Applied Network Science

A comparative study of overlapping community detection methods from the perspective of the structural properties

Vinícius da Fonseca Vieira[1]* , Carolina Ribeiro Xavier[1] and Alexandre Gonçalves Evsukoff[2]

*Correspondence:
vinicius@ufsj.edu.br
[1]Federal University of São João
del-Rei, São João del-Rei MG, Brasil
Full list of author information is
available at the end of the article

Abstract

Community detection is one of the most important tasks in network analysis. It is increasingly clear that quality measures are not sufficient for assessing communities and structural properties play a key hole in understanding how nodes are organized in the network. This work presents a comparative study of some representative state-of-the-art methods for overlapping community detection from the perspective of the structural properties of the communities identified by them. Experiments with synthetic and real-world benchmark Ground-Truth networks show that, although the methods are able to identify modular communities, they often miss many structural properties of the communities, such as the number of nodes in the overlapping region and the memberships of the nodes. This is a strong suggestion that a deeper comprehension of the overlapping properties of the communities is needed for the design of more efficient community detection methods.

Keywords: Overlapping community detection, Structural properties, Comparison of methods

Introduction

The organization of nodes in communities, i.e., groups of nodes with several internal connections and a few external connections, is one of the most prominent properties in complex networks. Communities have been studied in many fields, like social sciences, biology and economics, since it can reveal key functional roles of the elements in the complex system modeled (Newman and Girvan 2004a; Fortunato 2010; Palla et al. 2005; Ravasz et al. 2002).

A great number of works investigating communities in networks assume the division of nodes as a partition problem, i.e., a node must belong to one and only one community, thus, the most traditional methods for identifying communities are suitable for situations where a node cannot be part of multiple groups simultaneously. Although the investigation of the disjoint community structure is fundamental for a better comprehension of the networks, it is intuitive that, in many situations, elements can participate concomitantly in several groups. For instance, in a social context, a particular person is deeply connected

to her family, whilst also related to her coworkers and members of her sports club. From the perspective of network science, the organization of nodes in different groups should be considered by community detection methods and, in this sense, many authors have been dedicating efforts in characterizing communities with overlapping nodes in addition to developing methods that are able to automatically identify them, despite it is still an open task (Peel et al. 2017; Peixoto 2020; Jebabli et al. 2018; Harenberg et al. 2014).

The characterization and identification of community structure are intrinsically linked. Any method for community detection must precisely assume how a group of nodes is organized as a community and, even though intuitive and clear general concepts can be adopted, there is no consensual definition in this sense (Peixoto 2020; Fortunato 2010; Cherifi et al. 2019; Xie et al. 2011). More specifically, when the overlaps of nodes are considered, the problem becomes even more challenging since, besides the internal connection of the nodes in groups, assumptions on the way that the nodes are organized in overlapping regions must be made (Fortunato 2010; Harenberg et al. 2014). One point that is now clear is that there is algorithm that can generalize the problem of community detection in networks and obtain a canonical solution (Schaub et al. 2017; Peel et al. 2017; Leskovec et al. 2010). Moreover, the definition of what is a community is still far from a consensus as more questions are raised about what properties of complex systems a group of nodes should represent. Despite the difficulty in characterizing the limits of density for the definition of what is a community, some recent works show that structural properties of networks may show very low correlation with the metadata of the nodes, i.e, the functional communities presented in the network (Peel et al. 2017; Yang and Leskovec 2014; Harenberg et al. 2014; Peixoto 2020; Jebabli et al. 2018; Kudelka et al. 2019). Thus, when a particular criterion is defined to be optimized as an objective function by a community detection method, it can be considered as an implicit assumption about the characterization of an overlapping community structure (Schaub et al. 2017; Peel et al. 2017; Jebabli et al. 2018).

Particularly for the characterization of overlapping community structure, many definitions of communities can be found in the literature. Lancichinetti et al. (2009) assume that a community is a local structure that maximizes a fitness function controlled by a flexibility parameter. Rhouma and Romdhane (2014) propose the control of the embeddedness of communities by a fuzzy coefficient. Some authors propose variations of Newman's modularity (Newman 2006), which considers a group of nodes as a community if the fraction of internal edges overcomes the expected number of edges in a random version of the network with the same degree distribution, in order to adapt it to overlapping communities. This is the case of Nicosia et al. (2009) and Shen (2013). The work of Palla et al. (2005) also considers the identification of dense regions of the network in order to define communities, but, in a different perspective, the authors assume that a community comprises cliques of size k, which is a parameter of the method, connected by cliques of size $(k-1)$.

In order to classify and understand overlapping community methods, some works found in the literature present surveys and comparative studies and can be used as good references for studies in the area, such as the works of Orman et al. (2012), Xie et al. (2011), Hric et al. (2014), Amelio and Pizzuti (2014), Dao et al. (2020), Harenberg et al. (2014) and Jebabli et al. (2018). The present work also performs a comparative analysis of some of representative state-of-the-art methods for overlapping community detection in complex networks, however, the focus is not primarily on quality measures and hit rates

of the methods as performed in some works (Xie et al. 2011; Hric et al. 2014; Dao et al. 2020).

In a different perspective, the experiments conducted in this work investigate the structural properties of the overlapping communities identified by the studied methods and the similarity of patterns in the topology of the extracted communities with Ground-Truth communities. The investigation of some structural properties of communities is performed in some works in the literature (Clauset et al. 2004; Ravasz et al. 2002; Palla et al. 2005; Leskovec et al. 2010), but more recently it has become clearer the importance of a deep investigation in the topological properties of the community structure identified by community detection (Peel et al. 2017; Jebabli et al. 2018; Orman et al. 2012; Kudelka et al. 2019). A similar approach was previously followed by Hric et al. (2014), but only considering the disjoint community structure. The authors test a basic hypothesis for community detection methods: does the topological organization of the networks solely are able to reveal the communities of a network? The authors conducted experiments to assess if community detection methods are able to correctly classify the Ground-Truth communities (which they call metadata networks), disconsidering objective measures and verify that, on most cases, there is a substantial difference between identified and metadata groups. Harenberg et al. (2014) also investigate measures on the structure and the metadata of networks and verify that there is no clear relation between them. More recently, Peel et al. (2017) have proven that there are theoretical and practical problems in trying to uncover metadata from network structure, pointing out and discussing some reasons for that.

The present work is an invited extension of Vieira et al. (2019), published in the Proceedings of the Conference on Complex Networks 2019 and its main objective is to promote a wide comparison of some of the most successful methods for overlapping community detection found in the literature, particularly in works focusing the comparison of methods (Xie et al. 2013; Hric et al. 2014; Amelio and Pizzuti 2014; Jebabli et al. 2018), whose computational implementations are freely available in the web: Bigclam (Yang and Leskovec 2013), CFinder (Palla et al. 2005), Copra (Gregory 2010), Demon (Coscia et al. 2014) and SLPA (Xie et al. 2011).

This work prioritizes the use of methods related in the literature as good methods for overlapping community detection (which is the case of SLPA and COPRA). It also considers Demon and Bigclam due to their good results reported by Coscia et al. (2014) and their ability in dealing with large scale networks, which is very appropriate in real-world scenarios. CFinder was also considered, as it is one of the first methods proposed in the literature for overlapping community detection, considered in almost all works regarding this subject. It is worth to mention that after some preliminary experiments on the state-of-the-art methods for overlapping community detection, we have selected a subset of five of them that presented the lowest execution time and some state-of-the-art methods were not considered in this work.

The study regards objective measures and relate these results with some structural properties of the communities identified by the methods. Two main research questions are raised and will be investigated in this work: I) Is there an overlapping community detection method that stands out regarding from the others and can be adopted as a state-of-the-art method? II) Does the quality measures frequently considered in the literature

for the validation of overlapping community detection method are really appropriate and consistent with the real community structure?

The main objective of this extension is the same as the original work, however with more experiments to support it, considering a wider set of metrics and networks of different types. First, the overlapping community detection methods are compared regarding objective functions, namely: Overlapping Normalized Mutual Information (ONMI) (McDaid et al. 2011) based on Lancichinetti et al. (2009), Omega (Collins and Dent 1988), Modularity (Shen 2013) and F1-score. Then, we investigate if the obtained results are a consequence of the community structure or due to the sequence of steps of the community detection methods. This dependency extends the findings of Peel et al. (2017), going against a basic and consensual notion that the community structure must be encoded by the connections of a network and must be identified from them, and suggesting that the community structure may be a consequence of the working mechanisms of the methods. Thus, in many cases, the resulting community structure may lead to good results regarding objective measures, but this result may be biased by the method, since similar assumptions about what is a modular community are made by the method and the metrics. On the other hand, structural properties of the communities identified by the methods may differ a lot from the organization of the elements in real-world contexts. From the perspective of networks, the actual organization of nodes in communities will be treated as the Ground-Truth of the networks in this work, even aware of the limitations of this strategy, as pointed by Peel et al. (2017). The following structural properties of the communities are investigated: number of memberships of the nodes, sizes of the communities, number of nodes in the overlapping region and probability of edges in the overlapping region.

The experiments are conducted on synthetic and real-world networks. The set of synthetic networks are generated by the LFR benchmark (Lancichinetti et al. 2008), frequently considered for the evaluation of overlapping community detection algorithms. The set of real-world networks correspond to benchmark networks found in the literature, some of them with the metadata annotation of the association of the nodes to the communities, treated as Ground-Truth in this work, and some of them without this information. The investigation performed in this work allow us to verify that the methods are able to extract highly modular communities from small and medium sized networks when objective measures are considered, confirming the results obtained by other authors in the literature (Nicosia et al. 2009; Gregory 2010; Xie et al. 2013). It is also observed that most state-of-the-art methods are unable to identify the correct groups for nodes in networks with Ground-Truth. Moreover, the identified communities are often very distinct from those observed for the Ground-Truth networks when structural properties are considered. The investigation on synthetic and real-world networks, with and without Ground-Truth information also suggest that the structural properties of the communities obtained by the community detection methods are more related to the method itself and its way of operation than the topological organization of the networks.

Overlapping community detection
Problem statement
The division of a network in communities is traditionally defined as a partition, where each node belongs to a single community, or alternatively as a cover, where a node can

participate in different groups at the same time, which is the approach explored in the context of the present work.

The problem of overlapping community detection can be stated as follows. Consider a network $G(V, E)$ where V represents the set of nodes and E represents the set of edges, such that $n = |V|$ and $m = |E|$. In many contexts, it is appropriate to represent the weights and the directions of the links between pairs of nodes, however, this adds a level of complexity that can make the problem of community detection even harder to solve (Fortunato 2010). Although some methods can be applied to directed and weighted networks, such as SLPA (Xie et al. 2013), OSLOM (Lancichinetti et al. 2011) and Demon (Coscia et al. 2014), it is usual to omit the weights and the directions of the links when comparing community structure methods (Fortunato 2010). This strategy is widely adopted in many works that present community detection methods designed to unweighted and undirected networks (Yang and Leskovec 2013; Palla et al. 2005; Gregory 2010). However, the weights and directions are also disregarded in works that propose methods for community detection in directed and weighted networks, such as Demon (Coscia et al. 2014) and SLPA (Xie et al. 2013) and works that focus on a comparison of methods, as the work of (Hric et al. 2014), in order perform the comparison in a broad class of methods. In this work, the edges are considered in the simplest form, undirected and unweighted, allowing the investigation to be conducted in a wider set of methods and networks.

G can be represented by an adjacency matrix A, where an element $A_{uv} = 1$, if a node u is connected to a node v and $A_{uv} = 0$, otherwise, and $A_{uv} = A_{vu}$. A community structure in a network can be identified when there is a division of the network in groups with high density of internal edges and a low density of external edges. An overlapping community structure C can be defined as a cover of V in n_c communities $C = \{C_i, i = 1...n_c\}$ and a node u can participate in one or more community C_i with a belonging factor α_{uC_i} of node u in community C_i such that $0 \leq \alpha_{uC_i} \leq 1$, $\forall u \in V$, $\forall C_i \in C$ and $\sum_i^{n_c} \alpha_{uC_i} = 1$, $\forall u \in V$.

As well as in the partition problem, the sense of community becomes more evident as the difference between intra and inter edges increases and there is not an universal definition for a particular division of the network in communities, overlapped or not. The next section discusses some of the definitions of community structure explored in the literature and methods to identify them.

Characterization and identification of overlapping community structure

The organization of nodes in communities is traditionally seen as a partition of the network, each node belongs to only one community. The identification of network structure considering the definition of communities as a partition problem is very important to reveal functional roles of nodes and provides a clear picture of the network and many works in the literature present reviews and surveys of disjoint community detection (Fortunato 2010; Radicchi et al. 2004; Schaeffer 2007). However, it is very intuitive that, in practical contexts, communities may overlap and the same node can participate in several groups.

When a method for community detection is proposed, the authors assume, implicitly or explicitly, how in which extent a group of nodes can be organized as a community and, even though intuitive and clear general concepts can be adopted, there is no consensual definition in this sense (Fortunato 2010; Radicchi et al. 2004; Schaeffer 2007). When the

community detection problem is considered from the perspective of overlapping nodes, the problem becomes even more challenging since, besides the internal connection of the nodes in groups, assumptions on the way that the nodes are organized in overlapping regions must be made (Xie et al. 2011; Hric et al. 2014). Yang and Leskovec (2012) find that, in real-world networks, overlapping community structure shows highly dense overlapping regions and observe that this is the opposite to the traditional view of community structure. The authors point that, traditionally, communities emerge from the "strength of weak ties" theory (Granovetter 1973), meaning that communities are dense clusters connected by a small number of weak links. This is the basic assumption made by the min-cut partition problem (Schaeffer 2007) and followed by the traditional method for community detection proposed by Newman and Girvan (2004a), which works by removing bridges between dense groups. However, Yang and Leskovec argue that by making this assumption, a random pair of nodes in an overlapping region tends to be disconnected, since the overlapping region may be considered as a bridge between two communities. Moreover, the consequence of this traditional perspective is that, the more communities a pair of nodes share, the less likely is for them to be connected, which is unnatural and unintuitive.

A large number of methods can be found in the literature aiming on the identification of overlapping community structure and it is difficult to separate the methods for community detection from the assumptions made by the authors about how a community should be defined. Thus, when a particular criterion is defined to be optimized as an objective function by a community detection method, it can be considered as an implicit assumption about the characterization of an overlapping community structure.

Lancichinetti et al. (2009) define a local fitness function for communities based on the density of nodes inside the groups using a parameter that controls the resolution of the solution. The authors also propose an heuristic to optimize this quantity based on the analysis of an histogram of the fitness function.

The work of Rhouma and Romdhane (2014) proposes an agglomerative method for identifying overlapping nodes (DOCNet) which works by the defining of a core of a community and expanding it with fuzzy belonging coefficients that control the embeddedness of the communities.

Newman and Girvan (2004b) is one of the most famous measures for the assessment of community partition and considers a group of nodes as a community if the fraction of internal edges overcomes the expected number of edges in a random version of the network with the same degree distribution. Although it has been originally defined for the partition problem, it has been adapted by some authors in order to be appropriate for overlapping communities. This is the case of the works of Nicosia et al. (2009) and Shen (2013) which also define algorithms to optimize their extended versions of modularity.

Palla et al. (2005) assume an explicit definition of community structure in order propose a method for overlapping community detection based on the identification of dense regions of the network, called CFinder. The authors relax the strongest definition of a community — the clique — and assume that a community is a group of k-cliques connected by $(k-1)-$cliques, considering the overlapping regions as edges of a network in which the communities are the nodes.

Yang and Leskovec (2012) start from the observation that the overlapping regions are heavily connected and make more explicit assumptions about the structural properties of overlapping communities. Then, the authors consider these assumptions to define a model for overlapping community structure that simulates dense overlapping regions. Then the authors propose a method for maximizing the likelihood of networks to the model (Yang and Leskovec 2012) and a method to fit the community structure of an empirically observed network to the model following a Non-negative Matrix Factorization approach is proposed (Bigclam) (Yang and Leskovec 2013).

Other works are based on the Label Propagation method (RAK) by Raghavan et al. (2007) and consider a set of nodes as a community if they share the same label after a propagation process, which can be defined considering different updating rules. Copra, proposed by Gregory (2010) extends RAK Label Propagation by allowing the nodes to propagate not only one label, but a set of labels with a value above a threshold v, which controls the potential overlapping degree between the communities. Coscia et al. (2014) propose a local-first approach (Demon) for the RAK Label Propagation that allows the nodes to democratically vote in the communities surrounding them. Demon extracts the ego network, i.e., the network built around an ego node, for each node of the network and applies the RAK Label Propagation method, ignoring the ego node, which is, then, evaluated by its neighbors. SLPA, proposed by Xie et al. (2013) and also known in the literature as Ganxis, adapts the propagation rule of RAK Label propagation in order to restrict the updates only to nodes in the borders of the communities. According to the authors, internal vertices do not need to be updated and this adaptation significantly reduces the execution time of the method without compromising the quality of the results.

A game theory perspective is considered by Zhou et al. (2015) for the problem of overlapping communities in a work in which the authors propose a function to assess the quality of the union of hierarchical groups and a greedy algorithm to identify overlapping communities. In addition to the relationship of the elements, Xin et al. (2015) also incorporate text analysis in order to better identify semantic overlapping communities in social network contents. Li et al. (2018) also incorporate more information than just the relationship of the nodes but, instead of the context of the elements, the authors consider a priori information about which nodes must be in a certain community and which nodes must not be in the same community.

Evaluation of overlapping community structure

It is a common sense in network since that a community is a group of nodes with many internal edges and few external edges (Fortunato and Barthélemy 2007; Fortunato 2010; Schaeffer 2007). However, community detection is still an open problem mainly due to the difficulty in finding an universal definition of the limits of density that should be considered to characterize a community (Fortunato and Barthélemy 2007; Gregory 2010). When the perspective of overlapping communities is considered, community detection is an even more challenging problem since, beyond the characterization of the core of the groups, the structure of the overlapping region must also be assumed (Yang and Leskovec 2012).

