Computational Intelligence
AN INTERNATIONAL JOURNAL

Computational Intelligence, Volume 0, Number 0, 2016

CORRECTION FOR CLOSENESS: ADJUSTING NORMALIZED MUTUAL INFORMATION MEASURE FOR CLUSTERING COMPARISON

ALESSIA AMELIO[1] AND CLARA PIZZUTI[2]

[1]*DIMES, University of Calabria, Rende (CS), Italy*
[2]*National Research Council of Italy (CNR), Institute for High Performance Computing Networking (ICAR), Rende (CS), Italy*

Normalized mutual information (NMI) is a widely used measure to compare community detection methods. Recently, however, the need of adjustment for information theory-based measures has been argued because of the so-called selection bias problem, that is, they show the tendency in choosing clustering solutions with more communities. In this article, an experimental evaluation of these measures is performed to deeply investigate the problem, and an adjustment that scales the values of these measures is proposed. Experiments on synthetic networks, for which the ground-truth division is known, highlight that scaled NMI does not present the selection bias behavior. Moreover, a comparison among some well-known community detection methods on synthetic generated networks shows a fairer behavior of scaled NMI, especially when the network topology does not present a clear community structure. The experimentation also on two real-world networks reveals that the corrected formula allows to choose, among a set, the method finding a network division that better reflects the ground-truth structure.

Received 30 July 2015; Revised 5 July 2016; Accepted 10 July 2016

Key words: complex networks, community structure evaluation, information theoretic evaluation measures, normalized mutual information.

1. INTRODUCTION

Networks are a powerful formalism to represent many real-world systems, such as collaboration, biological, communication and transport, technological systems. Networks consist of a set of objects and a set of interconnections among these objects. Objects, generally, organize in densely connected groups to form a community structure. The detection of such structure is a challenging problem that, in the last few years, has been investigated in several research contexts because of the many potential applications it can be employed (Fortunato 2010).

The availability of effective criteria to evaluate whether a method finds significant groups of nodes that best fit the underlying community organization is an important issue in community detection, because it allows to compare clustering algorithms and choose the solution deemed more appropriate. Measures to empirically analyze the performance of methods have long been studied for clustering objects represented as a set of features (Halkidi et al. 2001, 2002). Among them, indices that assess the quality of a clustering by comparing it with a known division, called *reference* or *ground truth* solution, are particularly important because they allow an objective evaluation of methods. Several measures have been proposed, such as *set matching* based (van Dongen 2000; Meila and Heckerman 2001), *pair counting* based (Hubert and Arabie 1985; Ben-Hur et al. 2002), and *information theory*-based measures (Meilă 2007; Vinh et al. 2010). The *normalized mutual information* (NMI) is one of the most popular information theoretic measures for community detection methods, after Danon et al. (2005) proposed it for comparative evaluation of community detection methods. The authors, in fact, pointed out that the evaluation method proposed by

Address correspondence to Clara Pizzuti, National Research Council of Italy (CNR), Institute for High Performance Computing Networking (ICAR), Via P. Bucci 7/11,87036 Rende(CS), Italy; e-mail: clara.pizzuti@cnr.it

Newman (2004) to measure the number of nodes correctly classified by a method does not count some nodes which may be considered correctly clustered.

Recently, however, it has been pointed out the necessity to $adjust$ information theoretic measures because they show the so-called *selection bias* problem, that is, the tendency of choosing solutions with more clusters when compared with a ground truth division (Vinh et al. 2009, 2010; Romano et al. 2014). This is mainly because these indices, as discussed by Vinh et al. (2009, 2010), do not satisfy the *constant baseline property*, that is, the similarity between random partitions of a data set should be a constant, ideally a zero value. Proposals for corrections have been carried out by Vinh et al. (2009, 2010), and Romano et al. (2014). However, although these adjustments reduce the selection bias problem, we experimentally found that the *normalized mutual information* value is rather high for clusterings where the ratio between the number of nodes and the number of clusters is small, when compared with a reference one.

In this article, we deeply investigate this aspect of mutual information (MI) measures by performing an experimental evaluation that highlights their behavior on synthetic networks for which the ground truth division is known. We then propose a new property that such measures should satisfy, called *reference closeness*, consisting in considering pairs of clusterings having a closer number of communities as more similar. This feature corroborates our intuitive idea of similarity between two divisions. To this end, we suggest to *scale* information theoretic measures proportionally to the difference between the reference and predicted number of clusters. Two scaling functions are proposed and studied. Experimental results show that the scaled measures are able to better exploit the range [0,1] of values they can assume and dampen the undesirable behavior of considering a predicted clustering very similar to the ground truth one even when the former consists of a too few or too high number of communities with respect to the latter.

The article is organized as follows. In the next section, the notation used in the article is first introduced, then a review of the most popular measures proposed in the literature to compare community detection methods is reported. In Section 3, the information theory-based measures are recalled. In Section 4, the problem of selection bias, inherent to information theoretic-based measures, is described. In Section 5, the unfair behavior of these measures is investigated on artificial clusterings. Section 6 introduces the *reference closeness property* and proposes to multiply MI measures by a *scaling factor* that allows to reduce the selection bias problem. Section 7 shows the fair behavior of the scaled measures on the same clusterings of Section 5 and compares some well-known methods with respect to NMI and its adjusted versions by using synthetic generated networks and two real-world networks. Finally, Section 8 concludes the article and discusses the advantages of adopting the scaled NMI.

2. DEFINITIONS AND RELATED WORK

Let $G = (V, E)$ be the graph modeling a network $\mathcal{N}$, where V is the set of n nodes constituting the network, and E the set of m edges connecting couples of elements of V. A *community structure*, or *clustering*, on V is a partition $A = \{A_1, \ldots, A_R\}$ of V in a number R of subsets, such that $\cup_{i=1}^{R} A_i = V$ and $A_i \cap A_j = \emptyset$.

Given two partitions $A = \{A_1, \ldots, A_R\}$ and $B = \{B_1, \ldots, B_S\}$ of V, the overlap between A and B can be represented through the *contingency table* C (Table 1), also called *confusion matrix*, of size $R \times S$, where n_{ij} denotes the number of nodes that clusters A_i and B_j share.

Community detection methods, as pointed out in Fortunato (2010), generally discover different community structures because of the variety of adopted strategies and functions

TABLE 1. Contingency Table Between Clusterings A and B.

	B_1	B_2	...	B_S	Sums
A_1	n_{11}	n_{12}	...	n_{1S}	a_1
A_2	n_{21}	n_{22}	...	n_{2S}	a_2
...	...	...	...	...	...
A_R	n_{R1}	n_{R2}	...	n_{RS}	a_R
Sums	b_1	b_2	...	b_S	n

they optimize. Thus, an important issue is the availability of validity indices that assess the quality of the results obtained by an algorithm. As regards clustering objects represented as feature sets, there has been more research and a plenty of validity indices have been defined. They have been classified in Halkidi et al. (2001) and Halkidi et al. (2002) as internal, when they rely on characteristics inherent the data and evaluate the fit between the data and the expected structure, relative when the clustering structure is compared with other clustering schemes, and external, if they use additional domain knowledge to assess the clustering outcomes.

These indices have often been borrowed and modified to evaluate the many methods proposed for community mining. Rabbany et al. (2013), especially, considered the same classification scheme for community detection methods, investigated quality criteria for internal, relative, and external evaluation and modified some validity indices to make them apt for network data.

In the following, we consider only external criteria, that is, indexes that evaluate a community structure by comparing it with a so-called *gold-standard ground-truth* community partitioning for which nodes are explicitly labeled.

These kinds of external measures for comparing network clusterings have been classified as *set matching*, *pair counting*, and *information theoretic*-based measures (Meilă 2007; Vinh et al. 2010). Set matching measures compute the best match for each cluster and then sum up these contributions. As discussed in (Meilă 2007), this approach generates the so-called problem of matching because the unmatched part is completely disregarded, that is, two clusterings C_1 and C_2 could obtain the same evaluation because of the equivalence between matched parts, but be very different on the assignment of the remaining elements to clusters.

Pair counting-based measures count the pairs of nodes on which two clusterings overlap. Given two divisions A and B, these measures compute, among the possible pairs $\binom{n}{2}$ of nodes, the number n_{11} of pairs appearing in the same cluster in both A and B, the number n_{00} of pairs that appear in different clusters in both A and B, the number n_{10} of pairs appearing in the same cluster in A but in different clusters in B, and the number n_{01} of pairs that are in the same cluster in B and not in A.

The *Rand Index*, introduced by Hubert and Arabie (1985), is one of the most popular measures in this class, and it is defined as $RI(A, B) = (n_{11} + n_{00})/\binom{n}{2}$.

Information theoretic measures are based on information theory (Cover and Thomas 1991) and will be described in detail in the next section. In the following, the most recent proposals for new evaluation criteria or extensions of existing ones are reported.

Criteria that take into account topological and functional properties of communities have been proposed by Orman et al. (2012). The authors observed that two algorithms can obtain the same performance but on solutions having rather different link distribution. Thus,

they suggest that the comparison between community structures should consider these differences. To this end, they introduced a number of measures that take into account the neighbors of a node, the community size distribution and density, the reachability of nodes inside the same community, and the presence of hub nodes, that is, nodes connected with many others inside the same community. By comparing some of the most popular community detection methods with respect to these new topological indexes, the authors found that performance of algorithms and topological properties do not always agree, that is, divisions with high values of measures such as Rand Index sensibly differ from the reference one.

