure to be found, expected qualities, hypothesis models, or even theoretical model employed, etc. Moreover, many methods are not just some simple algorithms to resolve a specific problem but instead are combinations of many different approaches in order to leverage as much as possible algorithmic power provided from each one, which makes the problem more tricky. There is not a consensus on how different methods are similar and how they can be classified into different families whose functionality can be resumed in some simple words. (Porter *et al.*, 2009) uses centrality based, local techniques, *modularity* optimization⁴, spectral clustering to describe communities in networks. (Fortunato, 2010), (Fortunato & Hric, 2016) group community detection methods into traditional data clustering methods, divisive algorithms, modularity-based methods, spectral algorithms, dynamic algorithms, statistical inferences based methods. (Coscia *et al.*, 2011) summaries community discovering into feature distance based, internal density, bridge detection, diffusion process, closeness based, structural pattern based, link clustering, meta clustering. In a context of Social Media, (Papadopoulos *et al.*, 2011) compares methods

⁴ First introduced by (Newman & Girvan, 2004) to assess hierarchical clustering levels of a community detection algorithm, *modularity* has become the most popular objective function in the context of community detection.

in substructure detection, vertex clustering, community quality optimization, divisive and model-based. (Bohlin *et al.*, 2014) aggregates different approaches into three principle classes representing different network models: null models, block models and flow models. (Schaub *et al.*, 2017) classifies methods into four perspectives: cut based, clustering internal density based, stochastic equivalent based and dynamical based showing four different facets of community structure.

In the following sections, community detection methods are classified according to different theoretical approaches including edge removal, modularity optimization, spectral partitioning, dynamic process and statistical inference. Although every theoretical taxonomy can be questionable, this categorization is expected to support the empirical analyses in the next sessions to verify whether theoretical and conceptual closeness could engender quality closeness in practice.

2.1.1 Edge removal based methods

Edge betweenness (GN) by (Newman & Girvan, 2004) detects communities by removing edges progressively according to their betweenness centrality scores. This method is based on the intuition that dense zones in a graph are loosely connected by a few edges that contribute a high inclusion in the shortest paths between every pair of nodes. Removing these edges would reveals densely connected communities.

Edge clustering coefficient (RCCLP) by (Radicchi *et al.*, 2004) suggests to replace the edge betweenness centrality of Girvan-Newman's method by edge clustering coefficient, which requires less computation time and hence reduces the algorithm complexity. In this paper, we analyze two configurations of this method corresponding to triangular ($g = 3$ denoted by *RCCLP-3*) and quadrangular ($g = 4$ denoted by *RCCLP-4*) versions.

2.1.2 Modularity optimization methods

Greedy optimization (CNM) by (Clauset *et al.*, 2004) greedily maximizes the modularity function Q by aggregating iteratively connected communities which induce a maximum increase or smallest decrease in modularity ΔQ .

Louvain method by (Blondel *et al.*, 2008) adopts two-step agglomerative process similar to that of the greedy optimization method. However, in each iteration of the first step, it allows nodes to move between communities until no additional gain in modularity can be obtained due to local switch. Then, a new graph whose vertices are the communities resulting from the first step is build and the process is repeated on the new graph to reduce computation time.

Spectral method (SN) by (Newman, 2006) identifies community structure by finding leading eigenvectors corresponding to largest eigenvalues of a modularity matrix. In this method, the problem of modularity optimization is *translated* to the problem of vector partitioning of modularity matrix.

2.1.3 Dynamic process based methods

Walktrap by (Pons & Latapy, 2005) defines a pairwise *dynamic distance* between nodes of a graph and then applies traditional hierarchical clustering to detect community structure. The distance is formulated using the transition probability of a random walker based on the concept that nodes belonging to the same community tend to "see" other nodes in the same way.

Infomod by (Rosvall & Bergstrom, 2007) uses an information theoretic model where a *signaler* try to send the structure of a network over a limited capacity transmission channel to a *receiver*. The network must be encoded in community structure in a way that minimizes the transferred information and the information loss tat the side of receiver.

Infomap by (Rosvall *et al.*, 2009) represents networks by a two-level structure description. Analogically, each node in a network is encrypted by a unique codeword composed by two parts: a prefix representing the community to which it belongs and a suffix representing the local code. Detecting community structure becomes equivalent to searching the coding rule to minimize the average code length describing random walks on the network.

2.1.4 Statistical inference based methods

Stochastic Block Model (SBM) by (Riolo *et al.*, 2017) uses a Monte Carlo sampling scheme to maximize a Bayesian posterior probability distribution over possible divisions of the network into communities. This probability implies an expected network model to be fitted from the observed network data. In this block model variant, the authors employ a new prior on the number of communities based on a queueing-type mechanism to calculate posterior probability. We analyze in the following sections both traditional *SBM* and *degree-corrected* version *DCSBM*, which is proved to perform better in practice.

Order statistics local optimization (OSLOM) by (Lancichinetti *et al.*, 2011) measures the statistical significance of a community by calculating the probability of finding a similar one in a null model. Following this concept, nodes are gradually aggregated into communities to find significant communities. Then nodes are considered to be swapped between communities in order to increase significance level.

2.1.5 Other methods

Spin glass model (RB) by (Reichardt & Bornholdt, 2006) finds communities by fitting the ground state of a spin glass model. Instead of favoring only intra-community edges and penalizing inter-community edges like the traditional modularity, this model also favors inter-community non edges and penalizes intra-community non-edges.

Label propagation (LPA) by (Raghavan *et al.*, 2007) exploits the topology of networks to infer community structure. It is closely related to the context of message passing paradigms or epidemic spreading. The principled idea of this method is based on the concept that nodes should belong to the community of most of their neighbors. Hence, they gradually update their memberships according to their incident nodes.

Speaker-listener label propagation (SLPA) - of Xie and Szymanski (Xie & Szymanski, 2012) modifies the propagation mechanism above by a new label update strategy. Also,

Table 1: Community detection methods involved in the study.

Approach	Publication	Ref. label	Time order	Implementation
Edge removal	(Girvan & Newman, 2002)	GN	$\mathcal{O}(nm^2)$	igraph ^a
	(Radicchi <i>et al.</i> , 2004)	RCCLP	$\mathcal{O}(m^4/n^2)$	Authors ^b
Modularity optimization	(Clauset <i>et al.</i> , 2004)	CNM	$\mathcal{O}(m \log^2(n))$	igraph
	(Blondel <i>et al.</i> , 2008)	Louvain	$\mathcal{O}(n \log(n))$	Authors ^c
	(Newman, 2006)	SN	$\mathcal{O}(nm \log(n))$	igraph
Dynamic process	(Pons & Latapy, 2005)	Walktrap	$\mathcal{O}(n)$	igraph
	(Rosvall & Bergstrom, 2007)	Infomod	NA	Authors ^d
	(Rosvall <i>et al.</i> , 2009)	Infomap	$\mathcal{O}(m)$	Authors ^e /igraph
Statistical inference	(Lancichinetti <i>et al.</i> , 2011)	OsloM	$\mathcal{O}(n^2)$	Authors ^f
	(Riolo <i>et al.</i> , 2017)	(DC)SBM	Parametric	Authors ^g
Other methods	(Reichardt & Bornholdt, 2006)	RB	$\mathcal{O}(n^2 \log(n))$	igraph
	(Raghavan <i>et al.</i> , 2007)	LPA	$\mathcal{O}(m)$	igraph
	(Xie & Szymanski, 2012)	SLPA	$\mathcal{O}(m)$	Authors ^h
	(Meo <i>et al.</i> , 2014)	Conclude	$\mathcal{O}(n+m)$	Authors ⁱ

^a Published at <http://igraph.org/>

^b Published at <http://homes.sice.indiana.edu/filiradi/resources.html>

^c Published at <https://sourceforge.net/projects/louvain/>

^d Published at <http://www.tp.umu.se/~rosvall/code.html>

^e Published at <http://www.mapequation.org/>

^f Published at <http://www.oslom.org/>

^g Published at <http://www-personal.umich.edu/~mejn/>

^h Published at <https://sites.google.com/site/communitydetectionslpa/>

ⁱ Published at <http://www.emilio.ferrara.name/code/conclude/>

instead of keeping only hard membership information, each node is equipped by a memory to contain the labels that it receives. Then, in the update phase, nodes transmit the membership to their neighbors according to the membership frequency in the memories.

Mixing global and local information (*Conclude*) by (Meo *et al.*, 2014) combines a dynamic distance with a modularity optimization process to identify community structure. Firstly, the authors define a new pairwise proximity function using random and non backtracking walks of finite length to determine distances between vertices. Then, the multi-level modularity optimization strategy of *Louvain* method (Blondel *et al.*, 2008) is employed in combining with the defined distance to find community structure.

Table 1 summaries the methods presented previously grouped by different approaches. Since community detection is getting more and more attention in the network science community, there is a huge volume of work that has been published in the recent years to evaluate different methods including both theoretical and empirical approaches. However,

there is not any quantitative definition of community that is explicitly implemented inside algorithms, therefore it is challenging to distinguish the topological differences of community structures using different methods, even when the associated concepts are quite theoretically discernible. Additionally, it is still not clear yet whether a proximity in the assumption of community concept will engender a structural similarity of communities that could be detected. Our comparative analysis in the next sections will try to address these questions in more details.

2.2 Experimental dataset

In this section, we describe some statistical properties of networks that will be included in the following analysis. It is expected that networks in each category are spread in a wide range of structural measures. However, available biological networks that have been published and analyzed widely are relatively small in comparison to the other networks of the other families. Besides, due to the complexity of the analysis process, we limit the domains of interest at 5 categories which are commonly researched and where numerous networks are available. In this study, we consider 108 different networks, which is relatively large in comparison to many studies. Many notable related work where some of these networks are also employed to study community structure could be mentioned for a quick reference: Orman *et al.* use 6 networks to evaluate the structure of communities discovered by several detection techniques (Orman *et al.*, 2012); Lancichinetti *et al.* use 15 networks to characterize structural communities (Lancichinetti *et al.*, 2010); Hric *et al.* use 16 networks to reveal differences between structural communities and ground truth (Hric *et al.*, 2014); Leskovec *et al.* use over 100 networks to analyze network community profile (Leskovec *et al.*, 2008) and 230 networks to evaluate the goodness of ground-truth communities in social networks, within this number, 225 samples of the Ning online social networking platform's networks⁵ are aggregated (Yang & Leskovec, 2013). Table 2 resumes the composition of networks that have been analyzed in this section.

Some notable structural measures of networks in the dataset are illustrated in Figure 1. It is noticeable that apart from biological networks which are relatively small, the other classes cover quite a wide range of number of nodes, edges, mean degree, clustering coefficient and edge density. Since real world networks are relatively sparse, the number of edges increase linearly in function of the number of nodes and consequently, the edge densities decrease linearly by the number of nodes (since the number of possible connections increase quadratically by the number of nodes in a community). This sparsity property can be easily noticed from Figure 1(a,d). Specifically, the number of edges increases linearly in function of the number of nodes with equivalent rates among different network categories as can be deduced from the gradients of the linear estimates. From Figure 1(b), it can be seen that the average degree of the networks in the dataset varies principally between 1 and 100 edges per node except for 2 communication networks. Also, the majority of networks has an average degree of approximately from 10 to 20 connections. In a global point of view, networks in the dataset have a quite strong modular quality since most of them have relatively high clustering coefficient as shown in Figure 1(c).

⁵ <https://www.ning.com/>

Table 2: A summary of network dataset used in this analysis where “Size” is the number of networks analyzed in each category, “Nodes” and “Edges” indicates the average number of nodes and edges of networks in each category respectively. *The last row shows the total number of networks, nodes and edges in the whole dataset. This dataset is collected from several sources including: <http://networkrepository.com> (Rossi & Ahmed, 2015), <http://konect.uni-koblenz.de> (Jerome, 2013), <http://snap.stanford.edu> (Leskovec & Krevl, 2014)

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*	108	1.99M	9.08M	

3 Preliminary analysis of community detection methods

3.1 Computation time performance

Since computation time is a crucial factor to be considered in the selection of an algorithm, it is worth analyzing experimental performances to see how different community detection methods accomplish their task in real-world networks. By reusing the dataset summarized in Table 2, we proceed to assess official implementations of community detection methods introduced in Table 1. As listed in this table, the implementations are provided officially either from their own authors or popular network analysis tools, which can be easily accessed from a large public.