As discussed in "Characterization and identification of overlapping community structure" section, many methods can be found in the literature for overlapping community

detection and each method assumes the definition of communities that the authors consider more appropriate (Schaub et al. 2017; Cherifi et al. 2019). Moreover, when a method for community detection is proposed, the authors frequently compare it to other methods and evaluate its performance considering some metrics and measure, which can be directly related to the method or not (Yang and Leskovec 2013; Nicosia et al. 2009; Xie et al. 2013).

Many metrics have been proposed in the literature in order to evaluate the quality of community detection methods by assessing the community structure, overlapping or non-overlapping, identified by them (Dao et al. 2020). One of the most accepted measures for the evaluation of disjoint community structure in networks is Newman's modularity (Newman 2006). Nicosia et al. (2009) propose a more general definition of modularity in order to deal with overlapping community structures in directed networks. The authors consider a term $\beta_{(u,v),C_i}$ that can be defined as a function to weight the belonging of the nodes u and v to community C_i and define the overlapping modularity as

$$\hat{Q}^{Nicosia} = \frac{1}{m} \sum_{C_i \in C} \sum_{u,v \in V} \left[\beta_{(u,v),C_i} A_{u,v} - \frac{\beta_{(u,v),C_i}^{out} k_u^{out} \beta_{(u,v),C_i}^{in} k_v^{in}}{m} \right], \tag{1}$$

where k_u^{in} and k_u^{out} are, respectively, the in-degree and the out-degree of a node u. The belonging coefficient β makes the modularity definition as general as possible, since it can be defined from any appropriate function of that relates the participation of u and v in C_i, such as a multiplication. Thus, the definition of an aggregation function for the coefficients β is a key aspect of the extended modularity proposed by Nicosia et al.. Nicosia et al. (2009) use their extended modularity in order to evaluate a community detection method proposed by them and, as stated by the authors, after testing several different functions, the best results were obtained with a logistic function with a linear scaling. The same function proposed by Nicosia et al. for the definition of β is also considered by Gregory (2010).

A simpler extension of the modularity for overlapping community structure is proposed by Shen (2013). It relaxes the fact that a node can belong to only one community and defines a more flexible factor α_{u,C_i}, that denotes how much a node u belongs to the community C_i.

$$\hat{Q}^{Shen} = \frac{1}{2m} \sum_{C_i \in C} \sum_{uv} \left(A_{uv} - \frac{k_u k_v}{2m} \right) \alpha_{u,C_i} \alpha_{v,C_i}. \tag{2}$$

The factor α_{u,C_i} can be simply defined as the weight of the contribution of u to C_i, which is equal to the weights of contributions of u in any community that it participates.

Lancichinetti et al. (2009) propose a variation of the Normalized Mutual Information (NMI) (Danon et al. 2005), a measure that assesses the similarity between two partitions based on the information mutually contained on them, in order to adapt it to overlapping community structures. Given two sets of overlapping communities, McDaid et al. (2011) identified that the variation of the NMI proposed by Lancichinetti et al. overestimates the similarity of communities in some particular situations when all the communities of one cover have a perfect match in the other cover, regardless of the number of its communities. Thus, McDaid et al. propose a different normalization of the NMI, which they call Overlapping Normalized Mutual Information (ONMI), considered in this work. Overlapping adaptations of the NMI are frequently used in many works in the literature regarding

overlapping communities, such as Gregory (2010), Yang and Leskovec (2013) and Coscia et al. (2014). In this work, we consider the Overlapping Normalized Mutual Information proposed and implemented by McDaid et al., based on the definition of Lancichinetti et al. (2009).

Another quality measure for overlapping community structure frequently considered in the literature is Omega index (Collins and Dent 1988), an adaptation of the Rand index, defined for partition problem, which assesses, for each pair of nodes, the ratio of communities both are correctly placed simultaneously.

Considering the evaluation of the communities identified by the methods, a very common measure considered is F1-score, that can be interpreted as a weighted average of the precision and recall. As traditionally used for more general data classification, the precision can be viewed as the ability of a method in correctly classify some items and recall can be interpreted as the ability of the method in correctly identify nodes regarding a class of interest. Many definitions of what is a class of interest can be adopted, from the Ground-Truth communities to the nodes in the overlapping region.

Two significant difficulties when assessing a method for community detection are the reduced transparency in the results obtained by the authors when publishing their methods and the lack of a test bed for the methods. For instance, in Yang and Leskovec (2013), the quality of the methods are not presented as absolute values, but relatively to other competing algorithms and in Palla et al. (2005), only some structural properties of the communities are presented for a small set of networks. In Gregory (2011), the method is mostly evaluated with different parameterizations in the experiments and in Coscia et al. (2014), the synthetic networks are much less mixed than other synthetic networks explored in the literature.

Moreover, each evaluation measure captures an aspect of the community structure but inevitably misses many other, limiting the understanding of the methods and posing serious risks to their validation. For this reason, when evaluating an overlapping community detection method, it is very important to consider a wide set of networks, considering different sources and features. Many works consider real-world networks, even accepting that it is difficult to control and understand their properties. In this sense, the use of synthetic networks is a useful strategy, since it allows a more controlled environment. The LFR synthetic benchmark (Lancichinetti et al. 2008) have been widely used in the literature, since it generates networks with overlapping communities, allowing the definition of a set of parameters of the topology of the generated networks. In the end, it is important to be aware that many definitions of community structure can be considered and any benchmark network only cover some of them. For instance, it is known by now that real-world overlapping communities exhibit the property of dense overlapping regions (Yang and Leskovec 2012), which is not observed in the networks generated by the LFR benchmark. In this sense, Chykhradze et al. (2014) proposed a benchmark capable of generating networks with overlapping community structure that consider this property.

The evaluation of overlapping community structure can be considered a even more challenging task if we consider some fundamental aspects of the community detection problem. First, Yang and Leskovec (2012) point that it is difficult to obtain networks with Ground-Truth community structure and, even when they are possible to obtain, comprise a large number of nodes and edges, being hard to evaluate, due to the lack of scalable algorithms. Moreover, there is not a best and universal metric that is able to properly evaluate

community structure, since each one assume a different view of what a community should look like (Almeida et al. 2011; Jebabli et al. 2018; Harenberg et al. 2014). The work of Yang and Leskovec (2014) discusses that the actual functional communities may be different from the structural communities and Harenberg et al. (2014) empirically identifies that a clear relation between the quality of structural and functional communities can not be observed. In a more recent work, Peel et al. (2017) show that the metadata of the elements in a networked system are not the same as what is frequently assumed as the Ground-Truth of the communities and no algorithm can solve an general community detection problem.

Despite the many issues in the definition of communities and methods to identify them, community detection remains a very valuable tool to reveal important information about the organization of elements in complex systems (Peel et al. 2017). In this sense, many works have being emphasizing the importance of evaluating the structural properties of the communities, in addition of objective assessment of the division of nodes, in order to better evaluate the community detection methods and understand the organization of nodes (Leskovec et al. 2010; Harenberg et al. 2014; Jebabli et al. 2018; Yang and Leskovec 2014; Orman et al. 2012). This work conducts experiments, presented in the next section, to evaluate the properties of the community structure identified by community detection methods and enhance the understanding of the division of nodes in overlapping groups.

Experiments and discussion

In this work, we perform a broad comparison of some of the most successful methods for overlapping community detection reported in the literature. This section describes the results obtained by a set of experiments on synthetic and real-world networks in order to assess the quality of the methods considering some objective quality metrics frequently used to validate their performance and relate the values obtained with some structural properties observed for the communities identified. The main idea is to investigate whether the quality of the methods, measured from an objective perspective, can be somehow explained by the topology of the resulting networks and the operating mode of the methods.

Experimental setup

All the experiments were performed in the same computational environment: an Intel Core i9-9900K processor with 32Gb RAM running an Ubuntu 18.04 OS.

All the methods considered in this work were executed considering the code made available by the authors in the respective original works. Although it is interesting to investigate some variation on the parameters, due to the requirement of delimiting a scope for this work, it was necessary to define an arbitrary set of parameters for the considered methods. In an attempt to get the best possible results, the methods were parameterized as suggested by the authors, as follows: CFinder (Palla et al. 2005) (size of k-cliques $k = 4$), Bigclam (Yang and Leskovec 2013) (minimum number of communities $mc = 5$, maximum number of communities $xc = 100$, number of trials $nc = 10$), Demon (Coscia et al. 2014) (merging threshold $\epsilon = 0.3$, minimum community size $mc = 3$), COPRA (Gregory 2010) (maximum memberships of nodes $v = 8$) and SLPA (Xie et al. 2011) (propagation threshold $r = 0.45$).

It is important to mention that the parameterization of the methods is a fundamental step in the execution of community detection methods, since different parameters may lead to very different results. Moreover, different measures may reveal distinct properties of the methods and the networks. Each work that presents a community detection method considers a set of measures for the evaluation of the methods and a set of parameters, that may differ from those suggested by the authors. Thus, the experimental setup defined in this work can be distinct to those presented by the works in which the methods are presented and a direct comparison between the results can be limited.

Experiments on synthetic networks

In order to define the experiments on synthetic networks, we consider the LFR benchmark (Lancichinetti et al. 2008). In this work, we defined a standard set of parameters which we considered as a base for the experiments, as suggested by the authors (Lancichinetti et al. 2009). The parameters are set as follows: the number of nodes is $n = 1000$; the average degree of the nodes is $\hat{k} = 20$; the maximum degree is $k_{max} = 50$; the minimum and maximum sizes of communities are $c_{min} = 20$ and $c_{max} = 100$ respectively; the nodes degrees and community sizes are governed by power-laws, with exponents $\tau_1 = -2$ and $\tau_2 = -1$; the mixing parameter is $\mu = 0.4$; the number of memberships of the overlapping nodes is $O_m = 2$; the number of overlapping nodes is $\frac{O_n}{n} = 0.2$. In order to investigate the behavior of the methods in networks with different parameters, we defined four sets of synthetic networks, each one obtained by the variation of a particular parameter: μ varies from 0.1 to 0.8 with steps of 0.1; n varies from 1000 to 10000 with steps of 1000; O_m varies from 1 to 8 with steps of 1; $\frac{O_n}{n}$ varies from 0.1 to 0.6 with steps of 0.1. For each set of parameters, ten instantiations were generated by the LFR benchmark.

First, Fig. 1 shows the average execution time of the methods considering a variation of the number of nodes from 1000 to 10000. It is important to notice that this work only consider high performance methods, able to execute in a reasonable time. The LFR benchmark generates random networks. Thus, the number of edges may suffer from fluctuations, which can make it difficult to evaluate the computational efficiency of the methods in the small and medium networks considered in this work.

Quality measures on synthetic networks

After the application of the overlapping community detection methods on all the generated networks, the resulting community structures were investigated regarding some of the objective measures most frequently considered in the literature in order to assess the quality of the methods and validate them. The main idea is to obtain quality measures in a very similar fashion as mostly found in other works to enrich the discussion on the community structure.
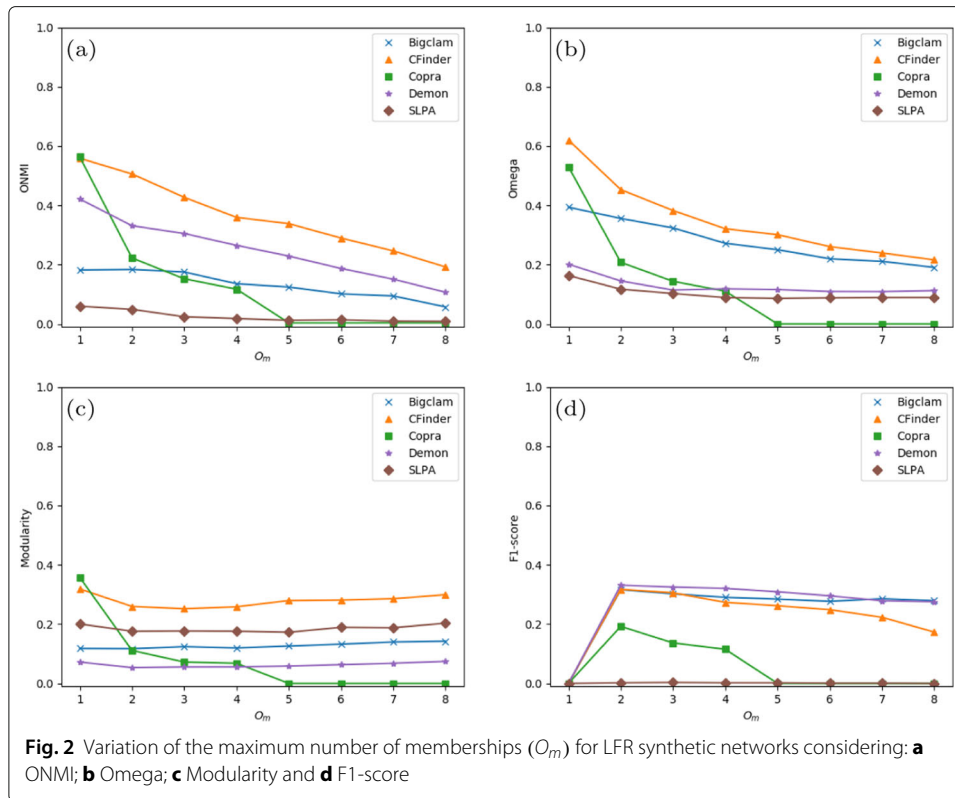
We investigated the following quality measures frequently considered by other authors in order to evaluate and validate community detection methods: Overlapping Normalized Mutual Information (ONMI), Omega index, Modularity and F1-score. For the sake of simplicity, Modularity refers to the overlapping modularity proposed by Shen (2013) (Eq. 2), which is referred by $\hat{Q}$ for the remainder of this work. F1-score assesses the ability of a method to identify the set of nodes in the overlapping regions. Each result plotted in the figures represents the average of the results observed for the ten instantiations of each parameter set.

Fig. 1 Real-world networks considered in this work
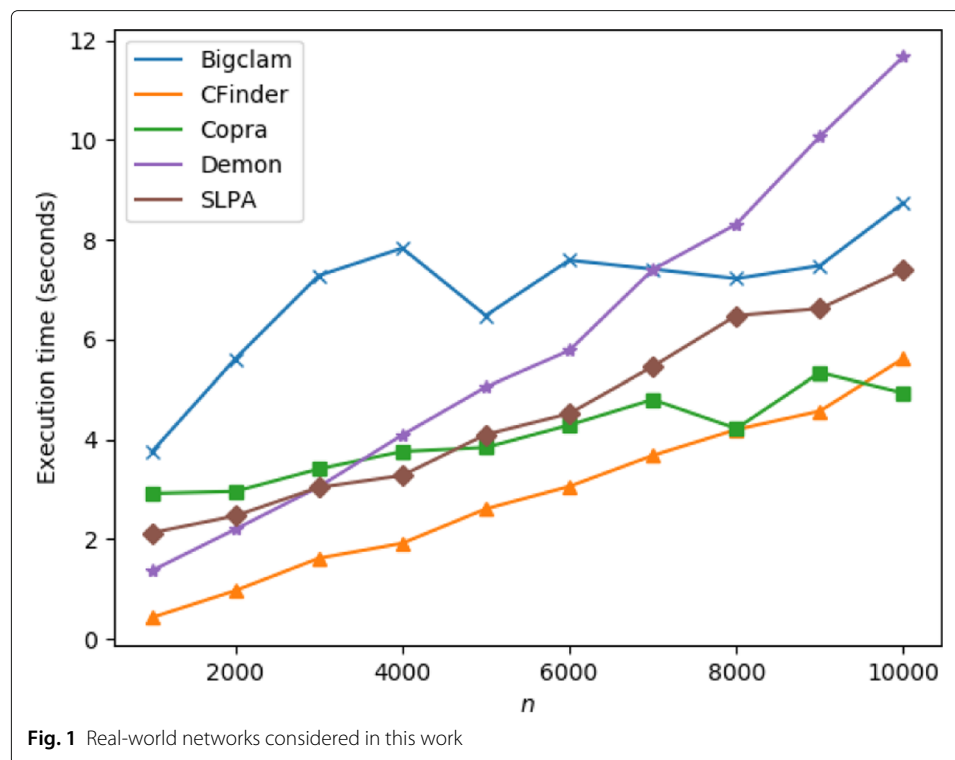
Considering the experimental setup described in the beginning of this section, four parameters were varied (μ, n, O_m and O_n) and four objective measures were assessed (ONMI, Omega, Modularity and F1-score) for the community structure identified by five methods (Bigclam, CFinder, Copra, Demon and SLPA). Therefore, a large amount of results were generated and, in order to preserve the organization and readability of the work, this section only reports the most remarkable and significant observations, together with general conclusions about taken from the experiments. A more detailed description of the experiments and a complete report of the figures obtained are exhibited in Appendix A.

Figure 2 presents the quality measures (ONMI, Omega, Modularity and F1-score) observed for the community structure obtained by the considered methods for the LFR benchmark networks with the basic parameter set and the number of memberships varying from $O_m = 1$ to $O_m = 8$ and illustrates well the general discussion in this section (for more detailed observations, please refer to Appendix A).