Labatut (2015) has investigated three well-known evaluation measures, *Purity* (Manning et al. 2008), *Rand Index*, and *Normalized Mutual Information*, and pointed out their limitations. The author experimentally found that both Purity and NMI favor algorithms that obtain many small communities. Moreover, because of the results of the study of Orman et al. (2012), he suggests to modify these indexes to take into account the role of each node in the network topology when computing the closeness between two partitions. This is obtained by assigning a weight to each node. The choice of the more appropriate weight is not an easy task. The author proposes a value that combines the node degree and the number of connections it has in its community. Experiments on the results of some well-known community detection algorithms showed to be consistent with the results found by Orman et al. (2012), and that, among the three measures, the modified NMI is able to assess the similarity between a reference and predicted clustering in terms of both membership and topological properties.

Zhang (2015) performed an analytic and experimental study showing that NMI has a systematic bias when evaluating methods obtaining different numbers of groups, because it prefers algorithms obtaining a large number of partitions. The author shows that this is due to the *finite size effect* of entropy, which is different when considering an infinite or finite number of nodes. To fix this problem, he proposes to consider the statistical significance of NMI by comparing it with the NMI of a null model. Given the ground truth partition A, Zhang chooses a random partition C having the same cluster size distribution of the predicted partition B and defines the *relative normalized mutual information (rNMI)* as $rNMI(A, B) = NMI(A, B) - \langle NMI(A, C)\rangle$, where $\langle NMI(A, C)\rangle$ is the expected NMI between A and C, averaged on different random configurations of C. A comparison between the results of different community detection methods on synthetic networks shows that a method deemed more accurate than another with respect to NMI can obtain lower $rNMI$.

Extensions of the NMI measure for hierarchical community structure have been proposed by Perotti et al. (2015) and for overlapping partitions by Lancichinetti and Fortunato (2009), McDaid et al. (2011), and Rabbany and Zaïane (2015).

Perotti et al. (2015) generalized the MI to compare hierarchical structures represented as trees. The MI of two subtrees is assumed to be a null value if one of the two subtrees is a leaf; otherwise, it is computed by recursively considering the descendants of the set of nodes shared between the two subtrees.

Lancichinetti and Fortunato (2009) proposed an extension of NMI to compare community structures where nodes can appear in more than one community. The measure is based on the normalization of the variation of information, introduced by Meilă (2007). McDaid et al. (2011) found that this kind of normalization often overestimates the similarity of two clusters; thus, they proposed a different normalization factor that avoids the problem.

Recently, Rabbany and Zaïane (2015) proposed a generalized clustering distance that encompasses agreement measures belonging to the pair counting and information theoretic families. This distance is based on two functions η and ϕ, where the former quantifies the similarity between two partitions by building their contingency table, and the latter computes the dispersion in each row of this table. The advantage of this generalized measure is

that it can derive other evaluation indexes by properly choosing the two functions. Moreover, it can be extended to deal with overlapping clusters and to take into account the data structure. Experiments on the clusterings obtained by some state-of-the-art community detection methods on synthetic networks show that the ranking of these algorithms with respect to agreement measures is consistent with the proposed distance.

In the next section, a detailed description of information theoretic measures is reported.

3. INFORMATION THEORETIC MEASURES FOR COMPARING COMMUNITY STRUCTURES

Information theoretic measures are based on the information theory concepts (Cover and Thomas 1991; Vinh et al. 2010) of *entropy* (formula (1)), *joint entropy* (formula (2)), and *conditional entropy* (formula (3)). These concepts are defined in terms of the elements of the contingency table as follows:

$$H(A) = -\sum_{i=1}^{R} \frac{a_i}{n} log \frac{a_i}{n} \tag{1}$$

$$H(A, B) = -\sum_{i=1}^{R} \sum_{j=1}^{S} \frac{n_{ij}}{n} log \frac{n_{ij}}{n} \tag{2}$$

$$H(A \mid B) = -\sum_{i=1}^{R} \sum_{j=1}^{S} \frac{n_{ij}}{n} log \frac{n_{ij}/n}{b_j/n} \tag{3}$$

The MI between two clusterings A and B is then defined as

$$I(A, B) = -\sum_{i=1}^{R} \sum_{j=1}^{S} \frac{n_{ij}}{n} log \frac{n_{ij}/n}{a_i b_j/n^2} \tag{4}$$

The MI of two clusterings is the amount of information that one clustering has about the other. This can be expressed as

$$I(A, B) = H(A) - H(A \mid B) = H(A) + H(B) - H(A, B) \tag{5}$$

In order to compare clusterings, the normalized version of MI in the range [0,1] is preferred, where 0 means no similarity between A and B, and 1 that A and B coincide. Several normalizations have been considered in the literature because MI is upper-bounded by the following:

$$\begin{aligned} I(A, B) \leq min\{H(A), H(B)\} \leq \sqrt{H(A)H(B)} \leq \\ \frac{1}{2}\{H(A) + H(B)\} \leq max\{H(A), H(B)\} \leq H(A, B) \end{aligned} \tag{6}$$

Depending on the type of normalization, different versions of the *Normalized Mutual Information* can be obtained. In Table 2, the most popular ones, as described in Vinh et al. (2010), are reported.

TABLE 2. Different Normalizations of Mutual Information.

NMI	Expression	
NMI_{sum}	$\frac{2I(A,B)}{H(A)+H(B)}$	Kvalseth (1987)
NMI_{joint}	$\frac{I(A,B)}{H(A,B)}$	Yao (2003)
NMI_{max}	$\frac{I(A,B)}{max\{H(A),H(B)\}}$	Kvalseth (1987)
NMI_{sqrt}	$\frac{I(A,B)}{\sqrt{H(A)H(B)}}$	Strehl and Ghosh (2002)
NMI_{min}	$\frac{I(A,B)}{min\{H(A),H(B)\}}$	Kvalseth (1987)

In particular, NMI_{sqrt} has been used in Kvalseth (1987) and Strehl and Ghosh (2002) for ensemble clustering, while NMI_{sum} has been proposed by Danon et al. (2005) as a reliable measure for evaluating the similarity of community structures obtained by community detection methods. NMI_{sum} is one of the most used normalizations of MI; thus, in the following, we refer to it as NMI, without the subscript.

Vinh et al. (2010) pointed out that the variants of NMI can be used as distance measures by subtracting them from 1 and showed that $d_{joint} = 1 - NMI_{joint}$ and $d_{max} = 1 - NMI_{max}$ are metrics because they satisfy the properties of positive definiteness, symmetry, and triangle inequality, while d_{sum}, d_{sqrt}, and d_{min} are not metrics. Thus, they suggest that d_{joint} and d_{max} are preferable with respect to the other measures. The *variation of information* (VI), defined by Meilă (2007) as $VI(A, B) = H(A) + H(B) - 2I(A, B)$, is an example of metric.

4. SELECTION BIAS OF MEASURES

Vinh et al. (2009, 2010) argued that a measure comparing two independent clusterings, such as, for example, clusterings sampled at random, should have the *constant baseline property*, which is their expected similarity value should be a constant, ideally equal to zero to indicate no similarity. Measures without this property have the so-called *selection bias*, that is, they have the tendency of selecting clusterings having a higher number of clusters. Vinh et al. (2010) showed that the information theoretic measures do not satisfy the constant baseline property, and that a correction for chance is needed in some particular situations, such as the number of clusters of the two partitions sensibly differs.

This problem was already pointed out by Hubert and Arabie (1985) for the Rand Index, and a proposal of index corrected for chance was formulated as follows:

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{7}$$

where the expected value is that obtained when two partitions are chosen at random with the constraints of having the same number of groups and the same number of elements in each group. They computed the expected value of the Rand Index and then proposed the *Adjusted Rand Index* as follows:

$$ARI(A, B) = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{(n_{00} + n_{01})(n_{01} + n_{11}) + (n_{00} + n_{10})(n_{10} + n_{11})} \tag{8}$$

Vinh et al. (2009, 2010) adopted the same form of correction of Hubert and Arabie and defined the *Adjusted Mutual Information (AMI)*, as follows:

$$AMI(A, B) = \frac{NMI(A, B) - E\{NMI(A, B)\}}{1 - E\{NMI(A, B)\}} \tag{9}$$

where $E\{NMI(A, B)\}$ is the expected MI between A and B.

Romano et al. (2014), however, showed that this corrected formula has the same selection bias of NMI, although in a much dampened form. The same authors, thus, proposed a standardization of NMI, named SMI, based on the variance of the MI. To standardize the measure, it is necessary to compute the number of standard deviations that the MI is away from the mean value. The standardized formula SMI does not present the bias, but its computational complexity is rather high, being $O\left(max\left\{RSn^3, S^2n^3\right\}\right)$, where R and S are the number of rows and columns of the confusion matrix, and n is the data set size. Although the authors suggest a parallel implementation of the formula, even for moderately low values of these parameters, comparing two partitions to obtain the standardized MI is very computing demanding, and unfeasible for thousands of nodes.