We employed all implementations stated above to identify community structures on all the networks contained in the dataset. We measured the time needed for each implementation to compute each partition on each network. The default parameters configured by the implementations are kept unchanged during the test. The calculations were executed on a server equipped by an Intel Xeon CPU E5-2650 with 32 cores of 2.60 GHz and a memory capacity of approximately 100 GBytes. However, due to the high complexity of some methods, only processes that finish in a practical amount of time (less than 4 hours) are taken into account. However, for a reference purpose, we let some of longer computations go on, for example, *Conclude* method took approximately 9 days to identify community structure on a network of 300 thousand vertices and 1 million edges; *GN* method did not finish its calculation for networks of more than 4 thousand nodes and 40 thousand edges within 2 days. Consequently, the experiments that theoretically require too much time are neglected in the test. It is also worth noting that the calculations of communities on large-scale networks are also restrained by limited memory, therefore calculations that are supposed to be finished within 4 hours but required too much memory can not be shown

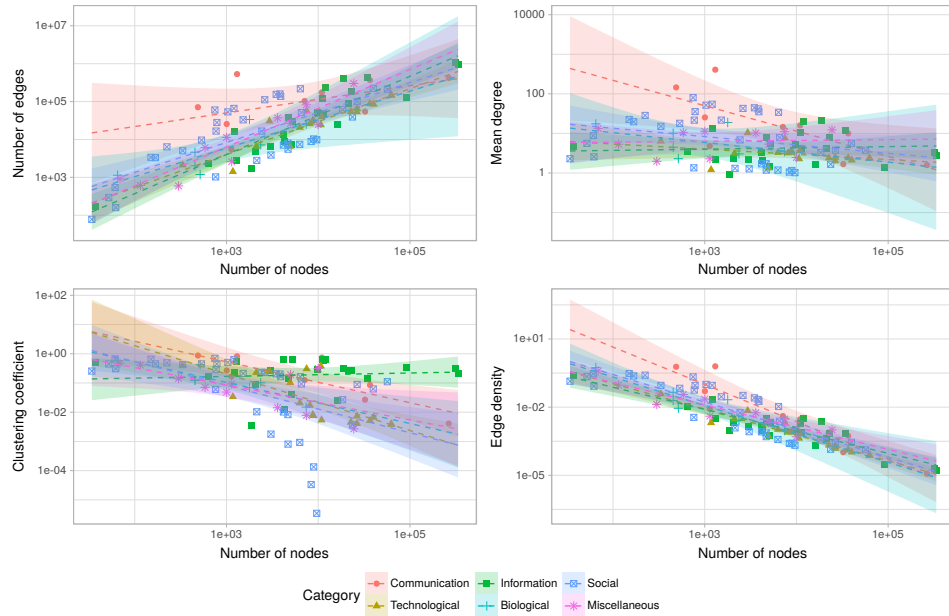


Fig. 1: From left to right, up to bottom, we illustrate structural measures of the 108 networks: (a) Number of edges as a function of the number of nodes, (b) Mean degree $\langle k \rangle$ as a function of the number of nodes, (c) Clustering coefficient in function of the number of nodes, (d) Edge density in function of the number of nodes. The colored backgrounds represent the 95% confidence intervals of the relations estimated from the dataset using a linear regression model for the corresponding variables on each network category.

here neither. We repeat the calculations 5 times on average for each pair graph/method to reduce the fluctuation impact. Eliminating all the cases that do not satisfy our requirements, the final successful rate (number of partitions identified over the number of possible tests) ended up at around 44.72%, mainly due to time/memory surpassing.

In the following figures (from Figure 2 to Figure 7) that illustrate the analyses on experimental time consumption, some conventions are commonly used. Points in the figures correspond to separated executions that have been measured. The solid lines with the same corresponding colors to the points are estimated relations between computation time and network size (number of vertices and number of edges) using a local regression model (Cleveland, 1979). The dark colored backgrounds around the regression curves represent 95% confidence intervals of the model parameters. Besides, we show the worst case theoretical execution time (number of calculation needed in this case) of associated algorithms are included for a comparative reference purpose. From the analysis of structural characteristics of the dataset as shown in Figure 1(a), it is noticeable that most networks are sparse, i.e the number of edges (m) increase in a linear function by the number of nodes (n). Hence, in our estimate, we plot theoretical execution time by assigning $n = m$. For the simplicity of illustration, we grouped the measures of the methods by their approaches (refer to our Classification on Table 1).

Community structure evaluation

11

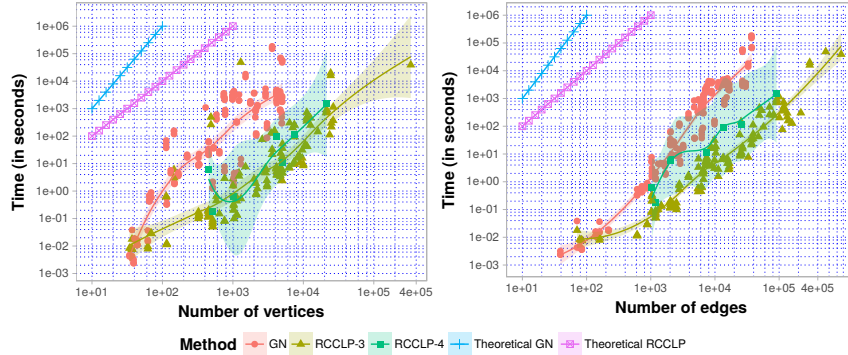


Fig. 2: The execution time needed by GN, RCCLP-3 and RCCLP-4 methods to identify community structures on networks of the dataset.

The first group of methods consists in centrality detection techniques to identify community structure. As we can see in Figure 2, the *GN* method can not be accomplished in our test for networks of more than 4 thousand of nodes or 30 thousand of edges. The outcome is quite reasonable since the theoretical estimation for this method is $\mathcal{O}(nm^2)$, which grows quickly in function of network size. Remind that one of the primary purpose of the *RCCLP* method is to reduce the time complexity of the *GN* method. We can easily observe that this objective is achieved since the *RCCLP-3* reduces an order of around 10^3 times for graphs from 3 hundred nodes. *RCCLP-3* can well function with graphs up to millions of edges. However, when we proceeded the same test with *RCCLP-4*, the method rarely reached its terminus for large graphs as well as small graphs. As we can see in the figure, there are few dots at the two sides. The reason is that there are not many (or even absent) 4-step close paths on real world networks. As it is not very probable that such structures exist in small graphs, finding them in large graphs also require a huge amount of time, *RCCLP-4* shows a poor performance in our tests. Therefore, this configuration of the method is not recommended, as well as versions with $g > 4$ would logically poorly perform. It is also worth noticing that *RCCLP-3* and *RCCLP-4* are extremely memory consuming and are not suitable for limited resource devices. Finally, theoretical and practical time seem to find a consensus as the increments of time in function of network size are quite consistent in the three cases.

The next group includes methods using modularity optimization processes whose experimental measures are shown in Figure 3. Practically, the three methods in this family require a reasonable time for calculating community structures. The most time consumed experiment took less than 2 hours for a graph of 1 million edges. *Louvain* method is the fastest in this group whose computation increases approximately in linear time. It took only 9 seconds for the largest graph. Among the three methods, the optimization using spectral approach is the most expensive. However, all of these three methods have higher performance than the methods in the edge removal group previously stated. The experimental results also justify theoretical estimates about the complexity of these methods.

Similarly to the two previous group, the computation time needed by methods in the dynamic process group is illustrated in Figure 4. In terms of time consumption, this group

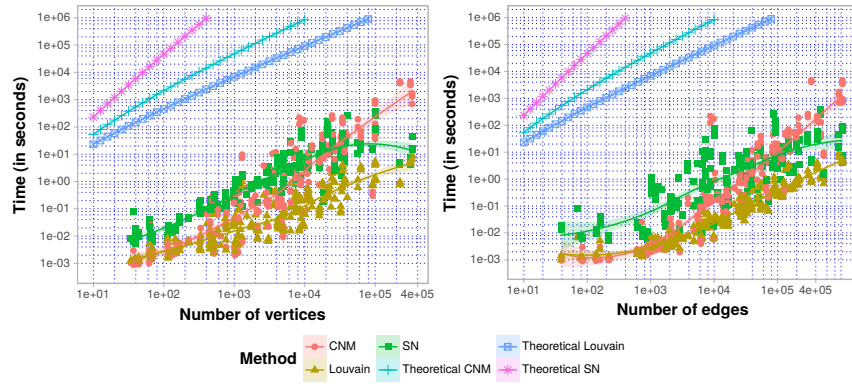


Fig. 3: The execution time needed by CNM, Louvain and SN methods to identify community structures on networks of the dataset.

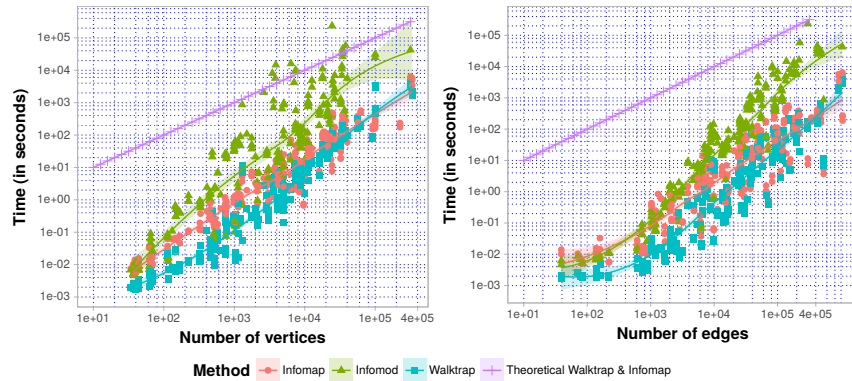


Fig. 4: The execution time needed by Infomap, Infomod and Walktrap methods to identify community structures on networks of the dataset.

shows a better performance with respect to the first group, but generally worse than the modularity optimization group (except for the *Walktrap* method for small and average size graphs). Among them, *Infomod* has the poorest performance. In the meanwhile, *Walktrap* and *Infomap* work asymptotically equally good with a slightly better rendition for *Walktrap* in small and average size graphs.

The same analyses for methods in the two final groups are shown in Figure 5 and Figure 6. We can easily see that *DCSBM* and *Oslom* have practically identical performance in terms of time consumption with a slightly less expensive on the side of *DCSBM*. In the last group, the results are quite discernible between different methods. The label propagation method *LPA* shows a clear distinctive curve indicating its out-performance over the other methods. Besides, *SLPA* works quite well, but less fast than *LPA* although it employs some additional techniques to reduce the number of necessary calculations (Xie & Szymanski, 2012). This difference in the performance is due to the more complicated mechanism that *SLPA* uses in comparison to *LPA*. The fact that *SLPA* has to reserve dedicated memories

Community structure evaluation

13

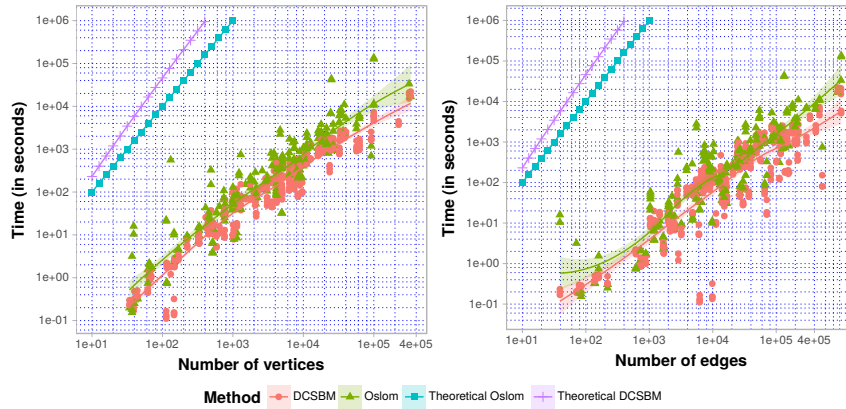


Fig. 5: The execution time needed by DCSBM and Osloom methods to identify community structures on networks of the dataset.

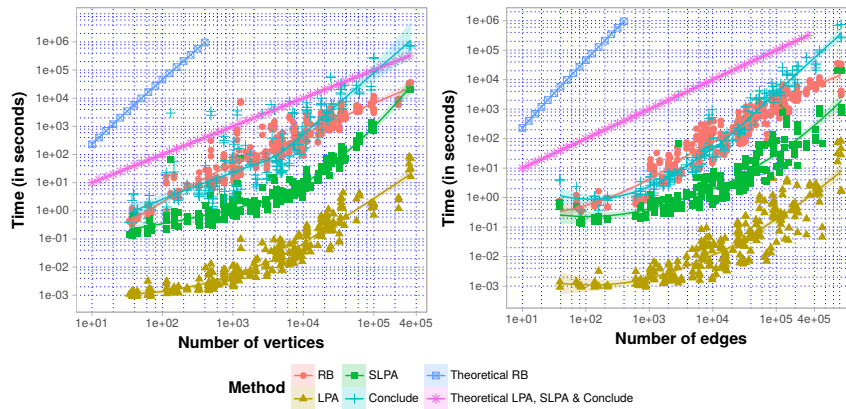


Fig. 6: The execution time needed by RB, LPA, SLPA and Conclude methods to identify community structures on networks of the dataset.

for all nodes of the network to stock the membership information that they received during the detection process and update them regularly to transfer into their neighbors makes it demanding. Therefore, despite of a 5 to 10 times of improvement in the label update strategy, the global performance can not surpass that of *LPA* method. In terms of scalability, *LPA* and *SLPA* seem to exhibit the same comportment which is nearly linear for small and medium graphs but accelerate in large graphs. The spin glass model *RB* manifests a better than expected presentation with an undeviating linear augmentation. The only unexpected behavior is spotted in *Conclude* method, as when the size of input graphs exceed some thousands, the required time has been inflated by a factor of n , making it very demanding for large graphs.

Finally, we aggregate all the analysis measures in the 5 previous groups into a common illustration as shown in Figure 7. At the same time, for a more convenient observation, we remove all the points corresponding to the experiments and keep only the regression

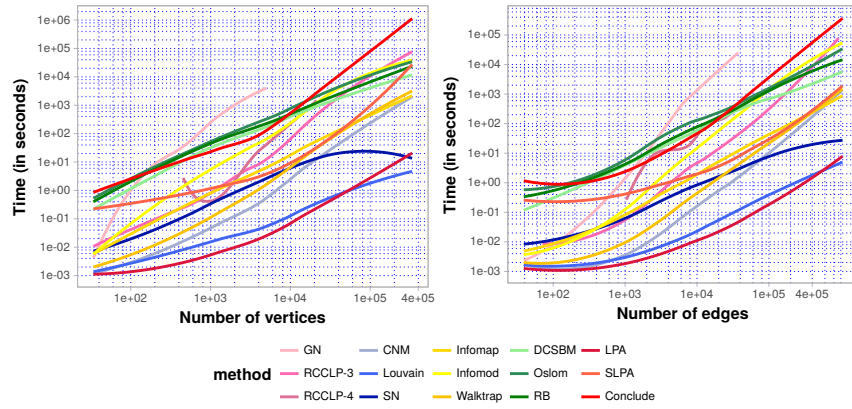


Fig. 7: The estimated execution time needed for each method to identify community structures on networks of the dataset using a local regression model. Methods of the same theoretical family (in the same group) are represented by chromatically similar color.

curves, which are the estimates of execution time for these methods in function of number of vertices on the left hand side and number of edges on the right hand side. At a first sight, it is easy to see that except for *GN*, the necessary execution time for all other methods are limited in a range that increases polynomially in function of network size, which reflect well theoretical estimates. This range is upper-bounded by *Conclude/Oslom* and lower-bounded by *LPA* which corresponds to worst and best performed method(s) respectively. Another important information which can be deduced from this figure is that, for most real world networks of size in the range up to 1 million edges, choosing a fast detection method could economize an order of 10^3 times to 10^5 times calculation effort. This is an important element to be considered in applications where time consuming is a serious problem.