First, a typical behavior that can be observed is that Copra performs relatively well, overcoming the other methods, when the parameters sets generate networks with a very well defined community structure, specially when the network comprises disjoint communities ($\mu = 0.1$, $O_m = 1$, $O_n = 100$). The same performance is observed when the number of nodes in the network is greater than 2000. However, more mixed communities ($\mu > 0.3$) or more overlapped communities ($O_m > 2$ or $O_n > 200$) lead to very poor results of the method. Copra is based on the Label Propagation method of Raghavan et al. (2007) and requires the definition of a parameter v that controls the potential overlapping degree between the communities and taking $v = 1$ considers the problem as a disjoint community detection. Gregory observes that more modular communities are

Fig. 2 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering: **a** ONMI; **b** Omega; **c** Modularity and **d** F1-score

found as v varies from 1 to 9 in large scale networks (with more than 5000 nodes). However, the quality of the community structure does not significantly increases from $v = 8$ to $v = 9$ and, for this reason, in a more general case, the author recommends $v = 8$, which is the parameter considered in this work. In the original paper (Gregory 2010), the author observes a significant reduction in the performance of Copra regarding ONMI, but with different parameters from those observed in this work ($O_n \geq 600$). One of the possible reasons for this is the difference in the experimental setup considered in this work and in Gregory (2010). In Gregory (2010), the author considers $\mu = 0.1$ or $\mu = 0.3$, but in this work $\mu = 0.4$. Although the author defines a basic parameter set, considered for many experiments in the original work, and recommends $v = 8$, for these experiments, particularly, the author considers $v = 1$ and $v = 4$. Such parameterization, reducing the potential degree of overlapping of the communities, could disfavor the identification of overlapping nodes and consider the overlapping regions as separate communities, dealing with the problem from the perspective of disjoint communities, even though benefiting ONMI values. This can be observed not only regarding ONMI, but Omega, that also considers the number of global hits or Modularity, which favors the identification of dense regions.

Even showing a typically worse performance than the other methods, the quality values observed for SLPA can be considered competitive, specially in experiments that do not investigate the variation of parameters regarding the overlapping structure of the communities (μ and n) and measures that do not consider the nodes in the overlapping region (ONMI, Omega and Modularity). On the other hand, the F1-score for the communities identified by SLPA, which evaluates the quantity of nodes in the overlapping region correctly guessed by the method, shows very poor results in all of the experiments

performed. When looking closer to the communities identified by SLPA, we can see that the method places few nodes in the overlapping region, which is illustrated and more discussed in "Structural properties of communities on synthetic networks" section. As observed, and previously reported, for Copra, SLPA also solves the community detection problem as a partition problem and its use for overlapping structures should be made carefully. SLPA also extends the Label Propagation Method of Raghavan et al. (2007), but it limits the number of updates only in the borders of communities, in order to improve the execution time of the method. This strategy actually improves the execution time and allows the method to identify communities with a reasonable performance, however, in the work where SLPA is presented (Xie et al. 2011), the authors only analyze it with metrics that does not evaluate the nodes in the overlapping region, being unable to identify this drawback of the method with overlapping communities.

In all experiments performed for the evaluation of objective measures, Bigclam shows clear patterns and the performance of the method is proportional to the variation of the parameters. Despite the fact that all the methods are very sensitive to definition of their parameters, this is particularly emphasized in Bigclam. Bigclam works with the definition of a network model that simulates an overlapping community structure and the fitting of this model to an empirically observed network. In order to do this, the maximum number of communities ($maxc$) must be set for the execution of the method. By default, the method considers $maxc = 100$, which clearly underestimates the number of communities in the LFR networks, specially when the number of nodes is increased to $n > 2000$ for the LFR networks. Of course, we were aware of this before the execution and we had the opportunity to set $maxc$ to a higher and more appropriate value. However, this is a very difficult task to be performed properly, specially considering the extensive batch of executions performed in this work. Moreover, any efforts in this sense would make the comparison of methods unfair, since it would be virtually impossible to place all the methods in equal conditions. For instance, if we decided to take investigate a more appropriate parameter for $maxc$ in Bigclam, we should proceed similarly for Copra and the choice of v. When the basic number of nodes is considered ($n = 1000$ nodes), Bigclam presents very competitive results, being able to identify relatively modular communities and guess the correct communities of the nodes.

Demon is also based on the Label Propagation Method of Raghavan, Albert and Kumara, as well as SLPA and Copra, and it is interesting to notice that even though the three methods are based on the same idea, their behaviors are remarkably different when comparing the results of the objective quality measures. As opposed to Copra (which typically shows a drastic performance reduction when the communities are worse defined) and SLPA (which shows competitive results for global measures as ONMI and Omega but poor results for F1-score), the results of Demon are very reliable and, in spite of not presenting outstanding quality results, follow clear patterns for the measures, which can be considered a positive feature of the method.
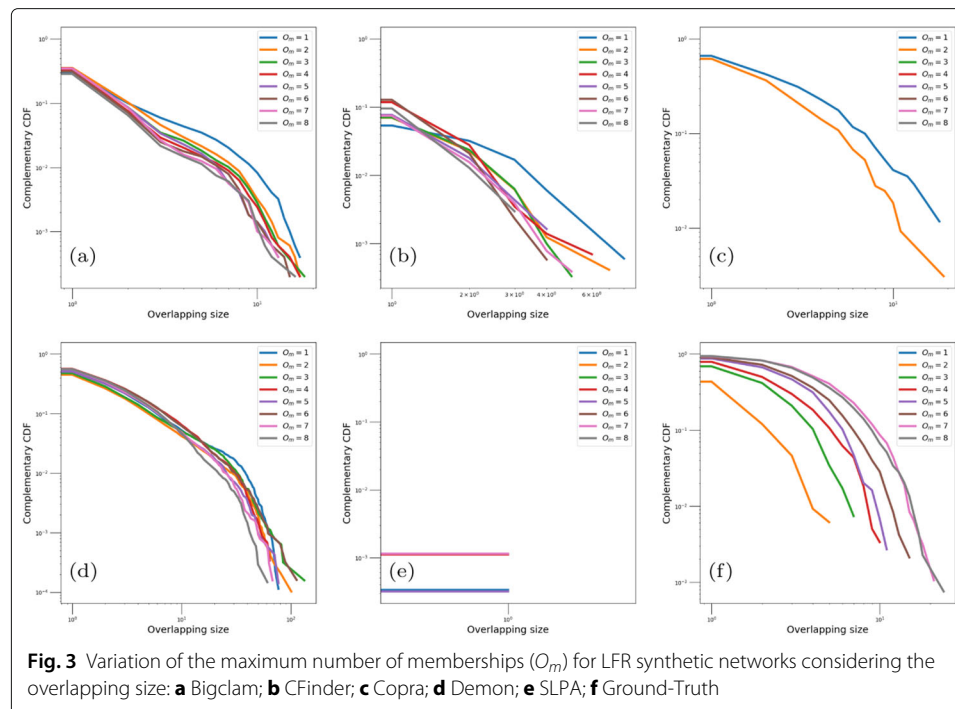
Structural properties of communities on synthetic networks

Currently, many works show that community structure and community detection methods can only be understood if the structural properties of the communities are investigated (Jebabli et al. 2018; Yang and Leskovec 2014; Harenberg et al. 2014; Kudelka et al. 2019), as discussed in "Introduction". One of the main objectives of this work is to

investigate if the community structures identified by the methods for overlapping community detection arise from the connections embedded in the network or if they are a mere consequence of the assumptions that the methods make about what is a community structure and the sequence of steps implemented by the methods to obtain this structure. Frequently, when a community detection method is proposed, the authors rely on traditional quality measures, such as those explored in "Quality measures on synthetic networks" sections, in order to compare the proposed method to others in the literature and validate it. However, the resulting community structure may provide good results for quality measures, even missing important properties of the community structure that are not captured by them. This section investigates the structural properties of synthetic networks with known overlapping community structure, treated as Ground-Truth, and compare them to the community structure extracted by the methods studied in this work. The experimental setup comprises the synthetic networks generated by the LFR benchmark with the same basic parameter set and variations of the mixing parameter μ, number of nodes n, maximum membership of nodes O_m and number of nodes in the overlapping region O_n as described in the beginning of this section. The following structural properties of the communities are investigated: number of memberships of the nodes, sizes of the communities, number of nodes in the overlapping region and probability of edges between two nodes in function of the number of communities they share. A huge amount of results and plots were generated, since each structural property for each parameter variation and each method was tested. This section discusses the outstanding and more general results from the experiments, in order to build a solid understanding of the methods and the community structure. A more detailed and complete presentation of the results, along with specific discussion are displayed on Appendix B.

The structures of the Ground-Truth networks are sensitive to the variations of parameters in all experimented scenarios (with less intensity when μ is varied). None of the methods considered in this work are able to perfectly identify communities that capture the shapes and the variations of the curves observed in the Ground-Truth networks. The complex nature of the community structure of the Ground-Truth networks is evident when we analyze the overlapping size, i.e., the number of nodes in the overlapping regions, since it is not a fixed parameter in the LFR benchmark. In this sense, more rich and revealing conclusions can be made in this experiment. Figure 3 presents the Complementary Cumulative Distribute Function (CCDF) of the overlapping size of the communities identified by the considered methods when the maximum number of memberships (parameter O_m) is varied. Even taking into account the limitation of the experiment, we believe that Fig. 3 summarizes well the ability of the methods in identifying the overlapping properties of the communities and can be considered to motivate most comments and discussions performed in this section in this sense (for a more detailed version of the results, please refer to Appendix B).
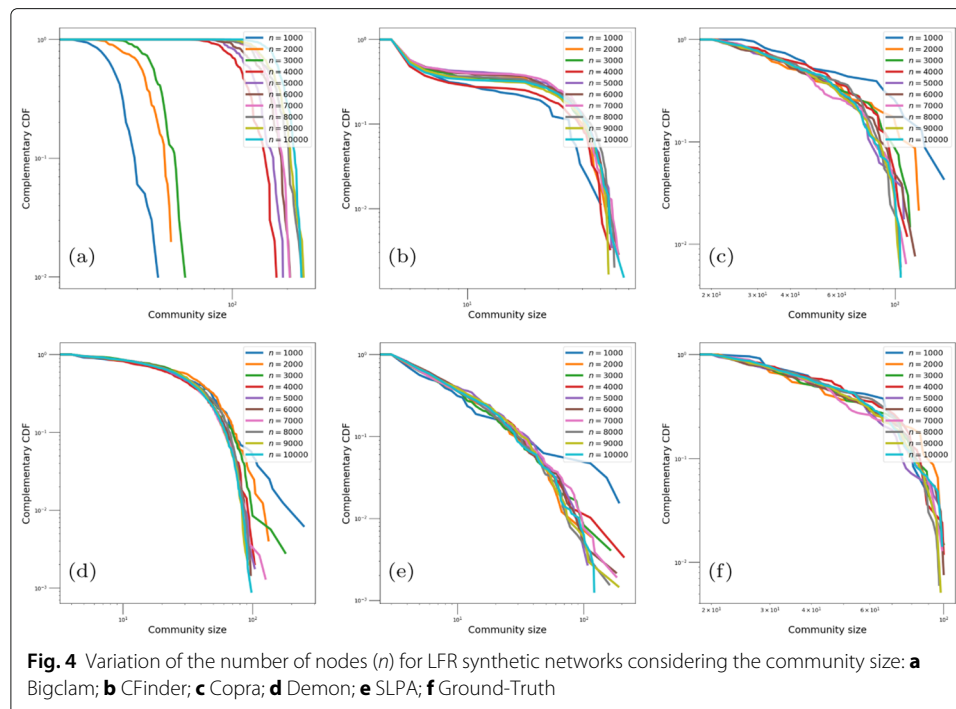
In this scenario, all methods overestimate the number of memberships, but most of them capture the order of magnitude. However, even though the methods identify nodes as part of multiple communities, all methods underestimates, in many degrees of magnitude, the probability of occurrence of overlapping nodes. Copra and SLPA place all nodes in a single community for networks, independent of the maximum number of memberships considered, and a plot can not be obtained on most cases. This behavior for Copra ans SLPA persists with the variation of other parameters and the methods do not capture

Fig. 3 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the overlapping size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

the belonging of nodes to more than one community with many combination of parameters, which corroborates with the observation of poor F1-score for SLPA and Copra, as reported in "Quality measures on synthetic networks" section and confirms that these methods treat the division of the network in communities as a partition problem. Still considering the memberships, for Bigclam, CFinder and Demon the curves are highly overlapped, which suggests that the methods are not sensitive to changes in the overlapping region. Specifically, when investigating the variation of parameters that control the overlapping of nodes (O_m and O_n), the curves observed for Bigclam, CFinder and Demon are almost coincident, meaning that the presence of nodes in the overlapping region is irrelevant to these methods and suggesting that, in this experimental setup, their suitability to overlapping community detection may be limited. This is a strong suggestion that the identified community structure does not reflect the Ground-Truth community structure, but a consequence of the way that the methods operate.

The distribution of the sizes of communities is the most consolidated and commonly investigated structural property of communities, frequently considered in the study of community detection methods (Schaeffer 2007; Fortunato 2010; Orman et al. 2012), less related to the overlapping characteristics of the networks than the other aspects considered in this study. In order to investigate this property, Fig. 4 shows the community size distribution observed for the synthetic networks when the number of nodes (parameter *n*) is varied.

In this work, the minimum and the maximum community sizes are fixed in the basic parameter set of the LFR benchmark. Two main remarkable results can be noticed regarding the sizes of the communities identified by the methods for the LFR networks. First, almost all shapes of curves follow a very clear tendency, distinct from those observed for the Ground-Truth networks, that resemble much more a signature of the methods than

Fig. 4 Variation of the number of nodes (*n*) for LFR synthetic networks considering the community size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

a property of the communities, even with small variations in scale. Second, the methods are able to identify community structures with up to about 10^2 nodes, which is the maximum size of communities set for the LFR benchmark. A particular behavior can be observed for Copra, Demon and SLPA, which are able to identify communities with up to 2×10^2 nodes when $n = 1000$, but with 10^2 nodes with $n > 1000$, suggesting that, as the networks become larger, the community structure becomes more clear and the methods are more prone to identify them. As this experiment allows us to investigate the methods regarding structural properties not necessarily related to the overlapping nature of the networks, but to their ability in identifying modular structures, it is worthy noticing that some particularities of the methods may directly affect the identified structures, as discussed in Appendix B.
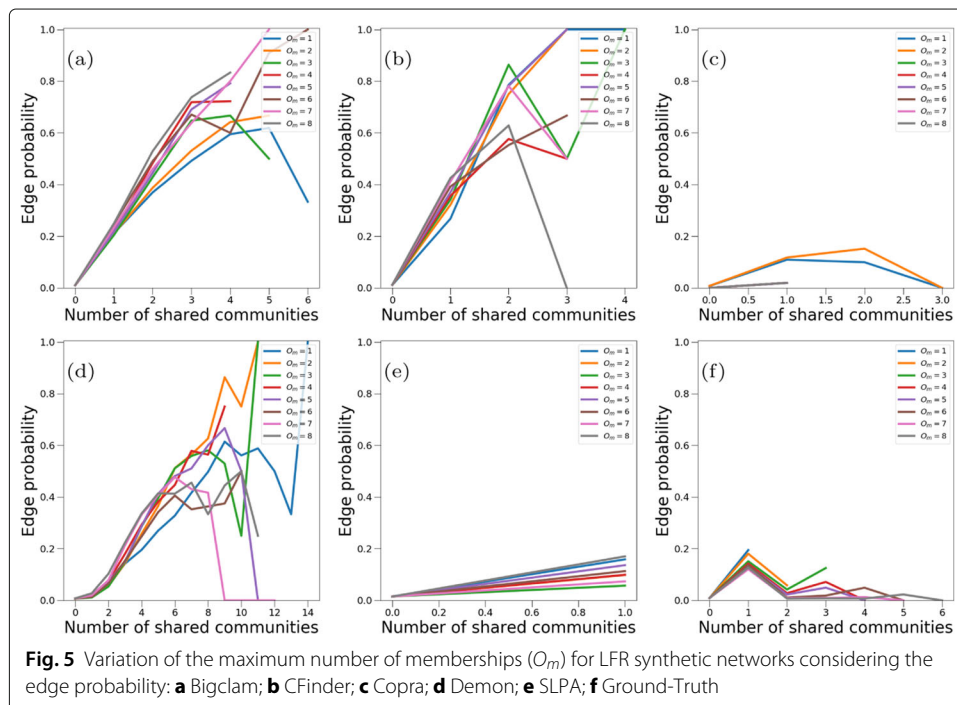
As discussed in "Overlapping community detection" section, Yang and Leskovec (2012) observed an important property in the structure of real-world networks with Ground-Truth communities that is usually neglected by community detection methods: the region of the network where communities overlap tend to be denser than the rest of the network. Moreover, there is a positive correlation between the number of shared communities of a pair of nodes and their probability of being connected, i.e. a pair of nodes that share two communities are more likely to be connected than a pair of nodes that share one community, a pair of nodes that share three communities are more likely to be connected than a pair of nodes that share two communities, and so on. Considering the LFR synthetic networks, we investigate the findings of Yang and Leskovec in two different perspectives: whether this dense overlapping regions actually exist in the synthetic networks generated by LFR and whether the methods are able to identify dense overlapping regions.

Figure 5 presents the probability of occurrence of an edge between a random pair of nodes as a function of the number of communities that they participate concomitantly

and illustrate well the following discussion (for a more detailed version of the results obtained in this experiment, please refer to Appendix B).

In this sense, first it is important to observe that the increasing in edge probability as a function of shared communities in Ground-Truth networks is not a universal behavior. The edge probability usually increases when a random pair of nodes shares one community, as expected even in disjoint communities, however, when the number of shared communities is increased, the edge probability does not increase in Ground-Truth communities (with few exceptions observed for some particular experiments and parameter sets). This indicates a clear limitation in the LFR benchmark, which is able to generate networks that simulate some structural properties of disjoint communities, but fail in reproducing an important feature of overlapping communities. From the studied methods, Bigclam, Demon and CFinder are those that better capture dense overlapping regions, even though it is not an universal behavior and only occurs in some scenarios. Copra and SLPA do not identify substantial overlapping regions in most experiments and the analysis of edge probability can not be made. Due to the limitation of LFR benchmark in reproducing communities with dense overlapping regions, it is fundamental to investigate the ability of the methods in real-world networks.

It is important to mention that, as reported on "Introduction" section, this work only considers a subset of the state-the-art-methods, which presented low execution time after some preliminary experiments. However, we believe that the subset of methods considered in this work are a good representation for the state-of-the-art, since the preliminary results did not affect the observations discussed from the experiments reported in this section.

Fig. 5 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the edge probability: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Experiments on real-world networks

In order to investigate the structural properties of communities from real world scenarios, two sets of networks were considered: a set of benchmark networks, frequently explored for the evaluation of community detection methods and sampled Ground-Truth networks.