In the next section, we investigate more in depth the selection bias problem, by considering in particular the measures NMI, NMI_{joint}, NMI_{max}, AMI, and SMI.

5. NORMALIZED MUTUAL INFORMATION UNFAIRNESS

Let A and B be the ground truth division in R communities of a network constituted by n nodes, and the partitioning in S communities obtained by a method, respectively. The NMI $NMI(A, B)$ of A and B can be written in terms of contingency table as follows:

$$NMI(A, B) = \frac{-2\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} log\left(n_{ij}n/a_i b_j\right)}{\sum_{i=1}^{R} a_i log(a_i/n) + \sum_{j=1}^{S} b_j log\left(b_j/n\right)} \tag{10}$$

Vinh et al. (2010) argued about the *unfair* behavior of information theoretic measures when the ratio n/S is small. In fact, if the similarity of two clusterings B and B' must be evaluated with respect to a true partitioning A, the measures of MI, RI, and NMI monotonically increase as the number of obtained clusters increases.

Here, we want to emphasize that, in the extreme case of a solution B having a number of clusters equal to the number n of nodes, the NMI value between A and B should be zero, because knowing B gives no information about A; thus, there is not any reduction of uncertainty about A. Instead, the NMI value depends only on A and n because the expression $NMI(A, B) = 2I(A, B)/\left(H(A) + H(B)\right)$ reduces to $2H(A)/\left(H(A) + nlog(1/n)\right)$, being $H(A \mid B) = 0$ and $H(B) = nlog(1/n)$ (Cover and Thomas 1991). In fact, suppose $A = \{A_1, \ldots, A_R\}$ be the ground truth division of a network in R communities, and $B = \{B_1, \ldots, B_n\}$ a division constituted by n singleton clusters. In Equation (10), we will have n_{ij} is either 1 or 0 $\forall i, j, b_j = 1$, thus

$$\sum_{j=1}^{S} n_{ij} log(n_{ij}n/a_i b_j) = \sum_{j=1}^{n} n_{ij} log(n/a_i) = a_i log(n/a_i) \tag{11}$$

TABLE 3. Contingency Table Between Two Clusterings A and B.

	B_1	B_2	
A_1	39	11	50
A_2	11	39	50
	50	50	100

$$\sum_{j=1}^{S} b_j log(b_j/n) = \sum_{j=1}^{n} log(1/n) = nlog(1/n) \tag{12}$$

$$NMI = \frac{-2\sum_{i=1}^{R} a_i log(n/a_i)}{\sum_{i=1}^{R} a_i log(a_i/n) + nlog(1/n)} \tag{13}$$

that is, it depends only on the clustering A and the number n of nodes. In these cases, the use of NMI to compare community detection results can lead to wrong conclusions. Consider the toy example in which the contingency table between the ground truth clustering A and a division B is that of Table 3. In this case, $NMI(A, B) = 0.2398$. If we now consider a division B' of 50 singleton clusters $NMI(A, B') = 0.2616$, thus B' is evaluated as a better partition than B, which is not intuitive because if B is given, for the 78 % of nodes, the knowledge that a node u is a member of a cluster in B gives the correct information regarding the true cluster u should belong to. Instead, the knowledge that u appears in a cluster of B' does not reduce the uncertainty about the membership to a reference community for all the n nodes.

In order to show the unfair behavior of information theoretic measures, we performed two types of experimentations. The former, analogously to Romano et al. (2014), evaluates the measures on randomly generated partitions, and the second one considers partitions of a network that either *refine* or *merge* ground truth communities.

5.1. Random Cluster Generation

The first experiment is analogous to that reported by Romano et al. (2014). We consider a reference clustering of $R = 10$ clusters of equal size for a network constituted by $n = 500$ nodes. Then, six random clusterings are generated with increasing number of communities $S = \{2, 6, 10, 14, 18, 22\}$, and the values of each measure are computed. The solution that obtains the highest value, for each evaluation index, is counted as a win. The winning frequencies for 5,000 trials are reported in Figure 1(a–e), while the average values of measures are displayed in Figure 1(f–j).

Figure 1 clearly points out the selection bias of *NMI, joint NMI*, *max NMI*, and AMI towards the solutions having the highest number of clusters, that is, 18 and 22, although AMI has a less pronounced bias than the others. SMI actually shows the constant baseline property because every random clustering has the same constant value to be chosen. The results confirm those obtained by Romano et al. (2014), but it is worth to observe that, on average, the values of SMI computed for community structures consisting of 18 and 22 communities are higher than those with a lower number of communities. However, the main

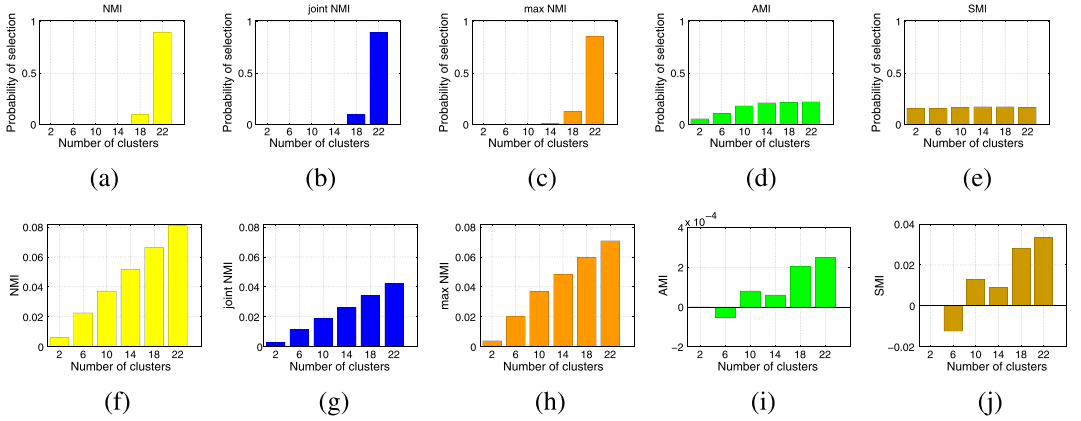


FIGURE 1. (a–e) Selection probability of random clusterings with increasing number of communities when compared with a reference community structure having 10 clusters with 50 nodes each, (f–j) values of measures.

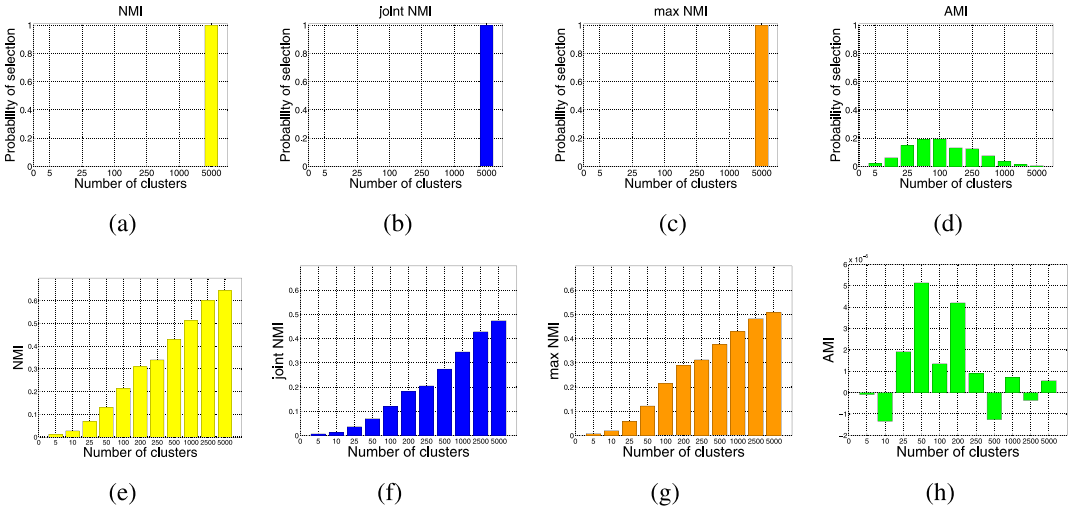


FIGURE 2. (a–d) Selection probability of random clusterings with increasing number of communities when compared with a reference community structure having 100 clusters with 50 nodes each, (e–h) values of measures.

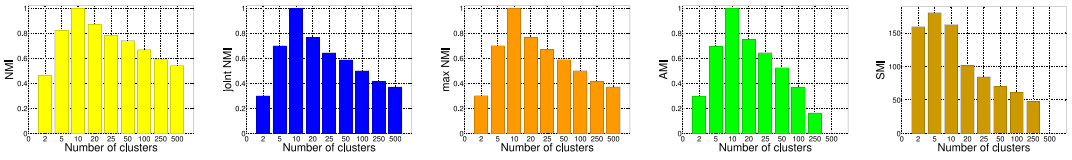
problem of SMI is that it cannot be practically used as a measure to compare clusterings because of the too high-computational resources it needs. To compute SMI for this example of 500 nodes, we needed 120 h on a cluster computer with 32 cores, 2.6 GHz, and 48 GB RAM, which is impractical for real situations.