We demonstrate in Table 3 the ranking of these methods according to our test for reference purpose. *GN* and *RCCLP-4* are not involved in this ranking since they failed to accomplish their tasks in large graphs, which also means they are the most time consumed methods within the methods that we analyze. We show both the ranking by the average and the median of time. Since the average-time ranking is heavily affected by the measures on large graphs, i.e. the methods that succeeded to discover communities on very large graphs are lower ranked than methods that were not able to do so. In these cases, the ranking by median is more accurate and it reflects well the relative performance on small and medium graphs between the methods. For large graphs, using the ranking by average would better fit.

3.2 Analysis on community size distribution

The number of latent communities that should be induced from a given network is one of the major question in community detection context (Fortunato & Hric, 2016), (Riolo *et al.*, 2017), equivalent to the subject of the expected number of clusters in classical clustering problem. Observing the number of communities discloses useful information about the mesoscopic structure of a network. The variation of the number of communities

Table 3: Ranking of analyzed methods according to their amount of time consumed to identify community structure on networks of the dataset.

Method label	Rank by average	Rank by median	Scalability
RCCLP-3	9	8	Low
CNM	5	3	Medium
Louvain	1	2	High
SN	3	5	High
Walktrap	4	4	High
Infomod	12	9	Low
Infomap	6	7	Medium
OsloM	11	14	Low
DCSBM	8	12	Low
RB	10	13	Low
LPA	2	1	High
SLPA	7	6	Medium
Concude	13	11	Low

in a network involves different level of resolutions. An analogous way to describe the concept of resolution is the distance from an object that we prefer in order to contemplate it. The closer we get to an object, the more its detailed micro-structures could be perceived, in the meanwhile the less information about the global organization tends to be clear. Although several multi-resolution approaches (Lambiotte, 2010; Pons & Latapy, 2011) incorporate resolution parameters into their solutions providing more flexible mechanisms and different modular scales of networks, it is not always obvious to regulate appropriately these parameters without ad-hoc cases. The inclusion of multi-resolutions parameters, of course, widen the possibility of understanding networks, but in the expense of the automatic aspect that is sometimes required in clustering problems.

In this section, we compare again the previously mentioned methods but this time according to their resolution abilities. We use the same network corpus and we keep again all default configurations of the implementations unchanged to ensure the consistency of future results. From the antecedent analyses, some modifications will be applied on our testing process as follows:

1. From the observation of the network size distribution in Figure 1(a) as well as the previous computation time analyses, the linear relation between number of vertices and number of edges of networks in our corpus becomes evidenced. As a consequence, it will be redundant to address the relation of dependent variables in respect of these two latter predictors. Therefore, only analyses in function of number of vertices will be introduced.

2. In community detection problem, showing only the numbers of communities discovered on networks or their statistical derivatives would not always be enough. Assume that the sizes of communities in an arbitrary network follow a negative power-law distribution, its means that the number of communities depends heavily on the number of tiny communities. Therefore, we also observe the distribution of community size to discern the differences between methods which could not be recognized by seeing solely the number of blocks.
3. Due to a huge number of required calculations and a limited hardware resource, discovering processes in the last section were interrupted unless they are finished in a few hours. Here, some more efforts have been flexibly given if a method is supposed to be finished in a reasonable amount of time.

For a given network in the dataset, we applied all of the presented methods to identify the set of communities predicted by each one and measured their volumes. Similarly to the last part, for the simplicity of observation, we group methods by different families depending on their approaches. We illustrate the obtained results of community repartition measures in Figure 8 to 12 by using some conventions as follows:

Conventions for Figure 8 to Figure 12

1. A figure (denoted *a*) on the top contains three following sub-figures:
 - (a) The central figure (a1): shows a scatter plot about the distribution of communities in function of the number of vertices of the network to which they belong. The solid lines in the figure represent the estimated average community size in function of number of vertices using a local regression model (Cleveland, 1979). Dark colored backgrounds around the lines are 95% confidence intervals of the estimates.
 - (b) The top figure (a2): exhibits marginal density distributions of communities found in each range of network sizes. They are rendered by a Gaussian kernel estimator.
 - (c) The right figure(a3): illustrates another type of marginal density distributions of communities in function of their sizes. They are also rendered by a Gaussian kernel estimator.
2. A figure on the bottom (denoted *b*) presents the number of communities in function of the number of vertices of different networks as well as the estimated relation between these variables using the regression model stated above. Dark colored backgrounds around the lines shows 95% confidence intervals of the estimate relations.

3.2.1 Edge removal approach: *GN*, *RCCLP-3* and *RCCLP-4*

From Figure 8, we can notice again that *GN* method can only be able to function on small and medium networks due to its high complexity, which is quite obvious from theoretical analysis. *RCCLP-3* and *RCCLP-4* can detect up to the largest networks in our corpus. By

Community structure evaluation

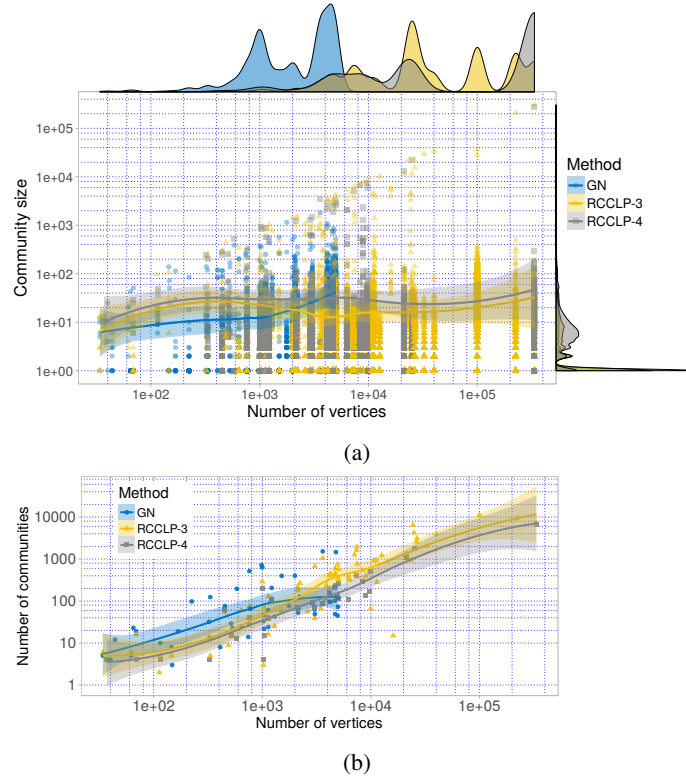


Fig. 8: Fitting quality of GN, RCCLP-3 and RCCLP-4 methods on the networks of the dataset.

observing the right marginal density distribution, surprisingly, all of these methods identify a huge number of singleton communities⁶. The average number of singleton communities is around 24% which can be up to 60% in some cases. The reason for this aberrant phenomenon is that in some dense and small networks, there exists too many high and equivalent central vertices and edges. The separating mechanism employed here keeps removing central nodes or edges until a large number of vertices are isolated, creating singletons or very small communities. Since *GN* only works on small graphs, it is highly impacted by this phenomenon in our experiment. Besides, in a global observation, we can see in the top figure that the majority of communities detected by these methods are very small for the same reason. From Figure 8(a), we can see that a large number of communities have only less than 10 vertices even in very large networks. This makes the number of communities increase rapidly as one can remark on Figure 8(b). Remind that the distributions of community size have right-skewed shapes, meaning that the majority of communities are small and most of them are found under the lines of average community sizes. Therefore, the three methods of this family have very high resolutions. Notwithstanding, this result need to be understand with caution due to two reasons:

⁶ Communities that contains only one vertex

1. The density function in Figure 8(a1) reveals that the successful rate on discovering community structures of the three methods are distinguished fundamentally. In fact, due to the high complexity of time and memory, many networks are not successfully resolved, which degrade importantly the comparison quality.
2. As a consequence of the first reason, there is a high fluctuation in the dependent variables which make the confidence intervals quite large. A deeper investigation on the quality on small and medium networks could partially palliate this problem.

Although the previously mentioned issues, this class of methods remain the one which conjectures the highest number of communities with a great consensus.

3.2.2 Modularity optimization approach: CNM, Louvain and SN

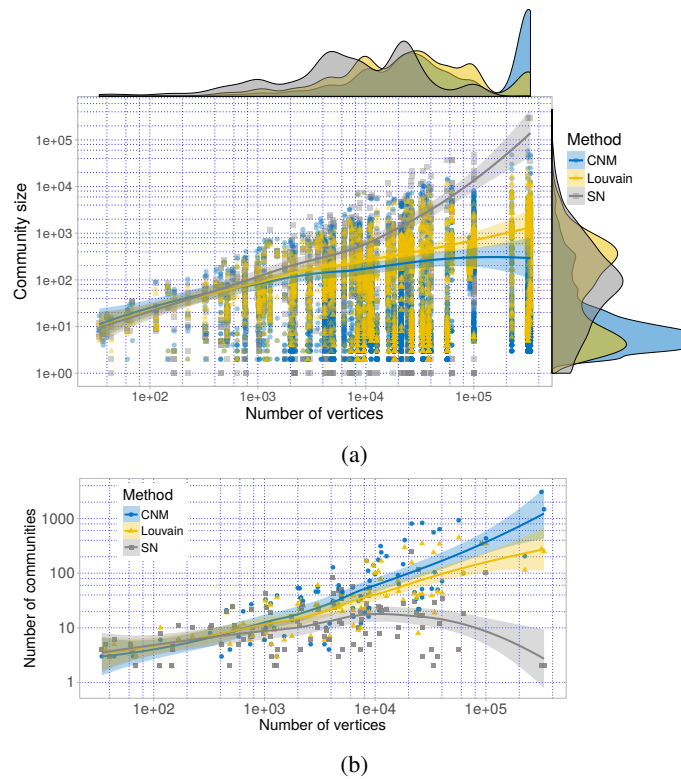


Fig. 9: Fitting quality of CNM, Louvain and SN methods on the networks of the dataset.

On this second group, our measures are more complete since all three methods succeeded to resolve large networks. From Figure 9(a2), it can be seen that there is a regularity between the distributions of communities over the whole range of networks except for the range of very large networks. Actually, in this range, the behavior is very different with the three methods. While *CNM* determines a very large number of medium and small communities, *Louvain* identifies less small communities and more medium and large

Community structure evaluation

communities. On the other hand, *SN* only proposes a partition of two giant communities. For instance, if we take the *Amazon* network (Leskovec & Krevl, 2014), while *CNM* detected 1480 clusters, this number is 249 for *Louvain* and only two for *SN*. The same phenomenon is also remarked for another example, the *DBLP* network (Leskovec & Krevl, 2014), the corresponding numbers are 3077, 275 and 2 in the same order for the three methods. This notice can also be remarked in smaller networks as can be seen in Figure 9(b), however gap between the number of communities reduces gradually from the right to the left of the figure. But in general, the order remains unaltered in as experienced in our observations, i.e. the average number of communities detected by *CNM* is larger than that of *Louvain* which is in its turn larger than that of *SN*. Consequently, the order of community sizes are inversed since the sizes of graphs are fixed as can be seen in Figure 9(a1). Another remark can be extracted from Figure 9(a3) about the diversity of community size, while *CNM* and *CN* consistently move towards small and medium communities respectively, *Louvain* on the other hand tends to propose both small and medium size communities.

3.2.3 Dynamic process approach: Infomap, Infomod and Walktrap

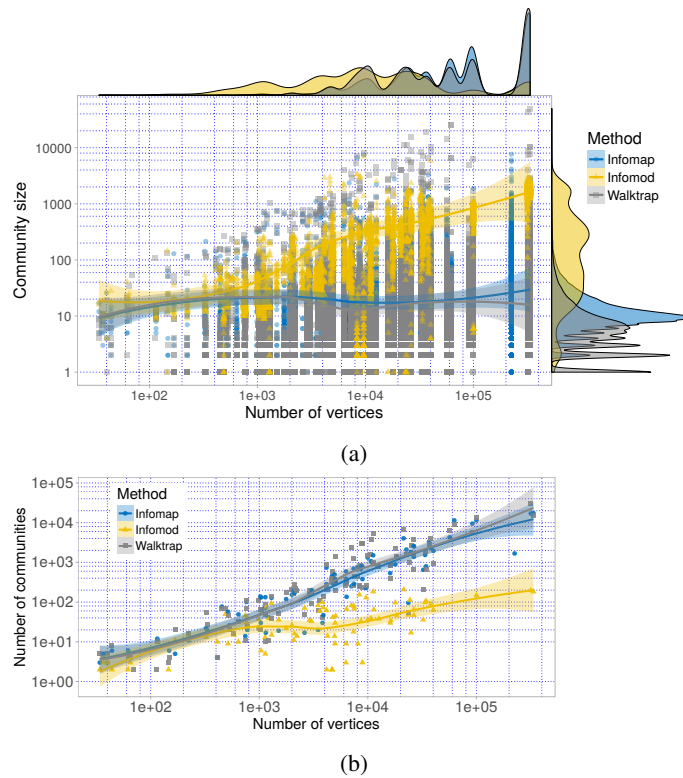


Fig. 10: Fitting quality of Infomap, Infomod and Walktrap methods on the networks of the dataset.