As discussed in the "Introduction" section, many authors show that the evaluation of community detection method is a hard task, since a definition of what is a community is not clear (Schaub et al. 2017; Jebabli et al. 2018; Almeida et al. 2011). Yang and Leskovec (2012) state that, traditionally, works that aim in studying overlapping communities lack of reliable Ground-Truth, what makes it even more difficult to evaluate overlapping community structure in networks and, later, the authors identified that functional communities may be different than structural communities (Yang and Leskovec 2014), what is corroborated by other authors, like Harenberg et al. (2014). More recently, Peel et al. (2017) have demonstrated that treating metadata as the Ground-Truth communities may induce to several problems, but enphasize the importance of the investigation of the community structure found in networks in order to better understand complex networked systems.

Some metadata about communities in networks, treated in this work as Ground-Truth may even be found in the literature, but since they are very large (frequently with millions of nodes), it is unpractical to consider them to evaluate community detecting methods. Thus, a sampling strategy was adopted in this work for Ground-Truth networks due to the impossibility of the methods in dealing with their complete versions, with millions of nodes. A variation of the sampling strategy proposed by Yang and Leskovec (2012) for Ground-Truth networks is adopted. The original strategy randomly chooses a seed node u from the network and identifies all the communities at which u participates. Then, it separates all the nodes from these communities and determines the induced subgraph considering these nodes. In this work, we do not select the seed node randomly. Instead, we first identify the node which participate in most communities and define it as the seed node. By doing this, we expect to obtain a sample network with denser overlapping regions and closer to a real world scenario.

The set of real-world networks investigated is composed of five benchmark networks and four sampled networks with Ground-Truth communities. The networks are presented in Table 1 with their number of nodes n, number of edges m and the type (benchmark or sampled).

Quality measures of real-world networks

Table 2 shows some results regarding objective measures of the communities extracted by the explored methods on real-world networks. As in the experiments with synthetic networks, the overlapping modularity $\hat{Q}$ considered implements the extension proposed by (Shen 2013) (Equation 2). The Overlapping Normalized Mutual Information (ONMI) (McDaid et al. 2011) (based on Lancichinetti et al. 2009) is presented only for networks with Ground-Truth information. The execution time for each method and each network is also presented.

First, it is possible to observe from Table 2 that, for most Ground-Truth networks, the methods were able to execute in a reasonable time, allowing them to be used in real world small and middle sized applications. For some benchmark networks, CFinder and Demon

Table 1 Real-world network considered in this work

Network	n	m	Type
CA-GrQc[1]	5242	14496	Benchmark
CA-HepPh[1]	12008	118521	Benchmark
Cit-HepTh[1]	27770	352324	Benchmark
Email[2]	1133	5451	Benchmark
Keys[2]	10680	24319	Benchmark
Amazon[1]	10308	21575	Sampled
DBLP[1]	7515	31784	Sampled
Live Journal[1]	5843	12543	Sampled
Youtube[1]	6792	28177	Sampled

[1] Downloaded from: http://snap.stanford.edu/data/
[2] Downloaded from: http://www-personal.umich.edu/~mejn/netdata/

presented a high execution time, making them prohibitive to be considered in real time applications. Bigclam and SLPA, specially, presented consistent results, executing almost all the networks in a short time. For slightly larger networks the methods start to show a high execution time and some considerations must be done. All the methods shown a high execution time for CitHepTh (with $\approx 27k$ nodes and $\approx 352k$ edges). It is important to highlight that CFinder was unable to run CitHepTh and CAHepph.

Some observations regarding quality measures are made with the purpose of complement the analysis performed for synthetic networks and better understand the structural properties of the community detection methods. First, it is possible to observe that the Ground-Truth community structure of the Youtube network is not modular $\left(\hat{Q} = 0.30 \right)$ if we consider that a modular community structure has $0.30 < \hat{Q} < 0.70$. The studied methods are able to identify non modular community structure, with $\hat{Q} \approx 0$, even if Copra slightly overestimates $\hat{Q}$ $\left(\hat{Q} = 0.36 \right)$. COPRA, Bigclam and SLPA were able to identify modular communities in almost all networks. When these results are compared to those observed for synthetic networks ("Quality measures on synthetic networks" section), it is possible to see that the Modularity of the real-world networks is consistent to the Modularity of the synthetic networks obtained when the parameter set generates a well defined community structure. This shows that the real-world networks actually exhibit a clear community structure and indicates the choice of these networks is good for the purpose of this work. It is interesting to notice that, for Amazon network, Bigclam, Demon, COPRA and SLPA were able to identify communities with much more modular structure than the Ground-Truth, specially SLPA, which obtained $\hat{Q} = 0.80$, while the Ground-Truth modularity is $\hat{Q} = 0.40$. Differently from the Modularity of the real-world communities, the results for Overlapping Normalized Mutual Information (ONMI) observed in Table 2 are much poorer than those observed for the synthetic networks and show that the methods were unable to identify the communities correctly although the results obtained by COPRA stand out from the other methods. Some hypothesis to explain why the methods mostly show a low ONMI, while presenting a high $\hat{Q}$ are discussed in the next sections.

Structural properties of communities on real-world networks
Considering the features of the community detection methods and the results for objective quality measures previously presented, it is possible to investigate other aspects of

Vieira *et al. Applied Network Science* (2020) 5:51

Page 21 of 42

Table 2 Execution time (in seconds), Overlapping modularity $\hat{Q}$ and Overlapping Normalized Mutual Information (ONMI), when applicable, for the studied networks
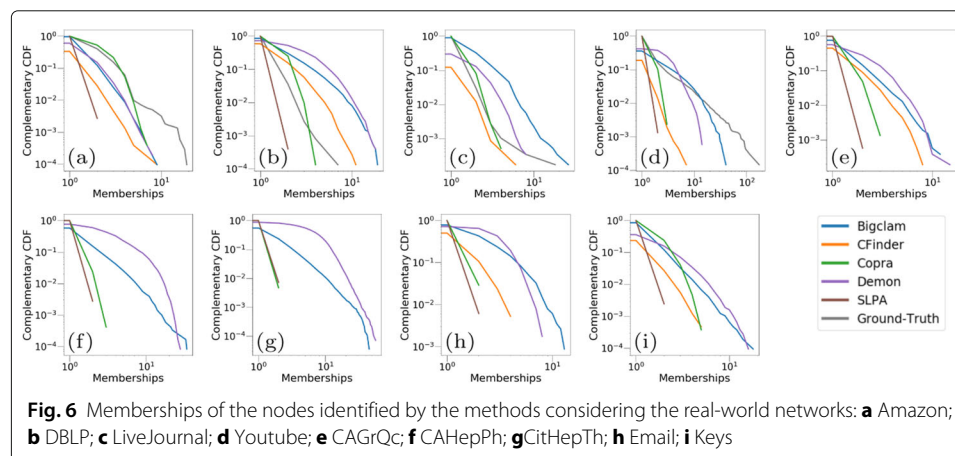
	GT	CFinder			Bigclam			Demon			COPRA			SLPA		
	$\hat{Q}$	Time	$\hat{Q}$	ONMI	Time	$\hat{Q}$	ONMI	Time	$\hat{Q}$	ONMI	Time	$\hat{Q}$	ONMI	Time	$\hat{Q}$	ONMI
Amazon	0.40	0.65	0.33	0.04	0.31	0.81	0.11	2.60	0.51	0.05	14.82	0.53	0.25	2.50	0.80	0.05
DBLP	0.57	2.43	0.38	0.04	1.11	0.43	0.03	6.54	0.20	0.04	2.00	0.50	0.25	1.77	0.52	0.07
LJ	0.59	0.35	0.15	0.00	0.58	0.42	0.01	1.03	0.21	0.05	0.88	0.52	0.12	1.06	0.58	0.04
Youtube	0.30	16.32	0.02	0.00	3.77	0.06	0.00	2.76	0.03	0.03	0.72	0.36	0.16	1.07	0.01	0.00
CAGrQc	-	0.71	0.46	-	0.29	0.61	-	2.13	0.41	-	0.87	0.47	-	1.14	0.66	-
CAHepPh	-	-	-	-	5.03	0.35	-	74.65	0.14	-	2.37	0.16	-	4.31	0.25	-
CitHepTh	-	-	-	-	35.37	0.16	-	86.18	0.04	-	9.67	0.01	-	15.02	0.14	-
Email	-	0.27	0.27	-	0.69	0.18	-	0.54	0.06	-	0.22	0.51	-	0.24	0.11	-
Keys	-	2111.34	0.38	-	0.93	0.63	-	2.98	0.31	-	1.64	0.68	-	2.50	0.71	-

the structural properties of the networks. As performed for the synthetic LFR bench-mark networks and presented in "Structural properties of communities on synthetic networks" section, we investigate four structural properties of the overlapping communities identified by the community detection methods considered in this work: node memberships, community size, overlapping size and edge probability in the overlapping region. The structural properties were investigated for all the networks and generated a large amount of results. The most relevant and general conclusions from this experiment are reported in this section. In order to avoid to harm the readability of the work, the plots are completely displayed in Appendix C, where more specific comments about the results can be also found.

Figure 6 shows the Complementary Cumulative Distribute Function (CCDF) of node memberships, i.e., the number of communities in which each node participate for all the identified communities and the Ground-Truth communities for each network and illustrates the discussion in this section (for more detailed results, please refer to Appendix C).

First, it is very important to observe that except for COPRA and SLPA, the methods do not necessarily assign a community to each node and some nodes can be left with no community and, thus, the probability of a node to belong to at least one community can be less than one. This behavior particularly stands out for CFinder, that severely underes-timates the number of nodes that belong to one community. It is also possible to observe that the patterns for curves observed for the Ground-Truth communities are not captured by any of the studied methods. It is also interesting to notice that, despite to differences in scale, some methods, like SLPA, Demon and Bigclam, although different among them, have shown a very similar behavior when the network is varied, in most cases, suggesting that the resulting community structure is not very dependent on the relations existing on the network, but mostly on the strategy of the algorithm. It is important to note that most methods depend on the definition of parameters to execute, which may strongly affect the community structure and a deeper instigation should be performed for each method in order to draw more conclusive observations.

The investigation of the size of communities obtained by the methods in the synthetic LFR benchmark allows us to conclude that the methods capture a great number of aspects of this property. However, this is not observed in the real-world networks. Except for

Fig. 6 Memberships of the nodes identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** CitHepTh; **h** Email; **i** Keys

COPRA, which is able to identify communities with very similar size to the Ground-Truth, this is atypical and such similarity can not be observed in the other scenarios. It is interesting to notice that the shapes of the curves observed for Demon and, more clearly, for Bigclam for the community size is very consistent for most networks, though very different from the Ground-Truth, suggesting that the structural properties of the communities may be highly affected by the mechanisms used by the methods and not by the network itself, as observed in other experiments on synthetic networks.

It is worth to notice that, for most methods, the occurrence of overlapping regions is very unlikely. The exception is Bigclam, which identifies communities with relevant overlapping regions, specially in the Ground-Truth networks. Yet considering the networks with Ground-Truth, the number of small overlapping regions is very high. Moreover, for three Ground-Truth networks (Amazon, DBLP and Live Journal), the methods underestimate the sizes of overlapping regions in all the range of the distribution. In the communities identified by COPRA, SLPA and CFinder the occurrence of nodes in the overlapping region is extremely rare. Therefore, we can argue that these methods are treating the community detection as a partition problem for a large number of nodes. CFinder and, specially, SLPA also identify very small overlapping regions for almost all networks. When considering the results from Table 2, it may suggest that the methods fail in identifying the overlapping nodes, concentrating the hits on the nodes that belong to only one community.

If we take into consideration the results presented in Table 2, we can see that, even though in some cases the modularity observed for community structure identified by the studied methods is lower than the modularity observed for the Ground-Truth communities, the methods are able to identify modular communities, while ONMI is very low (the exception is Youtube network, for which the Ground-Truth community structure is not modular). Apart from Youtube network, the probability of occurrence of large Ground-Truth communities are higher than the probability of occurrence of large networks in the community structure identified by the studied methods. In this cases, possibly, the methods are mistakenly disconsidering larger communities as they are prone to identify very high modular structures, even if they are very small compared to Ground-Truth communities. This may be causing the methods to fail in discovering the right communities, for instance, considering dense overlapping regions as separate communities. The tendency of the methods in underestimating the overlapping nodes, as observed in many experiments in this work, also contributes to the identification of modular communities, even if harming ONMI. A slightly different behavior from the other methods can be observed for Copra, that shows large communities more frequently for Amazon and DBLP networks and also presents higher values for ONMI in these cases.

We also investigate the findings of Yang and Leskovec (2013) regarding the density of edges in the overlapping region and evaluate whether this property is observed in the Ground-Truth networks, verifying if the methods investigated are able to capture this behavior even for the benchmark networks.

Figure 7 shows the probability of an edge between two nodes to exist in function of the number of communities shared by them.

Unlike the observation made for the synthetic LFR benchmark, in the real-world networks there is, in fact, a positive correlation between the number of shared nodes and the edge probability in Ground-Truth communities. This observation strongly suggests that
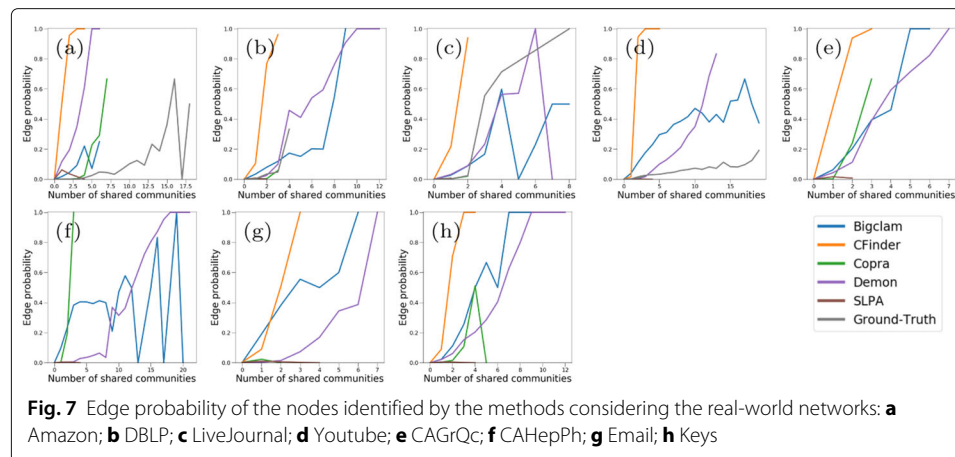
Fig. 7 Edge probability of the nodes identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** Email; **h** Keys

the LFR benchmark, which is considered by a great number of authors in order to evaluate and validate community detection methods, must be used carefully if one needs to identify communities that meet the properties of real-world communities. Regarding the considered methods, the increasing in edge probability is better captured by Bigclam and Demon, and CFinder usually overestimates it signifiantly. Overall, however, the methods fail in capturing the density of overlapping regions, as stated by Yang and Leskovec (2012). When the other results are considered, a clearer view of the results presented in Fig. 7 is possible. For instance, in SLPA and COPRA the occurrence of large overlapping regions and the presence of nodes in several communities is rare. Thus, it is expected that the probability of an edge to exist in the overlapping region does not show a clear pattern among the different networks.

Conclusions

This work presents a comparative analysis of the overlapping community structure identified by five of the state-of-the-art methods for overlapping community detection in complex networks focusing on their topological properties. A wide set of 435 networks, comprising of synthetic networks, generated by the LFR benchmark, real-world benchmark networks, frequently used in the literature and real-world sampled networks with Ground-Truth were considered. Objective quality measures, frequently considered in the literature (ONMI, Modularity, Omega index and F1-score) were also considered and related to the structural properties of the communities, regarding the features of the methods and assumptions made about the characterization of communities.

This works confirms some recent findings from other authors that highlight the importance of the analysis of the topology of the community structure in order to better understand the way that the elements are organized in a complex network (Peel et al. 2017; Jebabli et al. 2018) and other authors that state that there is not a best algorithm that solves a general community detection problem (Almeida et al. 2011; Peel et al. 2017; Schaub et al. 2017). Moreover, the notion that, despite the definition of what is a community, the way that the methods perform may lead to very different divisions of the network (Cherifi et al. 2019; Schaub et al. 2017), is extended by the results obtained from the experiments performed in this work, since the topological organization of the communities, in many cases, looks more like a signature of the method than the actual organization of the

Ground-Truth communities, which answers the Research Question II ("does the quality measures frequently considered in the literature for the validation of overlapping community detection method are really appropriate and consistent with the real community structure?"). The results of the experiments in this work also corroborate the views of Peel et al. (2017) and Harenberg et al. (2014) that the topological properties of the community structures are may be weakly correlated with the metadata of the elements in the system, making the functional communities, in many case, impossible to be recovered by community detection methods, that lay only in information encoded in the connections of the network.

Despite the difficulty in the definition of the communities and the possibility of recovering communities from the network, an important aspect that poses a great challenge in the analysis of overlapping community detection methods, which indirectly affects the correct evaluation of the applicability of each one, is the difference in the conditions in which the methods are compared, which arises from two main sources: the parameterization of the methods and the measures chosen for the evaluation of the methods. In the works where the methods are proposed, the authors perform a series of experiments in order to investigate the impact of the parameterization in the performance of their performance. However, in practical contexts it is virtually impossible to test different sets of parameters and one must rely on suggested or default parameters. Particularly in this work, where a very large number of experiments were performed and each method were executed many times, it was impracticable to investigate how the methods behaved when their parameters were varied. Moreover, even with a huge effort, it is unfeasible to define an experimental setup that put all the methods in the same conditions and, thus, the full potential of the methods were not explored. The results obtained in some experiment, thus, were directly affected by the poor parametrization of the methods. For instance, SLPA requires the definition of a parameter that controls the threshold limit and, by setting the default value, the method severely underestimated the number of nodes in the overlapping regions and the parametrization of Bigclam requires the definition of the maximum number of communities to be investigated, which is clearly smaller than the number of communities in real world scenarios. Besides the lack of a general test bed for community detection methods, one difficulty for a direct comparison between the results in this work and those shown in the works in which the methods are originally presented is, mainly due to a frequently reduced transparency in the presentation of the results. For instance, in Yang and Leskovec (2013), the quality measures are presented in a relative way, comparing to competing methods.

