Journal of Universal Computer Science, vol. 25, no. 8 (2019), 887-903
submitted: 26/12/18, accepted: 30/5/19, appeared: 28/8/19

A New Information-Theoretical Distance Measure for Evaluating Community Detection Algorithms

Mariam Haroutunian
(Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia
armar@ipia.sci.am)

Karen Mkhitaryan
(Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia
karenmkhitaryan@gmail.com)

Josiane Mothe
(Toulouse Institute of Computer Science Research, Université de Toulouse
UMR5505 CNRS, 31062, Toulouse, France
josiane.mothe@irit.fr)

Abstract: Community detection is a research area from network science dealing with the investigation of complex networks such as social or biological networks, aiming to identify subgroups (communities) of entities (nodes) that are more closely related to each other inside the community than with the remaining entities in the network. Various community detection algorithms have been developed and used in the literature however evaluating community structures that have been automatically detected is a challenging task due to varying results in different scenarios. Current evaluation measures that compare extracted community structures with the reference structure or ground truth suffer from various drawbacks; some of them having been point out in the literature. Information theoretic measures form a fundamental class in this domain and have recently received increasing interest. However even the well employed measures (NVI and NID) also share some limitations, particularly they are biased toward the number of communities in the network. The main contribution of this paper is to introduce a new measure that overcomes this limitation while holding the important properties of measures. We review the mathematical properties of our measure based on χ^2 divergence inspired from f-divergence measures in information theory. Theoretical properties as well as experimental results in various scenarios show the superiority of the proposed measure to evaluate community detection over the ones from the literature.
Key Words: Community detection, f-divergences, evaluation measure.

Categories: E.4

1 Introduction

The goal of community detection is to partition any network into communities to extract the subgroups of densely connected nodes [Fortunato, 2010]. Extraction of communities has many applications in different disciplines such as biology, medicine, social network analysis, information retrieval, machine learning, etc.

When a community detection algorithm is applied and the studied network is partitioned into communities, the output is an N dimensional random vector $X = (x_1, x_2, ..., x_N)$, where N is the number of nodes in the network and each $x_n \in \{1, ..., K\}$, $n \in \{1, ..., N\}$ element represents the community assignment of node n, where K is the number of communities in the network.

In order to quantitatively assess the goodness of the applied partitioning algorithm or its derived community structure, it can either be compared with other partitions of the network or with pre-known ground truth partition [Mothe et al., 2017, Malek et al., 2018]. In the literature of the domain it is mostly accomplished by employing evaluation measures based on counting pairs (adjusted rand index, Fowlkes-Mallows index, Jaccard index, etc.), set overlaps (F-Measure, Van Dongen-Measure, etc.) and mutual information (normalized mutual information, normalized variation of information, normalized information distance) [Mothe et al., 2017, Malek et al., 2018, Yang et al., 2016].

Existing measures based on pair counting and set overlaps share drawbacks that lead to prospect alternative means to compare community structure and clustering results. Information theoretic measures are worth investigating because of their strong mathematical foundation and ability to detect non-linear similarities [Vinh et al., 2010, Wagner and Wagner, 2007].

In community structure or clustering comparison problems it is desired that the applied measure satisfies the main properties of **metric** (that conforms to feeling of distance), **normalization** (requires that the measure lies within a fixed range) and **constant baseline property** (measure should be constant for communities sampled independently at random).

Calculating the similarity of two network partitions can be viewed as comparing two random variables which is typical to encoding/decoding problem from information theory. More specifically, let $X = (x_1, x_2, ..., x_N)$ and $Y = (y_1, y_2, ..., y_N)$ be two different partitions of the network, we assume that community assignments x_n and y_n are values of random variables X and Y respectively with joint probability distribution $P_{XY} = P(X = x, Y = y)$ and marginal distributions $P_X = P(X = x)$ and $P_Y = P(Y = y)$.

Mutual information (MI) is one of the measures that comes from information theory. It is a popular measure in information theory that measures the mutual dependence of two random variables X and Y. It measures how much information about one random variable is obtained through the other random variable [Cover and Thomas, 2006].

Considering random network partitions as random variables, mutual information can be viewed as a similarity measure when comparing community structures. Although the application of MI is pretty straightforward in the literature, in community structure or clustering comparison the use of a measure satisfying both the metric and normalization properties is of a high priority.

While MI is not a normalized measure, several normalized variants of MI called normalized mutual information (NMI) were introduced by Yao [Yao, 2003], Kvalseth [Kvalseth, 1987] and Strehl et al. [Strehl and Ghosh, 2002]. Later Meila [Meila, 2007] introduced variation of information (VI) which unlike NMI is a metric measure. Finally normalized variation of information (NVI) and normalized information distance (NID) were proposed by Kraskov et al. [Kraskov et al., 2005].