At a first glance, we can see a clear separation within the three methods. While *Infomap* and *Walktrap* display quite comparable evolution of average community size depicted by Figure 10(a1) as well as marginal distribution as depicted by Figure 10(a2-a3), *Infomod* is driven distinctly apart. Diving into the measures, we notice that in *Infomod*, there is a relatively uniform repartition of communities which is upper-bounded by the largest containing 6948 vertices. Unlike many other methods including *Infomap* and *Walktrap*, the number of medium and large communities discovered by *Infomod* does not outnumber the number of small communities as stipulated by heavy-tailed distributions. As a consequence, the total number of communities observed remains low and increases with a small constant pace.

Infomap and *Walktrap* tend to keep their average community size limited around 10 to 30 over the whole range of networks. This phenomenon keeps them away from the resolution limit issue. In both methods, the most popular community size can be found around 10 nodes or smaller. Our more specific measure on the median community size shows almost similar results for *Infomap* while this number decreases slightly for *Walktrap*. Above these values, the number of communities decreases profoundly. The biggest difference between these two methods can be easily observed at the spurious region on the marginal distribution of Figure 10(a3). In fact, unlike *Infomap* which produces very moderately small communities, *Walktrap* identifies a huge number of isolated nodes (around 10% according to the statistics) and small communities similarly to *RCCLP-3* and *RCCLP-4*. This problem may be due to the agglomerative hierarchical clustering employed by *Walktrap* to detect communities which engenders orphaned peripheral vertices which has been indicated by (Newman & Girvan, 2004) - Figure 3. This problem, however, is quite simple to be palliated since these peripheral vertices could be assigned to their closest neighbor's community. By removing this issue, we have got a quite similar result for *Infomap* and *Walktrap*.

In terms of average number of communities, *Infomap* and *Walktrap* show practically the same behavior. The evolutions are nearly coincided over the whole range of networks with small confidence intervals, especially in the middle range. For medium and large networks, as seen in Figure 10(b), it is very likely that *Infomod* identify much less number of communities. In fact, more than 75% of *Infomod*'s partitions have less communities than those of the other two methods.

3.2.4 Statistical inference approach: *SBM*, *DCSBM* and *Oslom*

In the case of statistical inference, we see a quite similar phenomenon previously experienced in the dynamic approach. Specifically, the distributions of community size of the two implementations *SBM* and *DCSBM* are nearly coincided with a slightly higher average community size for the former. In fact, in this Bayesian block model, it is necessary that the prior distribution of number of block is given. According to different block model variants, one could assume various hypotheses about underlying mechanisms that create observed network under the corresponding regulations of block structures and define a prior probability. In the implementation that has been employed, the authors initialize the community discovering process by assigning nodes randomly to groups according to a queuing-type mechanism and then use a Monte Carlo sampling process to maximize the posteriori probability. However, the calculation becomes extremely time consuming when

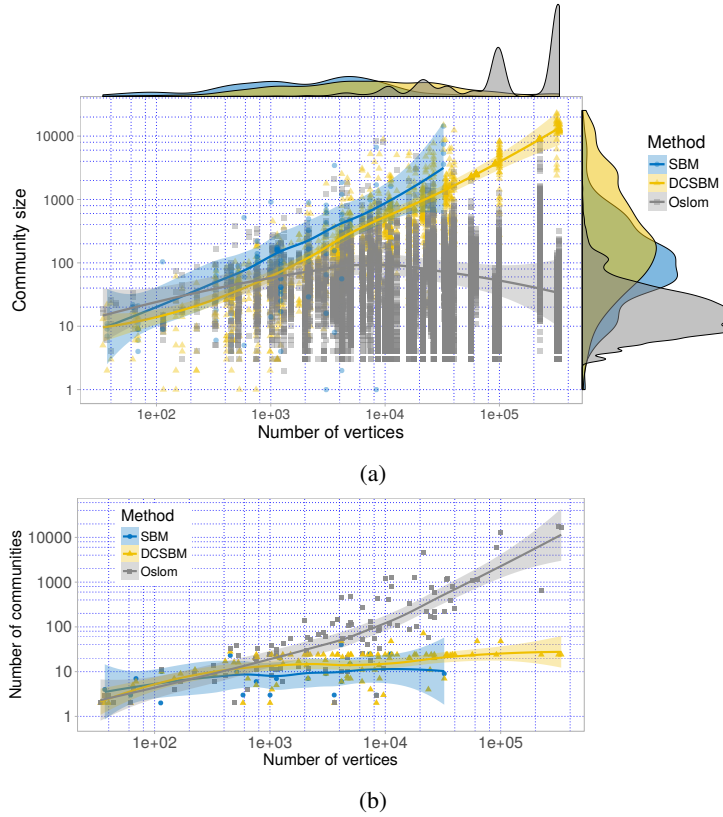


Fig. 11: Fitting quality of SBM, DCSBM and Osлом methods on the networks of the dataset.

the maximum number of communities is too large (Riolo *et al.*, 2017). Hence, by default, the maximum number of communities is configured at 25 as proposed, which leads to an underestimation of medium and large graphs as shown in Figure 11(b) as also be noticed by the authors. One can see the impact of this regulation as the number of communities approaches asymptotically 25 independently with the network size on the right hand side of the figure.

By observing the distribution of community size in Figure 11(a1), it is understandable that the average block size of *SBM* and *DCSBM* increases linearly in function of number of vertices. As the number of communities remains constant, the average community size must increase proportionately. Besides, the Figure 11(a3) also reveals that community sizes are well spread around their mean values, which makes the marginal distribution quite symmetric for both *SBM* and *DCSBM*. There is nearly no particular inclination towards small communities as acknowledged in some previous methods.

For the case of *Osлом*, the separation is quite clear. It unveils much more communities, making their sizes very small. Figure 11(a) shows that the majority of *Osлом*'s communities are found under the average values of the associated partitions of *SBM* and *DCSBM*.

Our demonstrations show that there is indeed a significant difference in the repartition strategies of these methods.

3.2.5 RB, LPA, SLPA and Conclude methods

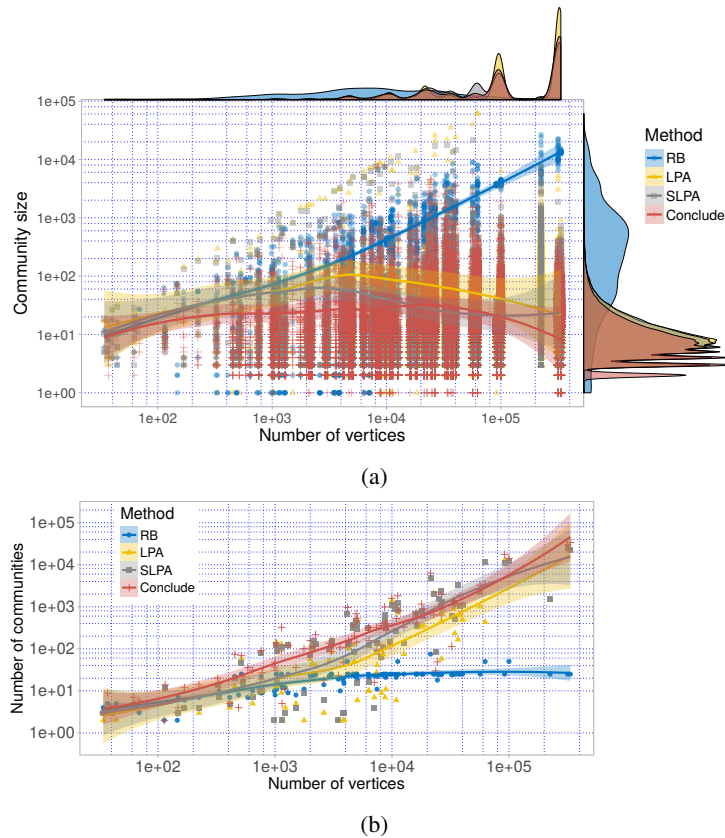


Fig. 12: Fitting quality of RB, LPA, SLPA and Conclude methods on the networks of the dataset.

In the last group, we discover that there is a remarkable coincidence in all distributions of the three methods *LPA*, *SLPA* and *Conclude*. In fact, the difference between them is nearly indistinguishable on the marginal measures. There is only a small discrepancy in the number of detected communities in very large networks as can be noticed from Figure 12(a2), such that *LPA* detected slightly more communities than *SLPA* and *Conclude*. From Figure 12(a3), one can see that the majority of communities are quite small in these three methods. Similarly to *CNM*, *Infomap* or *Walktrap*, the majority of communities are small, i.e. have less than 10 nodes.

On the three methods, one could see that the variation of the data is significantly large, which produce also a large variation in our estimates. Since the associated prediction

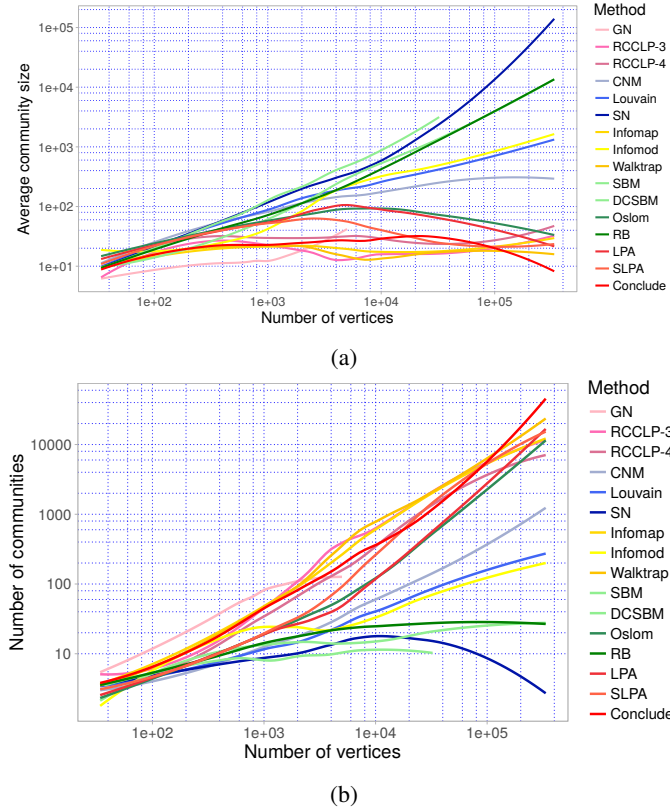


Fig. 13: A summary of community size estimation

intervals for the estimates are likely to be larger, predictions related to community size distribution are not expected to be accurate.

On the other hand, *RB* method shows a solid consistency with much less variations in our examination. Average community size increases regularly and number of communities becomes saturated from medium size networks. The behavior of *RB* method is very resembling to that of *DCSBM* observed in Figure 10. Consequently, it is supposed to suffer the resolution limit for large networks. Notwithstanding, since *RB* is provided with a resolution tune parameter, the method may escape from this effect if the parameter is correctly chosen.

3.2.6 Summary

For the final step of this part, in the same manner as the previously presented time computational analysis, we aggregate for all methods the estimates of average community size and the number of detected communities in function of number of vertices in the network in Figure 13(a) and 13(b) respectively. One can see that there exists several repartition strategies hidden in these methods. If we use the preference of theoretical number of recoverable communities in a k -planted partition model (Ames, 2013), being $O(\sqrt{n})$, the

studied methods could be considered to over-fit (create more than k clusters) or under-fit (create less than k clusters) as presented in Table 4, in the third column.

Table 4: Ranking of analyzed methods according to their number of detected communities. A method is considered to over-fit if it detects asymptotically more than \sqrt{n} clusters. The group numbers exhibit the estimated similarity based on fitting quality.

Method label	Size wrt. k -planted model	Fitting
GN	Bigger	Over-fit
RCCLP-3	Bigger	Over-fit
RCCLP-4	Bigger	Over-fit
CNM	Close	Over-fit
Louvain	Close	Under-fit
SN	Smaller	Under-fit
Walktrap	Bigger	Over-fit
Infomod	Close	Under-fit
Infomap	Bigger	Over-fit
Oslo	Smaller	Under-fit
SBM	Smaller	Under-fit
DCSBM	Smaller	Under-fit
RB	Smaller	Under-fit
LPA	Bigger	Over-fit
SLPA	Bigger	Over-fit
Concude	Bigger	Over-fit

We can see that, in a general view from the second and third column of Table 4, methods belonging to the same theoretical class which shares a common assumption about the definition of community have a tendency to show the same fitting quality, as also discovered by (Ghasemian *et al.*, 2018). However, although being useful to help practitioners to presume the expected number of clusters a method would detect with respect to the theoretical experience, it is still very embarrassing to know which method to use since the reference is based on an hypothesis about an underlying model. This also means that if the hypothesis about the partition model change (another model than k -planted model), the expected number of communities will be diversified, and hence the indicated fitting quality preference becomes disproved. As a consequence, we propose a novel technique to estimate the similarity of community detection methods based on community size distributions in the next section. Certainly, this is only one among interesting quality aspects that differentiate one method from the others. Nonetheless, we will demonstrate that it also allows to get more insight into the difference in terms of partitioning strategy.

3.3 Similarity based on community size distribution

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 (Dao *et al.*, 2018b). As such, two methods could be supposed to be similar if their corresponding density distributions expose a large intersection area as shown in Figure 14(a). From this notice, we can define our new similarity function as follows:

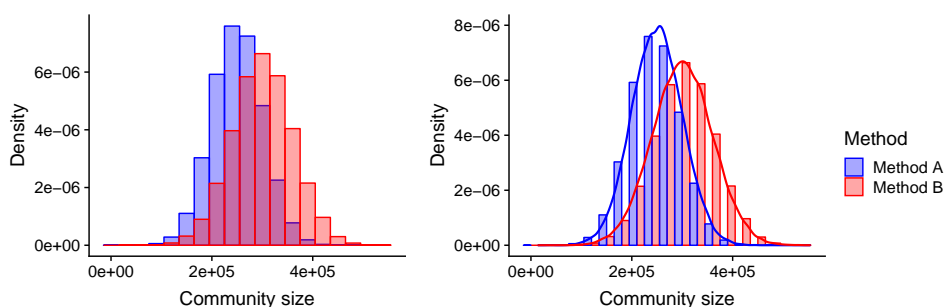


Fig. 14: The distribution of sizes of communities detected by two different methods. On the left (a) overlap fraction using histogram, on the right (b) when community sizes interlace, the similarity is better estimated using a kernel density estimator.

