Overcoming Bias in Community Detection Evaluation

Jeancarlo C. Leão, Alberto H. F. Laender and Pedro O. S. Vaz de Melo

[1] Instituto Federal do Norte de Minas Gerais (IFNMG), Brazil
jeancarlo.leao@ifnmg.edu.br
[2] Universidade Federal de Minas Gerais, Brazil
{laender,olmo}@dcc.ufmg.br

Abstract. Community detection is a key task to further understand the function and the structure of complex networks. Therefore, a strategy used to assess this task must be able to avoid biased and incorrect results that might invalidate further analyses or applications that rely on such communities. Two widely used strategies to assess this task are generally known as *structural* and *functional*. The structural strategy basically consists in detecting and assessing such communities by using multiple methods and structural metrics. On the other hand, the functional strategy might be used when ground truth data are available to assess the detected communities. However, the evaluation of communities based on such strategies is usually done in experimental configurations that are largely susceptible to biases, a situation that is inherent to algorithms, metrics and network data used in this task. Furthermore, such strategies are not systematically combined in a way that allows for the identification and mitigation of bias in the algorithms, metrics or network data to converge into more consistent results. In this context, the main contribution of this article is an approach that supports a robust quality evaluation when detecting communities in real-world networks. In our approach, we measure the quality of a community by applying the structural and functional strategies, and the combination of both, to obtain different pieces of evidence. Then, we consider the divergences and the consensus among the pieces of evidence to identify and overcome possible sources of bias in community detection algorithms, evaluation metrics, and network data. Experiments conducted with several real and synthetic networks provided results that show the effectiveness of our approach to obtain more consistent conclusions about the quality of the detected communities.

Categories and Subject Descriptors: H.2 [**Data Mining**]: Miscellaneous; G.2.2 [**Graph Theory**]: Miscellaneous; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

Keywords: Community Structure, Quality Evaluation, Bias, Ensemble Approach, Triangulation Method

1. INTRODUCTION

The community detection problem has been much studied in the context of social networks due to its wide application in many domains, giving rise to many methods to address it [Almeida et al. 2012; Fortunato 2010; Gandica et al. 2020; Yang and Leskovec 2015]. However, one of the major challenges related to this problem is the difficulty to evaluate the detected communities with respect to the various methods proposed in the literature. Part of this difficulty lies on the fact that there is still no universally accepted definition for the concept of community [Fortunato 2010], as well as for what we understand as being the quality of a community [Hric et al. 2014]. Besides, the evaluation of such communities is usually carried out by using experimental configurations greatly susceptible to biases, which are inherent to the algorithms, metrics and network data used in this task [Jebabli et al. 2018; Leão et al. 2019; Liu et al. 2020]. Moreover, this kind of evaluation is, in general, carried out without explicitly dealing with such biases, which may lead to inconsistent results.

In order to illustrate this problem, let us consider the example shown in Figure 1. Specifically, Figure 1a shows a social network formed by 34 members (vertices) of a karate club interconnected by edges representing interactions between them outside the club. Originally, this network was divided

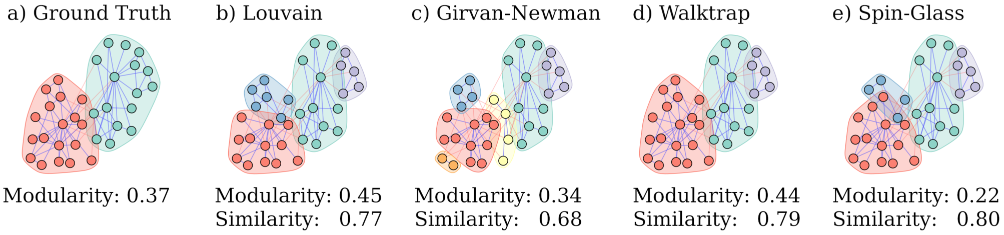

Fig. 1. Example of how bias can affect community detection in social networks.

into two non-overlapping communities labeled by Zachary [1977] with 16 and 18 members, respectively, each one supervised by a specific instructor. Figure 1b, on the other hand, shows the communities detected in this same network by the Louvain algorithm [Blondel et al. 2008], a well known and very effective community detection algorithm. Note that the community structure revealed by the Louvain algorithm is different from those presented in Figures 1c-e, which were respectively obtained by the Girvan-Newman [Newman and Girvan 2004], Walktrap [Pons and Latapy 2005] and Spin-Glass [Reichardt and Bornholdt 2006] algorithms, all of them also considered very effective. Here, we raise the possibility that the bias of each heuristic algorithm interferes with its final results, making them different from each other and from the absolute optimal theoretical result, not necessarily present in Figure 1.

Thus, let us check the pieces of evidence present in this example in order to reach a consensus on which algorithm produces the best quality communities. First, the modularity metric indicates that the communities shown in Figure 1b present a better quality with respect to their modular structure (i.e., they present the highest modularity value). On the other hand, when comparing the detected communities with the ground truth (Figure 1a) using the Rand Index similarity metric [Rand 1971], it indicates that the network in Figure 1d is the best one, since its communities are among the most modular ones, being also more similar to those shown in Figure 1a, even though there is not a perfect match. However, due to its own bias this similarity metric scored better the communities in Figure 1e, even though they show more visual differences with respect to the ground truth than the communities in Figure 1d. Finally, due to some specific bias in the original network data or in the ground truth data, the Girvan-Newman algorithm has not been able to identify good communities (Figure 1c), as shown by the values of the two metrics considered. Note that, although the modularity value of these communities is the one closest to the ground truth's, these two sets of communities are less structurally similar, since the number of subgraphs obtained by this algorithm is the largest among all networks.

In face of these pieces of evidence pointing to opposite directions with respect to the quality of the communities in our example, a question arises on which one presents the best structure and what causes this divergence. We could try to obtain a consensual result, but this would be inadequate, because hypotheses about bias were not tested. In addition, evidence obtained from visual inspection is only viable on very small networks such as the ones considered in Figure 1. Nevertheless, our assumption is that all pieces of evidence provide a significant amount of information about the network structure that, combined with other pieces of evidence on the main sources of bias, results in a consistent decision on the quality of the detected communities.

To the best of our knowledge, there is no comprehensive evaluation approach that, considering multiple strategies, is able to identify which one provides the best interpretation. More importantly, as we detail later, due to their own biases it is not always possible to find a consensus among different metrics and community detection algorithms on which community structure has the highest quality. This requires a cross-checking approach involving at least three distinct evaluation strategies to indicate a consensus and estimate a possible bias with respect to the quality of the revealed communities.

Table I. Methods for community detection.

