A Multi-Strategy Approach to Overcoming Bias in Community Detection Evaluation

Jeancarlo C. Leão[1], Alberto H. F. Laender[2], Pedro O. S. Vaz de Melo[2]

[1] Instituto Federal do Norte de Minas (IFNMG)
39600-000 – Araçuaí – MG – Brazil.

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
31270-901 – Belo Horizonte – MG – Brazil.

jeancarlo.leao@ifnmg.edu.br, laender@dcc.ufmg.br, olmo@dcc.ufmg.br

Abstract. *Community detection is key to understand the structure of complex networks. However, the lack of appropriate evaluation strategies for this specific task may produce biased and incorrect results that might invalidate further analyses or applications based on such networks. In this context, the main contribution of this paper is an approach that supports a robust quality evaluation when detecting communities in real-world networks. In our approach, we use multiple strategies that capture distinct aspects of the communities. The conclusion on the quality of these communities is based on the consensus among the strategies adopted for the structural evaluation, as well as on the comparison with communities detected by different methods and with their existing ground truths. In this way, our approach allows one to overcome biases in network data, detection algorithms and evaluation metrics, thus providing more consistent conclusions about the quality of the detected communities. Experiments conducted with several real and synthetic networks provided results that show the effectiveness of our approach.*

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, 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]. In general, this evaluation is done without explicitly dealing with bias on data, methods and metrics, which may lead to inconsistent results.

In order to illustrate this, 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 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

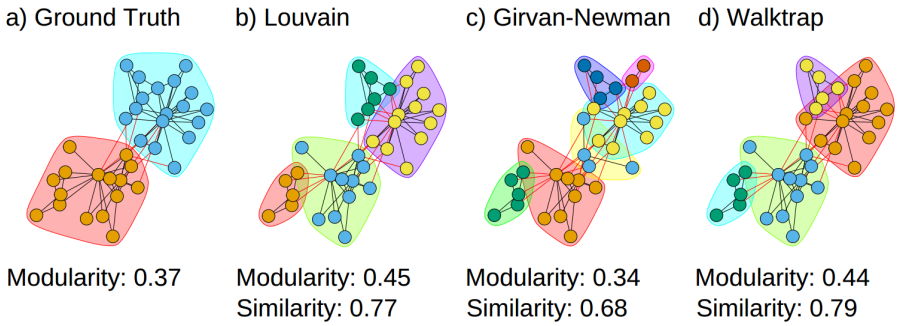

Figure 1. Example of how bias can affect community detection in social networks.

in this same network by the Louvain algorithm [Blondel et al. 2008]. Note that the community structure revealed by the Louvain algorithm is different from those presented in Figure 1c and in Figure 1d, which were respectively obtained by the Girvan-Newman [Newman and Girvan 2004] and Walktrap [Pons and Latapy 2005] algorithms, both of them considered very effective. In addition, the value of the modularity metric indicates that the communities shown in Figure 1b present a better quality with respect to their modular structure (i.e., it presents 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 more similar to the ones shown in Figure 1a, even though there is not a perfect match. Finally, due to some specific bias in the original network, such as sporadic interactions [Leão et al. 2018], the Girvan-Newman algorithm has not been able to identify good communities (Figure 1c), as shown by the values of the two metrics considered.

Thus, 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. One possible explanation to this fact is the lack of a comprehensive evaluation approach that considers multiple strategies and allows one to find out 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 methods on which community structure presents the highest quality. This requires a cross-checking approach involving more than two distinct evaluation strategies in order to indicate a consensus and estimate a possible bias with respect to the quality of the revealed communities.

A method that is generally employed to increase data reliability and validity is triangulation[1], which consists in using multiple methods to test a same hypothesis [Leão 2018]. Based on this idea, the main contribution of this paper 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,

[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."

Table 1. 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 node labeling	Label Propagation (LP)	N	[Raghavan et al. 2007]
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-Determinístic/N-Non determinístic).

detection methods 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 by considering multiple metrics, detection methods and distinct 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 datasets is not important, since we are not trying to identify the best existing community, but the best one among those being compared.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 describes our approach to community detection 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 on future work.

2. Related Work

Although community detection has become one of the most popular and best-studied research topics in network science [Zhao 2017], the problem of validating the quality of a community derived from a real network has not received the due attention, since there is no consensus on what is meant by a good community. For example, the methods listed in Table 1 usually extract distinct communities from a given network, which are usually considered of good quality by different metrics.