Despite the fact that information-theoretic measures such as NVI and NID are proper metrics, some experiments challenge their effectiveness and limit their use in certain applications [Vinh et al., 2010, Wagner and Wagner, 2007].

In [Vinh et al., 2010], Vinh et al. performed an organized study of information theoretic measures for clustering comparison. The authors mathematically proved that NVI and NID satisfy both the normalization and metric properties. Authors also highlighted the importance of correcting the measures for chance agreement, when the number of data points is relatively small compared with the number of clusters. They advocate NID as a "general purpose" measure for clustering comparison, possessing several useful properties such as using $[0, 1]$ range better than the other measures.

According to Amelio and Pizzuti [Amelio and Pizzuti, 2015], normalized mutual information has unfair behavior when the number of communities in the network is large. Authors experimentally showed that NMI reaches abnormal values when comparing a clustering of $5,000$ nodes into $5,000$ singleton communities with a reference clustering of $5,000$ nodes into 100 communities. The authors suggested to adjust the NMI by introducing a scaling factor which also compares the number of communities detected by an algorithm and the actual number of communities in the ground truth.

Another modification was suggested by Zhang [Zhang, 2015] who claims that NMI is affected by systematic errors as a result of finite network size which may result in wrong conclusions when evaluating community detection algorithms. Relative normalized mutual information (rNMI) introduced by Zhang takes into account the statistical significance of NMI by comparing it with the expected NMI of random partitions.

Considering the drawbacks that pair counting, set matching and information-theoretic measures share, we decided to search other alternatives among f-divergences which form an important class of information theoretic measures. These are measures of discrimination between two probability distributions. Their properties, connection inequalities and applications in information theory, machine learning, statistics and other applied branches were studied in many publications, see for example [Sason and Verdú, 2016, Csiszár and Shields, 2004, Sason, 2015, Topsoe, 2000].

Analyzing the properties of various f-divergences we propose a new measure based on χ^2-divergence from information theory. We demonstrate theoretically as well as experimentally that it could serve as an alternative in community detection evaluation or clustering comparison. Furthermore we show that unlike other regularly used measures (NID and NVI), our modification of χ^2-divergence satisfies the constant baseline property thus outperforming them; specifically in the cases when network size is relatively small compared to the number of communities.

This paper corresponds to a substantial extension of the workshop paper [Haroutunian et al., 2018]. The paper is organized as follows. We review the information-theoretic community structure comparison measures and their properties in Section 2. In Section 3 we survey alternative f-divergence measures and discuss their useful properties to consider them in community detection evaluation. In Section 4, we define a modified version of χ^2-divergence measure and provide the theoretical cues regarding its properties for community detection evaluation. Section 5 reports an experimental analysis that shows the advantages of χ^2-divergence measure over the other measures from the literature. Section 6 concludes this paper.

2 Information Theoretic Measures and Measure Properties

Information theoretic measures are applied in various fields such as coding theory, statistics, machine learning, genomics, neuroscience etc. [Cover and Thomas, 2006]. The same measures can be of paramount importance in community detection evaluation and clustering comparison for their strong mathematical foundation and the fundamental concepts they are based on.

One of the basic measures in information theory is the mutual information between two random variables, which tells how much knowing one of the random variables reduces the uncertainty about the other. **Mutual information (MI)** of two discrete random variables is defined as [Cover and Thomas, 2006]:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y),$$

where $H(X)$ is the entropy of X and $H(X|Y)$ the conditional entropy of X given Y.

$$H(X) = -\sum_{x} p(x) \log p(x), \qquad H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y).$$

Considering random network partitions as random variables, MI can be viewed as a similarity measure when comparing community structures. For evaluation of network partitions, it is highly desired that the used measure satisfies the following properties:

- **Metric property**
  A measure d is a metric if it satisfies the following properties:
  - Non-negativity: $d(X,Y) \geq 0$,
  - Identity: $d(X,Y) = 0 \Leftrightarrow X = Y$,
  - Symmetry: $d(X,Y) = d(Y,X)$,
  - Triangle inequality: $d(X,Z) + d(Z,Y) \geq d(X,Y)$.

- **Normalization property**

  A measure is normalized if the values it takes fall into a fixed interval. Normalized measures are easy to interpret and especially in community detection problems it is necessary to quantitatively assess the similarity of a given partition with other partitions or with ground truth. In community detection evaluation most of the measures fall into intervals $[0,1]$ or $[-1,1]$.