Main Method	Algorithm	ξ	References
Modularity maximization	Louvain Modularity (LM)	D	[Blondel et al. 2008]
	Greedy Optimization of Modularity (GM)	D	[Clauset et al. 2004]
	Leading Eigenvector (LE)	D	[Newman 2006]
Dynamic process	Label Propagation (LP)	N	[Raghavan et al. 2007]
	Spin-glass (SG)	N	[Reichardt and Bornholdt 2006]
Removal of edges between communities	Girvan–Newman (GN)	D	[Newman and Girvan 2004]
Node closeness given by random walks	Walktrap (WT)	N	[Pons and Latapy 2005]
	Infomap (IM)	N	[Rosvall and Bergstrom 2011]

ξ: State model (D-Deterministic/N-Non deterministic).

A strategy generally employed to improve methods, measures, and data reliability and validity is triangulation[1], which consists in using different approaches for measuring the same characteristic [Brender 2006]. Note that, with only one measure of a specific characteristic, the error and biases inherent in that measure are confounded with the characteristic itself. Thus, when measuring different aspects using different metrics, it is possible to decide which one can bring into better focus the characteristic of interest. Based on this idea, the main contribution of this article is a robust approach for community quality evaluation that allows one to obtain results less prone to bias when detecting communities in synthetic and real networks.

Thus, given a network, its set of ground truth communities and a set of its communities to be evaluated, our approach allows one to overcome biases in network data, detection algorithms and evaluation metrics by using distinct evaluation strategies when analyzing the quality of such communities. For this, each strategy must strongly highlight a distinct aspect of a community's quality in addition to considering multiple and diverse detection methods, metrics and network datasets. For example, in Figure 1 the structural and functional aspects of the communities are represented, respectively, by their modularity and similarity with the respective ground truths. Notice that, for our purpose, the choice of the best metrics, detection methods and ground truth data is not important, since we are not trying to identify the best existing community, but the best one among those being compared. Thus, when there is a large variety of metrics and methods available for this task, they can be pre-selected based on some criteria of relevance and diversification.

The rest of this article is organized as follows. Section 2 reviews related work. Section 3 describes our approach for community quality evaluation. Then, Section 4 analyzes the experimental results obtained by applying our proposed approach to real and simulated networks. Finally, Section 5 presents our conclusions and some considerations for future work.

2. RELATED WORK

Although community detection has become one of the most popular and best-studied research topics in network science [Fortunato 2010; Gandica et al. 2020; Hric et al. 2014; Kivelä et al. 2014; Leão et al. 2018; Vieira et al. 2020; Zhao 2017], the problem of validating the quality of a community derived from a real network has not received the due attention in the literature, since there is no consensus on what is meant by a good community [Yang and Leskovec 2015]. Moreover, there are several different

[1] According to O'Donoghue and Punch [2003], triangulation is a "method of cross-checking data from multiple sources to search for regularities in the research data."

definitions of what a community is, which has resulted in many distinct approaches to community detection [Coscia 2019; Ghasemian et al. 2020]. For example, the algorithms listed in Table I usually extract different communities from a given network, which are in general considered of good quality by distinct metrics.

In a previous work [Leão et al. 2018], we analyzed the similarity of communities detected by distinct methods in order to identify the set of algorithms that tend to produce more similar results and those that provide more distinct ones, when compared by multiple similarity metrics and considering a given network. Doing so, we aimed to increase the scope and the consistency of the evaluation process when detecting communities. More recently, Coscia [2019] proposed an analysis essentially similar to ours also aimed at classifying community detection algorithms according to the similarity of their results. For this, he considers how many times two algorithms provide similar communities. Besides the similarity of the results provided, other criteria have been used to distinguish community detection algorithms. For example, Abrahao et al. [2012], Yang and Leskovec [2015], and Ghasemian et al. [2019] consider the structural properties of the detected communities for such a distinction.

Although the analysis of the detected communities is commonly used to distinguish the detection algorithms, we have not identified any work that uses the variability of the characteristics of these communities to deal with bias. In fact, in general, such works show some bias when evaluating the quality of a community. Next, we describe some works that have proposed methods that make it possible to deal with some types of bias when assessing the quality of their detected communities.

Bias in Community Detection Algorithms. Heuristic algorithms for community detection often find communities that are systematically biased in the sense that their results might be different than the optimum objective function chosen for this specific task [Abrahao et al. 2012; Leskovec et al. 2010; Peel et al. 2017]. Moreover, distinct goals for community detection might lead to totally distinct objective functions [Coscia 2019; Gandica et al. 2020; Ghasemian et al. 2020]. In addition, the intensity of these differences might vary from a network to another due to detection methods being sensitive to different community structures, topologies, and types or instances of a network [Coscia et al. 2011; Ghasemian et al. 2020; Leskovec et al. 2010].

Note that detection algorithms are expected to introduce in their final results the same bias of the metrics they use in their optimization functions [Jebabli et al. 2018; Peel et al. 2017]. A popular case is that of the modularity, conductance and coverage metrics that have strong structural biases that make them favor smaller clusters and whose maximization is aimed at by most detection algorithms [Almeida et al. 2012; Jebabli et al. 2018].

In this context, different approaches have been proposed with the aim of reducing the effect of biases and improving the detection of communities. For instance, Lancichinetti et al. [2012] show how to combine the communities obtained from various detection methods into a consensual one, statistically more stable and with a better structure. It is worth noting that this approach seeks consensus only on the structural aspect of the communities and does not explore the different results produced by multiple methods to identify, analyze and consider any other kind of bias.

Bias in Evaluation Metrics. Existing works usually consider only specific aspects to assess the quality of a community, for example by measuring the structure derived from its connectivity (structural aspect) [Newman and Girvan 2004] or by measuring its similarity with a ground truth community (functional aspect) [Peel et al. 2017] to finally performing a comparison with a good baseline score [Hric et al. 2014]. Regarding the structural aspect, community detection algorithms are usually evaluated by correlated metrics or by the same metrics used by their optimization function, such as modularity [Fortunato 2010; Yang and Leskovec 2015], which can produce some biased results. Another type of bias associated with a metric may be an incorrect score systematically attributed according to a specific characteristic present in the data. For example, popular quality metrics present strong bias when applied to networks with different sizes or number of clusters [Almeida et al. 2012; Coscia et al. 2011; Pons and Latapy 2005]. In this context, there is no best metric to assess the quality of a community [Almeida et al. 2012].

In particular cases, it is possible to assess the functional aspect of a detected community by comparing it with its respective ground truths [Hric et al. 2014; Peel et al. 2017; Zaki and Meira Jr. 2014]. For Fortunato et al. [2010], this kind of evaluation involves the definition of a criterion to establish how "similar" is a community provided by an algorithm with respect to the ground truth. To address this, the authors adopt some specific metrics such as *Rand Index* and *Normalized Mutual Information.* Note that similarity metrics can provide different scores for a same network [Leão et al. 2019], in some cases even outliers. In addition, such metrics are also susceptible to producing biased results when comparing communities with specific characteristics [Amelio and Pizzuti 2015; Lei et al. 2017].

To deal with a specific type of bias associated with a metric, works like those carried out by Liu et al. [2019], Gösgens et al. [2020] and Labatut [2015] usually propose a new metric or enhance an existing one. However, we have not found in the literature any community quality or similarity metric that does not present any bias, i.e., that is capable of dealing with multiple types of bias or does not present any vulnerability to bias originating from other sources, such as detection algorithms or network data. In this sense, our contribution relies on mitigating bias in the evaluation of communities by using multiple existing metrics.