Regarding the results obtained in the experiments in this work, we can see that there is not a single method that performs clearly better than the other in all the scenarios and a rank of methods can not be defined, which answers the Research Question I ("is there an overlapping community detection method that stands out regarding from the others and can be adopted as a state-of-art method?"). The methods perform way better when the networks exhibit a very clear community structure. This characteristic is much evident for Copra, which shows outstanding results when the communities are less mixed and the overlapping has only few nodes. A remarkable result observed in the experiments is that all the methods perform relatively well when the number of memberships per node is set as one, i.e., when the community detection problem is solved as a partition problem. Many authors in the literature discuss that the community detection problem is still

an open problem (Peel et al. 2017; Jebabli et al. 2018; Peixoto 2020) and when overlapping communities are considered, the problem is even more challenging (Fortunato 2010). The traditional view is that a community is a dense group of nodes connected by a few number of edges. When this view is extended to overlapping community detection, it leads the methods to neglect the nodes in the overlapping region and, thus, treat the problem as a partition community detection, potentially identifying the overlapping regions as separate communities. Indeed, the community detection methods investigated in this work tend to strongly underestimate the number of nodes in the overlapping region, a fact that can be captured by the analysis of the structural properties of the communities, as originally pointed out by Yang and Leskovec (2012), and is recently discussed by many authors (Chykhradze et al. 2014; Jebabli et al. 2018) (also answering RQI). Opposing to the findings of Yang and Leskovec, the low density of nodes in the overlapping region can also be observed in the Ground-Truth communities generated by the LFR benchmark, which can arguably indicate a limitation in the experiments conducted in this work. More recently, Chykhradze et al. (2014) proposed the CKB benchmark in order to generate benchmark networks with dense overlapping regions, which should be considered in future works.

However, traditional measures for assessing the quality of overlapping methods (such as ONMI, Omega and Modularity) fail in identifying this behavior. When the F1-score is analyzed, for instance, it is easy to observe that the methods miss the overlapping nodes, even when identifying modular communities and placing the majority of nodes in the correct groups. This kind of investigation allows us to argue that it is virtually impossible to define an universal method that is able to identify satisfactory in all situations. Thus, if we need a method to be appropriate to overlapping communities, we must understand the properties of this kind of communities and assume these properties in the operation of the method. Bigclam is a good example in this sense, since it incorporates a recent finding of Yang and Leskovec (2013) that overlapping regions tend to be denser than the rest of the network in order to define a model to be fitted. In a more general view, it is important to deeply investigate and understand the structural properties of communities in particular contexts, in order to be able to propose methods that extract communities with those properties, aware that the applicability of the methods are limited to those scenarios.

Appendix A: Detailed comparison of quality measures on synthetic networks
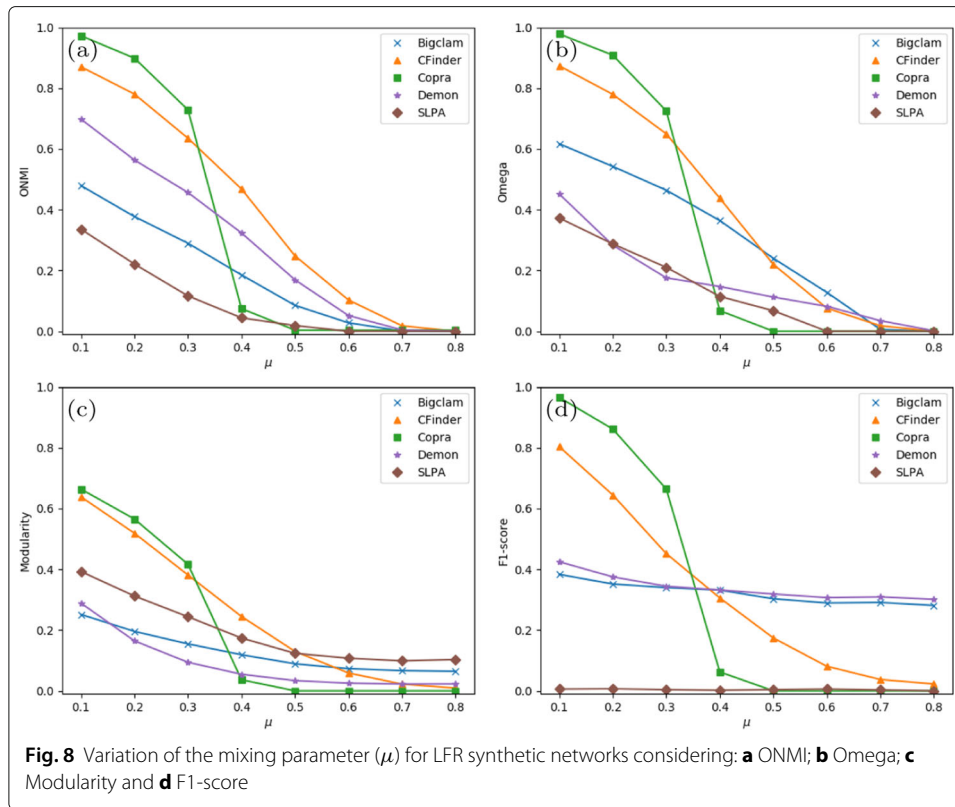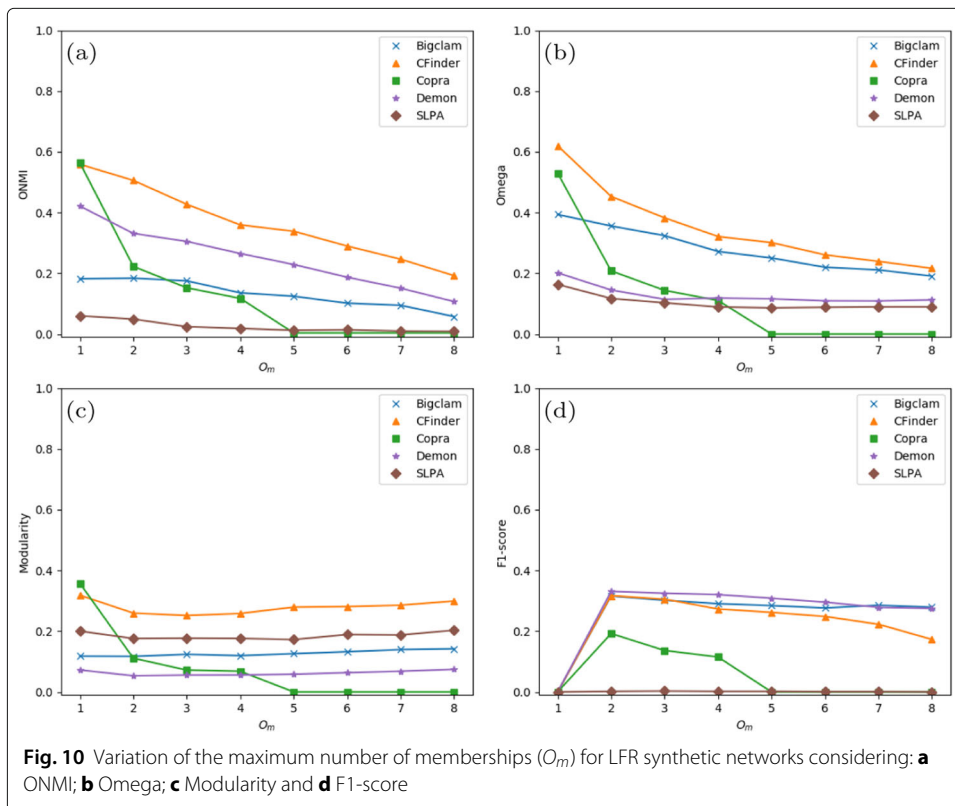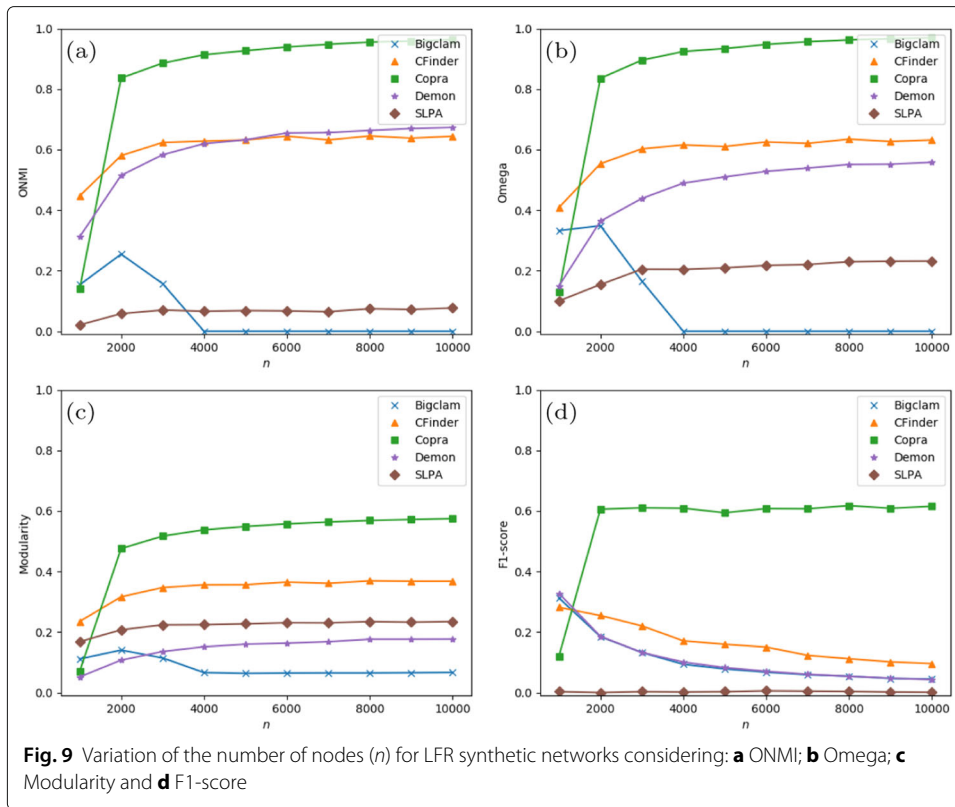
Figure 8 presents the quality measures (ONMI, Omega, Modularity and F1-score) observed for the community structure obtained by the considered methods for the LFR benchmark networks with the basic parameter set and the mixing parameter varying from $\mu = 0.1$ to $n = 0.8$.

First, it is important to observe that all methods are very susceptible to variations on μ. From Fig. 8, it is possible to see that, naturally, the quality of ONMI and Omega results drops as the mixing parameter μ is increased. The results observed for Copra stand out when compared to the other methods with $\mu = 0.1$, with an ONMI almost one, however, it exhibits a great decay, not only for ONMI and Omega, but for all quality measures, specially when $\mu = 0.4$. The F1-score observed for Bigclam and Demon remain almost constant for all values of μ. On the other hand, for Copra and Bigclam, high values of F1-score are observed for $\mu = 0.1$ (≈ 1.0 and ≈ 0.8 respectively), but when the value of μ is increased, the efficiency drops significantly.

Fig. 8 Variation of the mixing parameter (μ) for LFR synthetic networks considering: **a** ONMI; **b** Omega; **c** Modularity and **d** F1-score

Figure 9 presents the quality measures (ONMI, Omega, Modularity and F1-score) observed for the community structure obtained by the considered methods for the LFR benchmark networks with the basic parameter set and the number of nodes varying from $n = 1000$ to $n = 10000$.

When the number of nodes is increased, the values of ONMI, Omega and Modularity are improved until a certain value of n ($n \approx 4000$) but then the efficiency remains almost constant, as seen in Fig. 9. The only exception is Bigclam, which shows an elevation of these values from $n = 1000$ to $n = 2000$, but then it drops significantly. One possible explanations, as discussed in "Experiments and discussion" section is the parametrization of Bigclam, that requires the definition of the maximum number of communities to be extracted. One observation that stands out from the plots on Fig. 9 is that Copra does not show good results for $n = 1000$, but as the number of nodes increases, the results are rapidly improved.

Figure 10 presents the quality measures (ONMI, Omega, Modularity and F1-score) observed for the community structure obtained by the considered methods for the LFR benchmark networks with the basic parameter set and the number of memberships varying from $O_m = 1$ to $O_m = 8$.

It can be observed from Fig. 10 that, as the number of memberships increase, the least the methods are able to identify the correct communities. On the other hand, it is interesting to notice that the Modularity values remain almost constant in all the range of values of O_m, except for Copra, which rapidly decreases. It is also important to highlight that all values for F1-score are zero when $O_m = 1$, since the class of interest is the nodes in the overlapping region, absent with this configuration. One important observation that can
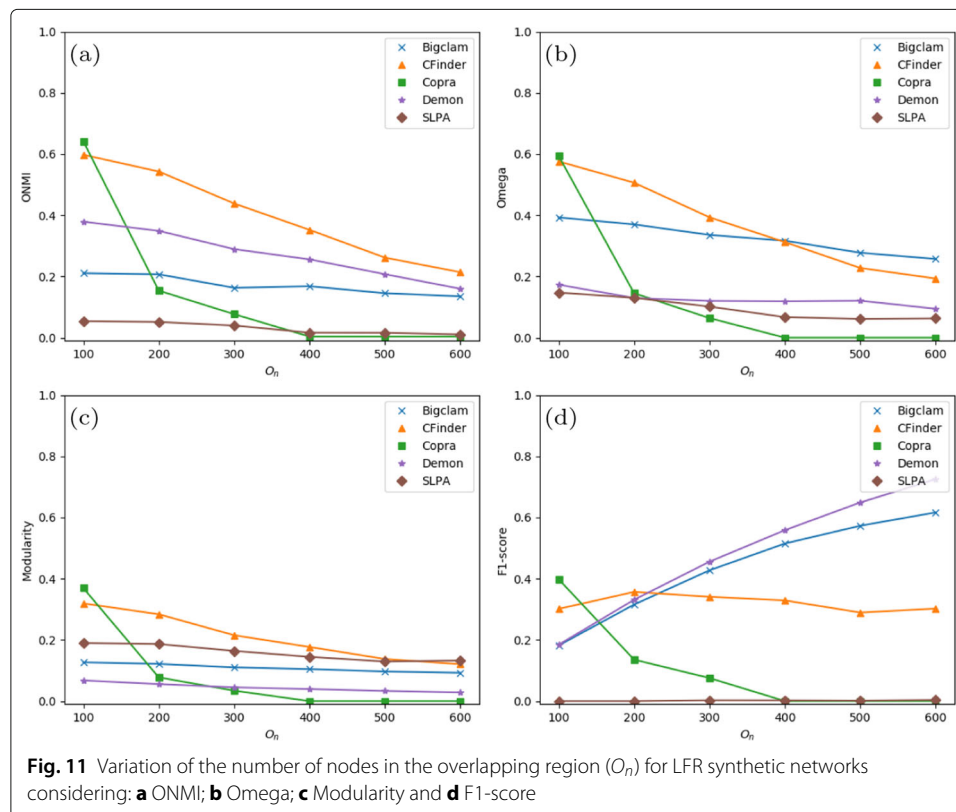
Fig. 9 Variation of the number of nodes (*n*) for LFR synthetic networks considering: **a** ONMI; **b** Omega; **c** Modularity and **d** F1-score

Fig. 10 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering: **a** ONMI; **b** Omega; **c** Modularity and **d** F1-score

be done from Fig. 10 is that all the methods perform way better when $O_m = 1$, suggesting again that the methods are more appropriate for the identification of disjoint communities than overlapping communities.

Figure 11 presents the quality measures (ONMI, Omega, Modularity and F1-score) observed for the community structure obtained by the considered methods for the LFR benchmark networks with the basic parameter set and the number of overlapping nodes varying from $O_n = 100$ to $O_n = 600 \left(\frac{O_n}{n} = 0.1 \text{ to } \frac{O_n}{n} = 0.6 \right)$.

A very similar behavior can be observed when the results of ONMI, Omega and Modularity are compared for the variation of memberships of nodes (O_m) and fraction of nodes in the overlapping region (O_n). The performance of the methods is notably harmed when O_n is increased. Again, Copra performs better for networks with a low density of nodes in the overlaps ($O_n = 100$), but the performance gets worse as O_n is increased. This results corroborate the observation that the methods are way more suitable to applications when the communities are not overlapped.

Appendix B: Detailed comparison of structural properties on synthetic networks
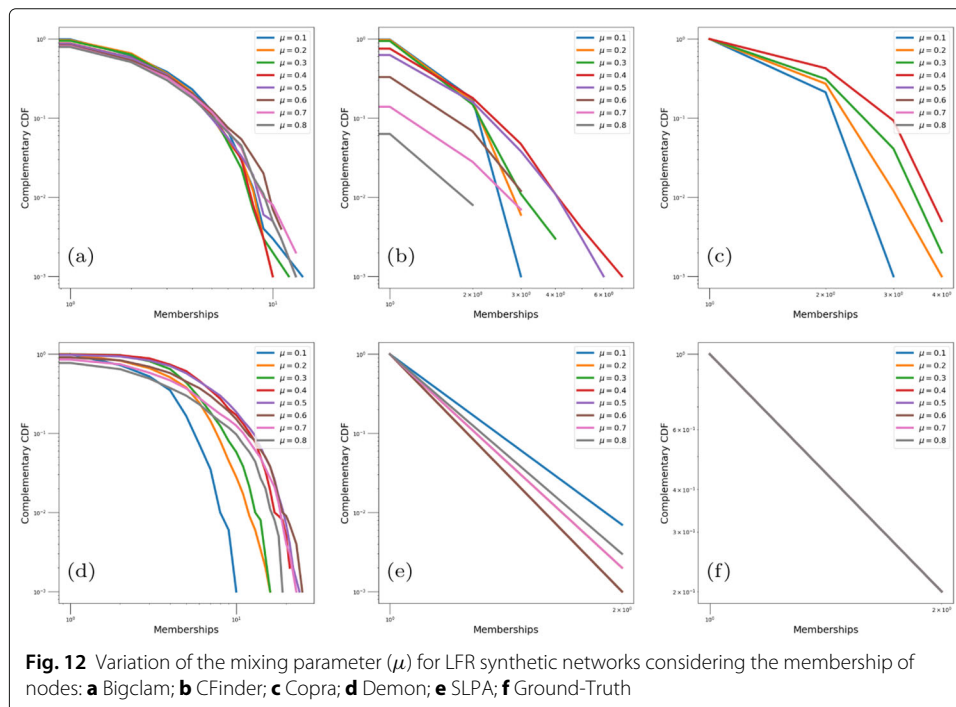
Figures 12, 13, 14 and 15 present the Complementary Cumulative Distribution Function of the memberships of nodes, i.e., the number of communities in which each node participate, observed on the community structure identified by each of the methods considered in this work on the LFR benchmark networks with the basic parameter set.