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 $\mathcal{A} = \{x_1^a, x_2^a, \dots, x_r^a\}$ and $\mathcal{B} = \{x_1^b, x_2^b, \dots, x_s^b\}$ being the ascending ordered sets of sizes of communities: $1 \leq x_1^a < x_2^a < \dots < x_r^a$ and $1 \leq x_1^b < x_2^b < \dots < x_s^b$. The multiplicity functions $n^a : \mathcal{A} \rightarrow \mathbb{N}_{\geq 1}$ and $n^b : \mathcal{B} \rightarrow \mathbb{N}_{\geq 1}$ 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:

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

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 Figure 14(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 in Figure 14(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 function as shown by the solid lines in Figure 14(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 rather a large portion of comparable-size ones. Hence, we consider the following estimator to take into account local information of community size x_0 :

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

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:

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

where

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

with $u \in \{a, b\}$. In the estimations of this paper, the bandwidth h is selected based on the normal reference rule (Silverman, 1986) to minimize the mean integrated squared error.

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 have obviously to be relativized. However, our large and representative corpus would help to reduce the dependency impact.

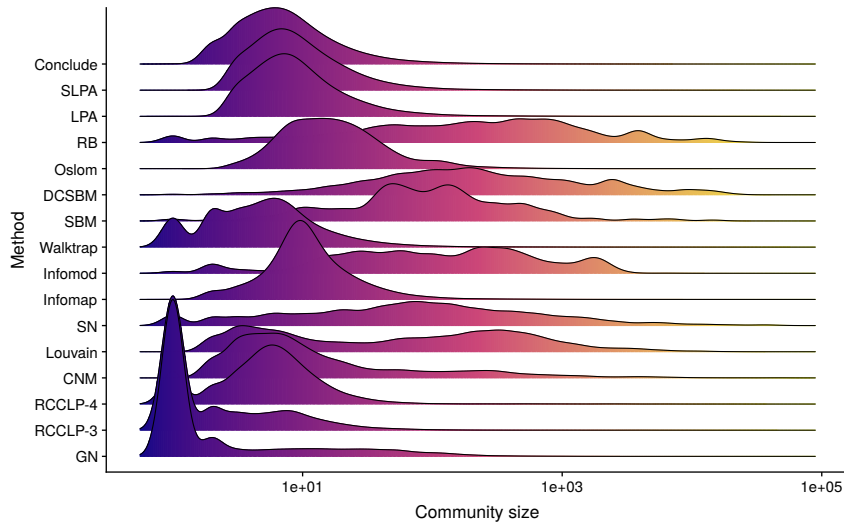


Fig. 15: The distributions of communities by sizes contained in the partitions detected on the networks of the dataset. They are smooth using a Gaussian kernel estimator. The illustrative gradient color is only for the ease of view purpose.

3.3.1 Experimental results

From the communities identified in the previous section, we proceed to measure the volumes of communities detected by each method to determine 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 experiments, only processes having a reasonable theoretical estimated time and memory consumption are maintained (less than a few days and at most 30 to 40 GBytes of memory). The outcome distributions are illustrated in Figure 15.

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 (Dao *et al.*, 2018a), they have a tendency to remove edges connecting hub and peripheral nodes and create singletons (single node community). This phenomenon is less distinguishable on *RCCLP-4* since there are much less quadrangular than triangular connections in networks.

Modularity optimization: Modularity is known to suffer from resolution limit phenomenon (Fortunato & Barthelemy, 2006), which often aggregates small communities in large scale networks. We can see from Figure 15 that *Louvain* and *SN* found very large communities as predicted. In the meanwhile, there are also a comparable number of small communities which are found on small graphs. However, the behavior is a little bit different on *CNM* method, which is an agglomerative clustering algorithm based on modularity optimization.

Dynamic process: Methods in this family show very discernible distributions although all based on dynamic processes. In fact, they make different assumptions about community structure and searching mechanisms. Therefore, belong to the same theoretical family 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 large networks. On the other hand, *Oslo* method use an agglomerative 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* method, 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 Figure 16. 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

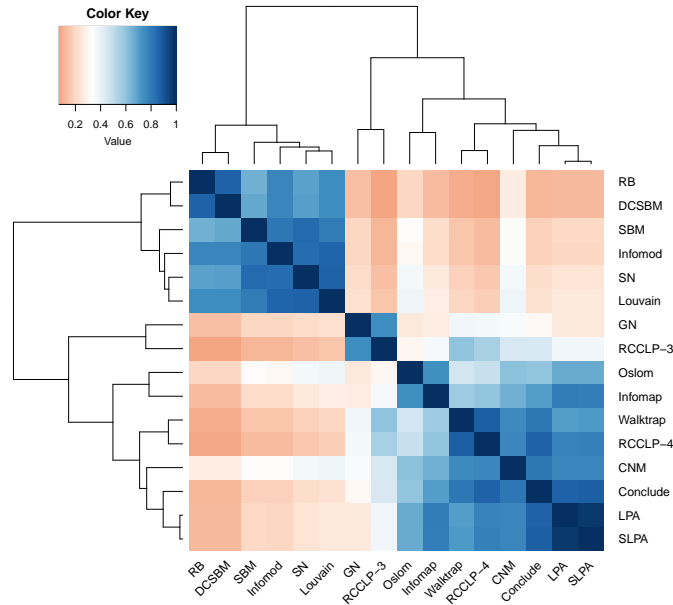


Fig. 16: The similarity between community detection methods in term of size fitting quality. Two methods are considered to be similar if they share a large fraction of same-size communities. Methods are ordered using hierarchical clustering (Joe H. Ward, 1963). The dendrogram proposes a hierarchical structure of the fitting closeness. Blue colors mean high similarity.

the distinction is very clear between groups. Therefore, we choose to characterize these methods by 3 (possibly 4) principle groups as follows:

1. **Group 1** - *RB, DCSBM, SBM, Infomod, SN, Louvain*: Methods in this group discover communities whose size vary in wide range of spectrum, from very small to very large communities. The characterized community size distribution is quite flat, meaning all sizes are nearly equally considered.
2. **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.
3. **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.

This characterization could help us to identify appropriate group of community detection methods according to different community size fitting strategies. Also, it helps to avoid brute-force tries when a method does not succeed to propose desired partitions by proposing substitute solutions. Moreover, by combining with the previous time computation analysis in Section 3.1, one could also choose a group of methods corresponding to size distinction criteria, and then select the fastest method that lead to a desired outcome.

The community distribution (or number of communities) is just one possible quality dimension, possibly one of the most intuitive and important information when choosing a clustering method. In the next part, we demonstrate some techniques that can be used to define other similarity aspects. We show that these notions of similarity can be combined to accentuate the distinction between different community detection methods.

4 Goodness profiling of community detection methods

4.1 Fitness functions

A popular way to evaluate the structure of communities is to design quality goodness metrics in order to measure different expected characteristics from subgraphs that we want to obtain. In practice, goodness metrics using network generative models are sometimes preferable as they reflect different assumptions about the underlying mechanisms that create community structure. One of the most widely used metric that quantifies the quality of community structure is the *modularity* function. The idea here is to reveal how the quality of an identified community structure is different from what would be expected. Although some unexpected phenomena known as resolution limit (Fortunato & Barthelemy, 2006), (Traag *et al.*, 2011) have been exposed when the scale of community size is too small, modularity remains to be the standard measure of quality.

The advantage of this approach is that one can “*embed*” the assumption of community structure inside quality functions, hence they provide better performance in some cases. However, community structure is quite an open question, such that according to different mechanisms that render the structure of networks, there will be models that are more suitable than others. Modeling networks hence contributes a great impact on the evaluation of network structure as well as community structure.

We present some quality metrics in this class to evaluate community structure. Many of them are initially or gradually employed as objective functions in some community detection methods since they expose good performance in searching processes. Firstly, we remind some notations that will be used to describe structural characteristic of communities. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of $n = |\mathcal{V}|$ vertices and $m = |\mathcal{E}|$ edges can be represented by an associated adjacency matrix A . Given a community C of n_C vertices being a subgraph of \mathcal{G} in an arbitrary partition P , a function $f(C)$ or $f(P)$ quantifies a structural goodness feature of community C or the whole partition P according to a particular expectation of community structure. Let m_C be the number of edges inside community C , $m_C = |(i, j) \in \mathcal{E} : i \in C, j \in C|$, l_C be the number of edges that connect C to other vertices outside of C , $l_C = |(i, j) \in \mathcal{E} : i \in C, j \notin C|$. Any vertex i belonging to community C has an *internal* degree k_{iC}^{int} and an *external* degree k_{iC}^{ext} satisfying $k_{iC}^{int} + k_{iC}^{ext} = k_i$, where k_i is the total degree of vertex i . The internal and external degree can be expressed via the adjacency matrix A as: $k_{iC}^{int} = \sum_{j \in C} A_{ij}$ and $k_{iC}^{ext} = \sum_{j \notin C} A_{ij}$. If vertex i in community C has $k_{iC}^{ext} > 0$ and $k_{iC}^{int} \geq 0$, i is called *boundary vertex* since i has neighbor(s) outside of C . Otherwise, if $k_{iC}^{ext} = 0$ and $k_{iC}^{int} > 0$, i is called *internal vertex* which only has connections with other vertex in the same community. In this paper, we employ the following functions to evaluate community structure:

4.1.1 Newman-Girvan Modularity

The standard version of modularity (Newman & Girvan, 2004) reflects the difference the fraction of intra community edges of a partition with the expected number of such edges if distributed according to a *null model*. In the standard version of modularity, the null model preserves the expected degree sequence of the graph under consideration. In other words, the modularity compares the real network structure with a corresponding one where nodes are connected without any preference about their neighbors. There are several ways to mathematically express the modularity, in order to compare the standard modularity with other variants, it is convenient to consider the modularity as a sum of contributions from pairs of vertices of the same community:

$$Q_{NG}(P) = \frac{1}{m} \sum_{c \in P} \left[m_c - \frac{(2m_c + l_c)^2}{4m} \right] \quad (5)$$

4.1.2 Erdős-Rényi Modularity

The Newman-Girvan modularity has attracted much attention in the research literature. Many alternative derivations have been proposed to adapt to different contexts. Some of them use different null models to quantify the modular structure of partitions. For example, one could assume that vertices in a network are connected randomly with a constant probability p as formulated in the Erdős-Rényi (ER) model (Erdős & Rényi, 1959). The connection probability is calculated as $p = \frac{2m}{n(n-1)}$ being the number of presented edges over the total number of edges that could be established. The expected number of edges in a community of size n_c becomes $\langle m_c \rangle = p \binom{n_c}{2}$. This null model leads us to the ER Modularity:

$$Q_{ER}(P) = \frac{1}{m} \sum_{c \in P} \left[m_c - \frac{mn_c(n_c - 1)}{n(n - 1)} \right] \quad (6)$$

4.1.3 Modularity Density (D-value or D-modularity)

The standard modularity is found to be impacted by resolution limits (Fortunato & Barthelemy, 2006), i.e. it is claimed that the sizes of detected modules depend on the size of the whole network such that optimizing standard modularity can not identify communities having a small number of vertices. The expected number of intra community edges is highly sensitive to the total number of edges in the whole network (Rosvall & Bergstrom, 2007) as can be observed in the second term of Equation (5). The modularity density (Li *et al.*, 2008) is one of several propositions that envisioned to palliate this issue. The idea of this metric is to include the information about community size into the expected density of community to avoid the negligence of small and dense communities. For each community C in partition P , it uses the *average modularity degree* calculated by $d(C) = d^{int}(C) - d^{ext}(C)$ where $d^{int}(C)$ and $d^{ext}(C)$ are the average internal and external degrees of C respectively to evaluate the fitness of C in its network. Finally, the modularity density can be calculated as follows:

$$Q_D(P) = \sum_{c \in P} \frac{1}{n_c} \left(\sum_{i \in c} k_{ic}^{int} - \sum_{i \in c} k_{ic}^{ext} \right) \quad (7)$$

4.1.4 Z-modularity

Z-modularity is another variant of the standard modularity proposed to avoid the resolution limit (Miyachi & Kawase, 2016). The concept of this version is based on an observation that the difference between the fraction of edges inside communities and the expected number of such edges in a null model should not be considered as the only contribution to the final quality of community structure. Specifically, the authors recommend that the statistical rareness of a community should be also taken into consideration. Such that an additive contribution amount of a community to the final modularity of a partition would be more important if its structure is less likely to be happen. Therefore, the variance of the probability distribution of the fraction of the number of edges within each community is included into the quality function throughout a standardization using Z-score. Following the null model of the standard modularity, the probability that an edge is placed inside community C is $p = \left(\frac{D_C}{2m}\right)^2$, where $D_C = 2m_C + l_C$ is the total degree of community C . The number of edges in each community follows a binomial distribution with the probability p and its normalized value approaches a normal distribution when the number of edges is sufficiently large. The statistical rarity of partition P in terms of the fraction of the number of intra-community edges using Z-score is hence translated into Z-modularity as follows:

$$Q_Z(P) = \left[\sum_{c \in P} \frac{m_c}{m} - \sum_{c \in P} \left(\frac{D_c}{2m}\right)^2 \right] \left[\sum_{c \in P} \left(\frac{D_c}{2m}\right)^2 \left(1 - \sum_{c \in P} \left(\frac{D_c}{2m}\right)^2\right) \right]^{-\frac{1}{2}} \quad (8)$$

4.1.5 Surprise

This statistical approach proposes a quality metric assuming that edges between vertices emerge randomly according to a hyper-geometric distribution (Aldecoa & Marín, 2011). Specifically, for a graph of n vertices and m edges, there are $M = \binom{n}{2}$ possible ways of drawing m edges. For a particular partition, there are $M^{int} = \sum_{C \in P} \binom{n_c}{2}$ possible ways of drawing an intra-community edge. Surprise metric computes the (minus logarithm of) probability of observing at least $m^{int} = \sum_{C \in P} \frac{k_C^{int}}{2}$ intra-community edges within m draws without replacement from the population of M possible choices in which consist precisely M^{int} possible intra-community edges. This probability is formalized as follows:

$$S(P) = -\log \sum_{k=m^{int}}^{\min(m, M^{int})} \frac{\binom{M^{int}}{k} \binom{M-M^{int}}{m-k}}{\binom{M}{m}}. \quad (9)$$

However, this formulation is not straightforward to work with in large-scale networks due to numerical computational problems. Hence, (Traag *et al.*, 2015) provides an asymptotic approximation for the metric which is a good alternative. By assuming that the relative number of intra-community edges $q = \frac{m^{int}}{m}$ and the relative number of expected intra-community edges $\langle q \rangle = \frac{M^{int}}{M}$ remain fixed, Surprise metric is approximated at:

$$S(P) \approx mD(q||\langle q \rangle), \quad (10)$$

where $D(q||\langle q \rangle)$ is the KullbackLeibler divergence (Kullback & Leibler, 1951):

$$D(q||\langle q \rangle) = p \log \frac{p}{\langle q \rangle} + (1 - q) \log \frac{1 - q}{1 - \langle q \rangle}. \quad (11)$$

According to the Surprise metric, the higher the score of a partition, the less likely it is resulted from a random realization, the better the quality of the community structure.