In order to better investigate the constant baseline property, we repeated the experiment with $n = 5{,}000$ nodes, a reference clustering of $R = 100$ communities of equal size, and generated random clusterings with number of communities $S = \{5, 10, 25, 50, 100, 200, 250, 500, 1{,}000, 2{,}500, 5{,}000\}$, thus taking into account also the limit case of clusters constituted by singleton nodes. The SMI values could not be computed because of its cubic complexity in n. Figure 2(a–c) shows that NMI, *joint NMI*, and *max NMI* always prefer the degenerate solution of 5,000 singleton communities, and

Figure 2(e–g) highlights the abnormal values of these measures for such a solution, which can reach the values 0.65, 0.48, 0.5, respectively. On the other hand, although AMI does not any more satisfy the constant baseline property, it does not present the selection bias problem; instead, it has a selection probability biased towards the ground truth solution, that is, 100 clusters, followed by the nearest one with 50 clusters, and then with decreasing probabilities for the other values.

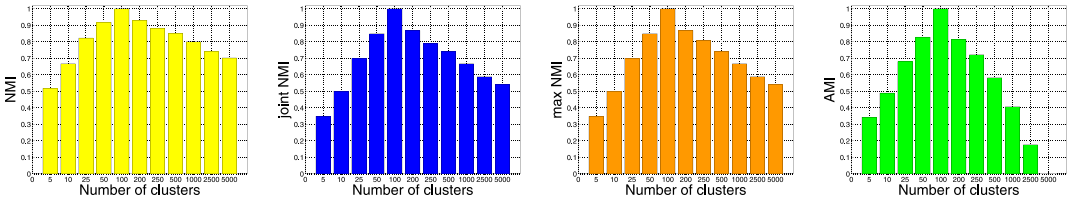
5.2. Join and Refine

In this experiment, we consider the set of clusterings that can be obtained from a reference clustering A by progressively splitting and merging its clusters. We first define the concept of clustering *refinement*, as introduced by Meilă (2007), and that of $merging$. A clustering B *refines* ($merges$) a clustering A if for each community $A_i \in A$ there is a unique community $B_j \in B$ such that $B_j \subseteq A_i$ ($B_j \supseteq A_i$). Thus, a refinement of A is obtained by splitting some clusters of A, while a merging by joining some of them. We considered the two reference clusterings of $R = 10, 100$ communities of equal size, with $n = 500, 5{,}000$ nodes, of the previous experiment and generated refined and merged clusterings by splitting or merging the original 10 and 100 clusters in $\{2, 5, 10, 20, 25, 50, 100, 200, 250, 500\}$ and $\{5, 10, 25, 50, 100, 200, 250, 500, 1{,}000, 2{,}500, 5{,}000\}$, respectively, communities of equal size. Figures 3 and 4 show the values of NMI, *joint NMI* and *max NMI*, AMI, and SMI (only for the $n = 500$ example). The figures highlight more clearly the counterintuitive behavior of the first three measures that assume very high values even when the original clusterings are either split or merged in a number of clusters rather far from the original one. For example, a clustering constituted by 2,500 communities with couples of nodes has an NMI value between 0.7 and 0.8. It is also worth to observe that NMI does not exploit the nominal range [0,1], as already observed by Vinh et al. (2010), but it ranges in the narrower intervals [0.43,1] and [0.5,1] when $n = 500$ and $n = 5{,}000$, respectively. In the next section, we focus only on NMI, considering that *joint* and *max NMI* have similar behavior.



(a) (b) (c) (d) (e)

FIGURE 3. Measure values of clusterings with merged/refined communities when compared with a reference community structure having 10 clusters with 50 nodes each.



(a) (b) (c) (d)

FIGURE 4. Measure values of clusterings with merged/refined communities when compared with a reference community structure having 100 clusters with 50 nodes each.

6. ADJUSTMENT FOR CLOSENESS

In this section, we propose to $correct$ the NMI measure to direct it toward the number of ground-truth clusters. We, thus, suggest to $adjust$ the NMI with respect to a new property, called the *reference closeness property*, that is, the closer the number of communities found by an algorithm to the number of clusters of the reference clustering, the higher the MI value should be. This desirable behavior of any evaluation criterion has been pointed out by Rabbany et al. (2013), where they state that the "ideal behaviour of an index should be that it gives low scores for partitionings/fragmentations in which the number of clusters is much higher or lower than what we have in a ground-truth." Moreover, Vinh et al. (2010) experimentally found that, in the context of consensus clustering Strehl and Ghosh (2002), a clustering having a number of clusters coincident with the true cluster number is more robust.

The extensive experimentation reported in the previous section has pointed out that NMI value is factitiously high when the clustering obtained by a method is constituted by many small groups, that is, as already argued by Vinh et al. (2010), the ratio n/S is small, where n is the number of nodes and S is the number of predicted clusters. A viable way to face this problem could be to dampen this value when S is far from the true number of clusters. We thus propose to $scale\ NMI$ by multiplying it with a *scaling factor* that diminishes its value as the difference between the true number R of clusters and the predicted number S of clusters increases. A desirable property of the scaling factor should be that, if we assume the number R of clusters of the reference clustering as the mean value of the differences between the predicted number S of clusters and the expected true number R, the shape of the function should follow the typical *bell* curve of a Gaussian distribution. In such a way, the *scaled* NMI should not differ too much from the not scaled NMI for those methods for which the difference $R - S$ in absolute value is low but punish methods that obtain a number of clusters either too higher or too lower than the true number. *Scaled* NMI would thus have a *fairer* behavior toward the former methods and reduce the selection bias problem.

The *scaling factor sf* we propose is defined as follows:

$$sf = \alpha e^{-\frac{|S-R|^{\beta}}{\gamma}} \tag{14}$$

Because formula (14) is defined in terms of three parameters α, β, γ, by varying their values, there can be infinitely many choices of scaling factors. We tested different combinations of these parameters. By setting $\alpha = \beta = 1$ and $\gamma = R$, we obtained a shape that resembles a normal distribution, although it never assumes a zero value. By combining different other values for α, β and γ, we obtained functions similar to the Gaussian function. Thus, we experimented two types of scaling factors. The first one is a Gaussian function having $\alpha = \frac{1}{\sqrt{2\pi}\sigma}$, $\beta = 2$, $\gamma = 2\sigma^2$, where σ is the standard deviation between the predicted number S of clusters and the expected true number R, the second one has $\alpha = 1$, $\beta = 1$, $\gamma = R$.

Thus, we have

$$Gauss\ NMI = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|S-R|^2}{2\sigma^2}} \times NMI \tag{15}$$

$$FNMI = e^{-\frac{|S-R|}{R}} \times NMI \tag{16}$$

It is worth to note that, as regards formula (15), when the predicted number S and the true number R of communities are the same, the standard deviation is zero. In this case, it

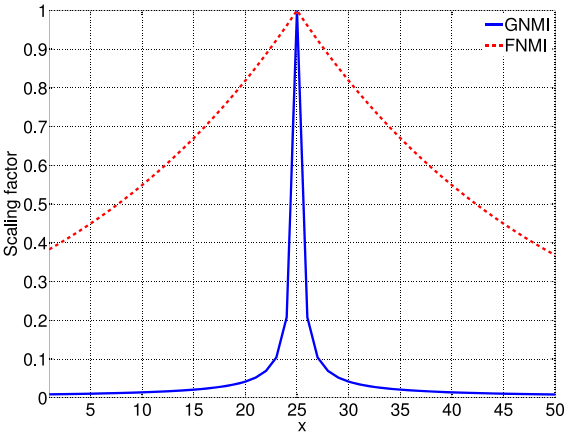


FIGURE 5. Scaling factors for $R = 25$ and S varying in the interval $[1, \ldots, 50]$.

is known that the Gaussian function reduces to the Dirak pulse of unit area (Parker 2009). Thus, we can then safely assume that the scaling factor $sf = 1$.

Analogously for formula (16), when $R = S$, the exponent of the exponential function is 0, and the scaling factor is 1. Thus, when $R = S$, the value of NMI is not changed. However, as the difference between R and S increases, both if either a lower or a higher number S of communities is obtained, the value of NMI proportionally decreases, because it is scaled by a factor sf.

The behavior of these scaling factors can be seen in Figure 5 for an example having $R = 25$ and S varying in the interval $[1, \ldots, 50]$. It is known that σ determines the width of the Gaussian function. Thus, the main difference between *Gauss NMI* and $FNMI$ is their amplitude, which is very sharp for the former and smooth for the latter. This implies that when the predicted number S of clusters deviates from the expected true number R, the NMI value is scaled either quickly or in a soft way.