Fig. 11 Variation of the number of nodes in the overlapping region (O_n) for LFR synthetic networks considering: **a** ONMI; **b** Omega; **c** Modularity and **d** F1-score
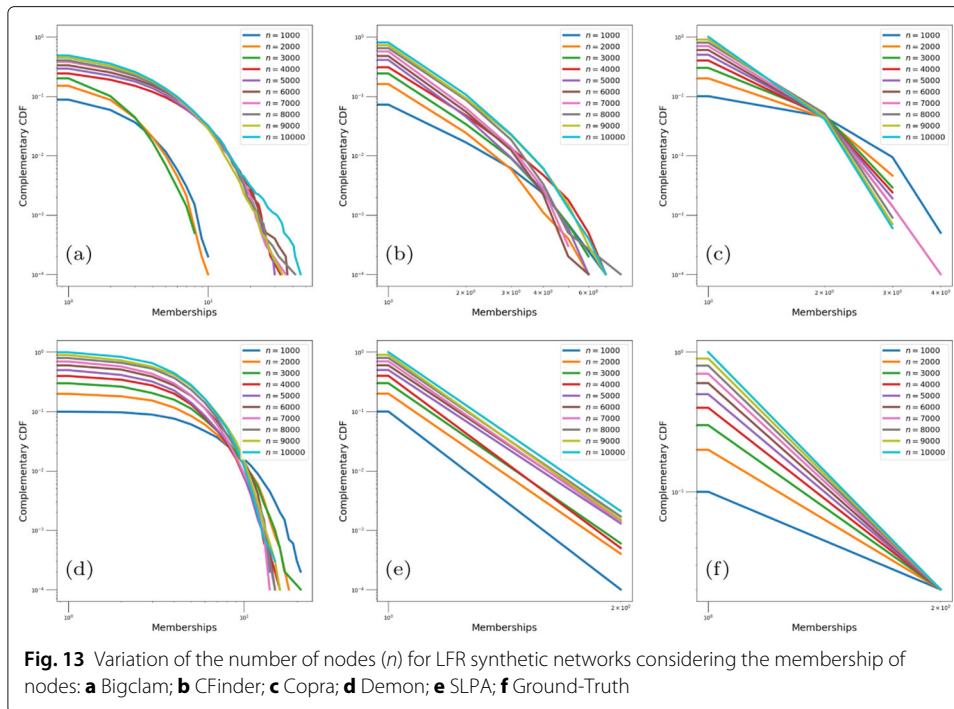
The distribution of membership on the Ground-Truth community structure (Fig. 12) allows us to see that, as defined by the parameter set, the nodes can participate in one or two communities and all the curves are overlapped. Only SLPA is able to capture this limitation of two memberships per node even though it underestimates the number of nodes with two memberships (for the Ground-Truth communities, this number is ≈ 0.2 and for SLPA communities this number is ≈ 0.001). For Bigclam and Demon, the number of memberships is quite overestimated, and these methods find communities in which the nodes participate in more than ten communities, even with a low frequency. A similar behavior can be observed in Figs. 13, 14 and 15, which illustrate the variation of n, O_m and O_n, respectively. In all these experiments, Bigclam and Demon overestimate the number of memberships of the nodes, even with a low probability. An interesting result can be observed from Fig. 13, where the memberships of the nodes in the communities observed by Bigclam are limited to 10 when $n = 1000$, $n = 2000$ and $n = 3000$ but limited to 20 when $n \geq 3000$.
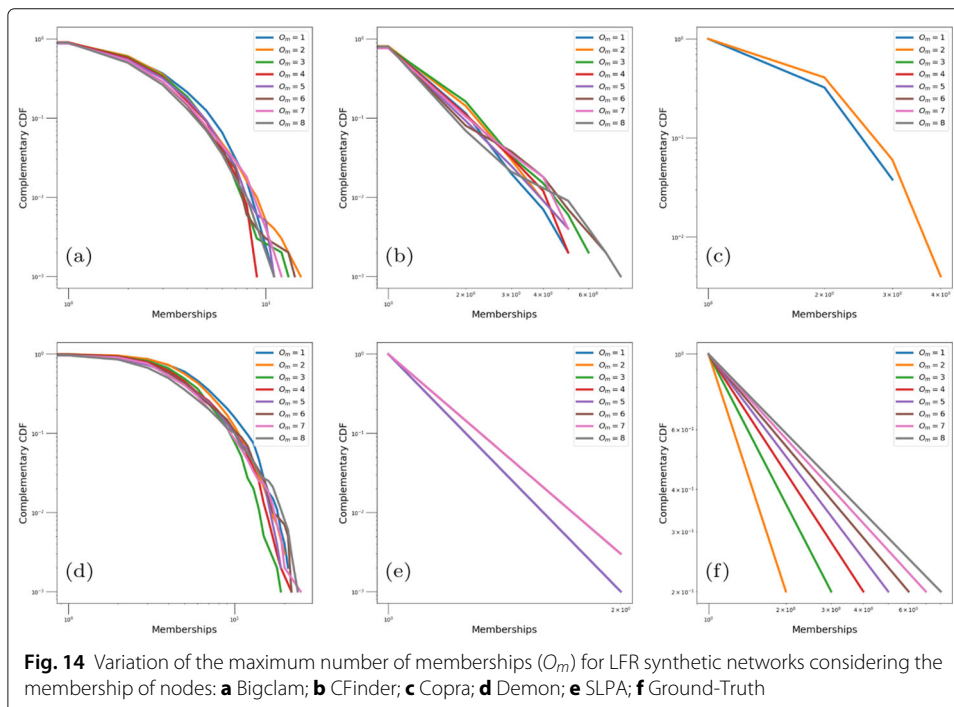
Figures 16, 17, 18 and 19 present the Complementary Cumulative Distribution Function of the sizes of the overlapping regions of the communities identified by each of the methods considered in this work on the LFR benchmark networks with the basic parameter set.
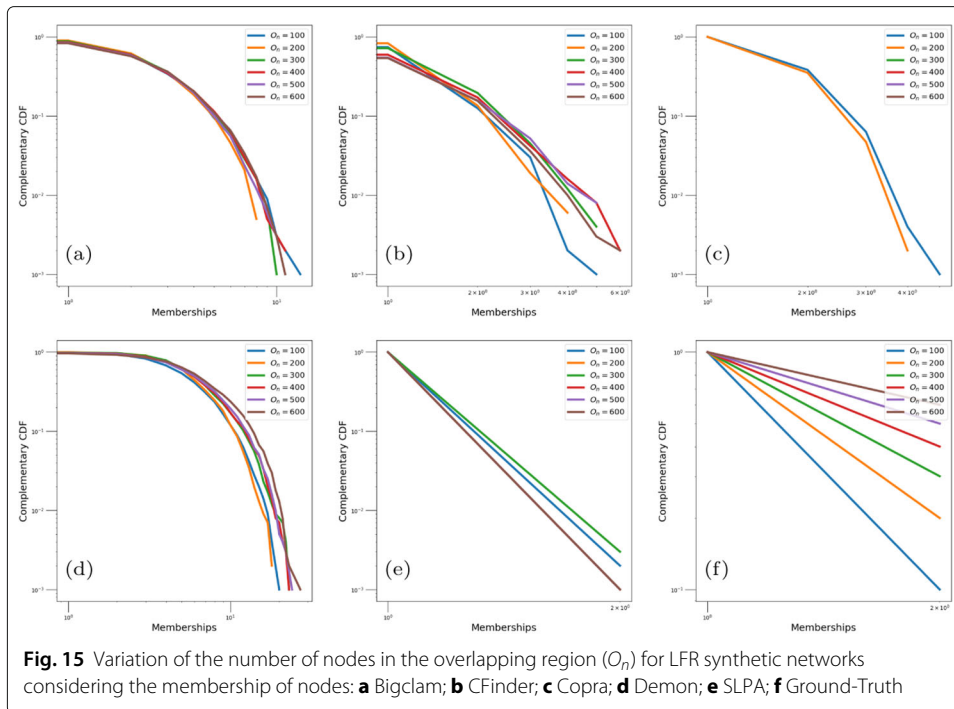
Figure 16 shows that the curves observed for the Ground-Truth communities regarding overlapping size have similar shapes when μ varies, with slight differences in scale. The number of nodes in the overlapping region is very small when compared to the number of nodes in the network in all the experiments considered varying n, O_m and O_n (Figs. 17, 18 and 19). The obvious exception is when O_n is varied (Fig. 19). It is interesting to notice that Bigclam shows a drastic variation on the shape of the distribution for all values of μ when the overlapping size is superior to ≈ 3. Another relevant result is

Fig. 12 Variation of the mixing parameter (μ) for LFR synthetic networks considering the membership of nodes: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 13 Variation of the number of nodes (*n*) for LFR synthetic networks considering the membership of nodes: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
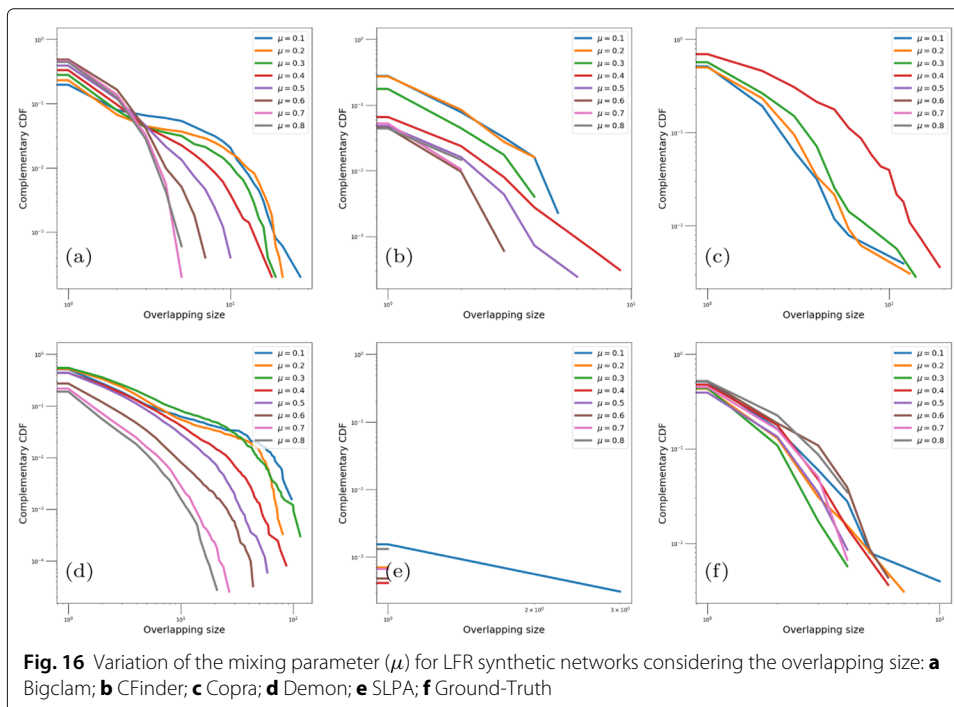
that, again, Bigclam shows very distinct behaviors when $n = 1000$, $n = 2000$ or $n = 3000$ and $n > 3000$. Regarding the overlapping size, except in the experiment when the number of nodes of the network varies, CFinder is able to capture relatively well the shape of the distributions obtained for the Ground-Truth communities. Bigclam and Demon show almost coincident curves when the number of memberships O_m and the number of nodes

Fig. 14 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the membership of nodes: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 15 Variation of the number of nodes in the overlapping region (O_n) for LFR synthetic networks considering the membership of nodes: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
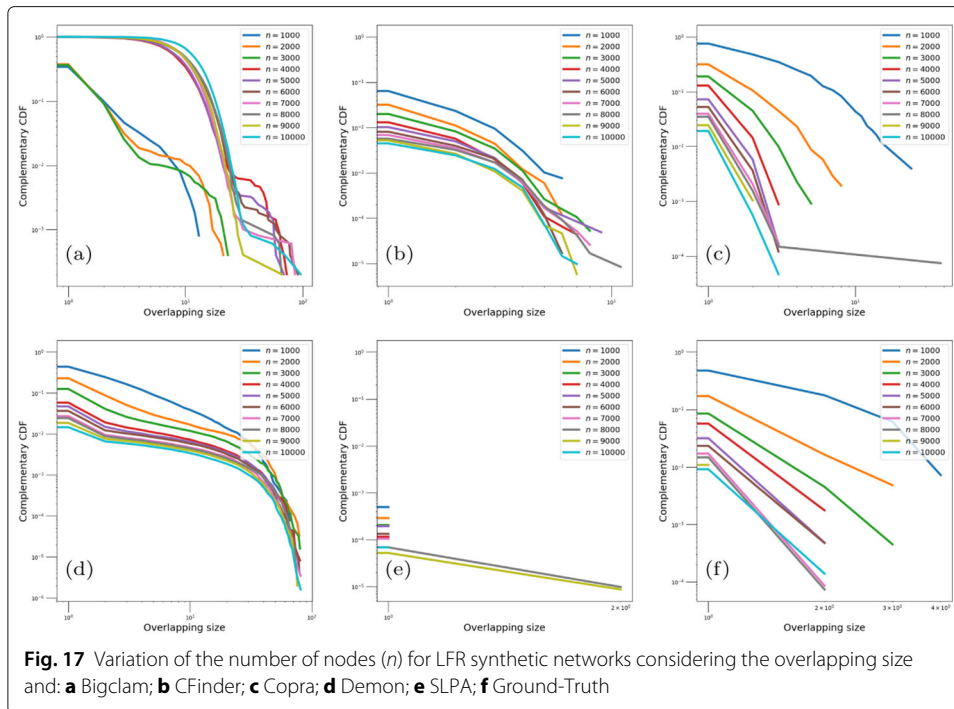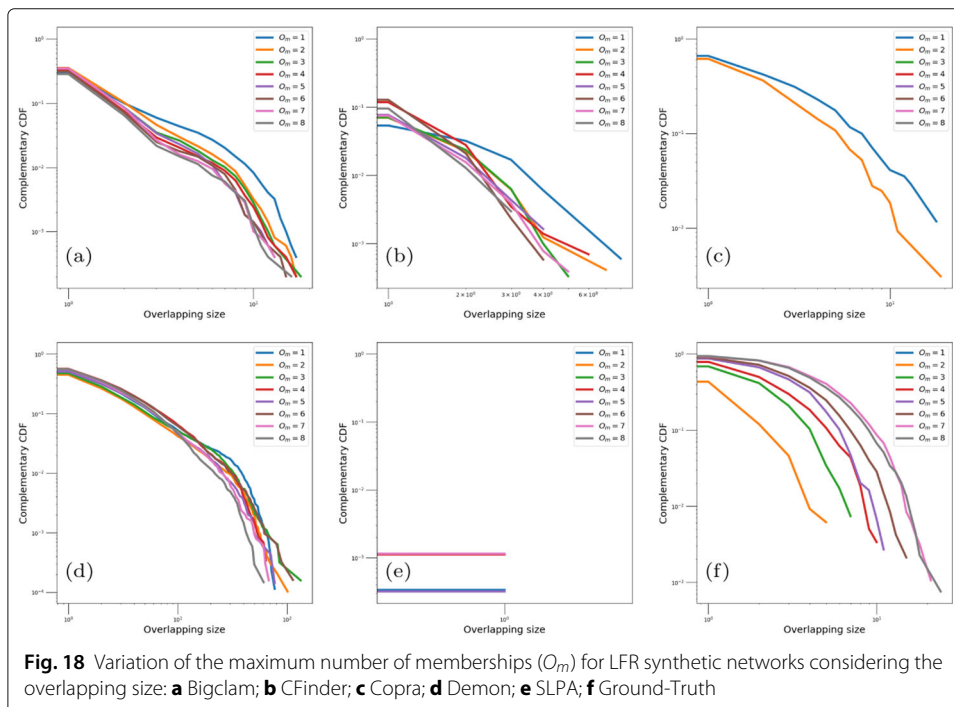
in the overlapping region O_n vary (Figs. 18 and 19), indicating that the methods are not sensitive to variations on overlapping properties of the network.
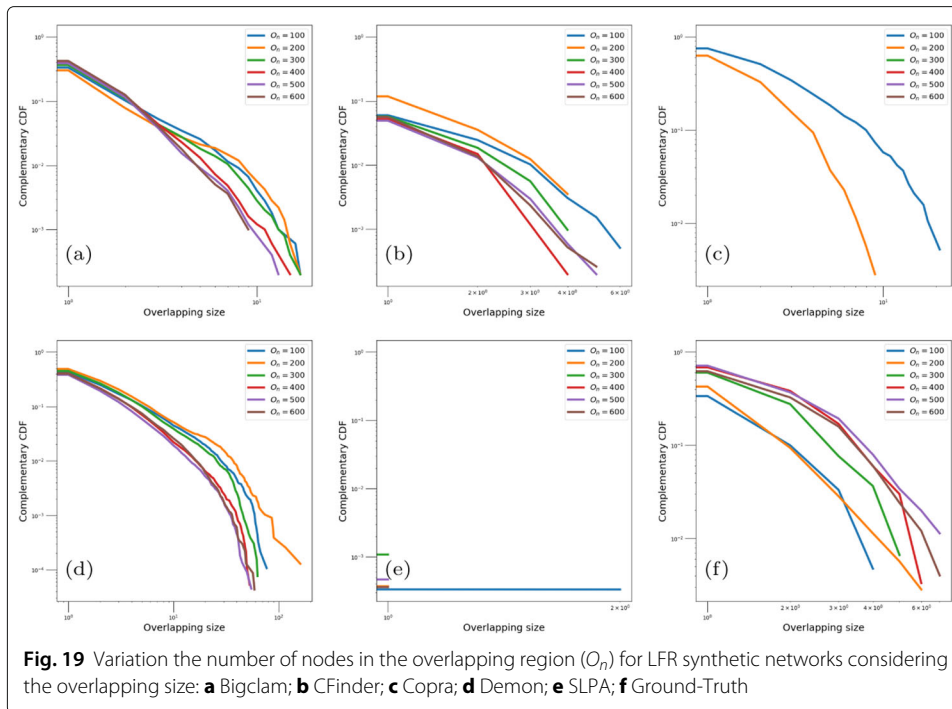
Figures 20, 21, 22 and 23 present the Complementary Cumulative Distribution Function of the sizes of communities identified by each of the methods considered in this work on the LFR benchmark networks with the basic parameter set.

