Evaluation Metrics for Overlapping Community Detection

Safa El Ayeb, Baptiste Hemery, Fabrice Jeanne, Estelle Pawlowski Cherrier, Christophe Charrier

Evaluation Metrics for Overlapping Community Detection

Safa El Ayeb*†, Baptiste Hemery*, Fabrice Jeanne*, Estelle Cherrier† and Christophe Charrier†
*Orange, Caen, France
Email: {safa.elayeb, baptiste.hemery, fabrice.jeanne}@orange.com
†Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
Email: {estelle.cherrier, christophe.charrier}@ensicaen.fr

***Abstract*—Networks have provided a representation for a wide range of real systems, including communication flow, money transfer or biological systems, to mention just a few. Communities represent fundamental structures for understanding the organization of real-world networks. Uncovering coherent groups in these networks is the goal of community detection. A community is a mesoscopic structure with nodes heavily connected within their groups by comparison to the nodes in other groups. Communities might also overlap as they may share one or multiple nodes. Evaluating the results of a community detection algorithm is an equally important task. This paper introduces metrics for evaluating overlapping community detection. The idea of introducing new metrics comes from the lack of efficiency and adequacy of state-of-the-art metrics for overlapping communities. The new metrics are tested both on simulated data and standard datasets and are compared with existing metrics.**

***Index Terms*—Social Network Analysis, Overlapping community detection, evaluation metric.**

I. Introduction

Social network analysis has received tremendous attention over the past decade. Its main objective is understanding individual behaviors, based on their interactions. Network analysis has attracted significant interest due to its potential to handle many real-world case studies [1], [2]. In particular, community detection has become a fundamental and highly relevant research area in network science [3]. Therefore, a substantial number of community detection algorithms have been developed, across varied disciplines such as statistics, physics, biology, sociology, etc.

The result of community detection is a partition with disjoint, overlapping, fuzzy, or hierarchical communities. To evaluate and compare community detection algorithms, the literature has given much attention to evaluation metrics [4], [5]. Evaluation metrics can be either quality metrics that assess structural quality of communities, or information recovery metrics that compare the result to a *gold standard*, also called *ground-truth*. Despite the number of evaluation metrics in the literature, very few are applicable to overlapping communities. Having a simple and easy to interpret metric is of importance when dealing with community detection algorithms.

In this paper, we propose four information recovery metrics for overlapping community detection results. Each of the proposed metrics considers a specific aspect of the network and is designed to provide a clear explanation. Our goal is to overcome the classical drawbacks of standard information recovery metrics, namely the difficulty to interpret the results.

This paper is organized as follows. Section II presents preliminary definitions about community detection and evaluation metrics. In section III, we illustrate the proposed metrics and their properties. Finally, section IV analyses several tests of the performance of proposed metrics both on synthetic and real-world networks.

The authors would like to thanks Orange and the ANRT for funding this work.

II. Background

A. Overlapping communities detection

One of the most important application of networks' analysis relies on the search for dense groups, also called communities. Community detection in networks has aroused a lot of interest during the last decade [2], [3], [5]. Although community is not an accurately defined concept, a general consensus implies that a community represents *a group of densely connected vertices, either sharing some properties or playing similar roles inside the network* as stated by the authors of [5]. Depending on the characteristics of the network, the result of community detection may lead to disjoint communities, overlapping communities, dynamic communities, etc.

Although most of the work in the literature is focused on disjoint communities, more efforts are oriented toward overlapping communities. In this paper, we are particularly focused on overlapping communities' detection. Unlike crisp communities, overlapping communities may share one or more nodes. A node can simultaneously be part of multiple communities of different scopes and levels, such as family, friends, work, city, etc. [6]. Overlapping communities were studied in the literature in various contexts such as biology [7], e-commerce [8], mobile networks [2], etc. For a complete study of overlapping community detection, we refer the reader to [9].

B. Evaluation Measures

One of the biggest challenges related to community detection is the ability to evaluate the generated results. Evaluation is a real issue for real networks where only little data are provided. Evaluation metrics in this area can be employed either to assess the performance of a community detection

978-1-6654-8001-7/22/$31.00 ©2022 IEEE

algorithm, or to compare the performances of different algorithms applied on the same set of nodes. A lot of work was done regarding evaluating metrics for community detection [5], [10]. Evaluation metrics can formally be classified into two categories: intrinsic and extrinsic metrics.

While intrinsic metrics evaluate structural properties of the identified communities, extrinsic quality metrics evaluate instead how the resulted communities are comparable to the *ground-truth*. These metrics are also called information recovery metrics because they measure the ability of algorithms to recover information from the ground-truth. For synthetic networks, the ground-truth communities are provided manually based on the network generation process.. However, for real networks this ground-truth is not always available. Therefore in this paper we only consider either synthetic networks, or real networks where the ground-truth is known.