A toy example that explains the effect of the scaling factors for a small data set clustered in two ground-truth clusters of 10 nodes each, where cluster membership is denoted by the plus (cluster 1 in the Figure) and triangle (cluster 2 in the Figure) symbols, is shown in Figure 6, on the left. If the data set is divided in a clustering B consisting of two clusters with mis-clustered nodes, or in a clustering C where the first cluster is split into two pure groups, or a clustering D with five clusters, then for clustering B, we have the following:

$$S = 2, R = 2, NMI(A, B) = 0.1187$$

$$FNMI(A, B) = e^{-\frac{|2-2|}{2}} \times 0.1187 = 0.1187$$

$$\sigma_{S,R} = \sigma_{2,2} = 0, \frac{1}{\sqrt{2\pi}0} \times e^{-\frac{|S-R|^2}{2\times 0^2}} \approx 1$$

$$GNMI(A, B) = \approx 1 \times 0.1187 = 0.1187$$

for clustering C, we have

$$S = 2, R = 3, NMI(A, C) = 0.8000$$

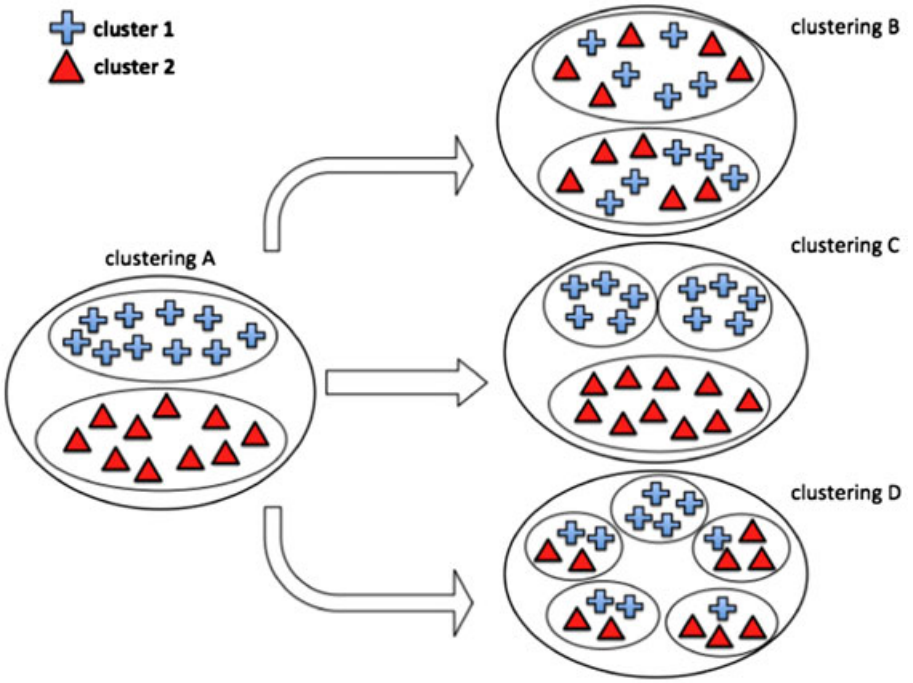


FIGURE 6. A toy example of 20 nodes divided in two groups. Three different clusterings are reported on the right: a clustering B with the same number of clusters but mis-clustered nodes, a clustering C with three pure groups, and a clustering D with four mixed groups and a pure group.

$$FNMI(A,C) = e^{-\frac{|2-3|}{3}} \times 0.8000 = 0.4852$$

$$\sigma_{S,R} = \sigma_{2,3} = 0.7071$$

$$GNMI(A,C) = \frac{1}{\sqrt{2\pi} \times 0.7071} \times e^{-\frac{|2-3|^2}{2\times 0.7071^2}} \times 0.8000 = 0.1660$$

Finally, for clustering D, we have

$$S = 2, R = 5, NMI(A,D) = 0.1659$$

$$FNMI(A,D) = e^{-\frac{|2-5|}{5}} \times 0.1659 = 0.0370$$

$$\sigma_{S,R} = \sigma_{2,5} = 2.1213$$

$$GNMI(A,D) = \frac{1}{\sqrt{2\pi} \times 2.1213} \times e^{-\frac{|2-5|^2}{2\times 2.1213^2}} \times 0.1659 = 0.0115$$

In the next section, we repeat the experimentation on the clusterings of the previous section and show that the modification proposed for MI measures sensibly dampens the selection bias problem and allows a better exploitation of the range [0,1] of values that the information theoretic-based measures can assume.

7. EXPERIMENTAL EVALUATION ON SCALED NORMALIZED MUTUAL INFORMATION

In this section, we reconsider the clusterings with 5,000 nodes of the Section 5 and show the values of the scaled measures. Moreover, the *constant baseline property*, investigated by Vinh et al. (2010), is studied also for these measures. Then, we perform a comparative evaluation among community detection methods on synthetic generated networks and two real-life networks. For the former data set, we also show the values of the $rNMI$ measure of Zhang (2015).

7.1. Random and Merged/Refined Communities

Figure 7 depicts the selection probability and the values of NMI, *joint NMI*, and *max NMI* scaled according to formulas (15) and (16) for the network of 5,000 nodes. Because the values of NMI, joint and max NMI are similar, only the values of *Gauss NMI* are reported. As can be observed from the figures, the selection probability is biased toward the true number of communities for all the measures. As regards the values of the MI, Figure 7(e–g) and (i–k) shows that the scaled values are now well distributed around the reference number of communities, both for the randomly generated clusterings and those obtained by merging/refining the 100 true clusters. In particular, it is worth to note that, for

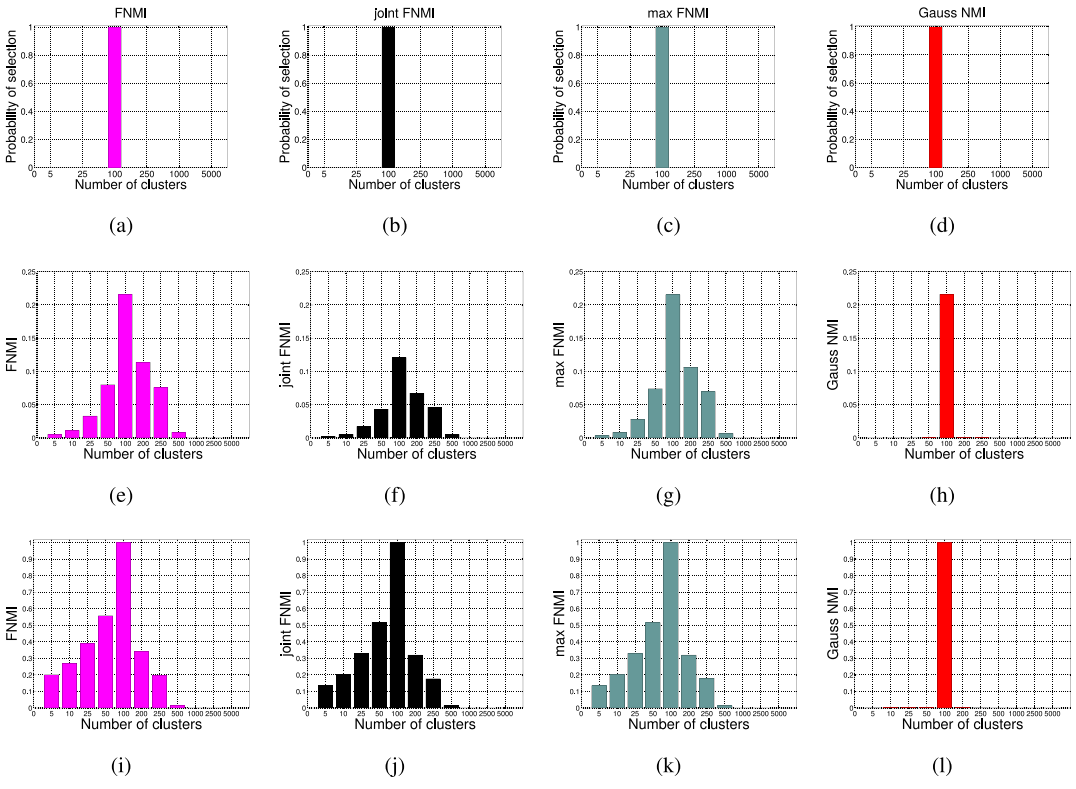


FIGURE 7. Selection probability (a–d) and values (e–h) of the adjusted measures for random clusterings with increasing number of communities when compared with a reference community structure having 100 clusters with 50 nodes each. (i–l) Adjusted measure values of clusterings with merged/refined communities.

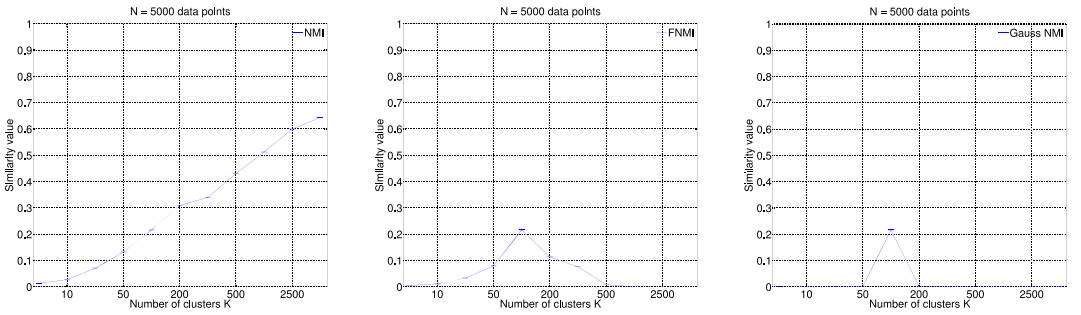
community structures with more than 500 communities, the values are near zero; thus, the not intuitive behavior of assigning high values of NMI to solutions with clusters of one or two nodes, and, in general, very few nodes with respect to the number of nodes of true clustering, is avoided. *Gauss NMI*, because of the very deep decrease it induces on NMI values when the difference between R and S increases, shows a severe penalty for solutions having S different from R, even for small differences.