- **Constant Baseline Property**
  When comparing two random network partitions, the expected value of the measure must be constant, preferably zero [Vinh et al., 2010, Romano et al., 2016].

The metric property conforms to the intuition of distance [Meila, 2007] and it is important in the case of complex space of clustering as many theoretical results already exist for metric spaces.

Based on the properties of MI, that is non-negativity and upper boundedness:

$$0 \leq I(X;Y) \leq \min\{H(X),H(Y)\} \leq \sqrt{H(X)H(Y)} \leq \frac{1}{2}(H(X)+H(Y)) \leq \\ \leq \max\{H(X),H(Y)\} \leq H(X,Y)$$

several normalized variants of MI can be considered as **similarity measures** [Vinh et al., 2010, Yao, 2003, Kvalseth, 1987, Strehl and Ghosh, 2002]:

$$\mathrm{NMI}_{\mathrm{joint}} = \frac{I(X;Y)}{H(X,Y)}, \qquad \mathrm{NMI}_{\mathrm{max}} = \frac{I(X;Y)}{\max(H(X),H(Y))},$$

$$\mathrm{NMI}_{\mathrm{sum}} = \frac{2I(X;Y)}{H(X)+H(Y)}, \qquad \mathrm{NMI}_{\mathrm{sqrt}} = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}},$$
$$\mathrm{NMI}_{\mathrm{min}} = \frac{I(X;Y)}{\min\{H(X),H(Y)\}}.$$

Based on the five upper bounds for $I(X;Y)$ also five **distance measures** are defined as follows [Vinh et al., 2010].

$$\begin{aligned}
D_{\mathrm{joint}} &= H(X,Y) - I(X;Y),\\
D_{\mathrm{max}} &= \max\{H(X),H(Y)\} - I(X;Y),\\
D_{\mathrm{sum}} &= \frac{H(X)+H(Y)}{2} - I(X;Y),\\
D_{\mathrm{sqrt}} &= \sqrt{H(X)H(Y)} - I(X;Y),\\
D_{\mathrm{min}} &= \min\{H(X)H(Y)\} - I(X;Y).
\end{aligned}$$

$D_{\mathrm{joint}} = 2D_{\mathrm{sum}}$ is known as **variation of information (VI)** introduced by Meila [Meila, 2007], satisfying the properties of metrics but not the one of normalization. In [Vinh et al., 2010] it was proved that D_{max} is a metric, while D_{min} and D_{sqrt} are not. Later Kraskov et al. [Kraskov et al., 2005] introduced normalized variant of variation of information called **normalized variation of information (NVI)** and **normalized information distance (NID)** which are both normalized and metric measures.

$$\begin{aligned}
\mathrm{NVI} &= \frac{H(X,Y) - I(X;Y)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)},\\
\mathrm{NID} &= \frac{\max(H(X),H(Y)) - I(X;Y)}{\max(H(X),H(Y))} = 1 - \frac{I(X;Y)}{\max\{H(X),H(Y)\}}.
\end{aligned}$$

An overview of popular information theoretic measures is given in Table 1.

Table 1: Overview of popular information theoretic measures.

Measures	Range	Normalization	Metrics
MI	$[0, min\{H(X),H(Y)\}]$	×	×
NMI	$[0,1]$	✓	×
VI	$[0,\log(N)]$	×	✓
NVI	$[0,1]$	✓	✓
NID	$[0,1]$	✓	✓

Although the above mentioned measures are very popular in community detection and clustering literature, many experiments challenge their effectiveness stating that they are biased to the number of communities i.e. violating the constant baseline property [Vinh et al., 2010] and being affected by systematic errors due to finite network size [Zhang, 2015]. For this reason we decided to search for alternatives among f-divergences.

3 Looking for Alternative Measures Among f-divergences

f-Divergences also known as Csiszár f-divergences are functions measuring the difference between two probability distributions introduced by Csiszár [Csiszár and Shields, 2004], Morimoto [Morimoto, 1963] and Ali & Silvey [Ali and Silvey, 1966].

Let $f : (0, \infty) \to R$ be a convex function with $f(1) = 0$ and let P and Q be two probability distributions. The **f-divergence** from P to Q is defined by

$$D_f(P \parallel Q) \triangleq \sum_x q(x) f(\frac{p(x)}{q(x)}).$$

Among others, f-divergences include well known notions from information theory listed below.

Kullback-Leibler divergence which is also known as relative entropy

$$D(P \parallel Q) = \sum_x p(x) \log(\frac{p(x)}{q(x)}),$$

is a f-divergence with $f(t) = t \log(t)$. Also $D(Q \parallel P)$ can be obtained from $f(t) = -t \log(t)$.