Bias in Data. In addition to the specific biases of each algorithm introduced in the data of the detected communities [Leskovec et al. 2010], there may also be some bias in data from the network [Jebabli et al. 2018; Leão et al. 2018] and from its ground truth data [Leão et al. 2019]. In a previous work, Rocha et al. [2017] described how the representation of real temporal interactions can result in biased data. More recently, Leão et al. [2018] proposed a solution that avoids the biased data problem by directly removing noise produced by sporadic relationships[2] found in a social network. They also showed that this kind of noise may cause errors when detecting communities. Note that datasets with good ground truths are rare [Yang and Leskovec 2015] and might not be always suitable for the type of community we want to analyze [Leão et al. 2018].

Final Comments. In addition to the need of overcoming the specific limitations discussed earlier, some studies report that it is also important to consider multiple strategies when assessing the quality of a community [Dao et al. 2020; Jebabli et al. 2018; Leão et al. 2019]. In addition, Dao et al. [2020] also systematize a process for obtaining a conclusion on the structural and functional quality of the communities, in this case, by applying the multiple criteria decision making process. Jebabli et al. [2018] also highlight the distinction between different community detection algorithms, however limited to a single comparison criterion that is based on the distribution of the number of communities.

Thus, by analyzing the above works, we identified the following main contributions of our proposed approach to assess the quality of a community: (i) it deals with different types of bias inherent in data, metrics and algorithms; (ii) it involves different aspects of the quality of a community through the use of multiple strategies; (iii) it uses multiple criteria to distinguish among different community detection algorithms and metrics used to assess them; (iv) it seeks a consensual and consistent decision; (v) it provides requirements for an evaluation maturity model; and, finally, (vi) it provides a framework to systematically assess the community detection task. These contributions are complementary to our preliminary results presented in our previous work [Leão et al. 2019].

[2] In the history of interactions of a social network there are those that represent a strong relationship between two people in a community (e.g., a teacher and a student in a school) and others, result of chance, that represent interactions between people from different communities and most likely will not occur in the future (e.g., a phone call from a telemarketer) [Leão et al. 2017; Leão et al. 2018].

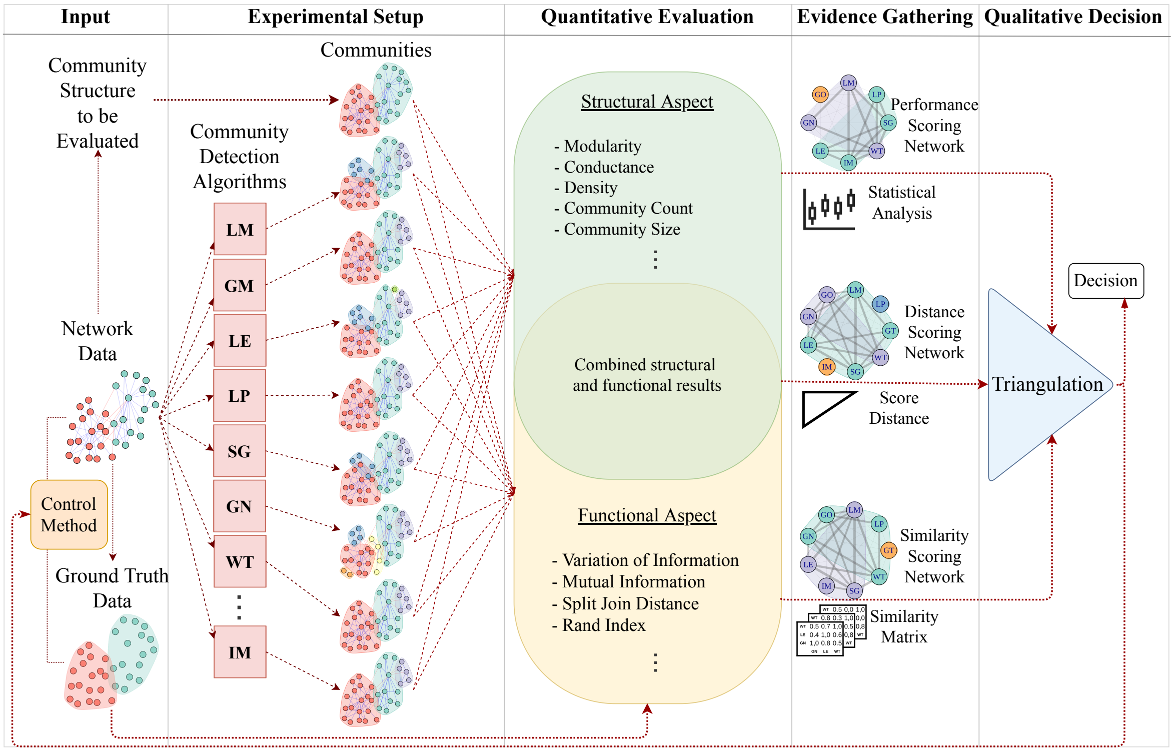

Fig. 2. Overview of the proposed approach to community quality evaluation.

3. PROPOSED APPROACH

Figure 2 summarizes our approach for community quality evaluation. First, in the *input* step, we provide a network, its ground truth communities and the set of its communities that we want to evaluate. Next, in the *experimental setup* step, in addition to the set of ground truth communities, we also consider as a further source of evidence the communities detected by distinct algorithms, for example, those listed in Table I. Then, in the *quantitative evaluation* step, all communities are assessed by multiple structural and functional metrics in order to be compared to each other to provide a set of combined evidence, whereas in the *evidence gathering* step we group the results produced by each algorithm in a new set of pieces of evidence to highlight structural and functional aspects related to the quality of each community. Finally, in the *qualitative decision* step, we compare all pieces of evidence to get a final decision on the quality of the communities. However, it is important to notice that in case of a robust decision based on the existing pieces of evidence is not available it is required to apply a control method[3] and start the whole process again to collect additional evidence, as we further describe next.

3.1 Collecting Evidence

By using structural metrics, we are able to quantify the connectivity of specific sets of nodes in the network in terms of structural characteristics that are typical of real-world communities [Yang and Leskovec 2015]. For this, we take into account multiple pieces of evidence on the quality of a

[3] A control method consists in raising a hypothesis about the effect of a given class of entities on an experimental result and testing such a hypothesis by comparing its result with that obtained without considering that class of entities on the data [Creswell and Creswell 2018].

community expressed by the results of statistical analyses of the scores obtained by metrics such as modularity, conductance and density [Newman and Girvan 2004; Yang and Leskovec 2015]. We also use specific statistics, such as the number and size of the detected communities, the distribution of the component sizes, the variance of these values, and other network metrics, such the ones considered in Table II (see Section 4), to help analyze the results.