The most popular information recovery metrics for overlapping communities are the overlapping Normalized Mutual Information (ONMI) [11], the omega index [12], and the average F1-score [13].

1) ONMI: The ONMI (Overlapping Normalized Mutual Information) is an adjustment of the normalized mutual information for overlapping communities [11]. The NMI metric has become one of the most popular metrics when it comes to evaluate the relevance of communities thanks to its reliability [14]. Based on information theory, the NMI measures the similarity between two partitions. Some of the drawbacks of NMI, noted by [15], is the finite size effect which implies that the average score would slide upward with the number of predicted communities, regardless of the number of the ground-truth communities.

2) Omega index: The Omega index [12] is the adaptation to overlapping communities of the Adjusted Rand Index (ARI) [16]. The ARI considers only disjoint partitions. Originally, the Rand Index (RI) is based on the agreement between all pairs of nodes in the graph: a pair of nodes are in agreement if they are assigned to the same communities. The ARI is then improved from of the RI. It considers both the observed and the expected agreement between partitions: an observed agreement is the fraction of pairs of nodes classified the same way in both partitions.

Omega index values are not affected by the number of communities (unlike NMI). However, it performs poorly with multi-resolution partitions and has a high computational complexity.

3) Average F1-score: The average F1-score is the mean of the F1-scores of the best matching ground-truth community to each detected community, and the F1-scores of the best-matching detected community to each ground-truth community [13]. Each F1-score is in fact the harmonic mean of Precision and Recall of considered communities.

One of the drawbacks of F1-score, is that it gives equal importance to precision and recall. It is also computed as an average of community-pairs F1-scores which can lead to high standard deviation.

III. Contribution

The main contribution of this paper relies on the proposition of a set of metrics for overlapping community detection evaluation, more specifically ground-truth based validation metrics. The need for a new metric arose from the inadequacy observed with the state-of-the art metrics. Adapted metrics, like ONMI, were proven to give different results than the underlying standard metrics (designed for disjoint communities) [4]. Another shortcoming observed with the available metrics, is that they compare partitions globally. While this approach may lead to good results, it generally misses information concerning similarities and dissimilarities between partitions.

In this paper, we propose four metrics for comparing communities that overlap. More specifically, the proposed metrics should take into consideration the structure of the obtained overlapping communities and compare them with the ground-truth. Therefore, our metrics combine features of both overlap-based metrics and structural-based metrics. The metrics should assess the match between the result of a community detection algorithm and a ground-truth.

These metrics are the *inclusion rate*, the *coverage rate*, the *overlapping rate*, and the *distribution rate*, which will be detailed below. Accordingly, we impose that a good set of result communities should have good scores for all the four metrics. All four metrics must be considered concurrently to fully understand the results, as some metrics may give good scores on some bad results.

In the following, let's consider two sets of communities: a ground-truth $G = G_1, G_2, \ldots, G_n$ of size n and a community detection result $R = R_1, R_2, \ldots, R_m$ of size m. Communities in R and G are overlapping. Essentially these two groups do not have the same number of communities but do contain the same set of nodes. We assume that a community does not contain duplicate nodes.

For all metrics, the input is two set of communities G, and R, and the output is a measure $d(R, G)$.

In order to be a substitute for conventional evaluation metrics, each of the proposed metrics should also fulfill some basic properties: it should be positive $d(R, G) \geq 0$, it should spread its score over its domain $[0, 1]$ (where 0 means the partitions are completely different and 1 means they are identical), and $d(G, G) = d(R, R) = 1$.

A. Inclusion rate

A simple way to define the similarity between two sets of communities R and G, is to consider how well R represents G, taking into account the inclusion of communities in both groups. In order to define the inclusion rate, and the coverage rate, we should start by defining the *precision* and *recall*.

Considering a result community R_i and a ground-truth community G_j, the *precision* defines the number of correctly classified nodes over the volume of the result R_i, and is defined by $\frac{|R_i \bigcap G_j|}{R_i}$. On the other hand, the *recall* defines the number of the correctly classified nodes over the volume of the ground-truth, defined by $\frac{|R_i \bigcap G_j|}{G_j}$. Precision does not account

for under-segmentation errors, while over-segmentation is not reflected in recall.