Total variational distance

$$V(P, Q) = \sum_x |p(x) - q(x)| = \sum_x q(x) |\frac{p(x)}{q(x)} - 1|,$$

is coming from the same f-divergence formula when $f(t) = |t - 1|$.

Hellinger distance defined by

$$H(P, Q) = \sum_x (\sqrt{p(x)} - \sqrt{q(x)})^2,$$

is a f-divergence with $f(t) = (\sqrt{t} - 1)^2$. The Hellinger distance is closely related to the total variational distance, but it has several advantages such as being well suited for the study of product measures.

Jeffrey divergence is the symmetrized Kullback-Leibler divergence

$$J(P \parallel Q) = D(P \parallel Q) + D(Q \parallel P) = \sum_x (p(x) - q(x)) \log(\frac{p(x)}{q(x)}),$$

that is obtained from $D_f(P \parallel Q)$ with $f(t) = \frac{1}{2}(t-1)\log(t)$.

Capacitory discrimination (similar to Jensen-Shannon divergence) is given by

$$C(P,Q) = D(P \parallel \frac{P+Q}{2}) + D(Q \parallel \frac{P+Q}{2}) = 2H(\frac{P+Q}{2}) - H(P) - H(Q)$$

which comes from $D_f(P,Q)$ with $f(t) = t\log(t) - (t+1)\log(t+1) + 2\log(2)$.

χ^2 ***divergence*** is a f-divergence measure,

$$\chi^2(P,Q) = \sum_x \frac{(p(x)-q(x))^2}{q(x)} = \sum_x q(x)(\frac{p(x)}{q(x)} - 1)^2,$$

where $f(t) = (t-1)^2$.

Bhattacharyya distance given by

$$B(P,Q) = \sqrt{1 - \sum_x \sqrt{p(x)q(x)}},$$

can be obtained from $D_f(P,Q)$, when $f(t) = 1 - \sqrt{t}$. An overview of discussed f-divergences is given in Table 2.

Table 2: Overview of properties of f-divergence measures.

f-divergence measures	Normalization	Metrics
Kullback-Leibler divergence	×	×
Total variational distance	✓	✓
Hellinger distance	✓	✓
Jeffrey divergence	×	×
Capacitory discrimination	✓	✓
χ^2 divergence	×	×
Bhattacharyya distance	✓	✓

We considered the properties of these measures to decide how they fit for comparing network partitions. To compare two community structures or network partitions X and Y we must consider the discrimination from P_{XY} to $P_X P_Y$, where P_{XY} is the joint probability distribution and $P_X P_Y$ the product of marginal distributions of X and Y partitions respectively. First note that there is a well known property

$$D(P_{XY} \parallel P_X P_Y) = I(X;Y)$$

and hence Kullback-Leibler divergence being very useful in information theory is not interesting for our task.

Although for all measures the output is zero when partitions are independent,

$$P_{XY} = P_X P_Y \implies D_f(P_{XY} \parallel P_X P_Y) = 0$$

the identity property of metric is violated for all discussed measures, when considering identical partitions, i.e. when $X = Y$, $D_f(P_{XY} \parallel P_X P_Y)$ is not zero. In the next section we suggest a new modified version of χ^2-divergence overcoming this issue.

4 Modified χ^2-divergence for Evaluating Network Partitions

Let K_{max} denote the maximum number of communities in X and Y respectively, i.e. $K_{max} = \max\{K_X, K_Y\}$, where K_X and K_Y are the number of communities in partitions X and Y respectively.

Consider the following measure which we suggest for comparison of two community structures or network partitions X and Y that we call modified χ^2-divergence and denote by $MD_{\chi^2}(X, Y)$:

$$MD_{\chi^2}(X,Y) = 1 - \frac{\chi^2(P_{XY}, P_X P_Y)}{K_{max}} = 1 - \frac{\sum_{x,y} \frac{(p(x,y)-p(x)p(y))^2}{p(x)p(y)}}{K_{max}-1} = \tag{1}$$
$$= 1 - \frac{\sum_{x,y} \frac{p^2(x,y)}{p(x)p(y)} - 1}{K_{max}-1}.$$

Theorem 1. MD_{χ^2} satisfies all metric properties except triangle inequality and is a normalized measure.