In this step, we also statistically estimate the consensus and the divergence between the detection algorithms with respect to the structure of the communities of a network. For this, we measure the similarity between communities detected by different algorithms, considering similarity metrics such as *Variation of Information* (VI), *Normalized Mutual Information* (NMI), *Split Join Distance* (SJD) and *Rand Index* (RI) to provide some functional evidence. By using such additional metrics, we collect evidence on the functional aspects of the detected communities by measuring the similarity between them and their respective ground truths.

Finally, in the last step (Evidence Gathering), we combine structural and functional aspects in order to make a final decision on the quality of the set of communities provided as input. For this, we measure the structural characteristics of the networks' ground truth communities. Then, we collect evidence about the agreement between such measures and the measures of the communities obtained by distinct community detection algorithms, by using the distance measure between two scores s_1 and s_2, defined by Equation 1.

$$d(s_1, s_2) = |s_1 - s_2|/(s_1 + s_2) \tag{1}$$

3.2 Diversification Criterion

In our experimental configuration, we estimate the diversification of community detection algorithms based on the distinction between the communities detected by them and the tendency of pairs of algorithms to detect very different communities. For this, we measure the similarity between the sets of detected communities and the distance d (see Equation 1) between the quality score of these communities. Then, we model each set of such measures as a multilayer network, by separating in each one of its layers the measures derived from the same metric. In this network, the nodes represent different algorithms and the edges correspond to a relation of similarity among them, weighted by the respective metric value (see Figure 2). Next, a clustering analysis is performed on each layer of the network to identify the algorithms that produce similar results (i.e., belong to the same group). Finally, the algorithms that are often assigned to different groups represent a greater diversification in the baseline repertoire.

Regarding this multilayer network, we also analyze the diversification of the quality measures. For this, we measure the similarity between different layers. Thus, each pair of layers with very similar clusters reinforces the respective pairs of evidence collected in those layers, when they correspond to uncorrelated metrics. On the other hand, very different clusters indicate diversity of measurements and can be an evidence of the effect of bias in some of the metrics. Then, this set of evidence is considered to decide on the final quality of the communities.

3.3 Decision by Triangulation

By analyzing all pieces of evidence considered (structural, functional and the two combined), we capture distinct aspects of the communities' quality and conclude on the quality of their structures by means of a consensual decision. For this, we first cross-check these pieces of evidence to estimate their validity and consistency. Thus, any agreed evidence on a specific aspect reinforces the internal consistency of that strategy on that specific aspect. On the other hand, the agreement on pieces of evidence collected considering different aspects allows our approach to validate the quality evidenced by the corresponding strategies.

The final decision on the quality of the community structures is obtained when a consensual consistency among all strategies is achieved. Possibly, the disagreement between different pieces of evidence explains the internal bias of a strategy (for example, a disagreement in one of its metrics) or even a bias of the strategy itself (when a set of its pieces of evidence disagrees with a set of pieces of evidence produced by another strategy). In this analysis, it is also carried out a cross-checking of the consistency and validity of the pieces of evidence with respect to the quality of the communities detected by a specific algorithm.

3.4 Controlling Data Bias

To strengthen the pieces of evidence on the quality of a community and allow a more consistent and consensual conclusion, we first verify the existence of bias in the network data and in its ground truth data. Then, we check the effect of data bias by controlling its source. More specifically, we estimate and minimize the effect of bias by removing from the network nodes and edges that systematically damage its structure. Note that here "data bias" is any error generated by a community detection algorithm that might be associated with some noise in the network being assessed.

For example, after collecting evidence about the structural and functional quality of the communities existing in a network, a low consensus among very convincing measures that express such a quality is an indication that the source of bias is the data. To verify this, we apply a control method, i.e., we deliberately remove from the network structure the part supposedly biased by the use of a specific network filter. Then, we apply the entire evaluation flow shown in Figure 2 and collect additional pieces of evidence for each evaluation strategy. If this procedure leads to a higher level of consensus (not necessarily indicating higher quality communities), we obtain the confirmation that the structure removed from the data influenced the assessment, possibly in a systematic way. Note that we adopt three different types of control filter: for noisy edges, noisy nodes and small components, which are described next.

To filter out noisy edges, we use the framework proposed in our previous work [Leão et al. 2018]. When identifying nodes in the network that belong to an entity class that violates the community structure, we use a filter per class, i.e., we remove from the network (and from its ground truth) all nodes labeled in the respective ground truth with that class identifier. With the removal of these noisy nodes and edges, we aim to extract communities with a better defined structure and, most importantly, that allow a more objective assessment and the collection of evidence to decide on the quality of the communities. Finally, the third filter, used as a control method, aims to remove very small and dense components from the network, since they contribute little to distinguish specific bias from the algorithms, but can distort the structure of the detected communities and influence their quality assessment due to the inherent bias of algorithms and metrics.

4. EXPERIMENTAL RESULTS

To evaluate our proposed approach, we run a series of experiments to assess the communities derived from twelve networks by applying a combination of eight algorithms based on state-of-the-art community detection methods. Note that in these experiments we analyze the communities generated by each algorithm separately, considering the other ones as their baselines. In addition, experiments involving non-deterministic algorithms (see Table I) were performed several times (at least 30 repetitions) to ensure the reliability of the results.

4.1 Networks

Initially, we modeled as temporal and aggregate edge graphs the following scientific collaboration networks (here identified according to their respective datasets): APS - coauthorship network of

Table II. Characterization of the networks.

Application Domain	Network	$\lvert V\rvert$	$\lvert E\rvert$	Δ	D	CC	C
Scientific Collaboration	APS [Brandão and Moro 2017]	181k	852k	305	0.5	0.33	5k
	PubMed [Brandão and Moro 2017]	444k	5.5M	4869	0.6	0.36	9k
	arXiv [Leão 2018]	33k	180k	424	3.3	-	3k
	SIC10 [Leão et al. 2019]	3k	12k	96	20	0.34	119
Contact in a Hospital	LH10 [Vanhems et al. 2013]	76	1k	65	4k	0.6	1
Contact in a High and Primary School	HSC [Génois and Barrat 2018]	327	5818	87	1k	0.44	1
	PSC [Stehlé et al. 2011]	242	8k	134	3k	0.48	1
Contact in a French Health Institute	IVS13 [Génois et al. 2015]	92	1k	44	2k	0.37	1
	IVS15 [Génois and Barrat 2018]	217	4k	84	2k	0.36	1
Contact in the Hypertext ACM Conference	ACM09 [Isella et al. 2011]	403	10k	189	1k	0.24	1
Simulated Networks	HSC(S) [Leão et al. 2019]	$\approx$327	$\approx$7k	$\approx$116	$\approx$1k	$\approx$0.4	1
	EEU(S) [Leão 2018]	$\approx$1k	$\approx$10k	$\approx$78	$\approx$267	$\approx$0.29	2

$\lvert V\rvert$: set of vertices; $\lvert E\rvert$: set of edges; Δ: max degree; D: density (x10^{-4}); CC: cluster coefficient; C: number of components. The min degree is 1 in all networks.