In this context, there is no best metric to assess the quality of a community [Almeida et al. 2012]. Moreover, 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]. Thus, existing work usually considers only specific aspects to assess the quality of a community, for example by measuring the structure derived from its connectivity [Newman and Girvan 2004], comparing its similarity with a ground truth community [Peel et al. 2017] or performing a comparison with a good baseline [Hric et al. 2014].

Regarding the structural aspect, 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 particular cases, it is possible to as-

sess 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 indexes such as *Rand Index* and *Normalized Mutual Information*.

In addition, community detection algorithms are sensitive to different community structures, topologies or instances of a network [Coscia et al. 2011]. 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. 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 to the biased data problem by directly removing noisy produced by sporadic relationships[2] found in a social network. In addition, they also showed that this kind of noise may cause errors when detecting communities.

To the best of our knowledge, the closest work to ours is the one by Yang and Leskovec [2015]. In their work, they use the correlation between distinct community definitions to evaluate their structure in large social networks. On the other hand, Lancichinetti and Fortunato [2012] seek consensus only on the structural aspect of the communities. In both works, the authors evaluate the quality of a community without aiming at a consensus involving distinct aspects or addressing any bias.

Thus, by analyzing the above works, we have not been able to identify any approach that deals with different types of bias for assessing the quality of a community. Moreover, differently from our work, the above ones do not provide a systematic and consistent strategy to produce a robust conclusion about a community's quality.

3. Proposed Approach

Figure 2 summarizes our approach. First, we provide as input a network, its ground truth communities and the set of its communities that we wish to evaluate. Next, in addition to a set of ground truth communities, we consider as further evidence the communities detected by multiple algorithms, for example, those listed in Table 1. Then, in the quantitative evaluation step, all communities are assessed by multiple structural and functional metrics, and then compared to each other to provide a set of combined evidence. Finally, we group the results produced by each algorithm in a new set of pieces of evidence in order to highlight structural and functional aspects related to the quality of each community, and compare them with those of the communities obtained by the other algorithms, as described next.

[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 intercations 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. 2018].

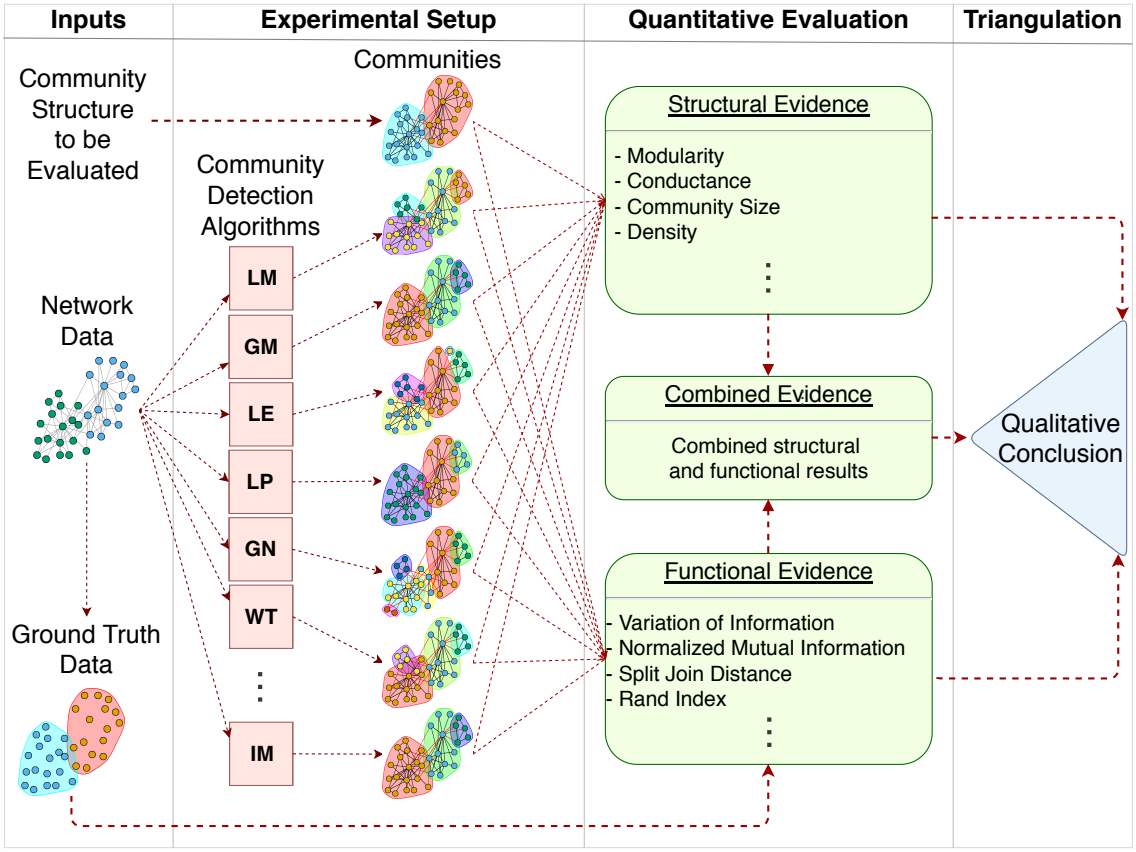

Figure 2. Overview of the proposed approach to community detection evaluation.