Proof:

The following properties of modified χ^2-divergence are obtained:

- **Non negativity**

$$\sum_{x,y} \frac{p^2(x,y)}{p(x)p(y)} \leq \min\{\sum_{x,y} \frac{p(x,y)}{p(x)}, \sum_{x,y} \frac{p(x,y)}{p(y)}\} = \min\{\sum_{x} \frac{p(x)}{p(x)}, \sum_{y} \frac{p(y)}{p(y)}\} = \tag{2}$$
$$= \min\{K_X, K_Y\} \leq \max\{K_X, K_Y\} = K_{max} \implies MD_{\chi^2}(X,Y) \geq 0.$$

- **Symmetry**
  It is obvious that $MD_{\chi^2}(X, Y) = MD_{\chi^2}(Y, X)$.

- **Identity**
  When X and Y partitions are identical, hence

$$MD_{\chi^2}(X,X) = 1 - \frac{\sum_x \frac{p^2(x,x)}{p(x)p(x)} - 1}{K_x - 1} = 1 - \frac{K_x - 1}{K_x - 1} = 0.$$

  The inverse is also correct.

$$MD_{\chi^2}(X,Y) = 0 \implies \sum_{x,y} \frac{p^2(x,y)}{p(x)p(y)} = K_{max} \implies$$
$$\sum_{x,y} p(x|y)p(y|x) = K_{max} \implies X = Y$$

- **Normalization**
  From the non-negativity property of modified χ^2-divergence,

$$MD_{\chi^2}(X,Y) \geq 0.$$

  To obtain the upper bound, we use the direct consequence of Cauchy-Bunyakovsky-Schwarz inequality, the Sedrakyans inequality [Sedrakyan, 1997]

$$\sum_i \frac{a_i^2}{b_i} \geq \frac{(\sum_i a_i)^2}{\sum_i b_i}. \tag{3}$$

  Using the inequality (3) we obtain

$$\sum_{x,y} \frac{p^2(x,y)}{p(x)p(y)} \geq \frac{(\sum_{x,y} p(x,y))^2}{\sum_{x,y} p(x)p(y)} = 1. \tag{4}$$

  Substituting (4) into (1) the following is obtained

$$MD_{\chi^2}(X,Y) \leq 1.$$

  Hence $MD_{\chi^2}(X,Y) \in [0,1]$ is a normalized measure.

- **Triangle inequality**
  It would be perfect if we could proof also the triangle inequality. Unfortunately, modified χ^2-divergence **does not obey triangle inequality**. It is sufficient to point out a singular counter example where triangle inequality is violated. For example, let $X = (2,1,2,2,1,2,2,2,2,2)$,

$Y = (1, 2, 2, 1, 2, 1, 1, 1, 1, 1)$ and $Z = (2, 1, 1, 2, 2, 2, 2, 2, 2, 2)$ partitions of $N = 10$ nodes into two communities. It can be easily checked that

$$MD_{\chi^2}(X, Y) + MD_{\chi^2}(Y, Z) < MD_{\chi^2}(X, Z).$$

Nevertheless a huge number of experiments show that in the majority cases the triangle inequality comes true. In any case, this is not the most important property, as usually the detected structures are compared with the ground truth or with each other.

The theorem is proved.

As we saw the modified χ^2-divergence satisfies all properties of metrics (non-negativity, symmetry, identity) except triangle inequality, it is also a normalized measure and can be considered as an alternative evaluation of network partitions. The next objective is to verify the third desirable property, that is, constant baseline property mentioned in the introduction. For this purpose we conduct experimental analysis in the next section that also justifies the use of the modified χ^2-divergence along with NID and NVI with respect to that property. We show that modified χ^2-divergence has huge advantage over NVI and NID being unbiased to the number of communities or clusters in the network.

5 Experimental Analysis

In order to see how modified χ^2-divergence performs and how it fits evaluation tasks in community detection, we implemented an experimental study and compared the outputs obtained by MD_{χ^2}, NVI and NID using artificially generated community structures. Criteria selected for comparison were the performance of measures on random community structures with different number of nodes and communities in the network, satisfaction of measures to the constant baseline property and the results obtained by applying community detection algorithms on a network having ground-truth community structure. For the experiments, the following notations are used:

– N, is the number of nodes in the network.

– K_{gt}, is the number of communities in the ground truth.