The inclusion rate we propose is a metric that is meant to measure the embeddedness of result communities into ground-truth communities. The basic idea behind this metric was the need of a measure that estimates the well classified nodes in the result communities compared to the ground-truth communities. For each result community R_i, the individual inclusion rate is given by the maximum precision rate.

$$Inclusion\,rate(R_i) = \max_j(precision(R_i, G_j)) \quad (1)$$

The overall inclusion rate is then defined by the ratio of a weighted sum of the individual inclusion rates divided by the sum of the resulting communities sizes.

$$Inclusion\,rate = \frac{\sum_i Inclusion\,rate(R_i) \times |R_i|}{\sum_i |R_i|} \quad (2)$$

B. Coverage rate

While the inclusion rate regards similarity from the result communities perspective, the coverage rate considers it from the ground-truth angle. Our purpose was to identify two complementary metrics, that account for analogous similarity perspectives. As the inclusion rate is based on the maximum of the precision, the coverage rate is a function of the maximum recall. For a given ground-truth community G_j, the individual coverage rate is given by:

$$Coverage\,rate(G_j) = \max_i(recall(R_i, G_j)) \quad (3)$$

The overall coverage rate is then defined by the ration of a weighted sum of the individual coverage rates by the sum of the ground-truth communities sizes.

$$Coverage\,rate = \frac{\sum_j Coverage\,rate(G_j) \times |G_j|}{\sum_j |G_j|} \quad (4)$$

C. Overlapping rate

The overlapping rate is a metric that relies on the number of overlapping nodes between communities. For every pair of communities, the overlapping rate is given by the ratio of the number of common nodes they share by the smallest size between the two communities:

$$Overlapping\,rate(R_i, R_j) = \frac{|R_i \bigcap R_j|}{min(|R_i|, |R_j|)} \quad (5)$$

For a given partition, the overlapping rate is given by the mean of the overlapping rates of all pairs of communities it contains. Since our metric is an information recovery metric, the overlapping rate between R and G, is given by the equation 6 below. It is a measure of ideal overlap, and ideally, we have a score close to 1.

$$Overlapping\,rate(R, G) = 1 - |Overlapping\,rate(R) - Overlapping\,rate(G)| \quad (6)$$

D. Distribution rate

The distribution rate compares the number of communities each node belongs to in the result communities and in the ground-truth. For each node i, the distribution rate is given by:

$$Distribution\,rate_i = |n_g(i) - n_r(i)| \quad (7)$$

Where n_g and n_r are the number of communities in the ground-truth and the result to which the node i belongs. The distribution rate is given by the mean of all nodes' individual distribution rates.

$$Distribution\,rate = \exp\left(-\frac{\sum_i Distribution\,rate_i}{|V|}\right) \quad (8)$$

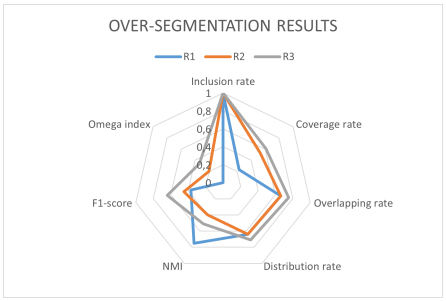
IV. Experiments

A. Data and results

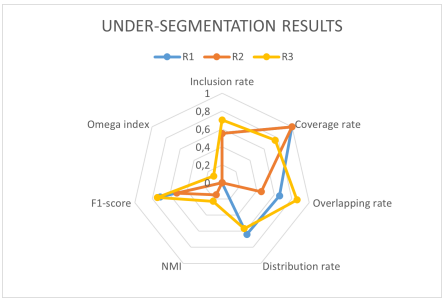
In order to assess the efficiency and test the practical limits of the proposed metrics, we applied them on small manual synthetic network data, and on a classical network of the literature: the US college football network [17]. For the synthetic data, our goal was to generate synthetic result partitions that impact specific aspects such as inclusion, overlap, or number of communities, and test how the metrics behave accordingly. When creating the synthetic partitions, our aim was to lay stress on extreme examples, such as over-segmentation and under-segmentation. While over-segmentation implies there is an over-partitioning of the reference communities, and having a higher number of small communities, under-segmentation refers to grouping multiple communities into one community and having fewer extensive communities than the ground-truth. As for the US college football network, we applied three overlapping community detection algorithms: Angel [18], BigClam [13], and Walkscan [19]. It is to note that our goal is not to compare the performances of the introduced algorithms but to consider implemented metrics on the different outputs they generate. Finally, we computed our proposed metrics along with classic information recovery metrics such as ONMI, F1-score and omega index on the obtained communities. Results are shown in figure 1.

B. Discussion

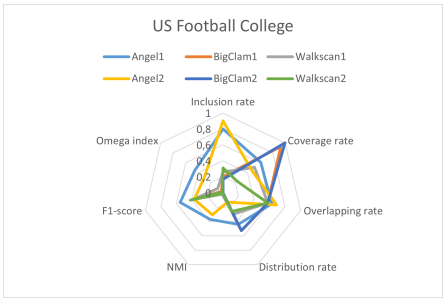
What can be deduced from these results, is that the inclusion rate and the coverage rate are highly complementary. While the first is an indicator of how similar the result communities are to the ground-truth, the second metric informs on how well the communities of the ground-truth are represented in the result. For under-sergmentation cases, we observe that coverage rate values are rather high. Due to their big size, result communities would tend to cover to the utmost ground-truth communities. This subsequently explains high coverage rates. The inclusion rate in the other hand varies from 0.55 to 0.77. For over-segmentation cases, we observe maximal values of the inclusion rate (1). This result confirms that the inclusion rate provides information on how well the result communities