4.1.6 Significance

This metric use a similar approach to Surprise metric. It estimates how likely a partition of dense communities appear in a random graph (Traag *et al.*, 2015). However, Significance estimates the unlikeness of dense communities in a random graph in a different way. While Surprise uses global quantities q and $1 - \langle q \rangle$, Significance compares each community density $p_C = \frac{m_C}{\binom{n_C}{2}}$ to the average graph density $p = \frac{m}{M}$. The asymptotic form of Significance can be written as:

$$Z(P) = \sum_{C \in P} \binom{n_C}{2} D(p_C || p). \quad (12)$$

Similarly, $D(x||y)$ is the KullbackLeibler divergence defined in Equation (11). Generally, if the number of communities is relatively large or the graph is relatively dense, Significance is more discriminative than Surprise. On the other hand, in case that $\langle q \rangle > p$, Surprise can be better than Significance (Traag *et al.*, 2015).

4.2 Co-performance index

We devise a new comparative approach using a matrix called community detection *co-performance* matrix. The idea is that, given an expected quality function, one could investigate whether there exist a correlation in the efficiency of enhancing (or aggravating) its scores between different methods. The co-performance matrices reveal how understanding the performance of a method in optimizing a quality would allow us to predict the performance of other methods on the same quality. Therefore, an exhaustive analysis of co-performance matrices on many qualities allows to profile the characteristics of community detection methods in a comparative way. The index could be calculated as follows:

Let methods A and B divide a graph $G_i = (V_i, E_i)$ of dataset $\mathcal{G} = \{G_i | i = 1..N\}$ into α and β communities described by partitions $P_{G_i}^a = \{C_{1G_i}^a, C_{2G_i}^a, \dots, C_{\alpha G_i}^a\} \in \mathcal{P}_{G_i}$ and $P_{G_i}^b = \{C_{1G_i}^b, C_{2G_i}^b, \dots, C_{\beta G_i}^b\} \in \mathcal{P}_{G_i}$ respectively, we consider solely hard clustering methods, meaning $C_{uG_i}^a \cap C_{vG_i}^a = \emptyset : 1 \leq u < v \leq \alpha$ and $C_{uG_i}^b \cap C_{vG_i}^b = \emptyset : 1 \leq u < v \leq \beta$. A function $Q : \mathcal{P}_{G_i} \rightarrow \mathbb{R}$ quantifies a quality of a partition of graph G_i according to a particular goodness aspect (or model).

We define a *co-performance index* of two methods A and B on \mathcal{G} by their mutual capacity in discovering community structures showing a particular quality Q . In other words, each couple of methods should be assigned a high index according to a quality Q if knowing the

performance of one method reveals significantly the information about the performance of the other. A straightforward solution for defining the index is using Pearson correlation. Denoting $q_i^a = Q(P_{G_i}^a)$ and $q_i^b = Q(P_{G_i}^b)$, the co-performance index can be calculated as follows:

$$I_{\mathcal{G}}(A, B, Q) = \frac{N \sum q_i^a q_i^b - \sum q_i^a \sum q_i^b}{[N \sum (q_i^a)^2 - (\sum q_i^a)^2]^{1/2} [N \sum (q_i^b)^2 - (\sum q_i^b)^2]^{1/2}}, \quad (13)$$

where $0 \leq I_{\mathcal{G}}(A, B, Q) \leq 1$. A high positive (negative) score implies that two methods often find a strong consensus (disagreement) in discovering communities having a particular quality. In other words, given a co-performance index, knowing the quality scores of one method could provide predictive information about the outcomes of the other method on the same dataset. This information in fact could be very useful in a context where alternative solutions must be deployed while maintaining an assumed quality is expected. We present in the following part the mutual performance of the presented detection methods by the previously presented quality functions.

Figure 17 illustrates the co-performance matrices according to six different quality goodness criteria. Again, similarly to the previous section, goodness functions with a *close* concept are placed together. For instance, *NG* modularity and *ER* modularity are both based on null models whose concept use an expected fraction of intra-community edges. While the hypothesis of *NG* version is to keep the expected degree sequence of the graph in question, the *ER* version redistributes edges randomly with a constant average degree for every nodes. *D*-modularity and *Z*-modularity attempt to penalize large communities by including community sizes and significance level respectively. One can notice a very slight similarity in the experimental results of the co-performance indexes between different quality functions. Also, it seems that the assumption about the quality of community structure has an impact on the co-performance outcome.

As shown in Figure 17, there is a class of methods (*Louvain*, *GN*, *CNM*, *RB*, *Infomod*, *Infomap*, *Walktrap*, *Oslom*, *LPA*, *SLPA*, *Conclude*) in which all methods show very consistent results, except for the case of *D*-modularity⁷. Besides, there is also a strong relation between *SBM* and *DCSBM*. For the other methods (*RCCLP* and *SN*), no clear tendency could be observed from this experiment. The similarity of a large number of methods by many quality functions imply that, globally, if a method performs well on a given network, there is a signal that the others (from the same group) could also reach good results. In other words, if the community structure in a network is clear, most method will be able to detect it with more or less accuracy and inversely. However, as the co-performance indexes also vary significantly (0.2 to 0.3) inside each group, there will be always a remarkable difference if one goes from one method to another.

Within the case of density modularity shown in Figure 17(c), we discover that the sizes of detected communities have a great impact on the co-performance. Since density is a measure that penalizes heavily large size communities, especially in sparse networks, *D*-modularity gives very small values of giant communities and very high values for small

⁷ In fact, density modularity is somehow apart from other traditional ways to define the modularity, as it is not defined based on a null model but solely on edge density. The term *D*-modularity is abused in this sense

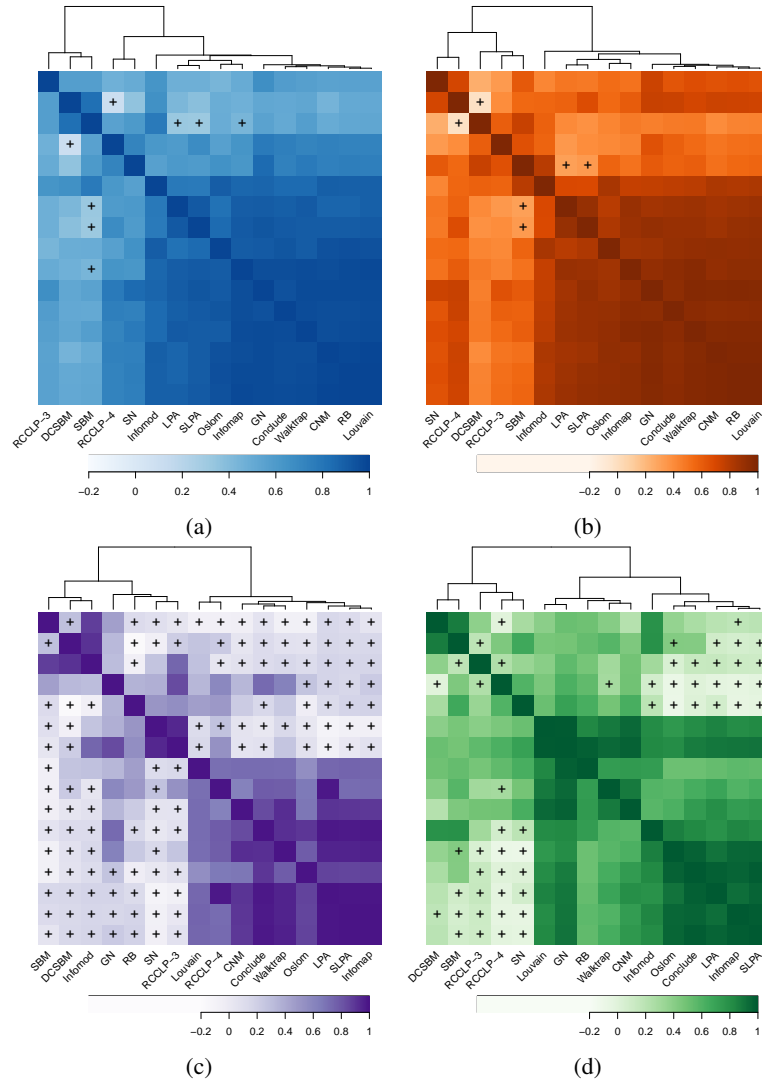


Fig. 17: The *co-performance* matrices of different methods. The "+" marks indicate cases where p-values are larger than 0.05. (a) Newman-Girvan modularity, (b) Erdős-Rényi modularity, (c) Density modularity and (d) Z modularity.

ones. Concretely, the methods *SBM*, *DCSBM*, *Infomod*, *RB*, *SN* discover very large communities (as shown in Section 3.3) and their co-performances in terms of *D*-modularity are very weak, showing that internal densities of communities detected by these methods are not linearly correlated. The reason is that the corresponding densities fluctuate unpredictably around zero. Similarly, *GN* and *RCCLP-3* found many tiny communities making the density either very high or zero (if internal degree is equal to external degree), consequently the co-performance index can not show significant information. On the other hand, we notice a consistency between the similarity of community size and the co-

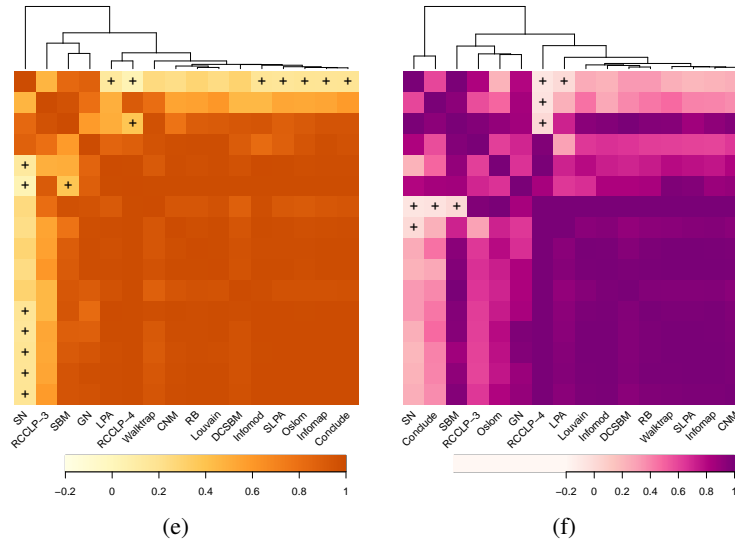


Fig. 17: The *co-performance* matrices of different methods (cont.) (E) Surprise and (F) Significance.

performance when methods identify medium size communities. Specifically, we find high co-performance indexes between *CNM*, *Conclude*, *Oslom*, *Walktrap*, *LPA*, *SLPA*, *Infomap* methods in most of the cases of the six quality fitness functions. This finding exposes a global agreement with our categorization determined by community size distributions.

The co-performance matrices also disclose interesting information about quality functions. As we can see in Figure 17(a,b,d), the matrices imply a similarity between *NG* modularity, *ER* modularity and *Z*-modularity in the assumptions of quality. In the same way, Surprise and Significance are quite close in practice as illustrated in 17(d,e). This experiment shows again another proof about the closeness between the theoretical assumption of community structure and the practical outcome. Moreover, although being based on different aspects of goodness, the performance of many methods tends to reach agreement on the modular structure of networks in general. This is to say, methods in the same group identify roughly and globally comparable results although there are always significance differences. In order to strengthen and validate our conclusion, we are interested in using other popular approaches in the literature to compare these community detection algorithms, which will be presented in the next section.

5 Partitioning strategy comparison

This section is dedicated to using conventional clustering validation metrics from the literature to verify the previous similarity analyses. We employ some popular metrics in the traditional clustering context (and also widely used in community detection context), which measure directly the likeliness of partitions using their corresponding contingency tables. These metrics do not take into consideration the structural information of commu-

nity structures, but only use the common numbers of nodes that are shared by pairs of communities in two partitions.

5.1 Validation metrics

The consensus of two partitions $P_1 = \{c_1^{(1)}, c_2^{(1)}, \dots, c_R^{(1)}\}$ and $P_2 = \{c_1^{(2)}, c_2^{(2)}, \dots, c_S^{(2)}\}$ can be more practically observed using a contingency table (sometimes called confusion matrix or association matrix) whose elements $n_{ij} = |c_i^{(1)} \cap c_j^{(2)}|$ corresponds to the number of common vertices between the i -th community of P_1 and the j -th community of P_2 as shown in Table 5.