7.2. Similarity Between Random Divisions

In this section, we investigate the corrected measures with respect to the *constant baseline property* that should be satisfied by any similarity index, that is, two clusterings sampled independently at random, should have a constant baseline value, ideally equal to zero. Figure 8 reports the values of NMI, $FNMI$, and *Gauss NMI* for the experiment with 5,000 nodes, described in Section 5.1, where also the reference clustering is generated at random. Bars denote standard deviation. Figure 8(a) shows that, as already pointed out by Vinh et al. (2010), the NMI measure does not satisfy the property because the similarity between two random clusterings increases as the number of communities augments. Figure 8(b) and (c) highlights that, although $FNMI$ and *Gauss NMI* have a small increase around the number of clusters of the reference random clustering, the average values are rather close to zero; thus, the reference closeness property of these measures implicitly induces the almost satisfiability of the constant baseline property.

7.3. Comparing Methods on Synthetic Networks

In this section, six very popular methods for community detection are compared on the LFR benchmarks proposed by Lancichinetti et al. (2008). The characteristics of the networks are the same of those reported in Lancichinetti et al. (2008) and extensively used in several papers for comparing the performances of methods (Lancichinetti and Fortunato 2009; Orman and Labatut 2010; Orman et al. 2013). The benchmark networks consist of 5,000 nodes, average node degree 20, maximum node degree 50, minimum community size 10, maximum community size 50, exponent of degree distribution -2, community size distribution -1, and mixing parameter μ varying in the range [0.1,0.8]. μ expresses the ratio between the external and total degree of nodes; thus, the higher its value the more difficult to find community structure because it is less well defined. The algorithms we considered are *Fast Greedy* (Newman 2004) and *Louvain* (Blondel et al. 2008) that are based on modularity



(a) (b) (c)

FIGURE 8. Average similarity for random clusterings. Bars denote standard deviation.

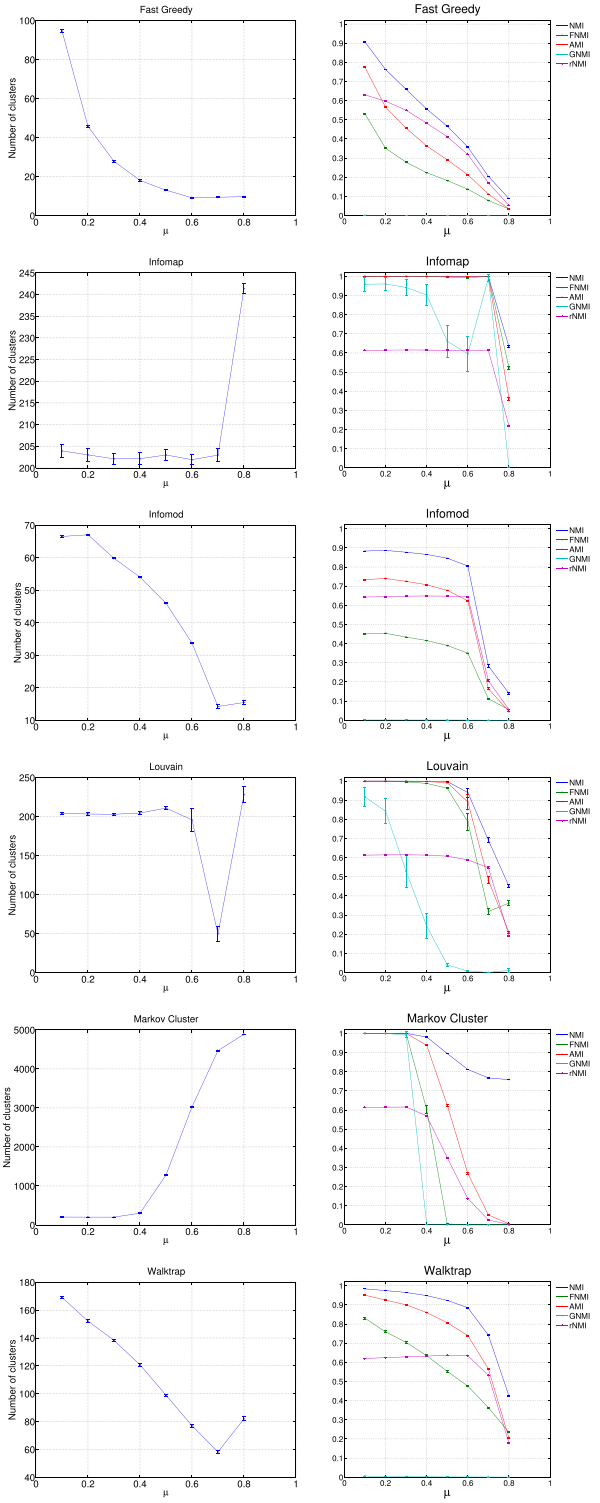


FIGURE 9. Average number of clusters obtained by the methods (left) and corresponding values of measures for synthetic networks. Bars denote standard deviation.

optimization, *InfoMap* (Rosvall and Bergstrom 2008) and *InfoMod* (Rosvall and Bergstrom 2007) that are based on data compression, *WalkTrap* (Pons and Latapy 2005) and *Markov Cluster* (van Dongen 2008) based on the concept of random walk. The former exploits random walks to define the distance between nodes, and the latter simulates random walks, also called flows, within a graph.

We generated 100 different instances of the benchmark and obtained networks with average number of communities 202 for $\mu = \{0.3, 0.4, 0.7, 0.8\}$ and 203 for the other values of μ. Figure 9 shows on the left the average number of clusters obtained by each method, while on the right, the average values of NMI, AMI, $rNMI$, $FNMI$, and *Gauss NMI* (denoted GNMI). Bars denote standard deviation of each measure.

Fast Greedy finds a number of clusters much lower than the true number. This number reduces as the mixing parameter increases; thus when the community structure becomes more difficult to uncover, this algorithm has the tendency to join communities. Actually, it merges too much, because also with $\mu = 0.2$ it gives, on average, only 45 communities. Information theoretic measures diminish as μ increases, as expected, but the NMI values are above 0.4 until $\mu = 0.5$. In such a case, however, *Fast Greedy* obtains only 13 communities out of the 203 it should find. AMI and $rNMI$ values are lower. $FNMI$ values are still lower than AMI and $rNMI$, which seems reasonable, because a predicted division with 13 communities and a true one with 203 do not seem so similar. *Gauss NMI* in this case is too low to be considered as a reliable similarity value.

InfoMap is a very accurate method that is capable to find the correct community structure until $\mu = 0.7$; thus, any evaluation measure does not change its superiority with respect to the other methods. When $\mu = 0.8$, it finds 240 communities, with an NMI value of 0.63; thus, values below 0.5 of the other measures are more coherent.

InfoMod obtains an NMI value above 0.8 until $\mu = 0.6$, although the number of communities it finds is much lower than 203. Also in this case, while $GNMI$ reduces too much because $\mu = 0.1$, $FNMI$ has a deeper decrease than NMI, AMI, and $rNMI$. For instance, when $\mu = 0.4$, *InfoMod* returns 54 communities, $FNMI = 0.41$, while $NMI = 0.92$, $AMI = 0.70$, and $rNMI = 0.65$. This algorithm, thus, merges communities, and it is not able to obtain a partition of nodes close to the ground truth division, while 0.92 and 0.70 values indicate high similarity.

Louvain performs quite well. It obtains a number of communities very close to the true number, except for $\mu = 0.7$. In such a case, it finds an average of 49 communities. Thus, it merges communities, and a value of $NMI = 0.69$ is too high. $AMI = 0.48$, $rNMI = 0.56$, and $FNMI = 0.31$ can be actually considered more consistent.

Markov Cluster clearly shows the selection bias problem discussed in Section 4. It finds the correct partitioning for $\mu \leq 0.3$, but, for $\mu = 0.4$, the number of clusters is 302, and then it drastically splits the networks by returning an average of 1,279, 3,027, 4,469, and 4,890 communities for $\mu = 0.5, 0.6, 0.7, 0.8$ with NMI values of 0.89, 0.81, 0.76, 0.75, respectively. These very high values are implausible and could be misleading when, for example, we have to choose between another method against *Markov Cluster* on the base of the performances measured by an assessment criterion such as NMI. In fact, *Markov Cluster* algorithm outperforms *Fast Greedy*, *InfoMod*, and *WalkTrap* for $\mu = 0.4$, and the first two for $\mu = 0.5, 0.6$ which is plausible, but it is considered better than all the other methods except *InfoMod* for $\mu = 0.7$, and the best for $\mu = 0.8$. $FNMI$ and *Gauss NMI* avoid the biased evaluation of *Markov Cluster* in a safe way by reducing its very good evaluation, due to a too high and inconsistent NMI value, when the mixing parameter is above 0.4.