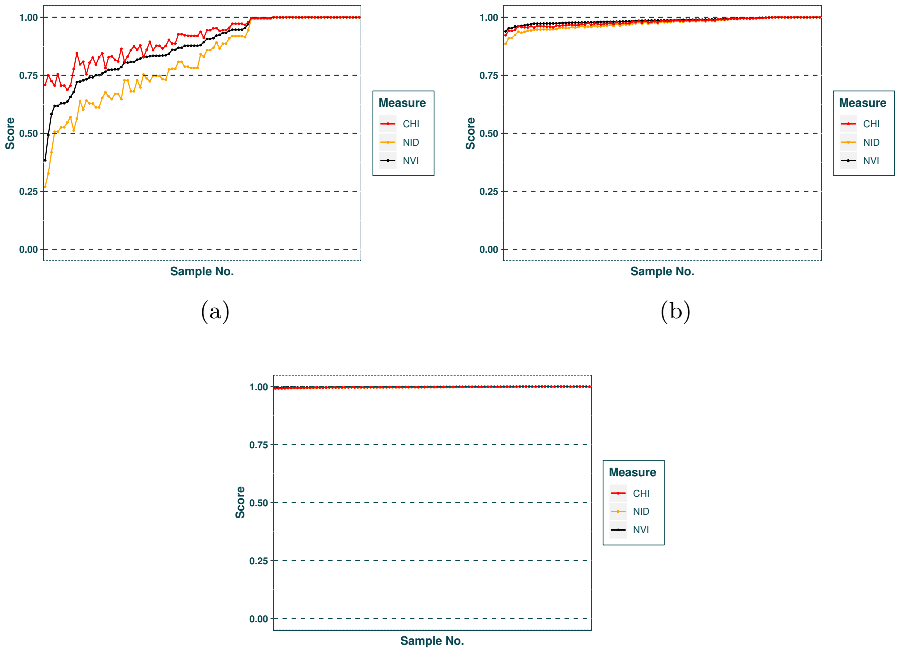
– K_{part}, is the number of communities in the partition detected by an algorithm.

– V_{gt}, is the community membership vector of ground truth.

– V_{part}, is the community membership vector of a partition detected by an algorithm.

– V_{gt}, is the community membership vector of ground truth.

We considered a scenario where a particular community detection method detected five communities and the ground truth also contains five communities.



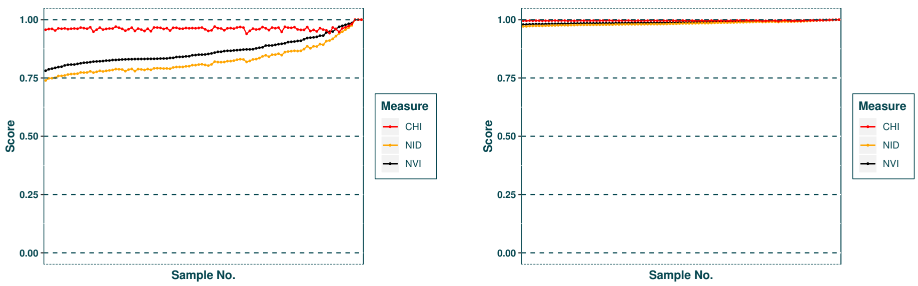
(a) (b)

(c)

Figure 1: Similarity between 100 random partitions and a ground truth ordered by increasing order of NVI, where $K_{gt} = 5$, $K_{part} = 5$ for $N = 10$ (a), $N = 100$ (b) and $N = 1000$ (c).

Different 100 random community membership vectors (V_{part}) were generated and compared with ground truth considering for $N = 10$ (Figure 1a), $N = 100$ (Figure 1b) and $N = 1000$ (Figure 1c) nodes. From the figures we see that for the three measures the results are similar to each other. Figures 1b and 1c clearly show that when number of nodes in the network increases, the output has a less chance to be closer to the ground truth indicating almost independent partitions, whatever the measure is.

In the next scenario we considered that the number of communities in the ground truth and clustering may vary. For this reason we set the number of communities a random number in the range $[1, 50]$ and generated again 100 random partitions for $N = 100$ (Figure 2a) and $N = 1000$ (Figure 2b). We see that the behaviour is again similar for all three measures.

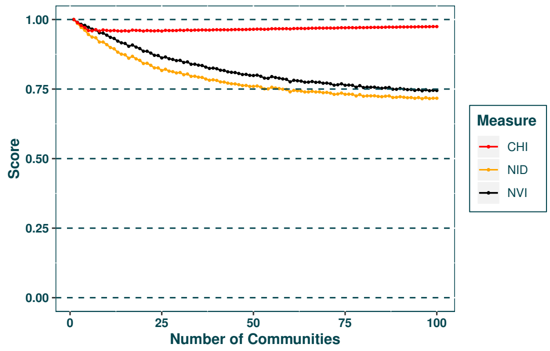


(a) (b)