By using structural metrics, we quantify how much the connectivity of specific sets of nodes in the network expresses 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 community expressed 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 variance of these values and network measures to help analyzing the results. In addition, we consider similarity (or functional) metrics such as *Variation of Information* (VI), *Normalized Mutual Information* (NMI), *Split Join Distance* (SJD) and *Rand Index* (RI) to provide some functional evidence.

We then combine structural and functional measures to compare the quality of the communities obtained by different community detection algorithms applied to the same network. For this, we assess the agreement between these measures by using the standard one-dimensional Euclidean distance. By crossing all types of evidence considered (structural, functional and combined), we capture distinct aspects of the communities' quality and conclude on the quality of the their structures by means of the consensus among all pieces of evidence. In addition, we check the effect of bias in data by controlling its source. More specifically, to minimize and estimate the effects of bias caused by noisy data, we filter the networks by using the framework proposed in our previous work [Leão et al. 2018]. Note that here "data bias" is any error generated by a community detection algorithm that might be associated with noise in the network.

Table 2. Caracterization of the networks.

Appl. Domain	Network	$\|V\|$	$\|E\|$	Δ	D	CC
Scientific Collaboration	APS	181k	852k	305	0.5	0.33
	PubMed	444k	5.5M	4869	0.6	0.36
	arXiv	33k	180k	424	3.3	-
Diseace Propagation	High School	327	5818	87	1.1k	0.44
Simulated Nets	Sinthetic	$\approx$1k	$\approx$13k	$\approx$78	$\approx$267	$\approx$0.29

$|V|$: set of vertices; $|E|$: set of edges; Δ: max degree; D: density (x10^{-4}); CC: cluster coeficient. The min degree is 1 in all networks.

4. Experimental Results

To evaluate our proposed approach, we run a series of experiments to assess the structural and functional aspects of the communities derived from five networks by applying a combination of seven algorithms based on state-of-the-art 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 1) were performed several times to ensure the reliability of the results.

4.1. Networks

Initially, we modeled as aggregate edge graphs the scientific collaboration networks (here identified by their respective datasets, namely APS, PubMed and arXiv)[3] and the contact network of secondary school students[4], which were used in previous works by Gemmetto et al. [2014] and Leão et al. [2018], respectively. Table 2 presents a general characterization of these networks.

Notice that these networks represent distinct social relationships. Thus, in the scientific collaboration networks, vertices represent researchers and there is an edge connecting two researchers if they are coauthors of a same article. In the contact network, vertices represent members of a school (for example, students or teachers) and there is an edge between two individuals if they are close to each other. We also used synthetic networks for which we have created their respective ground truth communities. These synthetic simulated networks are based on the GRM model [Nunes et al. 2017], which allows the representation of mobility networks with group (community) characteristics.

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

[3]Datasets obtained from http://homepages.dcc.ufmg.br/˜mirella/projs/apoena/. APS: coauthorship network of members of the American Physical Society; PubMed: coauthorship network derived from scientific articles available on MEDLINE; arXiv: coauthorship network derived from scientific articles obtained from https://www.kaggle.com/neelshah18/arxivdataset/

[4]Datasets obtained from http://www.sociopatterns.org/datasets/.