Fig. 16 Variation of the mixing parameter (μ) for LFR synthetic networks considering the overlapping size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 17 Variation of the number of nodes (n) for LFR synthetic networks considering the overlapping size and: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
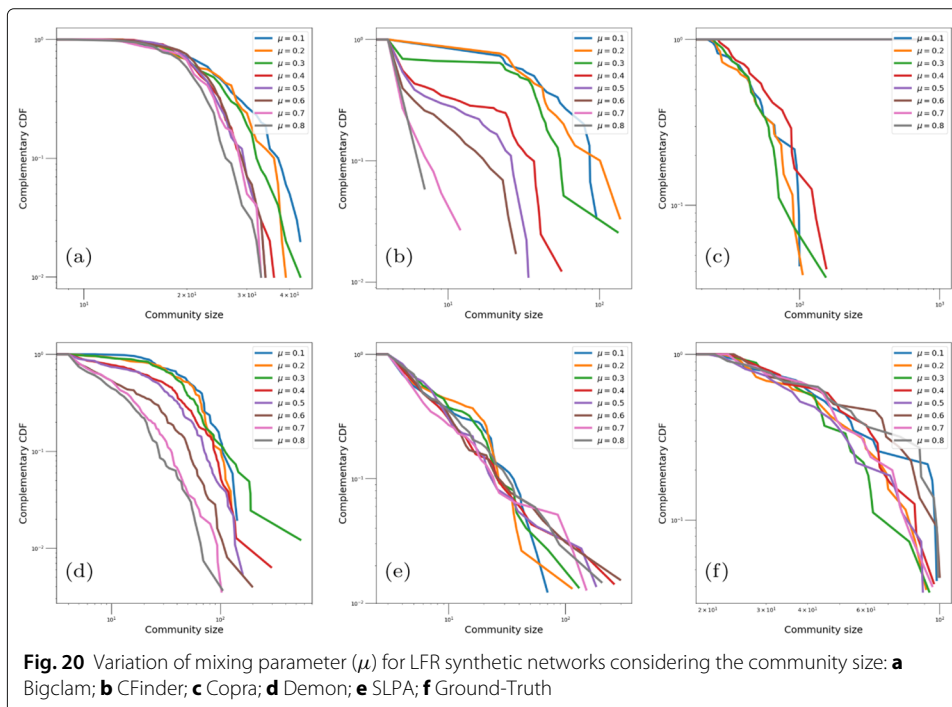
The sizes of the Ground-Truth communities range from 20 and 100, as defined by the parameter set on LFR benchmark. Figures 20, 21, 22 and 23 show that the considered methods are able to identify communities in this range of magnitude. Moreover, the majority of the observed distribution curves are almost overlapped, capturing a remarkable property of the Ground-Truth communities. A recurrent behavior of Bigclam can be highlighted: a drastic change in the distribution when the number of nodes exceeds

Fig. 18 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the overlapping size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 19 Variation the number of nodes in the overlapping region (O_n) for LFR synthetic networks considering the overlapping size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
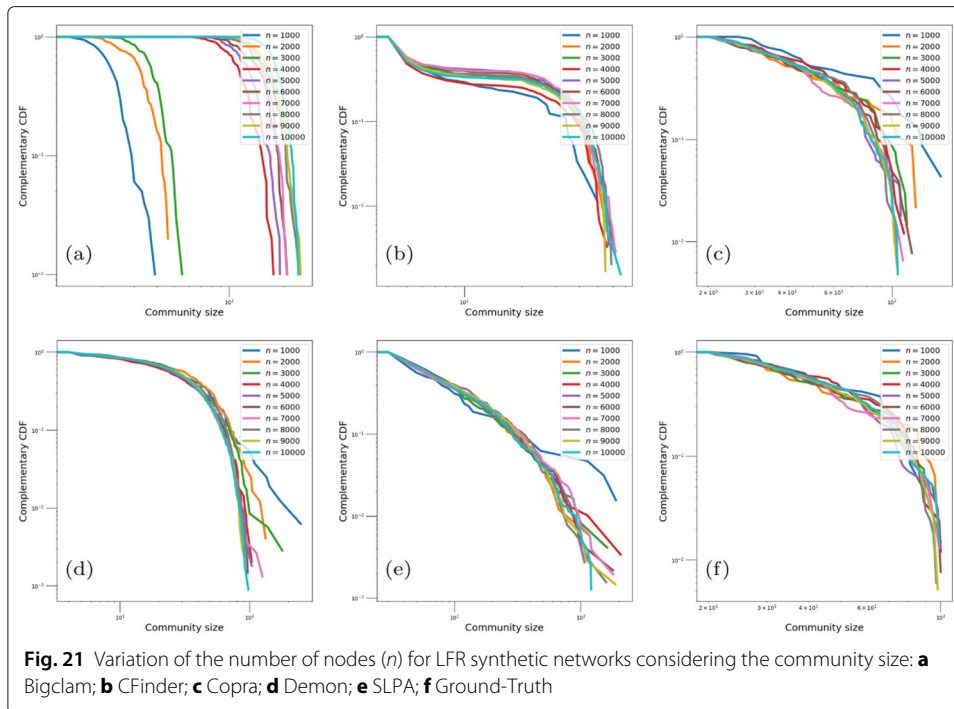
3000. It is interesting to observe that the distribution of communities sizes is one of the most consolidated and studied structural properties in works comprising communities in network.
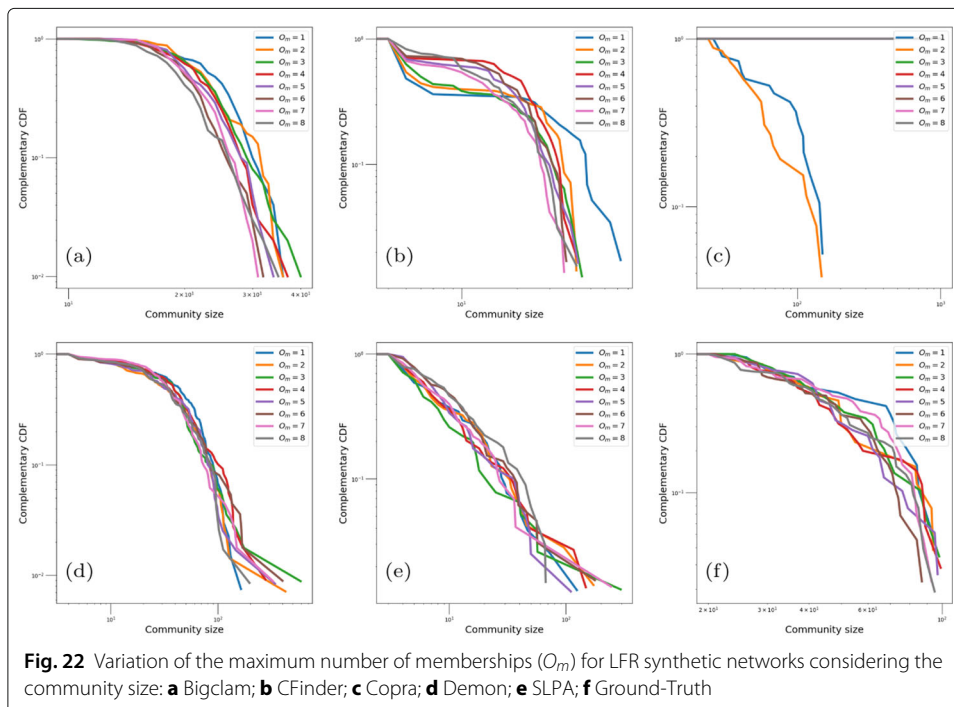
Figures 24, 25, 26 and 27 present the probability of occurrence of an edge between a random pair of nodes as a function of the number of communities that they participate

Fig. 20 Variation of mixing parameter (μ) for LFR synthetic networks considering the community size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 21 Variation of the number of nodes (*n*) for LFR synthetic networks considering the community size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
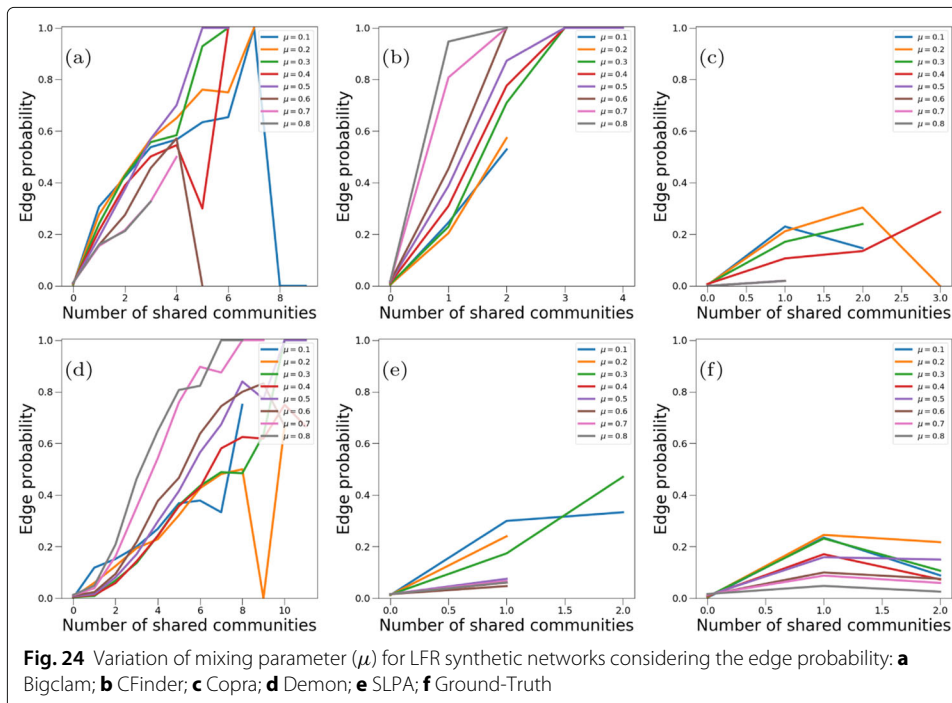
concomitantly. The results also consider each of the methods considered in this work on the LFR benchmark networks with the basic parameter set.

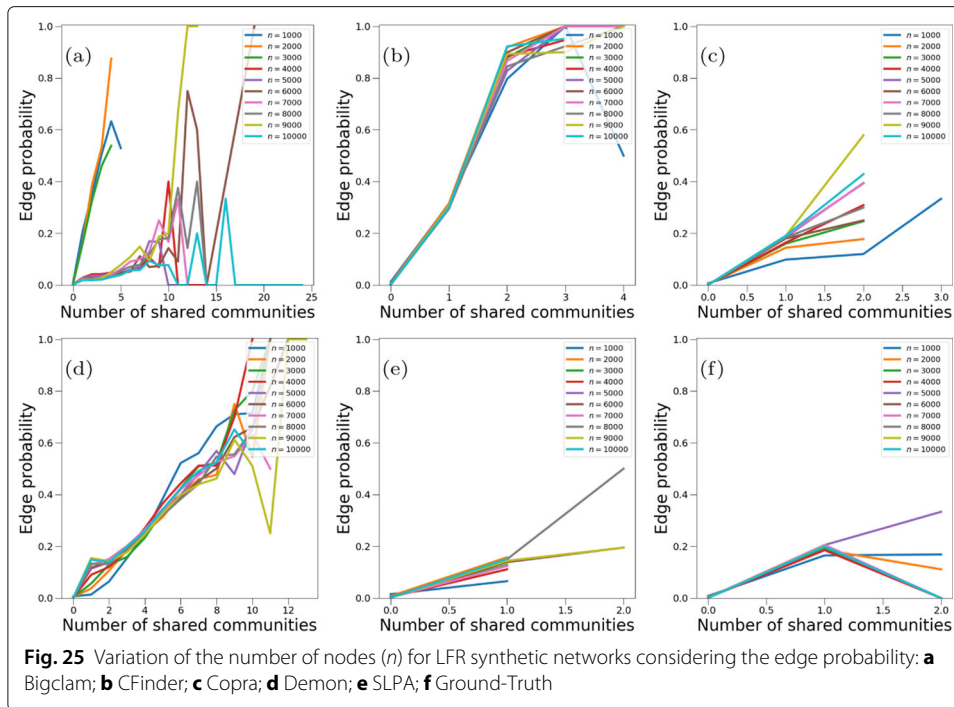From Figs. 24, 25, 26 and 27 it can be observed that the property reported by Yang and Leskovec (2012) that the probability of edges increases with the number of shared communities can not be observed in the Ground-Truth communities

Fig. 22 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the community size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 23 Variation of the number of nodes in the overlapping region (O_n) for LFR synthetic networks considering the community size: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
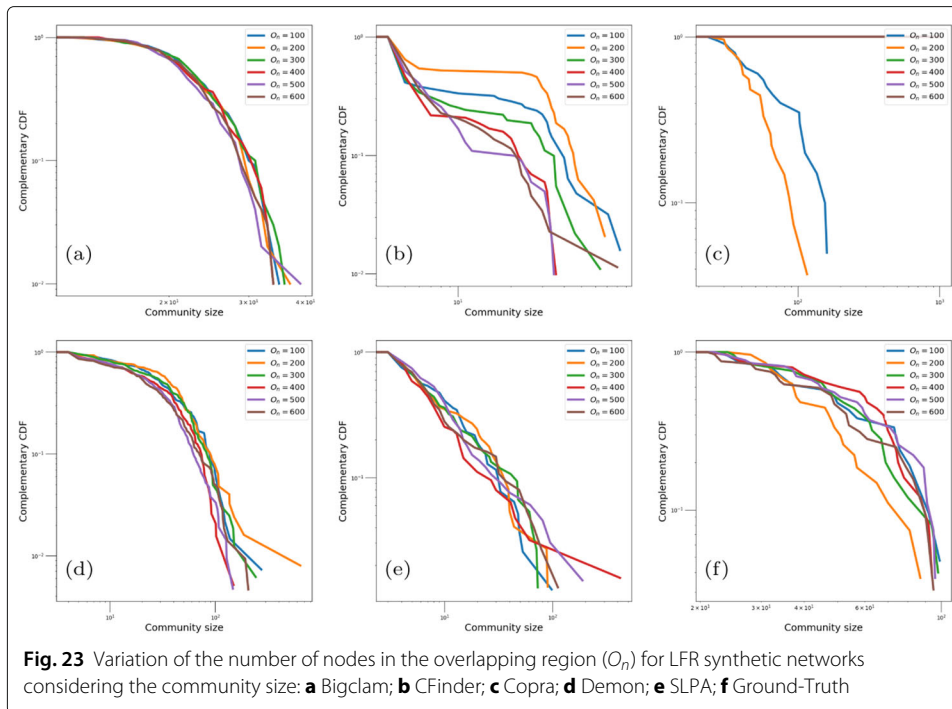
generated by the LFR benchmark. It is particularly observed with specific sets of parameters, but can not be taken as a rule. The increasing in the edge probability is observed for some methods with most parameter sets, such as Bigclam, CFinder and Demon.

Fig. 24 Variation of mixing parameter (μ) for LFR synthetic networks considering the edge probability: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 25 Variation of the number of nodes (*n*) for LFR synthetic networks considering the edge probability: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Appendix C: Detailed comparison of structural properties on real-world networks

Figure 28 shows the Complementary Cumulative Distribute Function (CCDF) of node memberships, i.e., the number of communities in which each node participate for all the identified communities and the Ground-Truth communities for each network.