Table 5: Contingency table of P_1 and P_2 on the same graph provides information about the similarity between the two partitions.

		Partition P_2				Σ
		$c_1^{(2)}$	$c_2^{(2)}$	\dots	$c_S^{(2)}$	
Partition P_1	$c_1^{(1)}$	n_{11}	n_{12}	\dots	n_{1S}	$n_{1\cdot}$
	$c_2^{(1)}$	n_{21}	n_{22}	\dots	n_{2S}	$n_{2\cdot}$
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	$c_R^{(1)}$	n_{R1}	n_{R2}	\dots	n_{RS}	$n_{R\cdot}$
	Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot S}$	n

In the evaluation of community structure using a validation metric, some following validation metrics are often used in the context of community detection to define the matching coefficient between two arbitrary partitions of a network:

5.1.1 Rand Index (RI)

The rand index is a pair-counting based measure, defined as the ratio of the number of vertex pairs correctly classified (either in the same community or in different communities) by the total number of pairs (Rand, 1971). The RI penalizes both false positive and false negative decisions of the clusterings. When the false positive need to be neglected, we can refer to the *Jaccard index* (Kuncheva & Hadjitodorov, 2004). The rand index value of two partitions can be calculated by the following:

$$RI(P_1, P_2) = \frac{\binom{n}{2} + 2\sum_i \sum_j \binom{n_{ij}}{2} - [\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}]}{\binom{n}{2}} \quad (14)$$

The value varies between 1 (meaning the two partitions are identical) and 0 (indicating that the two partitions do not agree on any pair of vertices). However, this value is only observed in the scenario when one partition consists in one community and the other consists in n community of 1 vertex, which has little practical value. Another shortcoming of the rand index is that its expected value for two randomly chosen partitions does not take a constant value which is normally expected for a good matching index (Vinh *et al.*, 2010). Therefore, a modified version of RI has been suggested, taking into consideration

the expected value of randomness (Hubert & Arabie, 1985), which is introduced in the following.

5.1.2 Adjusted Rand Index (ARI)

The corrected version of rand index takes the form:

$$Adjust_index = \frac{Index - Expected_Index}{Max_Index - Expected_Index} \quad (15)$$

It quickly becomes a replacement recommended for measuring agreement between two partitions in the analysis of clusterings. Its values ranges from -1 to 1 indicating completely different and identical partitions respectively. It is known to be less sensitive to the difference of the number of communities between two partitions. An ARI value of 0 indicates that the similarity is equal to the expected value from randomly chosen partitions. It can be calculated as:

$$ARI(P_1, P_2) = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}] - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}] / \binom{n}{2}} \quad (16)$$

5.1.3 Normalized Mutual Information (NMI)

Information theoretic based metrics constitute another approach for validating community structure with a given reference partition. Using the same notations as previously presented, the Mutual Information (MI) between two partitions quantifying the mutual dependence is calculated as:

$$I(P_1, P_2) = \sum_{ij} p(c_i^{(1)}, c_j^{(2)}) \log \frac{p(c_i^{(1)}, c_j^{(2)})}{p(c_i^{(1)})p(c_j^{(2)})} = \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij}n}{n_i n_j} \quad (17)$$

It measures how much knowing a repartition of vertices in one way would reduce the uncertainty about the other way. In order words, it could be considered as an indicator of information *closeness* expressing by the joint distribution between two variables. Therefore, the mutual information can be used as similarity measure between two partitions. However, it need to be normalized to reflect a consistency between different measures. The normalization is applied by using the entropy of each partition as:

$$H(P) = - \sum_k \frac{n_k}{n} \log \frac{n_k}{n} \quad (18)$$

Several variants of normalization can be considered, for instance taking the average, the root or the maximum of entropy of the two partitions as the denominator (Ana & Jain, 2003). In this document, we use the average version which is widely used in the context of community analysis (Danon *et al.*, 2005), (Chakraborty *et al.*, 2017). The closed form of NMI is hence defined from Equation (17) and (18) as follows:

$$NMI(P_1, P_2) = \frac{2I(P_1, P_2)}{H(P_1) + H(P_2)} = \frac{-2 \sum_{ij} n_{ij} \log \left(\frac{n_{ij}n}{n_i n_j} \right)}{\sum_i n_i \log \left(\frac{n_i}{n} \right) + \sum_j n_j \log \left(\frac{n_j}{n} \right)} \quad (19)$$

Likewise, the NMI similarity between two partitions varies between 0 corresponding to independent relation and 1 when two partitions are identical. The NMI does not follow triangle inequality.

5.1.4 Adjust Mutual Information (AMI)

Similarly to the Rand Index, the Mutual Information is also subject to the effect of randomness, i.e. there is not a constant baseline value between random partitions of a graph. This issue raises many difficulties in the comparison mechanism since it is expected that a comparative index should preserve the relativity between different clusterings and enhance intuitiveness about the mutual agreement. For that reason, the traditional Mutual Information is proposed to be normalized with a supplementary correction for chance and recently attracted attentions for comparing graph partitions. It is calculated as follows (Vinh *et al.*, 2010):

$$AMI(P_1, P_2) = \frac{I(P_1, P_2) - E\{I(M)|n_{i \cdot}, n_{\cdot j}\}}{\frac{1}{2}(H(P_1) + H(P_2)) - E\{I(M)|n_{i \cdot}, n_{\cdot j}\}}, \quad (20)$$

where $I(P_1, P_2)$ and $H(P)$ are introduced in equations (17) and (18) respectively. $E\{I(M)|n_{i \cdot}, n_{\cdot j}\}$ is the expected mutual information value of all feasible contingency tables constructed from the actual table M with the same marginals $n_{i \cdot}, n_{\cdot j}$.

5.1.5 Normalized Variation of Information (NVI)

Another popular metric that is often used in the context of comparing community partition similarity is the Variation of Information (VI) (Meilă, 2003), which is defined as:

$$VI(P_1, P_2) = H(P_1) + H(P_2) - 2I(P_1, P_2) \quad (21)$$

The VI metric can be interpreted as an index of shared information distance between two partitions. Its lower bound is 0 and is occurred when the two partitions are identical whether the upper bound $\log(n)$ happens when they are completely different. It is also preferable to use a normalized version with chance corrected to avoid the effect of randomness. Similarly to the construction of the Adjusted Mutual Information, with the same notation, the Normalized Variation of Information is calculated as follows:

$$NVI(P_1, P_2) = \frac{H(P_1) + H(P_2) - 2I(P_1, P_2)}{H(P_1) + H(P_2) - 2E\{I(M)|n_{i \cdot}, n_{\cdot j}\}}. \quad (22)$$

However, it turns out that NVI discloses the same information with AMI since from Equation (20) and (22), one has $NVI(P_1, P_2) = 1 - AMI(P_1, P_2)$. By consequent, calculating VI and NVI is unnecessary. We will be interested in using RI , ARI , NMI and AMI in our experiment. A summary of these validation metrics are shown in Table 6.

Validation metrics are often used in the context of community structure evaluation to measure the difference between the partition identified by a method with an expected partition of the network under consideration (*ground-truth*). The more similar the discovered partition to the ground-truth, the higher the performance of the method. However, in this

Table 6: Some popular validation metrics for comparing community partitions

Label	Range	Measure
RI	[0, 1]	Fraction of commonly grouped and separated vertices in two partitions.
ARI	[0, 1]	Rand index with a chance correction, less sensitive to differences of community sizes.
NMI	[0, 1]	Information theoretic approach, indicate how much information knowing one partition will help to guess the other.
AMI	[0, 1]	Similar with NMI, with a chance correction to set a constant baseline for random partitions.
VI	[0, log(n)]	Shared information distance measures the amount of mutual information. The higher the value, the less resembling the two partitions.
NVI	[0, 1]	Normalized version of shared information distance with chance correction.

section, validation metrics are exploited as a tool to compare community structures of different detection methods. They estimate the practical proximity of different algorithms through detected partitions, which constitutes a supplement source of information for evaluating their performance in a comparative approach.

5.2 Empirical results

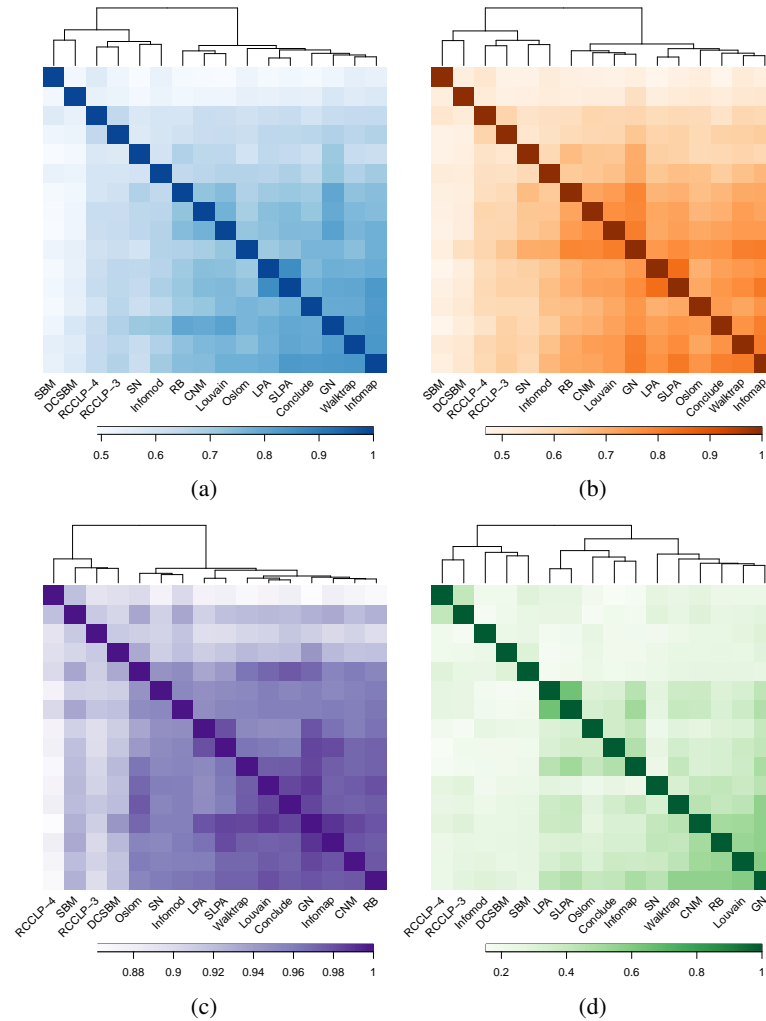
Once again, the experimental process is the same as those of the previous sections. From the partitions detected by the methods on the dataset, we calculate pairwise scores quantified by each validation metric on each network. Figure 18 illustrates pairwise average scores of the 4 metrics over the networks of the dataset⁸.

Again, by observing the dendrograms in Figure 18, one can see that all of the 4 metrics classify methods into two principle groups in a similar way that the co-performance matrices exposed in the previous section. The group of methods *CNM*, *Conclude*, *Osлом*, *Walktrap*, *LPA*, *SLPA*, *Infomap* mentioned in the last section also show very strong similarities in this experiment. Especially, *LPA* and *SLPA* being based on label propagation mechanism show nearly identical results in many cases. Besides, one could also discern another group including *RB*, *CNM*, *GN* and *Louvain* (modularity based), which show a high consistency in general. Additionally, even with weaker scores, *SBM* and *DCSBM* are often found in the same group as well as *RCCLP-3* and *RCCLP-4*. In a global view, it seems that methods with a close theoretical approach tend to provide more similar results, which is also noticed in the previous sections.

Another information that could be extracted from this experiment is that *RI* should not be used as validation metrics for evaluating detection performance. Since its average values vary generally in a small range (0.9 to 1.0), it is more difficult to see the different between partitions. On the other hand, *NMI* and *AMI* shows very close results in our experiment, which are between 0.5 and 1.0 meaning that structural communities detected by different

⁸ Where the corresponding methods are able to finish using a reasonable amount of time and memory as mentioned in the previous experiments.

Fig. 18: The similarity between community detection methods quantified by different validation metrics based on partitions discovered on networks of the dataset. Rows and columns are ordered according an hierarchical clustering method (Joe H. Ward, 1963). In the order, the average score of (a). NMI, (b). AMI, (c). RI, (d). ARI.



methods are quite comparable as concluded in the previous section. Finally, *ARI* seems to magnify the differences between methods, however there is no major difference in the similarity evaluation in comparison with the other metrics.

6 Related work

Orman *et al.* publish a comparative evaluation of eight community detection algorithms which most of them are also studied in this paper (Orman *et al.*, 2012). Different validation metrics are also used to compare detection performance and they also find that these

metrics (*RI*, *ARI*, *NMI*) “agree with each other with small differences when considering the way they rank algorithms”, as also illustrated in Section 5. Beside, the authors also focus on analyzing many topological aspects of community structure including also community size, transitivity, density, etc. These topological qualities are then used to inspect community structures detected by different algorithms. The analyses allow the authors to conclude that these two approaches (topological metrics and validation metrics) to evaluate community structures are “complementary and needed to perform a relevant and complete analysis of community detection results”. They also propose that the “traditional approach (*RI*, *ARI*, *NMI*) is much faster and easier to apply”, and hence is proposed to be used first. However, in practice, reference community structures (ground-truths) are not usually available⁹. Therefore, from these above notices, our analyses in this paper could be an important support dispensing additional information about the closeness between methods both in terms of topological aspect and partition-based aspect.

Agreste *et al.* evaluate different community detection algorithms in a empirical and comparative approach, especially for the context of web data analytic (Agreste *et al.*, 2017). The authors find that “time complexity is a crucial factor in the selection of a community detection algorithm” and recommend that the label propagation method (LPA) “has outstanding performance in scalability on both artificial and real graphs”, which is also in a global agreement with our analysis in Section 3.1 providing predictions about required time of each method in function of network size. They also conclude that “Infomap algorithm showcased the best trade-off between accuracy and computational performance” based on *NMI* score. The conclusion could be valid in some specific cases when the expected ground-truth community structure is well understood. Otherwise, some additional analyses should be done to determine whether *ground-truth* information aligns the final objective of community detection algorithms¹⁰ (Peel *et al.*, 2017).