members of the American Physical Society [Brandão and Moro 2017; Leão et al. 2018]; PubMed - coauthorship network derived from scientific articles available in MEDLINE [Brandão and Moro 2017; Leão et al. 2019]; arXiv - coauthorship network derived from scientific articles deposited in arXiv [Leão 2018]; SIC10 - coauthorship network of papers presented at the 2019 Seminar on Scientific Initiation held at the Federal Institute of the Northern of Minas Gerais [Leão et al. 2019]; LH10 - contact network between people in a hospital [Vanhems et al. 2013], HSC - contact network in a high school [Gemmetto et al. 2014; Génois and Barrat 2018]; PSC - contact network in a primary school [Stehlé et al. 2011]; IVS13 and IVS15 - contact networks in a French health institute in two different years [Génois et al. 2015; Génois and Barrat 2018]; ACM09 - contact network in the ACM 2009 Hypertext Conference [Isella et al. 2011]; HSC(S) - simulated network based on the HSC network [Leão 2018]; EEU(S) - simulated network based on an e-mail exchange network [Leão 2018].

In addition, we modeled the metadata of these networks to use them as ground truths for their respective communities, where each community is identified by the predominant research area of their respective researchers in the collaboration networks, by the class of the students in the high/primary school contact networks, and by the department of the workers in the the French Health Institute contact networks. With respect to the Hypertext ACM Conference network, we notice that it represents a single community. Finally, the ground truths of the simulated networks are the same of the real networks from which they were generated [Nunes et al. 2017].

Table II presents a general characterization of these networks. Note that, it was important to consider in our experiment synthetic and real-world networks that have many variations of type, size and complex structural features. In addition, the number of networks used is justified by the need to collect sufficient evidence to obtain the consistent results.

4.2 Evidence Considered

The combination of functional and structural evidence in our experiments allowed us to corroborate the quality of the ground truths as well as of the communities detected in all networks. This also made it possible to indicate the algorithm that identified the best communities in the networks. For this, we first analyzed the results of each strategy individually, providing hypotheses about the quality of the communities. Then, we combined these results, verifying the consensus among the quality of the communities. In this way, we verified which hypotheses were refuted, as well as the biases identified.

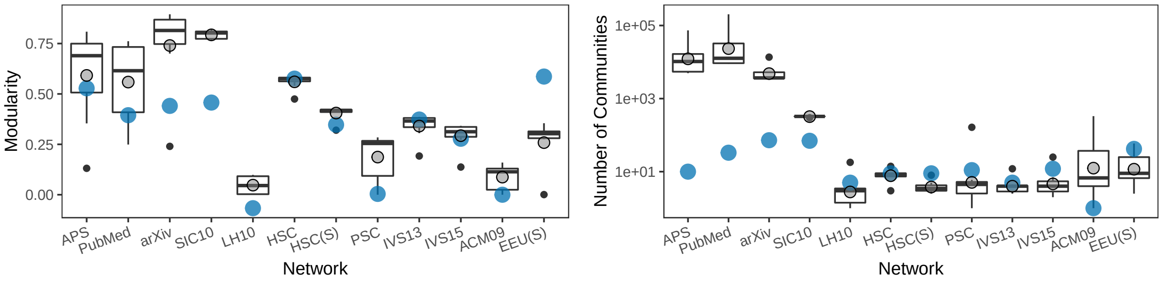

Fig. 3. Details of some of the pieces of evidence considered in the structural evaluation strategy: (left) modularity values for the communities detected by all considered algorithms in each network (boxplot) and for the respective ground truths (blue dot); (right) number of communities detected by all considered algorithms in each network (boxplot) and in the respective ground truths (blue dot).

4.2.1 *Identifying the Best Communities.* First, we analyze the structure of the communities detected by the different algorithms. Here, we note that the communities derived from the HSC, SIC10 and arXiv networks present the best defined characteristics. For this, we considered the following pieces of evidence: high average modularity (Figure 3), greater consensus on the structure of the communities (interquartile of the similarity between them, presented in Figure 4), greater confidence of the modularity value obtained in different experiments with the same non-deterministic algorithm (coefficient of variation less than 0.1 of the modularity values between repetitions of detection experiments) and small variation in the number of communities detected by these algorithms (Figure 3). However, as we shall see below, although such pieces of evidence indicate that the communities from these three networks have the same characteristics, we have not come to the same conclusion about their quality.

From a functional viewpoint, unlike the High School network, in the arXiv and SIC10 networks there is no convergence of evidence to confirm the quality of their communities when compared with their ground truths (Figure 4 (right)). This can be considered as a disagreement with respect to the structural aspect when we compare the distance between the structural measure values, such as modularity values or number of communities of the arXiv and SIC10 networks with those of most other networks.

Note that in Figure 3 (left), for example, there is a large difference between the modularity values of the ground truths and those estimated for the detected communities in the arXiv and SIC10 networks. In addition, according to Figure 3, the number of communities in the ground truths is far from the number of communities actually detected in the networks[4]. Therefore, the strength of these initial pieces of evidence has led us to the conviction that the communities detected in the actual networks are the correct ones. In addition, the confidence intervals of the measures and the structural evidence that strongly disagree with the functional one corroborate the interpretation that the detected communities are the real ones and not those shown by the ground truths. This means that the two sets of evidence, structural and functional, contradict each other on which communities in the arXiv and SIC10 networks are the best ones. Thus, we need to make a decision on which set of evidence is the strongest one, the structural quality of the ground truth or the detected communities. Based on our approach, the possibility of bias being the cause of these divergences makes it necessary to evaluate them by using a third set of evidence (on a new and independent particular aspect) to support one of the two contradictory sets of evidence. To do so, we have analyzed these contradictory sets of evidence in order to raise some hypotheses about the main source of bias in the convergence of the results on the quality of the communities of the arXiv and SIC10 networks.

[4] The distance between these two values (number of communities) is measured by the metric given by Equation 1.