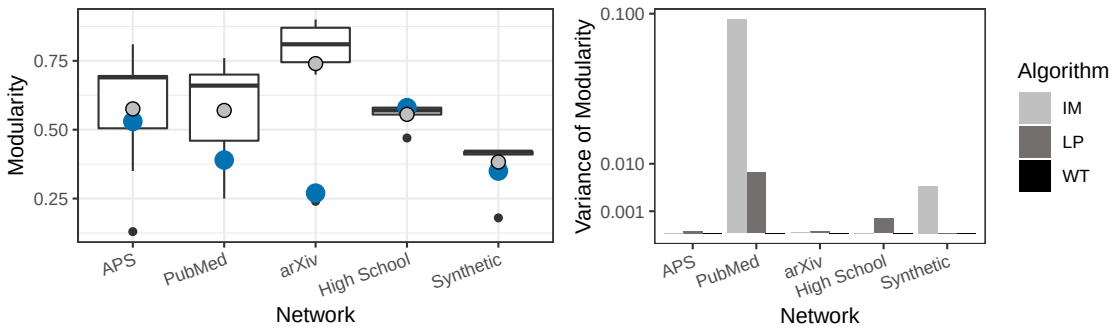

Figure 3. Modularity values for communities detected by all considered algorithms (boxplot) and for the ground truth (blue dot). $Var(Modularity)$: Variance of the modularity between replications of the detection experiments by non-deterministic algorithms (deterministic algorithms have zero variance and therefore are not presented in the right graph).

communities on 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 communities. In this way, we verified which hypotheses were refuted, as well as the biases identified.

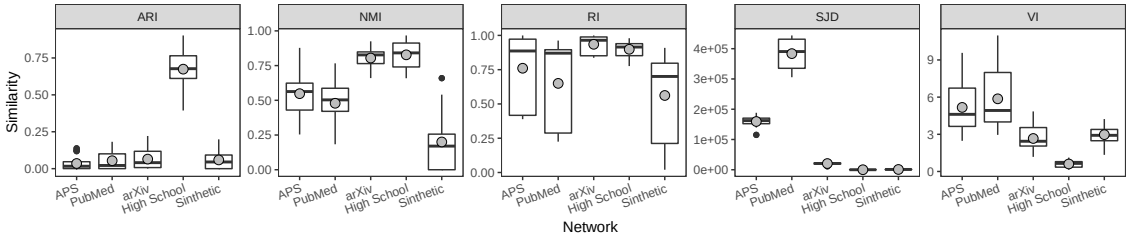

Figure 4. Similarity values between communities detected by different algorithms and measured by different metrics. Note that, especially for the metrics VI and SJD, the lower their values, the greater the similarity indicated.

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 High School and arXiv networks have the best well defined characteristics. For this, we consider the following evidence: high average modularity (Figure 3, left), greater consensus on the structure of the communities (interquartil 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 (Figure 3, right) and small variation in the number of communities detected by these algorithms (Figure 6). However, as we shall see below, although such pieces of evidence indicate that the communities from these two 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 one there is no convergence of evidence to confirm the quality of its communities when compared with the ground truth (Figure 5). This can be considered as a disagreement with

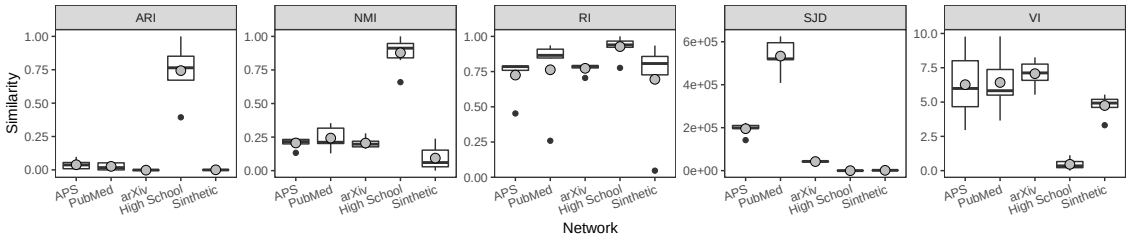

Figure 5. Values of the similarity between the ground truth and the communities detected by distinct algorithms and expressed by different metrics.

respect to the structural aspect when we compare the distance between the modularity values of the arXiv network with those of the other networks.

Note in Figure 3 (left), for example, that there is a large difference between the modularity values of the ground truths and those estimated for the detected communities. In addition, according to Figure 6, the number of communities in the ground truths is far from the number of communities actually detected in the networks. Therefore, the strength of this initial evidence has led us to the conviction that the communities detected in the actual networks are the correct ones. In addition, the confidence 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.

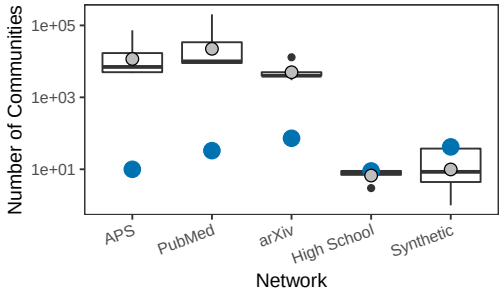

Figure 6. Number of communities detected by all considered algorithms in each network (boxplot) and in the respective ground truth (blue dot).