Ghasemian *et al.* present in a recent publication that an evaluation of overfitting and underfitting of several community detection models (Ghasemian *et al.*, 2018). The authors study the number of communities detected in practice by many methods and the maximum number of detectable clusters according to a theoretical model. Some conclusions are drawn about fitting qualities of methods in comparison to theoretical estimates. This study provides evidences that help to choose an appropriate method in function of fitting quality. Community detection methods are also grouped in distinct families based on their outputs on many real-world networks (similarly to our analysis in Section 5.1) using *AMI* metric. The authors also find that “what an algorithm finds in a network depends strongly on the assumptions it makes about what to look for”, which is aligned with our results through several analyses.

⁹ In the context where a new algorithm is invented, one normally uses networks whose community structures are well known in order to validate the proposed method. In reality, since community detection is often employed to discover structures of new networks, hence it is not likely that reference community structures always exist.

¹⁰ In fact, metadata information of nodes are usually used in practice as ground-truth community structure. However, it has been found that metadata communities are sometimes very sparse (Dao *et al.*, 2017).

7 Conclusion

Finally, it is quite challenging to recommend which method is better in which scenario. It is at least as much demanding as defining all possible scenarios in the reality that could happen. Our experiments in this paper provide several experiments demonstrating different aspects of community structure quality, which can be combined together in a flexible way to assist network analysts to find appropriate methods according to their context. Some questions could be sequentially asked during decision making processes:

1. What is the size of the network of interest and what is the acceptable computation time for the task of community detection?
2. What are the expectations about the number of communities as well as the community size distribution?
3. Is there any fitness function that should be optimized?
4. In circumstances where the targeted method can not be deployed, is there any alternative solution?

The experiments and results in this paper could help to identify quickly suitable method(s) if one is able to response the previous questions.

The consideration of computation time is very crucial in the process of choosing a community detection method for a problem at hand. Even if theoretical estimate of time complexity is important and reveals the scalability of a community detection method, practical computation time is worth being studied in practice. Our estimates provide detailed information of practical time required by many popular community detection methods in function of network size. In particular the most scalable methods that we tested (*Louvain*, *LPA* and *SN*) reduce by approximately 10^4 times the required computation time compared with most of the other methods, which is not only significant but also crucial for a large network. Given a network size our results help in filtering non suitable methods.

In addition, the expected number of communities to be obtained is another important criterion in choosing a community detection method. According to the context, one would prefer different granularity levels. Our study shows that there are globally three main strategies that community detection methods decompose a network. Specifically, some identify communities whose size vary regularly in a wide range of values from very small to very large communities, some others divide networks into a huge number of very small communities and very few large communities, and finally the last ones distribute nodes into similar medium-size communities (around 10 members). Therefore, knowing how a network should be broken down is very useful in order to end up with an appropriate community discovery method.

In cases where (advanced) network analysts can determine a targeted objective function, designing new algorithms (or employing existing algorithms) that optimize the function would be the most evident. Since improving an objective function usually means expending more computation time, a compromise between getting higher fitness score and using less time needs to be considered. However, finding a good method to optimize an objective function satisfying a time constraint condition in the problem of community detection is not straightforward and needs many investigations. Our approach presented in the co-performance analysis provides network practitioners a quick glance about how different

methods perform in improving some widely-used quality functions. This predictive information about the effect of using alternative methods in achieving good fitness scores would suggest network analysts multiple solutions for a certain objective function. This scenario is specifically useful when the desired method is too expensive in terms of computation time. Therefore, a combination of our empirical analyses about scalability and/or community size distribution with the co-performance index can identify eligible alternatives for specific cases.

Finally, we find that using some validation metrics to estimate the similarity between community detection methods could also provide interesting information that could help the decision process of network analysts. In situations where one knows exactly, or have clues about, what should be found (ground-truth information), studying the way nodes are allocated to communities is important as it provides useful information about how a method is able to reach the desired clusters. However, this scenario is generally not popular in practice since community detection is often used to discover the structures of networks when no a priori information is available. In these cases and when performing several methods is possible, validation metrics are used to compare the results and identify different types of partitions. From our empirical study, we noticed significant differences in the way that nodes are distributed into communities. Especially, methods such as *SBM* or *RCCLP-4* seem to detect partitions which are very discernible from that of the others. Hence, the use of these methods needs to be examined, and we shall recommend to use them along with other methods as they might bring totally different and probably complementary insights in the data.

References

- Agreste, Santa, Meo, Pasquale De, Fiumara, Giacomo, Piccione, Giuseppe, Piccolo, Sebastiano, Rosaci, Domenico, Sarne, Giuseppe M. L., & Vasilakos, Athanasios V. (2017). An empirical comparison of algorithms to find communities in directed graphs and their application in web data analytics. *IEEE transactions on big data*, **3**(3), 289–306.
- Aldecoa, Rodrigo, & Marín, Ignacio. (2011). Deciphering network community structure by surprise. *PLoS ONE*, **6**(9), e24195.
- Ames, Brendan P. W. (2013). Guaranteed clustering and biclustering via semidefinite programming. *Mathematical programming*, **147**(1-2), 429–465.
- Ana, L.N.F., & Jain, A.K. (2003). Robust data clustering. *IEEE computer society conference on computer vision and pattern recognition. proceedings*.
- Arifin, S, Zulkardi, Putri, R I I, Hartono, Y, & Susanti, E. (2017). Developing ill-defined problem-solving for the context of south sumatera. *Journal of physics: Conference series*, **943**(dec), 012038.
- Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: Theory and experiment*, **2008**(10), P10008.
- Bohlin, Ludvig, Edler, Daniel, Lancichinetti, Andrea, & Rosvall, Martin. (2014). Community detection and visualization of networks with the map equation framework. *Pages 3–34 of: Measuring scholarly impact*. Springer International Publishing.
- Chakraborty, Tanmoy, Dalmia, Ayushi, Mukherjee, Animesh, & Ganguly, Niloy. (2017). Metrics for community analysis: A survey. *Acm comput. surv.*, **50**(4), 1–37.
- Clauset, Aaron, Newman, M. E. J., & Moore, Christopher. (2004). Finding community structure in very large networks. *Physical review e*, **70**(6).

- Cleveland, William S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the american statistical association*, **74**(368), 829–836.
- Coscia, Michele, Giannotti, Fosca, & Pedreschi, Dino. (2011). A classification for community discovery methods in complex networks. *Statistical analysis and data mining*, **4**(5), 512–546.
- Danon, Leon, Díaz-Guilera, Albert, Duch, Jordi, & Arenas, Alex. (2005). Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, **2005**(09), P09008–P09008.
- Dao, Vinh-Loc, Bothorel, Cécile, & Lenca, Philippe. (2017). Community structures evaluation in complex networks: A descriptive approach. *The 3rd International Winter School and Conference on Network Science, NetSciX*, 11–19.
- Dao, Vinh-Loc, Bothorel, Cécile, & Lenca, Philippe. (2018a). An empirical characterization of community structures in complex networks using a bivariate map of quality metrics. *Arxiv e-prints*, June.
- Dao, Vinh-Loc, Bothorel, Cécile, & Lenca, Philippe. (2018b). Estimating the similarity of community detection methods based on cluster size distribution. *Pages 183–194 of: The 7th international conference on complex networks and their applications*. Springer International Publishing.
- Erdős, P., & Rényi, A. (1959). On random graphs, I. *Publicationes mathematicae (debrecen)*, **6**, 290–297.
- Fortunato, S., & Barthelemy, M. (2006). Resolution limit in community detection. *Proceedings of the national academy of sciences*, **104**(1), 36–41.
- Fortunato, Santo. (2010). Community detection in graphs. *Physics reports*, **486**(3-5), 75–174.
- Fortunato, Santo, & Hric, Darko. (2016). Community detection in networks: A user guide. *Physics reports*, **659**(Nov.), 1–44.
- Ghasemian, A., Hosseinmardi, H., & Clauset, A. (2018). Evaluating overfit and underfit in models of network community structure. *Arxiv e-prints*, Feb.
- Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**(12), 7821–7826.
- Hric, Darko, Darst, Richard K., & Fortunato, Santo. (2014). Community detection in networks: Structural communities versus ground truth. *Physical review e*, **90**(6).
- Hubert, Lawrence, & Arabie, Phipps. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.
- Jerome, Kunegis. (2013). The koblenz network collection. *Pages 1343–1350 of: Proceedings conference on world wide web companion*.
- Joe H. Ward, Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the american statistical association*, **58**(301), 236–244.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**(1), 79–86.
- Kuncheva, L.I., & Hadjitodorov, S.T. (2004). Using diversity in cluster ensembles. *Ieee international conference on systems, man and cybernetics*.
- Lambiotte, R. (2010). Multi-scale modularity in complex networks. *8th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks*, May, 546–553.
- Lancichinetti, Andrea, Kivela, Mikko, Saramaki, Jari, & Fortunato, Santo. (2010). Characterizing the community structure of complex networks. *PLoS ONE*, **5**(8), e11976.
- Lancichinetti, Andrea, Radicchi, Filippo, Ramasco, José J., & Fortunato, Santo. (2011). Finding statistically significant communities in networks. *PLoS ONE*, **6**(4), e18961.
- Leskovec, Jure, & Krevl, Andrej. (2014). SNAP Datasets: Stanford large network dataset collection. June.

- Leskovec, Jure, Lang, Kevin J., Dasgupta, Anirban, & Mahoney, Michael W. (2008). Statistical properties of community structure in large social and information networks. *Proceeding of the 17th international conference on world wide web - WWW 08*.
- Li, Zhenping, Zhang, Shihua, Wang, Rui-Sheng, Zhang, Xiang-Sun, & Chen, Luonan. (2008). Quantitative function for community detection. *Physical review e*, **77**(3).
- Meilă, Marina. (2003). Comparing clusterings by the variation of information. *Learning theory and kernel machines*, 173–187.
- Meo, Pasquale De, Ferrara, Emilio, Fiumara, Giacomo, & Provetti, Alessandro. (2014). Mixing local and global information for community detection in large networks. *Journal of computer and system sciences*, **80**(1), 72–87.
- Miyauchi, Atsushi, & Kawase, Yasushi. (2016). Z-score-based modularity for community detection in networks. *PLOS ONE*, **11**(1), e0147805.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, **103**(23), 8577–8582.
- Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review e*, **69**(2).
- Orman, Günce Keziban, Labatut, Vincent, & Cherifi, Hocine. (2012). Comparative evaluation of community detection algorithms: a topological approach. *Journal of statistical mechanics: Theory and experiment*, **2012**(08), P08001.
- Papadopoulos, Symeon, Kompatsiaris, Yiannis, Vakali, Athena, & Spyridonos, Ploutarchos. (2011). Community detection in social media. *Data mining and knowledge discovery*, **24**(3), 515–554.
- Peel, Leto, Larremore, Daniel B., & Clauset, Aaron. (2017). The ground truth about metadata and community detection in networks. *Science advances*, **3**(5), e1602548.
- Pons, Pascal, & Latapy, Matthieu. (2005). Computing communities in large networks using random walks. *Pages 284–293 of: Yolum, pInar, Güngör, Tunga, Gürgen, Fikret, & Özturan, Can (eds), Computer and information sciences - iscis 2005*. Springer Berlin Heidelberg.
- Pons, Pascal, & Latapy, Matthieu. (2011). Post-processing hierarchical community structures: Quality improvements and multi-scale view. *Theoretical computer science*, **412**(8-10), 892–900.
- Porter, M. A., Onnela, J.-P., & Mucha, P. J. (2009). Communities in Networks. *Notices of the american mathematical society*, Feb.
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the national academy of sciences*, **101**(9), 2658–2663.
- Raghavan, Usha Nandini, Albert, Réka, & Kumara, Soundar. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review e*, **76**(3).
- Rand, William M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the american statistical association*, **66**(336), 846–850.
- Reichardt, Jörg, & Bornholdt, Stefan. (2006). Statistical mechanics of community detection. *Physical review e*, **74**(1).
- Riolo, Maria A., Cantwell, George T., Reinert, Gesine, & Newman, M. E. J. (2017). Efficient method for estimating the number of communities in a network. *Physical review e*, **96**(3).
- Rossi, Ryan A., & Ahmed, Nesreen K. (2015). The network data repository with interactive graph analytics and visualization. *Proceedings of the twenty-ninth aai conference on artificial intelligence*.
- Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the national academy of sciences*, **104**(18), 7327–7331.
- Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *European physical journal special topics*, **178**(Nov.), 13–23.

- Schaub, Michael T., Delvenne, Jean-Charles, Rosvall, Martin, & Lambiotte, Renaud. (2017). The many facets of community detection in complex networks. *Applied network science*, **2**(1).
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London New York: Chapman and Hall.
- Traag, V. A., Dooren, P. Van, & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical review e*, **84**(1).
- Traag, V. A., Aldecoa, R., & Delvenne, J.-C. (2015). Detecting communities using asymptotical surprise. *Physical review e*, **92**(2).
- Vinh, Nguyen Xuan, Epps, Julien, & Bailey, James. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. mach. learn. res.*, **11**(Dec.), 2837–2854.
- Wasserman, Stanley. (1994). *Social network analysis : methods and applications*. Cambridge New York: Cambridge University Press.
- Xie, Jierui, & Szymanski, Boleslaw K. (2012). Towards linear time overlapping community detection in social networks. *Pages 25–36 of: Advances in knowledge discovery and data mining*. Springer Berlin Heidelberg.
- Yang, Jaewon, & Leskovec, Jure. (2013). Defining and evaluating network communities based on ground-truth. *Knowledge and information systems*, **42**(1), 181–213.