Over segmentation results



Under segmentation results



Us Football College results

Fig. 1: An illustration of the results of the inclusion rate, the coverage rate, the overlapping rate, the distribution rate, NMI, omega index, and F1-score for the synthetic and real data.

are contained in the ground-truth. Especially for the cases of over-segmentation where communities are rather of small sizes, they are more likely to be embedded in one of the ground-truth communities. In contrast, we note that coverage rates are lower than those observed with under-segmentation. Our results show roughly an agreement between the proposed metrics and standard information recovery metrics. For partitions where both the inclusion and coverage rates are high, ONMI, F1-score and omega index are also high.

Results also illustrate how algorithm parameters could affect the metrics to variant degrees. Therefore, our proposed metrics could be employed to tune algorithms based on the type of the intended result. Also, the proposed metrics could be employed to help the user pick the most suitable algorithm for a particular use-case and an intended community structure. Although we noticed the variation of standard extrinsic metrics, the challenging task is to know what it does mean when one result scores 0.1 higher than another. In this regard, our proposed metrics will enable more meaningful perception of the results. The four metrics offer a complete overview of the result partition with respect to the ground-truth.

V. Conclusion

Evaluating the accuracy of community detection results is an important issue, which has received much attention. Several evaluation metrics, specifically information recovery metrics are proposed in the literature. However, real world networks are generally composed of overlapping communities and require adequate metrics. In this paper, we made a step in this direction by presenting four metric definitions for evaluating the similarity of potential overlapping partitions. We tested our metrics on different synthetic data, and on a real-world network. Results show that besides being more meaningful, the proposed metrics present the advantage of being comprehensible. Future work includes further improvement of proposed measures as well as potential correlation with intrinsic metrics.

References

[1] S. Tang, L. Jin, and F. Cheng, "Fraud detection in online product review systems via heterogeneous graph transformer," *IEEE Access*, vol. 9, 2021.

[2] P. Kim and S. Kim, "A detection of overlapping community in mobile social network," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, Mar. 2014, pp. 175–179.

[3] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, Feb. 2010.

[4] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, "Analysis of network clustering algorithms and cluster quality metrics at scale," *PLoS ONE*, vol. 11, no. 7, Jul. 2016.

[5] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Computing Surveys*, vol. 50, no. 4, pp. 54:1–54:37, Aug. 2017.

[6] S. E. Ayeb, B. Hemery, F. Jeanne, and E. Cherrier, "Community detection for mobile money fraud detection," in *Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Dec. 2020, pp. 1–6.

[7] K. Wu, Y. Taki, K. Sato, Y. Sassa, K. Inoue, R. Goto, K. Okada, R. Kawashima, Y. He, A. C. Evans, and H. Fukuda, "The overlapping community structure of structural brain network in young healthy individuals," *PLoS ONE*, vol. 6, no. 5, May 2011.

[8] S. Y. Bhat and M. Abulaish, "Overlapping social network communities and viral marketing," in *International Symposium on Computational and Business Intelligence*, Aug. 2013, pp. 243–246.

[9] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Computing Surveys*, vol. 45, no. 4, pp. 1–35, Aug. 2013.

[10] M. K. Goldberg, M. Hayvanovych, and M. Magdon-Ismail, "Measuring similarity between sets of overlapping clusters," in *IEEE Second International Conference on Social Computing*, Aug. 2010, pp. 303–308.

[11] A. Lancichinetti, S. Fortunato, and J. Kertesz, "Detecting the overlapping and hierarchical community structure of complex networks," *New Journal of Physics*, vol. 11, no. 3, Mar. 2009.

[12] L. M. Collins and C. W. Dent, "Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions," *Multivariate Behavioral Research*, vol. 23, no. 2, pp. 231–242, Apr. 1988.

[13] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *ACM International conference on Web search and data mining*, Feb. 2013, pp. 587–596.

[14] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 9, Sep. 2005.

[15] P. Zhang, "Evaluating accuracy of community detection using the relative normalized mutual information," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015, no. 11, Nov. 2015.

[16] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[17] J. Park and M. E. J. Newman, "A network-based ranking system for US college football," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 10, oct 2005.

[18] G. Rossetti, "Exorcising the Demon: Angel, Efficient Node-Centric Community Discovery," in *International Conference on Complex Networks and Their Applications*, 2019, pp. 152–163.

[19] A. Hollocou, T. Bonald, and M. Lelarge, "Improving PageRank for local community detection," *arXiv:1610.08722 [physics]*, Nov. 2016.