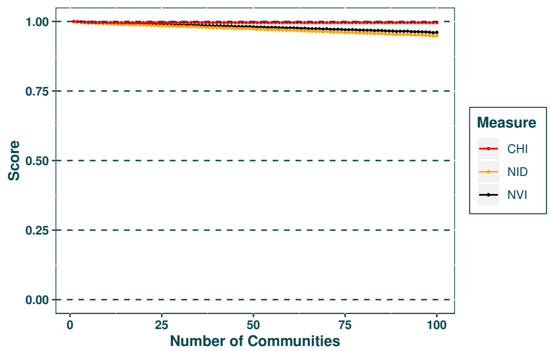
Figure 2: Similarity between 100 partitions with random number of communities from the range $[1, 50]$ and the ground truth ordered by increasing order of NVI, where $K_{gt} = 5$ for $N = 100$ (a) and $N = 1000$ (b).

Despite the similarity of results, both NVI and NID are sensitive to the number of communities in the network which might affect the results [Vinh et al., 2010]. Therefore we analyzed the performance of measures in order to see whether they are biased or not. We fixed the ground truth to be a random partition with 5 communities. Then we generated 1000 random partitions (V_{part}) and averaged the similarity with ground truth for each number of communities from the interval $[1, 100]$. The experiment was done for networks with $N = 100$ (Figure 3a) and $N = 1000$ (Figure 3b) nodes. From Figure 3a we can see that when number of communities increases, both NVI and NID scores decreases and show more similarity with the ground truth, although partitions are completely random. The same pattern was obtained when considering more communities in the ground truth. Interpretation to this is that NVI and NID are biased to the number of communities in the network and may give wrong results when network size is relatively small compared with the number of communities. The same experiment with partitions containing 1000 nodes (Figure 3b) shows less bias as NVI and NID become less biased when number of nodes in the network is relatively large compared with the number of communities in the network.

Finally the performance of measures were experimented on synthetic networks, generated using the concept of stochastic block model (SBM) [Abbe, 2017].



(a)



(b)

Figure 3: Similarity between V_{part} and V_{gt}, where $K_{gt} = 5$, $K_{part} \in [1, 2, ..., 100]$ for $N = 100$ (a) and $N = 1000$ (b). For each K_{part} result is averaged on 1000 random clusterings.

SBM takes the following parameters:

- $(C_1, C_2, ..., C_k), k \in \{1, ..., K\}$, vector of community sizes, where K is the number of communities in the network.
- A symmetric $K \times K$ matrix M of edge probabilities, where M_{ij} element represents the probability of edge between nodes from communities i and j.

By taking number of nodes in network, $N = 1000$, 10 communities with random sizes, $(C_1, C_2, ..., C_k)$ that sum up to 1000, probability of edges inside communities, $M_{i,i} \in [0.5, 1]$ and between communities $M_{i,j} \in [0, 0.5]$, where $i \neq j$, large number of networks was built where ground truth is predefined. Later six community detection algorithms [Fortunato, 2010] that are fast greedy modularity optimization (FG), infomap (IM), leading eigenvector (LE), label propagation

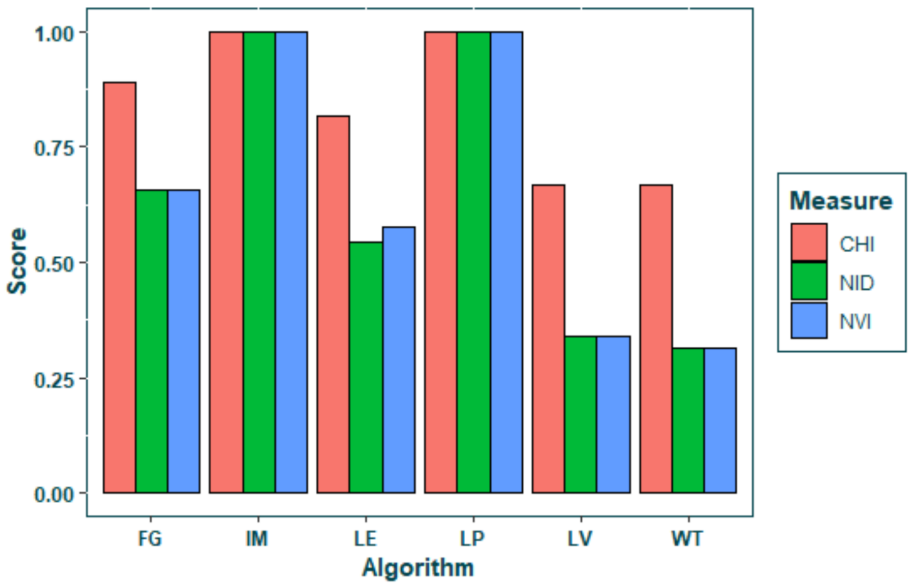


Figure 4: Similarity between the derived partition by the algorithm (V_{part}) and ground truth (V_{gt}), where $K_{gt} = 10$ and $N = 1000$. Modified χ^2-divergence is consistently showing less similarity with ground-truth being unbiased to the number of communities in the network.