Fig. 26 Variation of the maximum number of memberships (O_m) for LFR synthetic networks considering the edge probability: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth

Fig. 27 Variation of the number of nodes in the overlapping region (O_n) for LFR synthetic networks considering the edge probability: **a** Bigclam; **b** CFinder; **c** Copra; **d** Demon; **e** SLPA; **f** Ground-Truth
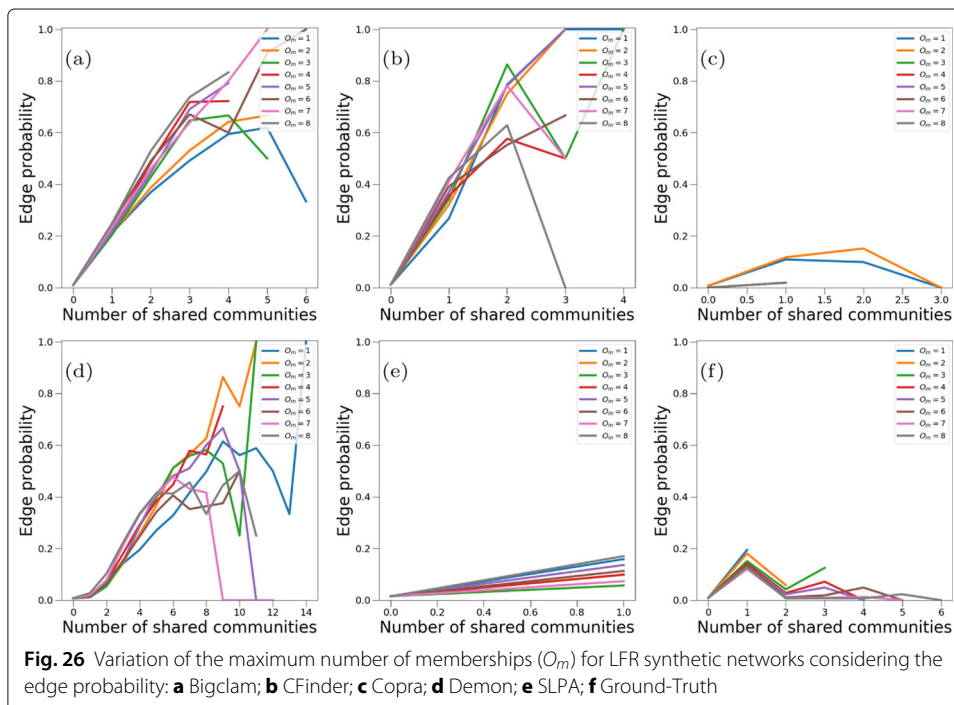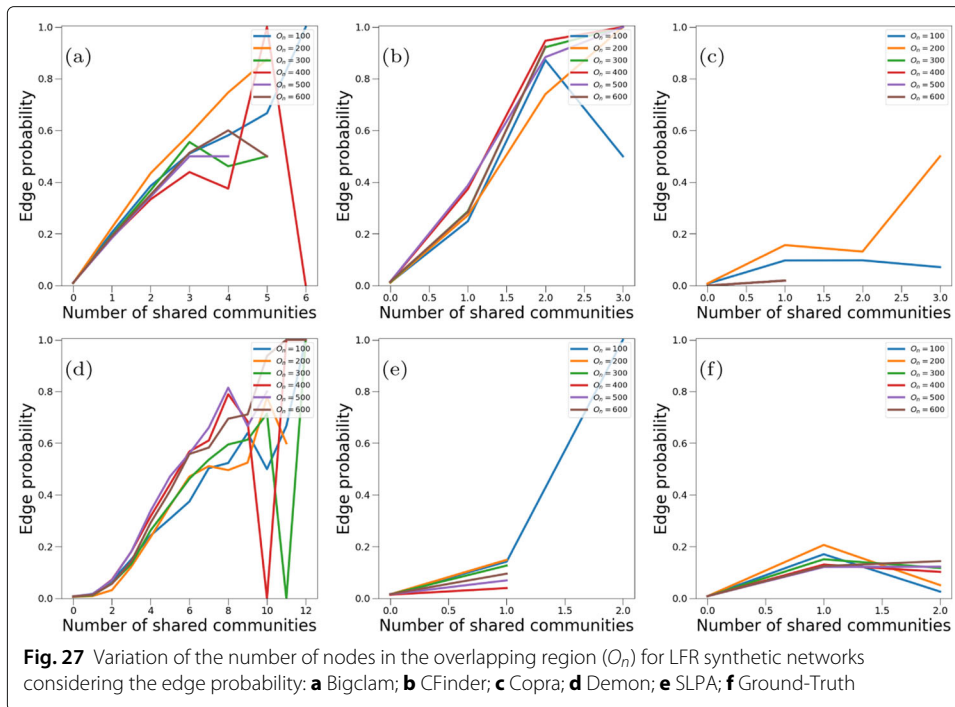
Figure 28 allows us to clearly see that except for COPRA and SLPA, all the considered methods do not necessarily assign a community to each node and some nodes can be left with no community and, thus, the probability of a node to belong to at least one community can be less than one. This behavior particularly stands out for CFinder, that severely underestimates the number of nodes that belong to one community. It is also possible to observe that the Complementary Cumulative Distribute Functions (CCDFs) for the Ground-Truth communities show an almost linear behavior (in $log - log$ scale), which was not captured by any of the studied methods. It is important to note that most methods depend on the definition of parameters to execute, which may strongly affect the community structure and a deeper investigation should be performed for each method in order to draw more conclusive observations.

Fig. 28 Memberships of the nodes identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** CitHepTh; **h** Email; **i** Keys

Fig. 29 Community size identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** CitHepTh; **h** Email; **i** Keys
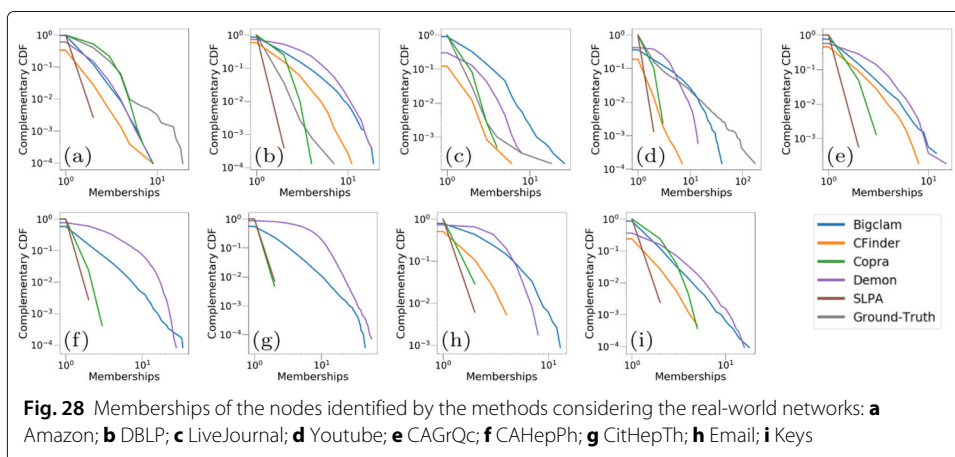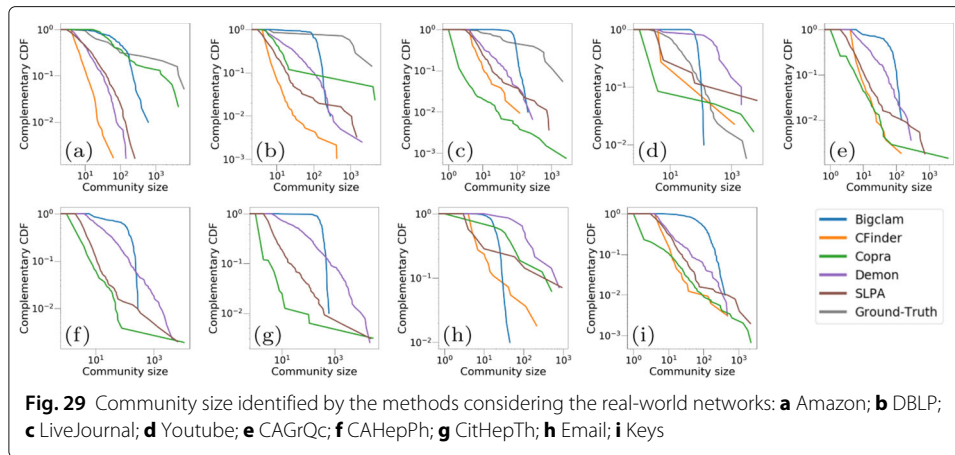
The CCDF for the sizes of communities identified by the methods were also considered and the results are presented in Fig. 29.

Figure 29 also shows that, despite for differences in scale, the Ground-Truth communities shows an almost linear behavior (in $log - log$ scale) in a wide range of the distribution, which was not captured by any of the methods. First, it is worth to note that the communities obtained by COPRA are very similar in size to the Ground-Truth. However, this is atypical and such similarity can not be observed in the other scenarios. It is interesting to notice that the shapes of the curves observed for Demon and, more clearly, for Bigclam is very consistent for most networks, though very different from the Ground-Truth, suggesting that the structural properties of the communities may be highly affected by the mechanisms used by the methods and not by the network itself, which must be better understood with further investigation on the methods.

The size of the overlapping region, i.e., the number of nodes at each region on the intersection of communities, was also investigated and the CCDFs for these quantities are presented in Fig. 30.

The methods do not show a clear and constant pattern for the distribution of overlapping size and the curves observed for a certain method are very distinct from one network to another. It is worth to notice that, for most methods, the occurrence of overlapping regions is very unlikely. The exception is Bigclam, which

Fig. 30 Overlapping size identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** CitHepTh; **h** Email; **i** Keys
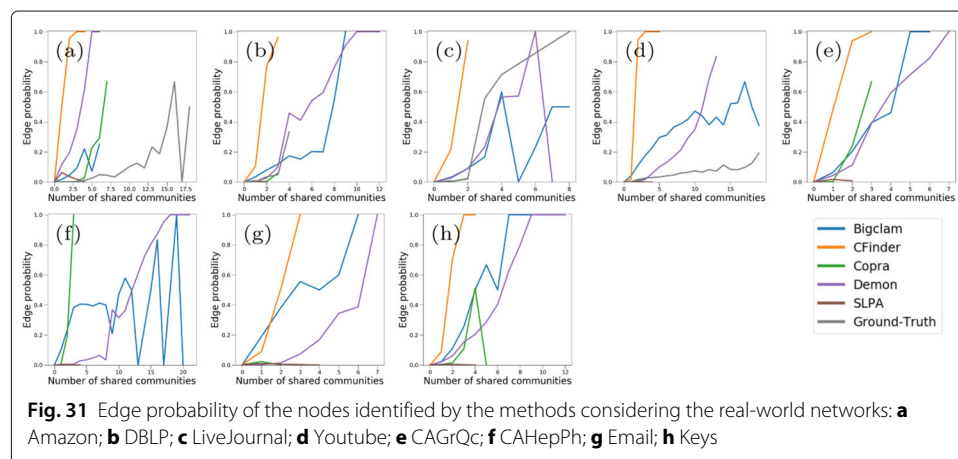
identifies communities with relevant overlapping regions, specially in the Ground-Truth networks. Yet considering the networks with Ground-Truth, the number of small overlapping regions is very high. Moreover, for three Ground-Truth networks (Amazon, DBLP and Live Journal), the methods underestimate the sizes of overlapping regions in all the range of the distribution. In the communities identified by COPRA, SLPA and CFinder the occurrence of nodes in the overlapping region is extremely rare.

Figure 31 shows the probability of an edge between two nodes to exist in function of the number of communities shared by them.

First, it is important to mention that this experiment is very memory consuming and it could not be performed for CitHepTh, which occurs because the number of communities that each pair of nodes share must be computed in order to compute the probability of the existence of an edge. Particularly in a large community c with, say, n_c nodes, each node shares at least one community with $n_c - 1$ other nodes and a data structure with at least $n_c \times n_c$ entries must be stored.

From Fig. 31 it is possible to notice that, in fact, there is a positive correlation between the number of shared nodes and the edge probability in Ground-Truth networks. This behavior is captured by Bigclam, CFinder and Demon although the edge probability does not show a clear pattern in all. Copra and SLPA mostly fail in capturing dense overlapping regions, as stated by Yang and Leskovec (2012) and it is important to notice that CFinder and Demon are able to identify dense overlapping regions but largely overestimate the probability of an edge to exist. Overall, it is difficult to make more conclusive observations. But when the other results are considered, a clearer view of the results presented in Fig. 31 is possible. For instance, in SLPA, COPRA and CFinder the occurrence of large overlapping regions and the presence of nodes in several communities is rare. Thus, it is expected that the probability of an edge to exist in the overlapping region does not show a clear pattern among the different networks.

Fig. 31 Edge probability of the nodes identified by the methods considering the real-world networks: **a** Amazon; **b** DBLP; **c** LiveJournal; **d** Youtube; **e** CAGrQc; **f** CAHepPh; **g** Email; **h** Keys

Author details
[1]Federal University of São João del-Rei, São João del-Rei MG, Brasil. [2]COPPE, Federal University of Rio de Janeiro, P.O. Box 68506 Rio de Janeiro, RJ, Brasil.

References
Almeida H, Guedes D, Meira W, Zaki Mj (2011) Is there a best quality metric for graph clusters? In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science, vol 6911. Springer, Berlin. pp 44–59

Amelio A, Pizzuti C (2014) Overlapping community discovery methods: A survey. In: Ögüdücü SG, Etaner-Uyar AS (eds). Social Networks: Analysis and Case Studies. Lecture Notes in Social Networks. Springer, Vienna. pp 105–125

Cherifi H, Palla G, Szymanski BK, Lu X (2019) On community structure in complex networks: challenges and opportunities. Appl Netw Sci 4(117):1–35

Chykhradze K, Korshunov A, Buzun N, Pastukhov R, Kuzyurin NN, Turdakov D, Kim H (2014) Distributed generation of billion-node social graphs with overlapping community structure. In: Contucci P, Menezes R, Omicini A, Poncela-Casasnovas J (eds). Complex Networks V - Proceedings of the the 5th Workshop on Complex Networks CompleNet 2014, Bologna, Italy, March 12-14, 2014. Studies in Computational Intelligence. vol. 549. Springer, Cham. pp 199–208

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111

Collins LM, Dent CW (1988) Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. Multivar Behav Res 23 2:231–42

Coscia M, Rossetti G, Giannotti F, Pedreschi D (2014) Uncovering hierarchical and overlapping communities with a local-first approach. ACM Trans Knowl Discov Data 9(1):6–1627

Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theory Exp 2005:09008

Dao V-L, Bothorel C, Lenca P (2020) Community structure: A comparative evaluation of community detection methods. Netw Sci 8:1–41

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. PNAS 104:36

Granovetter MS (1973) The Strength of Weak Ties. Am J Sociol 78(6):1360–1380

Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 12(10):103018

Gregory S (2011) Fuzzy overlapping communities in networks. J Stat Mech Theory Exp 2011(02):02017

Harenberg S, Bello G, Gjeltema L, Ranshous S, Harlalka J, Seay R, Padmanabhan K, Samatova N (2014) Community detection in large-scale networks: a survey and empirical evaluation. WIREs Comput Stat 6(6):426–439

Hric D, Darst RK, Fortunato S (2014) Community detection in networks: Structural communities versus ground truth. Phys Rev E 90:062805

Jebabli M, Cherifi H, Cherifi C, Hamouda A (2018) Community detection algorithm evaluation with ground-truth data. Physica A Stat Mech Appl 492:651–706

Kudelka M, Ochodkova E, Zehnalova S, Plesnik J (2019) Ego-zones: non-symmetric dependencies reveal network groups with large and dense overlaps. Appl Netw Sci 4(81):1–49

Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure of complex networks. New J Phys 11(3):033015

Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E Stat Nonlinear Soft Matter Phys 78 4 Pt 2:046110

Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLOS ONE 6(4):1–18. https://doi.org/10.1371/journal.pone.0018961

Leskovec J, Lang KJ, Mahoney M (2010) Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web. WWW '10. ACM, New York, NY, USA. pp 631–640

Li W, Xie J, Xin M, Mo J (2018) An overlapping network community partition algorithm based on semi-supervised matrix factorization and random walk. Expert Syst Appl 91:277–285

McDaid AF, Greene D, Hurley N (2011) Normalized mutual information to evaluate overlapping community finding algorithms. https://doi.org/1110.2515. Accessed 06 Jan 2020

Newman MEJ (2006) Modularity and community structure in networks. PNAS 103(23):8577–8582

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlinear Soft Matter Phys 69(2):026113

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):1–16

Nicosia V, Mangioni G, Carchiolo V, Malgeri M (2009) Extending the definition of modularity to directed graphs with overlapping communities. J Stat Mech Theory Exp 2009(03):03024

Orman GK, Labatut V, Cherifi H (2012) Comparative evaluation of community detection algorithms: a topological approach. J Stat Mech Theory Exp 2012(08):08001

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818

Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. Sci Adv 3(5):e1602548

Peixoto TP (2020) Revealing consensus and dissensus between network partitions. https://doi.org/2005.13977

Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci U S A 101(9):2658–2663

Raghavan N, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Rhouma D, Romdhane LB (2014) An efficient algorithm for community mining with overlap in social networks. Expert Syst Appl 41(9):4309–4321

Schaeffer S (2007) Graph clustering. Comput Sci Rev 1(1):27–64

Schaub MT, Delvenne J-C, Rosvall M, Lambiotte R (2017) The many facets of community detection in complex networks. Appl Netw Sci 2(1):4

Shen H-W (2013) Detecting the Overlapping and Hierarchical Community Structure in Networks. Springer, Berlin, Heidelberg

Vieira V, Xavier C, Evsukoff A (2019) Comparing the community structure identified by overlapping methods. In: Proceedings of Complex Networks and Their Applications 2019. Complex Networks 2019, vol. 281. Springer, Cham. pp 262–273

Xie J, Kelley S, Szymanski BK (2011) Overlapping community detection in networks: the state of the art and comparative study. CoRR 1110.5813

Xie J, Kelley s, Szymanski BK (2013) Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study. ACM Comput Surv 45(4):35. ISSN:0360-0300, Association for Computing Machinery, New York

Xin Y, Yang J, Xie Z-Q, Zhang J-P (2015) An overlapping semantic community detection algorithm base on the arts multiple sampling models. Expert Syst Appl 42(7):3420–3432

Yang J, Leskovec J (2012) Community-affiliation graph model for overlapping network community detection. In: Proceedings of the 2012 IEEE 12th International Conference on Data Mining. ICDM '12. IEEE Computer Society, Washington, DC, USA. pp 1170–1175

Yang J, Leskovec J (2013) Overlapping community detection at scale: A nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM '13. ACM, New York, NY, USA. pp 587–596

Yang J, Leskovec J (2014) Structure and overlaps of ground-truth communities in networks. ACM Trans Intell Syst Technol 5(2):1–35

Zhou L, Lü K, Yang P, Wang L, Kong B (2015) An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory. Expert Syst Appl 42(24):9634–9646

Publisher's Note