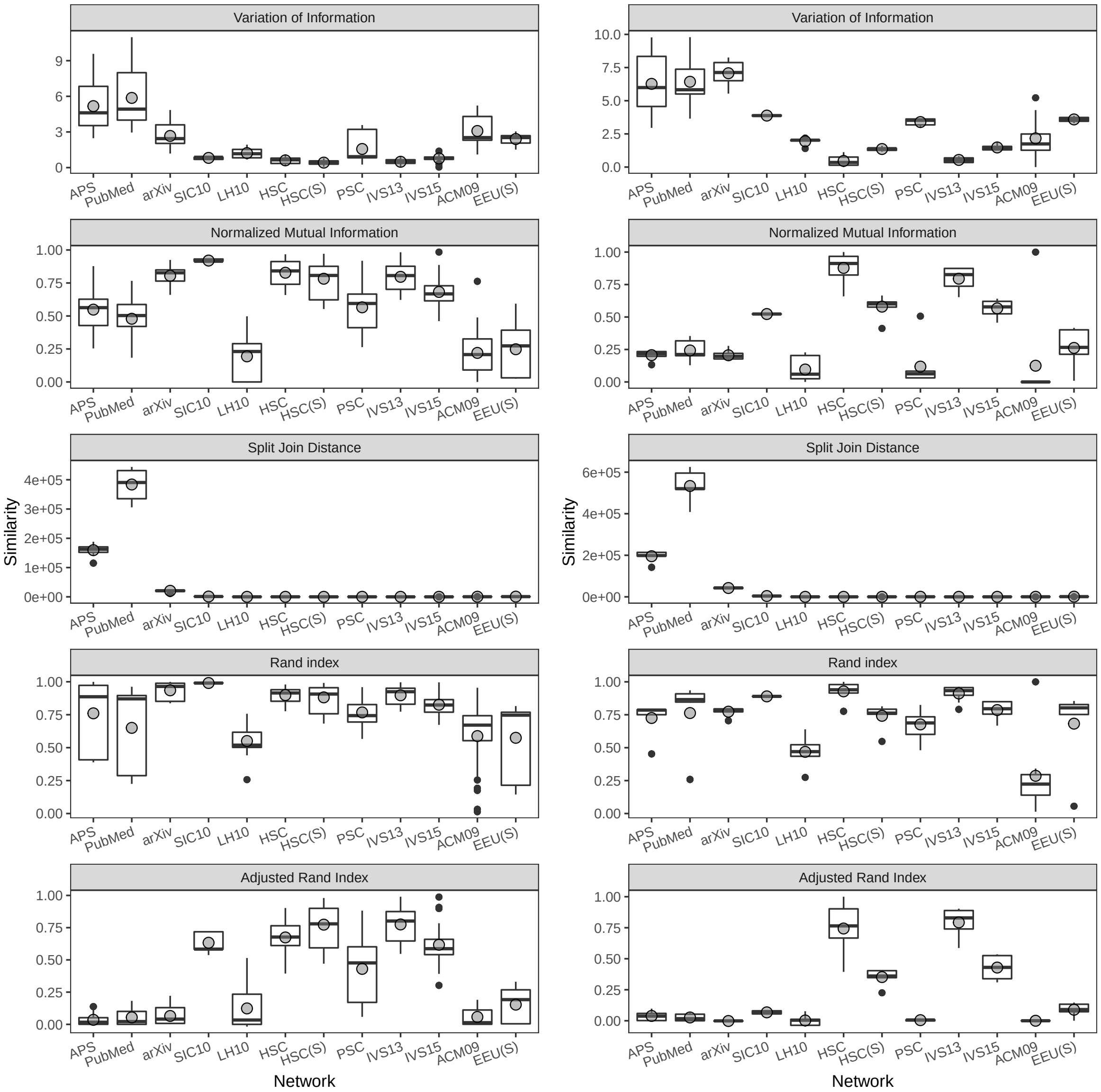

Fig. 4. Details of some of the pieces of evidence considered in the functional evaluation strategy: (left) similarity values between the communities detected by different algorithms and expressed by different metrics; (right) similarity values between the ground truth and the communities detected by distinct algorithms and expressed by different metrics. Note that, especially for the metrics VI and SJD, the lower their values, the greater the similarity indicated.

As we can see, in the arXiv and SIC10 networks the bias caused by the detection algorithms does not considerably interfere in their results, since the communities suggested by them are structurally similar. In addition, this was evidenced in these networks by all structural metrics considered, whose values corroborate a high-quality community structure, as already shown in Figures 3 and 4. Thus, we have hypothesized that the bias interference is predominantly in the data, provoking a disagreement between pieces of structural and functional evidence, as well as the identification of false communities of high quality.

For this, we first analyze the ground truth communities of the scientific collaboration networks and then consider the meaning of these communities in those networks, i.e., they are groups of researchers that publish together and predominantly in the same area of knowledge. However, it should be noted that this definition is not absolute, since there may be a multidisciplinary community with sporadic co-authorships or a community of researchers that work in the same area, but do not collaborate with each other. In both possible cases, ground truth communities are not very well captured by detection algorithms that rely on network connectivity. Thus, we consider the hypothesis that the bias that obscures the real community structure of the arXiv and SIC10 networks is a consequence of the existence of edges and nodes that represent, respectively, sporadic collaborations and researchers that work in the same knowledge area, but do not significantly interact with each other. Then, to test our hypothesis, we run the following bias control experiment: we removed such skewed edges and nodes using the filtering framework proposed by Leão et al. [2018], and then, as a new iteration, followed again the steps of our evaluation approach as shown in Figure 2, using as input the filtered version of that network[5]. From this new iteration, we obtained new communities in which the structural and functional metrics of the arXiv network converged, as indicated by its greater similarity with the ground truth communities, and showing a better structural aspect, as indicated by all metrics used for this purpose (Figure 3).

However, this same convergence was not observed in networks such as SIC10, which, therefore, was also inspected by a new iteration of the evaluation flow of our approach. This time, the second control method used was the filtering of the smallest connected components, due to the considerable number of them in this network. As a result, it was found that the measurement of the structural quality of their communities is biased by the presence of these components. Finally, with these findings, a consensual decision was reached among the evaluation strategies, thus confirming that the communities of both networks are of *High Quality.* Note that the components that distorted the structural quality of both networks to *Very High* correspond only to noise in the SIC10 network (5% of its structure) and in the arXiv network itself (90% of its structure).

Table III. Decision on the quality of the communities.

Network	Premature Decision		Data Quality		Control Method	Main Source of Significant Bias	Consensual Decision
	Structural	Functional	Net	Ground Truth			
LH10	Very Low	Very Low	Low	Medium	NF,EF	Net and GT	Low
PubMed	Medium	Low	Medium	Very Low	EF	GT, Alg and Met	Low
APS	High	Low	Medium	Low	EF	GT, Alg and Met	Low
ACM09	Very Low	Very Low	Low	Very Low	EF	Net and GT	Medium
EEU(S)	Low	Medium	Very Low	High	EF	Net	High
arXiv	Very High	Low	Medium	Low	EF	Net and GT	High
SIC10	Very High	Medium	High	Medium	EF,CF	GT and Met	High
PSC	Very Low	Very Low	Low	Medium	EF	Net	Very High
HSC(S)	Low	High	Medium	Very High	EF	Net	Very High
InVS15	Low	High	Medium	Very High	EF	Net	Very High
InVS13	Low	High	Medium	Very High	EF	Net	Very High
HSC	High	Very High	High	Very High	EF	None	Very High

Main source of significant bias: GT (ground truth data); Alg (algorithms); Met (metrics); Net (network data). Control method: NF (node filter); EF (edge filter); CF (component filter).