(LP), Louvain modularity oprimization (LV) and walktrap (WT) were applied on these networks which partitioned network into community structure. Finally using NVI, NID and modified χ^2 distance, derived community structure was compared with the predefined ground truth. One such result is shown in Figure 4. In all experimental cases measures had very similar behavior, but considering the fact that NVI and NID are biased to the number of communities, modified χ^2 outperforms them when number of nodes in the network is relatively small compared with the number of communities.

6 Conclusion

Existing evaluation measures that are widely employed in community detection to compare different community structures have some drawbacks that need to be addressed. Compared with measures based on pair counting and set overlaps, information-theoretic measures have strong mathematical foundation and are able to detect non-linear similarities. For that reason in recent years mostly information-theoretic measures were employed as comparison tools in community detection. However, information-theoretic measures also have their limitations.

Our investigations on normalized mutual information, normalized variation of information and normalized information distance showed that they are biased in favour of large number of communities or clusters in the network i.e. give unfair results when the number of nodes is relatively small compared with the number of communities. Admitting the drawbacks that existing information-theoretic measures share, we suggest a new measure, namely modified χ^2-divergence for comparing community structures based on χ^2-divergence from information theory. We mathematically proved that our modified χ^2-divergence satisfies all metric properties (except triangle inequality) and is a normalized measure. We also show experimentally that compared with NMI, NVI and NID, modified χ^2-divergence admits constant baseline, not being affected by the number of communities in the network which guarantees fair comparison in scenarios where number of nodes is relatively small compared with the number of communities.

References

[Abbe, 2017] Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

[Ali and Silvey, 1966] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Royal Statistical Society*, 28(1):131–142.

[Amelio and Pizzuti, 2015] Amelio, A. and Pizzuti, C. (2015). Is normalized mutual information a fair measure for comparing community detection methods? In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Paris, France. IEEE.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.

[Csiszár and Shields, 2004] Csiszár, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528.

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

[Haroutunian et al., 2018] Haroutunian, M., Mkhitaryan, K., and Mothe, J. (2018). f-divergence measures for evaluation in community detection. In *Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA)*, pages 137–144, Yerevan, Armenia.

[Kraskov et al., 2005] Kraskov, A., Stgbauer, H., Andrzejak, R. G., and Grassberger, P. (2005). Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):2837–2854.

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

[Malek et al., 2018] Malek, J., Hocine, C., Chantal, C., and Atef, H. (2018). Community detection algorithm evaluation with ground-truth data. *Physica A: Statistical Mechanics and its Applications*, 492:651–706.

[Meila, 2007] Meila, M. (2007). Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

[Morimoto, 1963] Morimoto, T. (1963). Markov processes and the h-theorem. *J. Phys. Soc. Jpn*, 18:328–331.

[Mothe et al., 2017] Mothe, J., Mkhitaryan, K., and Haroutunian, M. (2017). Community detection: Comparison of state of the art algorithms. In *Proc. of Intern. Conf.*

Computer science and information technologies, Reprint in IEEE Revised selected papers, pages 252–256.

[Romano et al., 2016] Romano, S., Vinh, N. X., Bailey, J., and Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666.

[Sason, 2015] Sason, I. (2015). Tight bounds for symmetric divergence measures and a refined bound for lossless source coding. *IEEE Transactions on Information Theory*, 61(2):701–707.

[Sason and Verdú, 2016] Sason, I. and Verdú, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.

[Sedrakyan, 1997] Sedrakyan, N. (1997). About the applications of one useful inequality. *Kvant Journal*, 97(2):42–44.

[Strehl and Ghosh, 2002] Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617.

[Topsoe, 2000] Topsoe, F. (2000). Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609.

[Vinh et al., 2010] Vinh, N. X., Apps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

[Wagner and Wagner, 2007] Wagner, S. and Wagner, D. (2007). Comparing clusterings - an overview. In *Internal report. Faculty of Computer Science, University of Karlsruhe*, volume 2. Karlsruhe.

[Yang et al., 2016] Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(30750).

[Yao, 2003] Yao, Y. (2003). Information-theoretic measures for knowledge discovery and data mining. *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136.

[Zhang, 2015] Zhang, P. (2015). Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment*, 2015.