WalkTrap, analogously to *Fast Greedy*, has the tendency to merge communities as μ increases. Nonetheless, its NMI values are above 0.9 until $\mu \leq 0.6$, which is misleading about its goodness. The lower values of $FNMI$ are more reliable.

It is worth to point out that when the number of communities obtained by a method, such as *Louvain* and *Infomap*, is very close to the ground truth and, consequently, the NMI values are high, the corresponding $rNMI$ values are sensibly lower than NMI. For example, $Infomap$ gives $NMI = 1$ for $\mu \leq 0.7$ and the corresponding $rNMI$ is 0.61. This rather high reduction rate is smoother for lower values of NMI. Neither $FNMI$ nor AMI present this behavior.

7.4. Comparing Methods on Real-World Networks

In this section, we consider two real-world networks studied by Yang and Leskovec (2015) for evaluating quality functions with respect to the ability of obtaining community divisions that resemble the ground-truth partitioning. The first network is the $Amazon$ product co-purchasing network. It consists of 334,863 nodes, 925,872 edges, and 75,149 ground-truth communities. The other network is the $DBLP$ collaboration network where nodes are paper authors and edges connect authors that coauthored a paper. It consists of 317,080 nodes, 1,049,866 edges, and 13,477 ground-truth communities. Both networks can be downloaded from http://snap.stanford.edu/data/. Because it is very difficult to have large size data sets with ground-truth divisions, even if Amazon and DBLP have overlapping communities, we made them nonoverlapping by choosing one of the partitions for those nodes belonging to more than one cluster.

Figure 10 reports the complementary cumulative distribution function of the size of the ground-truth communities in logarithm scale. As observed in Yang and Leskovec (2015), the distributions show the presence of many small communities, but also, large communities can exist. The six methods described in the previous section have been executed on the Amazon and DBLP networks, and the values of NMI, $FNMI$, and $GNMI$ are showed in Tables 4 and 5, respectively, along with the number of communities each method obtains.

The tables point out that $Infomap$, $Infomod$, and $Louvain$ detect a number of communities much lower than the ground-truth on both networks; thus, these methods merge almost all the small communities in bigger ones. The NMI values, however, are never less than 0.34. $FNMI$ properly reduces these values, although $GNMI$, because of the too high difference between the true and predicted number of communities, decreases them too much. $Infomap$, for the Amazon network, finds only 15 communities out of 75,149. Nonetheless, the NMI value is 0.3418, $FNMI$, instead, is 0.1258, which is a more realistic value. Analogously, $Louvain$ obtains 242 communities and an $NMI = 0.5565$, while

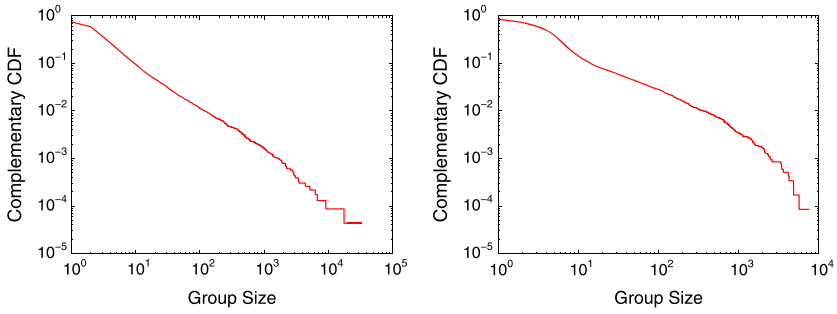


FIGURE 10. Complementary cumulative distribution function of the number of nodes belonging to the ground truth partitioning of (a) Amazon and (b) DBLP networks.

TABLE 4. Comparing Methods on the Amazon Network: 334,863 Nodes, 925,872 Edges, and 75,149 Ground-Truth Communities.

	Measure			
Algorithm	NMI	FNMI	GNMI	nc
Fast Greedy	0.4975	0.1869	$1.4037 \cdot 10^{-6}$	1,595
Infomap	0.3418	0.1258	$9.4413 \cdot 10^{-7}$	15
Infomod	0.5524	0.2037	$1.5293 \cdot 10^{-6}$	184
Louvain	0.5565	0.2054	$1.5421 \cdot 10^{-6}$	242
Markov Cluster	0.6956	0.4754	$5.0491 \cdot 10^{-6}$	46,557
Walktrap	0.6174	0.2537	$1.9170 \cdot 10^{-6}$	8,307

NMI, normalized mutual information; GNMI, Gauss NMI.

TABLE 5. Comparing Methods on the DBLP Network: 317,080 Nodes, 1,049,866 Edges, 13,477 Ground-Truth Communities.

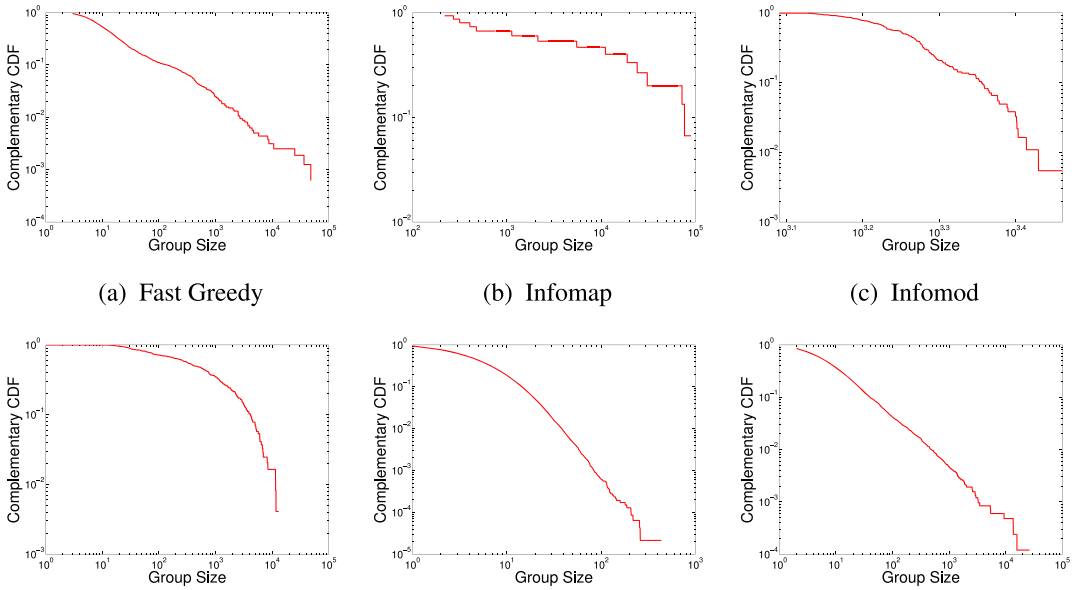
	Measure			
Algorithm	NMI	FNMI	GNMI	nc
Fast Greedy	0.3278	0.1516	$6.5442 \cdot 10^{-6}$	3082
Infomap	0.4298	0.1641	$6.8735 \cdot 10^{-6}$	499
Infomod	0.3726	0.1391	$5.8245 \cdot 10^{-6}$	198
Louvain	0.3622	0.1355	$5.6735 \cdot 10^{-6}$	225
Markov Cluster	0.6577	0.1069	$5.5744 \cdot 10^{-6}$	37964
Walktrap	0.5626	0.3908	$2.3783 \cdot 10^{-5}$	18387

NMI, normalized mutual information; GNMI, Gauss NMI.

$FNMI = 0.2054$. On the other hand, *Markov Cluster* returns 46,557 communities, which is more than the half of true number, with $NMI = 0.6956$, while $FNMI = 0.4754$. For this network, both NMI and $FNMI$ measures consider *Markov Cluster* as the best among the considered methods. This is a plausible result because the other methods find too few communities, including *Fast Greedy* (1,595 clusters out of 75,149) and $Walktrap$ (8,307 communities).

As regards the $DBLP$ network, *Markov Cluster* returns 37,964 clusters, a number much higher than that of ground-truth partitioning which is 13,477, and an $NMI = 0.6577$; thus, it is considered to outperform the other methods. However, while again $Infomap$, $Infomod$, and $Louvain$ merge too much thus are unable to detect small communities, $Walktrap$ in this case obtains 18,387 clusters, which is the nearest value to the ground-truth, with an NMI value of 0.5626, lower than that obtained by *Markov Cluster*. In this case, it is very clear the selection bias of NMI that evaluates this latter method better than $Walktrap$, in spite of the excessively high number of clusters it obtains. $FNMI$ decreases the NMI values from 0.6577 and 0.5626 to 0.1069 and 0.3908, respectively, thus scoring $Walktrap$ better than the other methods, which seems a reasonable conclusion.