Note that the two sets of evidence, structural and functional, contradict each other on which communities in the arXiv network are the best ones and also on which evidence in the first set is the strongest one. 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 particular aspect) to support one of the two contradictory pieces of evidence. For this, we identified the predominant source of bias and showed that it interferes with the final conclusion.

It occurs that in the arXiv network the bias caused by the detection methods does not considerably interfere in their results, since the communities suggested by them are structurally similar. In addition, this was evidenced in this network by all structural metrics considered, whose values corroborate a high-quality community structure, as already shown in Figures 3, 4 and 6. Thus, we hypothesized that the interference bias is predominantly in the data, provoking a disagreement between 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 arXiv network and then consider the meaning of these communities in this network, 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 cases, ground truth communities are not very well captured by detection methods that rely on network connectivity. Thus, we consider the hypothesis that the bias that obscures the real community structure of the arXiv network 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 applied the same detection algorithm on the filtered version of that network[5]. This time we obtained new communities in which the convergence between the structural and functional metrics occurred, as indicated by its greater similarity with the ground truth communities, and having a better structural aspect, as indicated by all metrics used for this purpose. This way, we have been able to identify that the most significant source of bias in the arXiv network was its data, which shows that considering few pieces of evidence can lead to apparently convincing results, but unreliable.

In the APS and PubMed networks, data bias is also the main source of divergence between the respective ground truths and the detected communities. However, two characteristics of these two ground truths are among those that most interfere in the quality of their respective communities: the high overlap among the communities [Leão et al. 2018] and the large number of communities formed by multiple components (see Figure 6). This shows how pieces of structural evidence, such as those obtained by modularity, are insufficient to characterize this disagreement, even though the modularity of the ground truth and the communities detected in these networks have relatively high values and are close to each other. We also verify that the bias in the data caused by sporadic relationships does not considerably interfere in the detection of the communities since there was no significant improvement after filtering the networks. On the other hand, as shown in Figure 4, there was little consensus among the communities detected in these networks by the different algorithms.

In this context, in addition to identifying significant bias in the structure of the arXiv and High School networks, this phenomenon was also verified in a smaller scale among the detection methods. For example, as demonstrated for the arXiv network, its consensual community, despite its high structural quality, presents results that are considerably skewed. On the other hand, in the PubMed network and more clearly in the APS network, the average quality of the communities detected by different algorithms stands out when compared with the ground truths. In the synthetic networks and in the High School one bias had no significant interference on the convergence of the structural, functional and combined pieces of evidence of the communities' quality.

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

Table 3. Best detection algorithm 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 but HS

*Metrics with the best overall value.

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 (there is a greater agreement) to those of the ground truth. In addition, Infomap provided a number of 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. In addition, 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 (Figure 4), we verified a higher consensus among them than with the ground truth communities, as shown in Figure 5.

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 and the one that obtained the best values for most of the structural metrics, particularly modularity. We also identified some algorithms that presented the best score when considering a specific metric. 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 3). Notice that our proposed approach is able to analyze distinct alternative solutions for the task at the hand, thus being able to identify those that provide the best trade-off.

5. Conclusions and Future Work

The main contribution of this paper is an approach to identify and reduce effect of biases when assessing the quality of communities detected by distinct algorithms. 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 a set of experiments using five networks (four real ones and one synthetic) and compared the results obtained by seven community detection algorithms considered the state-of-the-art in the area. In addition, we also evaluated the quality of the communities by using different strategies. In this context, our evaluation evidentiated the bias of each strategy, thus providing some consensus among them.

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, be it structural or functional, makes the results biased and unreliable. On the other hand, our multi-strategy evaluation approach made it possible to explain extreme values for some of the metrics considered. For example, we were able to verify the existence of bias in modularity metrics, some ground truths and network data, and some detection algorithms.

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 platform in which such features could be defined according to specific characteristics of the networks being considered. Thus, as future work, we intend to conduct a study to characterize the diversity of algorithms and metrics usually used for community detection, in order to provide insights for improving our approach. Finally, it is worth noting that the approach proposed in this paper can be adapted to other tasks besides community detection. Thus, 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

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

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

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

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

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

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

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

Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, 2:336.

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

Leão, J. C., Brandão, M. A., Vaz de Melo, P. O. S., and Laender, A. H. F. (2018). 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.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Nat. Acad. Sci.*, 103(23):8577–8582.

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

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

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

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

Pons, P. and Latapy, M. (2005). *Computing Communities in Large Networks Using Random Walks*, pages 284–293. Springer Berlin Heidelberg, Berlin, Heidelberg.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):1–12.

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

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

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

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

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4):452–473.

Zaki, M. J. and Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, UK.

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