This way, we have been able to identify that the most significant source of bias in the arXiv network was in its data and in the SIC10 network in the structural metrics used, such as modularity. This influenced the premature results of assessing the quality of the communities in these networks, respectively underestimating and overestimating their values. Furthermore, these two situations exemplify how considering few pieces of evidence can lead to apparently very convincing results, but unreliable. Note also that the divergence between consensus conclusions obtained by different strategies is somewhat recurrent in the premature conclusions of evaluating communities in all the networks

[5] The generated datasets are available by request at http://cnet.jcloud.net.br repository.

used. More precisely, in 60% of the networks, the quality measured by the structural or functional strategies showed underestimated values and in 30% of the networks, they presented overestimated values in relation to the decision considering the influence of biases. Table III shows all premature decisions about the quality of communities, as well as the consensual decision obtained by identifying and considering biases. It is worth highlighting some of these cases, as follows.

In the APS and PubMed networks, data bias is also the main source of divergence between sets of evidence, beyond bias in structural metrics and detection algorithms. The communities of the LH10, ACM09 and PSC networks presented themselves with very low quality, indicated individually by the scores in all strategies. Despite this, the low confidence of these estimates (also widely evidenced as detailed in Figures. 3 and 4) made us investigate and check for bias in their data, in part, caused by the presence of noisy edges, as defined by Leão et al. [2018]. In addition, in the LH10 network data, we verified a class of nodes that tend to violate the community structure, and that, when isolated, allowed to test its effect on the evaluation result. Finally, our assessment approach revealed communities with different quality factors in these networks: from low to very high. The PSC network stands out with communities with very high quality revealed by our approach and which demonstrates that the premature decision was very wrong and influenced by the noisy edges. Highlighting milder mistakes in their premature decisions, the identification and isolation of biases in the EEU(S), HSC, InVS15, HSC(S) and InVS13 networks, allowed us to conclude that these networks have high quality communities. In general, the final decision on the quality of the networks was obtained with two iterations of our assessment approach, the first always made on the data of the skewed network and the other controlling the bias of these data by filtering noisy edges (except on LH10 and SIC10, which required a third iteration and specific control methods).

4.2.2 *Best Detection Algorithms.* Although the most modular communities are those detected by the Louvain algorithm (upper bounds shown in Figure 3, left), the modularity values of the communities detected by the Infomap algorithm are generally closer to those of the ground truth (there is a greater agreement between them). In addition, Infomap provided several cases in which there was an agreement between the modularity of the detected communities and that of the ground truth. On the other hand, these same metrics achieved smaller values for the Louvain algorithm. Moreover, the modularity of the communities extracted by different algorithms and that of the functional communities have considerably varied for most networks. We also verified how different community detection algorithms agree with each other and with the network ground truths with respect to their communities. Despite the variation in the structure of the detected communities, as shown in Figure 4 (left graph), there was a higher consensus among them than with respect to their ground truth communities (Figure 4 (right graph)).

Table IV. Best detection algorithms according to distinct experiments.

Best Algorithm	LM	GM	LE	LP	WT	IM
Metrics	ARI, SJD	ARI	SJD*, VI*	RI, SJD, VI	ARI, RI	ARI, NMI*, RI, SJD, VI
Networks	PubMed, arXiv	APS	APS, PubMed	APS, Sinth.	arXiv, Sinth.	All

*Metrics with the best overall value.

In addition to obtaining a consensus among different algorithms, our approach also identified some algorithms with distinct behavior, such as Infomap, that detected less modular communities, but in general more similar to their ground truths. Despite such divergences among the strategies, most pieces of evidence indicate the Louvain algorithm as the least biased among those that are based on the modularity maximization and the one that obtained estimates with higher values for most of the structural metrics, particularly modularity. We also identified some algorithms that presented the best score on some specific metrics. This is the case of the Louvain algorithm (LM) for modularity, and of the Infomap (IM) and Leading Eigenvector (LE) algorithms for the similarity metrics Normalized Mutual Information (NMI), and Split Join Distance (SJD) and Variation of Information (VI), respectively (see Table IV). Notice that our proposed approach is able to analyze distinct alternative solutions for the task at the hand, thus being able to identify those algorithms that provide the best trade-off.

5. CONCLUSIONS AND FUTURE WORK

The main contribution of this paper is an approach to identify and reduce the effect of biases when assessing the quality of a set of communities. Specifically, we use multiple and diversified measurement strategies designed to capture different aspects of the quality of a community structure. For its evaluation, we carried out a set of experiments using twelve networks (ten real ones and two synthetic ones) and compared the results obtained by eight community detection algorithms considered the state-of-the-art in the area. In addition, we also used distinct metrics, each one providing a piece of evidence from a specific point of view on the quality of the assessed communities.

In this context, we consider the consensus and the divergence between such pieces of evidence to hypothesize about the influence of bias coming from metrics, algorithms or network data when assessing the quality of such communities. Thus, outliers observed by statistically analyzing the resulting measurements allowed us to test hypotheses of specific biases with respect to metrics and detection algorithms. To test a hypothesis on bias in network data or in its ground truth metadata, we use control methods, such as node filters or network edges. These methods allowed us to verify that, by removing the supposed biased structures from the data, a greater consensus in the evaluation of their communities can be verified by the multiple metrics and strategies used by our approach.

By doing so, we were able to sustain our hypothesis by showing that the quality evaluation of communities detected from a network must be supported by multiple pieces of evidence. That is, given the discrepancy between the quality indicated by distinct evaluation strategies, we evidentiate that the use of a single quality metric or strategy, be it structural or functional, makes the results biased and unreliable. On the other hand, our multi-strategy evaluation approach makes it possible to explain extreme values for some of the metrics considered and decide which strategies lead to a more consistent conclusion about the quality of a community. For example, we were able to verify the existence of bias in some metrics, network data and detection algorithms, which allowed us to reach a consensual decision very different from the premature decision, individually suggested by the metrics used by one of the strategies.

A current limitation of our proposed approach is the use of a predefined set of evaluation metrics and community detection algorithms. However, this limitation can be easily overcome by providing a configurable framework in which such features could be defined according to specific characteristics of the networks being considered. It is also worth noting that the approach proposed in this article can be applied to other types of algorithm (such as those for clustering tabular data, backbone extraction, core-periphery analysis, detection of dynamic or overlapping communities, etc.), as well as adapted to other tasks besides community detection (such as system modeling or simulation, supervised machine learning techniques, missing data prediction, etc.). Finally, another line of future work could be, for example, adapting this approach to assess the task of link prediction in social networks in order to provide more robust results.

Acknowledgements

Work supported by project MASWeb (FAPEMIG/PRONEX grant APQ-01400-14) and by the authors' individual grants from CNPq and FAPEMIG. Particularly, the first author would like to thank LBD/UFMG, JCLoud.net.br and LabSiCCx - Laboratório de Sistemas Computacionais Complexos (PROPPI/IFNMG, project Nr. 209/2019) for the infrastructure provided.

REFERENCES

Abrahao, B., Soundarajan, S., Hopcroft, J., and Kleinberg, R. On the Separability of Structural Classes of Communities. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA, pp. 624–632, 2012.