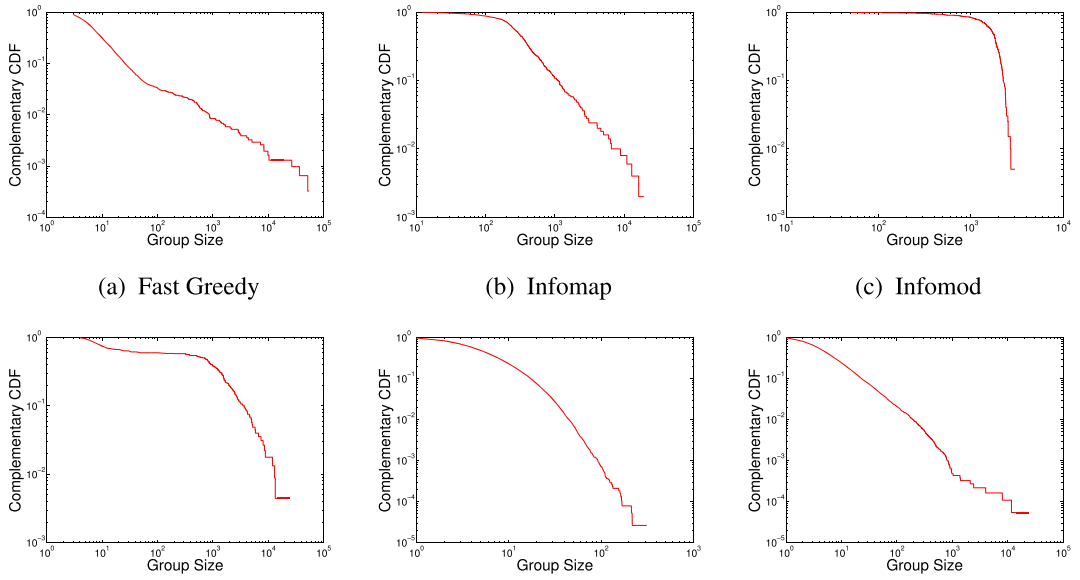
Finally, Figures 11 and 12 display the complementary cumulative distribution functions of the predicted community sizes. The figures confirm that the distributions of the predicted community structures are rather different from those of the true partitioning, except for *Markov Cluster* on the $Amazon$ network and $Walktrap$ on the $DBLP$ network.



(a) Fast Greedy (b) Infomap (c) Infomod

(d) Louvain (e) Markov Cluster (f) Walktrap

FIGURE 11. Complementary cumulative distribution function of the number of nodes belonging to the partitioning obtained by each method for the Amazon network.



(a) Fast Greedy (b) Infomap (c) Infomod

(d) Louvain (e) Markov Cluster (f) Walktrap

FIGURE 12. Complementary cumulative distribution function of the number of nodes belonging to the partitioning obtained by each method for the DBLP network.

8. CONCLUSIONS

In this article, we highlighted unfairness of information theoretic measures for clustering comparison, and the importance of the reference closeness property when there is not a clear community structure. In fact, in this case, some methods, although find either a too few or a too high number of clusters, have an NMI value rather high when compared with the ground truth division. This may be misleading in situations where a method must be chosen among others, and the criterion adopted for selection is based on NMI. The scaled NMI, as experiments showed, gives a more intuitive idea of clustering similarity and guarantees a fair comparison among methods. Thus, in situations where the network structure is difficult to uncover, a scaled NMI can be more useful when performing comparative analysis of methods to assess the superiority of an approach with respect to another. The reference closeness property is a contribution to better discriminate performances of algorithms. Further investigation, however, is necessary to determine other scaling functions that could improve the reliability of information theoretic measures. The experimentation reported in the article shows that, in general, NMI values alone are not sufficient to affirm the superiority of an algorithm. Authors should discuss about topological features of the obtained clusterings, at least reporting the number of communities, because, as also argued in Orman et al. (2011), often high NMI values do not correspond to topological properties similar to those of the ground truth division.

REFERENCES

BEN-HUR, A., A. ELISSEEFF, and I. GUYON. 2002. A stability based method for discovering structure in clustered data. *In* Pacific Symposium on Biocomputing, pp. 6–17.

BLONDEL, V. D., J. L. GUILLAUME, R. LAMBIOTTE, and E. LEFEVRE. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics, **P10008**.

COVER, T. M., and J. A. THOMAS. 1991. Elements of Information Theory. Wiley: Hoboken, NJ.

DANON, L., A. DÍAZ-GUILERA, J. DUCH, and A. ARENAS. 2005. Comparing community structure identification. Journal of Statistical Mechanics, **P09008**.

FORTUNATO, S. 2010. Community detection in graphs. Physics Reports, **486**: 75–174.

HALKIDI, M., Y. BATISTAKIS, and M. VAZIRGIANNIS. 2001. On clustering validation techniques. Journal of Intelligent Information Systems, **17**(2–3): 107–145.

HALKIDI, M., Y. BATISTAKIS, and M. VAZIRGIANNIS. 2002. Cluster validity methods: Part I. SIGMOD Record, **31**(2): 40–45.

HUBERT, L., and P. ARABIE. 1985. Comparing partitions. Journal of Classification, **2**: 193–218.

KVALSETH, T. O. 1987. Entropy and correlation: some comments. IEEE Transactions on Systems, Man and Cybernetics, **17**(3): 517–519.

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

LANCICHINETTI, A., and S. FORTUNATO. 2009. Community detection algorithms: a comparative analysis. Physical Review E, **80**(056117).

LANCICHINETTI, A., S. FORTUNATO, and F. RADICCHI. 2008. Benchmark graphs for testing community detection algorithms. Physical Review E, **78**(046110).

MANNING, C. D., P. RAGHAVAN, and H. SCHUTZE. 2008. Introduction to Information Retrieval. Cambridge University Press: New York.

MCDAID, A. F., D. GREENE, and N. HURLEYG. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. *In* arXiv:1110.2515 [physics.soc-ph].

MEILĂ, M. 2007. Comparing clusterings – an information based distance. Journal of Multivariate Analysis, **98**: 873–895.

MEILA, M., and D. HECKERMAN. 2001. An experimental comparison of model-based clustering methods. Machine Learning, **42**(1/2): 9–29.

NEWMAN, M. E. J. 2004. Fast algorithms for detecting community structure in networks. Physical Review, **E69**: 066133.

ORMAN, G. K., and V. LABATUT. 2010. The effect of network realism on community detection algorithms. *In* International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010, pp. 301–305.

ORMAN, G. K., V. LABATUT, and H. CHERIFI. 2011. Qualitative comparison of community detection algorithms. *In* International Conference on Digital Information and Communication Technology and its Applications DICTAP 2011, pp. 265–279.

ORMAN, G. K., V. LABATUT, and H. CHERIFI. 2012. Comparative evaluation of community detection algorithms: a topological approach. Journal of Statistical Mechanics: Theory and Experiment, **2012**(08): P08001.

ORMAN, G. K., V. LABATUT, and H. CHERIFI. 2013. Towards realistic artificial benchmark for community detection algorithms evaluation. IJWBC, **9**(3): 349–370.

PARKER, M. A. 2009. Solid State and Quantum Theory for Optoelectronics. CRC Press: Bocan Raton, FL.

PEROTTI, J. I., C. J. TESSONE, and G. CALDARELLI. 2015. Hierarchical mutual information for the comparison of hierarchical community structures in complex networks. *In* arXiv:1508.04388v2 [physics.soc-ph].

PONS, P., and M. LATAPY. 2005. Computing communities in large networks using random walks. *In* 20th International Symposium on Computer and Information Sciences - ISCIS 2005, pp. 284–293.

RABBANY, R., M. TAKAFFOLI, J. FAGNAN, O. R. ZAÏANE, and R. J. G. B. CAMPELLO. 2013. Communities validity: methodical evaluation of community mining algorithms. Social Network Analysis and Mining, **3**(4): 1039–1062.

RABBANY, R., and O. R. ZAÏANE. 2015. Generalization of clustering agreements and distances for overlapping clusters and network communities. Data Mining and Knowledge Discovery, **29**(5): 1458–1485.

ROMANO, S., J. BAILEY, V. NGUYEN, and K. VERSPOOR. 2014. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. *In* Proceedings of the 31st International Conference on Machine Learning JMLR W&CP 32 (1), pp. 1143–1151.

ROSVALL, M., and C. T. BERGSTROM. 2007. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences, **104**(18): 7327.

ROSVALL, M., and C. T. BERGSTROM. 2008. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, **105**(4): 118–1123.

STREHL, A., and J. GHOSH. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, **3**: 583–617.

VAN DONGEN, S. 2000. Performance criteria for graph clustering and Markov cluster experiments. Technical Report. Amsterdam: CWI (Centre for Mathematics and Computer Science).

VAN DONGEN, S. 2008. Graph clustering via a discrete uncoupling process. SIAM Journal of Matrix Analysis Applications, **30**(1): 121–141.

VINH, N. X., J. EPPS, and J. BAILEY. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? *In* Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pp. 1073–1080.

VINH, N. X., J. EPPS, and J. BAILEY. 2010. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. Journal of Machine Learning Research, **11**: 2837–2854.

YANG, J., and Y. LESKOVEC. 2015. Defining and evaluating network communities based on ground-truth. Knowledge Information Systems, **42**: 181–213.

YAO, Y. Y. 2003. Information-theoretic measures for knowledge discovery and data mining. *In* Entropy Measures, Maximum Entropy Principle and emerging Applications. *Edited by* KARMESHU, Volume 119 of Studies in Fuzziness and Soft Computing. Springer: Berlin Heidelberg, pp. 115–136.

ZHANG, P. 2015. Evaluating accuracy of community detection using the relative normalized mutual information. In arXiv:1501.03844v2 [physics.soc-ph].