Almeida, H., Guedes, D., Meira Jr, W., and Zaki, M. J. Towards a Better Quality Metric for Graph Cluster Evaluation. *Journal of Information and Data Management* 3 (3): 378–393, 2012.

Amelio, A. and Pizzuti, C. Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. Paris, France, pp. 1584–1585, 2015.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008, 2008.

Brandão, M. A. and Moro, M. M. The strength of co-authorship ties through different topological properties. *Journal of the Brazilian Computer Society* 23 (1): 5, 2017.

Brender, J. Framework for Meta-Assessment of Assessment Studies. In *Handbook of Evaluation Methods for Health Informatics*. Burlington, pp. 253–320, 2006.

Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks. *Physical Review E* vol. 70, pp. 066111, 2004.

Coscia, M. Discovering Communities of Community Discovery. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA, pp. 1–8, 2019.

Coscia, M., Giannotti, F., and Pedreschi, D. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4 (5): 512–546, 2011.

Creswell, J. W. and Creswell, J. D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Thousand Oaks, California, USA, 2018.

Dao, V. L., Bothorel, C., and Lenca, P. Community structure: A comparative evaluation of community detection methods. *Network Science* 8 (1): 1–41, 2020.

Fortunato, S. Community detection in graphs. *Physics Reports* 486 (3–5): 75–174, 2010.

Gandica, Y., Decuyper, A., Cloquet, C., Thomas, I., and Delvenne, J.-C. Measuring the effect of node aggregation on community detection. *EPJ Data Science* 9 (1): 6, 2020.

Gemmetto, V., Barrat, A., and Cattuto, C. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases* 14 (1): 695, 2014.

Génois, M. and Barrat, A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science* 7 (1): 11, 2018.

Génois, M., Vestergaard, C. L., Fournet, J., Panisson, A., Bonmarin, I., and Barrat, A. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science* 3 (3): 326–347, 2015.

Ghasemian, A., Hosseinmardi, H., and Clauset, A. Evaluating Overfit and Underfit in Models of Network Community Structure. *IEEE Transactions on Knowledge and Data Engineering* 32 (9): 1722–1735, 2020.

Gösgens, M., Tikhonov, A., and Prokhorenkova, L. Systematic analysis of cluster similarity indices: How to validate validation measures. *CoRR* vol. arXiv:1911.04773, 2020.

Hric, D., Darst, R. K., and Fortunato, S. Community detection in networks: structural communities versus ground truth. *Physical Review E* 90 (6): 62805, 2014.

Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and den Broeck, W. V. What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* 271 (1): 166 – 180, 2011.

Jebabli, M., Cherifi, H., Cherifi, C., and Hamouda, A. Community detection algorithm evaluation with ground-truth data. *Physica A: Statistical Mechanics and its Applications* vol. 492, pp. 651 – 706, 2018.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. Multilayer networks. *Journal of Complex Networks* 2 (3): 203–271, 2014.

Labatut, V. Generalised measures for the evaluation of community detection methods. *International Journal of Social Network Mining* 2 (1): 44–63, 2015.

Lancichinetti, A. and Fortunato, S. Consensus clustering in complex networks. *Scientific Reports* vol. 2, pp. 336, 2012.

Lei, Y., Bezdek, J. C., Romano, S., Vinh, N. X., Chan, J., and Bailey, J. Ground truth bias in external cluster validity indices. *Pattern Recognition* vol. 65, pp. 58 – 70, 2017.

Leskovec, J., Lang, K. J., and Mahoney, M. Empirical Comparison of Algorithms for Network Community Detection. In *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA, pp. 631–640, 2010.

Leão, J. C. *An Approach for Detecting Communities from Sequences of Social Interactions*. M.S. thesis, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil (in Portuguese), 2018.

Leão, J. C., Brandão, M. A., Vaz de Melo, P. O. S., and Laender, A. H. F. Mineração de Perfis Sociais em Redes Temporais. In *Anais do 32º Simpósio Brasileiro de Bancos de Dados, SBBD 2017*. Uberlândia, MG, pp. 264–269, 2017.

Leão, J. C., Brandão, M. A., Vaz de Melo, P. O. S., and Laender, A. H. F. Who is really in my social circle? Mining social relationships to improve detection of real communities. *Journal of Internet Services and Applications* 9 (1): 20:1–20:17, 2018.

Leão, J. C., Cardoso, R. J. S., and Santos, A. B. Uma análise temporal da rede de colaboração cientı́fica do IFNMG: 10 anos de iniciação cientı́fica e orientação acadêmica. *Anais dos Simpósios de Informática do IFNMG - Campus Januária* vol. 11, pp. 7, 2019.

Leão, J. C., Laender, A. H. F., and Vaz de Melo, P. O. S. A Multi-Strategy Approach to Overcoming Bias in Community Detection Evaluation. In *Anais do 34º Simpósio Brasileiro de Bancos de Dados, SBBD 2019*. Fortaleza, CE, pp. 13–24, 2019.

Liu, X., Cheng, H., and Zhang, Z. Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering* 32 (9): 736 – 1746, 2020.

Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103 (23): 8577–8582, 2006.

Newman, M. E. J. and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E* 69 (2): 26113, 2004.

Nunes, I. O., Celes, C., D., M. S., Vaz de Melo, P. O. S., and Loureiro, A. A. F. GRM: Group Regularity Mobility Model. In *Proceedings of the 20th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. New York, NY, USA, pp. 85–89, 2017.

O'Donoghue, T. and K., P. *Qualitative Educational Research in Action: Doing and Reflecting*. Routledge, Abingdon, UK, 2003.

Peel, L., Larremore, D. B., and Clauset, A. The ground truth about metadata and community detection in networks. *Science Advances* 3 (5): 1–8, 2017.

Pons, P. and Latapy, M. Computing communities in large networks using random walks. In *Computer and Information Sciences - ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings*, p. Yolum, T. Güngör, F. Gürgen, and C. Özturan (Eds.). Berlin, Heidelberg, pp. 284–293, 2005.

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

Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (336): 846–850, 1971.

Reichardt, J. and Bornholdt, S. Statistical mechanics of community detection. *Physical Review E* vol. 74, pp. 016110, 2006.

Rocha, L. E. C., Masuda, N., and Holme, P. Sampling of temporal networks: Methods and biases. *Physical Review E* 96 (5): 52302, 2017.

Rosvall, M. and Bergstrom, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLOS ONE* 6 (4): 1–10, 2011.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., and Vanhems, P. High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE* 6 (8): 1–13, 08, 2011.

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., and Voirin, N. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLOS ONE* 8 (9): 1–9, 2013.

Vieira, V. F., Xavier, C. R., and Evsukoff, A. G. Comparing the Community Structure Identified by Overlapping Methods. In *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha (Eds.). Cham, pp. 262–273, 2020.

Yang, J. and Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42 (1): 181–213, 2015.

Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33 (4): 452–473, 1977.

Zaki, M. J. and Meira Jr., W. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

Zhao, Y. A survey on theoretical advances of community detection in networks. *Computational Statistics* 9 (5): e1403, 2017.