scientific reports

Check for updates

OPEN

Proof of biased behavior of Normalized Mutual Information

Amin Mahmoudi✉ & Dariusz Jemielniak

The Normalized Mutual Information (NMI) metric is widely utilized in the evaluation of clustering and community detection algorithms. This study explores the performance of NMI, specifically examining its performance in relation to the quantity of communities, and uncovers a significant drawback associated with the metric's behavior as the number of communities increases. Our findings reveal a pronounced bias in the NMI as the number of communities escalates. While previous studies have noted this biased behavior, they have not provided a formal proof and have not addressed the causation of this problem, leaving a gap in the existing literature. In this study, we fill this gap by employing a mathematical approach to formally demonstrate why NMI exhibits biased behavior, thereby establishing its unsuitability as a metric for evaluating clustering and community detection algorithms. Crucially, our study exposes the vulnerability of entropy-based metrics that employ logarithmic functions to similar bias.

Background

Community detection (CD) within social networks has emerged as a pivotal area of research, given its potential to unravel intricate patterns of interaction and group dynamics. It is used in many disciplines, including biology[1], criminology[2], economics[3], and urban planning[4], to mention just a few examples. In particular, this topic has also emerged as a critical field in the battle against disinformation. Social networks are often the primary conduits for the spread of disinformation, with communities within these networks playing a significant role in the dissemination and amplification of misleading content. By identifying and understanding these communities, we can gain valuable insights into the dynamics of disinformation spread, enabling more effective interventions[5,6]. It is essential, as the rapid spread of medical[7,8], political[9,10], social[11–13], as well as scientific[14–16] misinformation and disinformation are among the greatest civilization challenges of our times[17–19].

Research has shown that disinformation tends to spread rapidly within tight-knit communities and is often characterized by homogenous beliefs and high levels of trust among members[20]. These communities can act as echo chambers, reinforcing and amplifying disinformation and using more and more sophisticated strategies for masking their agendas[21]. By employing CD algorithms, we can identify these communities and understand their structure and behavior, providing a basis for targeted, community-specific strategies to combat disinformation. As such, CD in social networks is not only a theoretical exercise but also a practical tool in the fight against disinformation. However, even though many new emerging approaches have been tested[22–25], we are still far from an optimal approach.

A plethora of algorithms have been developed to enhance the accuracy of CD, yet comprehensive and diverse sets of metrics for evaluating these algorithms are lacking. Predominantly, metrics such as modularity, conductance, pairwise F-measure (PWF), NMI, variation of information (VI), purity, and adjusted rand index (ARI) have been employed to assess the performance of CD algorithms. These metrics, which were originally designed for evaluating clustering techniques, have been adapted for CD due to the conceptual similarities between clustering and community detection[26].

Certain metrics, such as modularity, operate based on the internal structure of communities identified by a specific algorithm, independent of the availability of ground truth[27]. Conversely, metrics such as the NMI, ARI, VI, purity, and F-measure necessitate the availability of ground truth for deployment[28]. Regardless of the accessibility of ground truth, each metric is subject to a resolution limit[29], which is a factor that has been highlighted and well established in existing research[30–34].

Motivation

Predominantly, external metrics (those reliant on ground truth) construct a contingency matrix (table), where each cell represents the intersection of members between actual classes and detected communities. In some

Management in Networked and Digital Societies (MINDS) Department, Kozminski University, Warsaw, Poland. ✉email: amahmoudi@kozminski.edu.pl

instances, the accuracy of CD algorithms is evaluated solely based on the number of true positive members. However, an effective CD algorithm should take into account two primary aspects of communities: distribution and joint membership. Therefore, the evaluation metric for CD algorithms must be capable of discerning the distribution of communities and joint members in relation to the ground truth. This implies that the critical factor is not merely identifying joint members but also accurately determining the number of communities relative to the ground truth[35].

Despite the identification of biased behavior in these measures by several researchers[32–34], a mathematical explanation for this issue has not been adequately addressed. Most reports suggest that this bias is due to the finite size effect. In this study, we aim to formally demonstrate why the NMI metric exhibits bias. To achieve this, we first present the results of NMI across 40 scenarios (representing different community numbers assumed to be detected by a specific algorithm) and compare them with five other well-established measures. Subsequently, we dissected the NMI formula and ultimately proved that this formula inherently leads to biased behavior[36].

Contribution

Community detection is a critical concern spanning diverse scientific disciplines, including biology, health, social networks, politics, targeted marketing, recommender systems, link prediction, and criminology[37]. As such, the accuracy of community detection algorithms is of paramount importance, and the evaluation metric for these algorithms assumes even greater significance. Given the widespread use of the NMI in evaluating community detection algorithms, illuminating its biased behavior contributes significantly to fields that employ community detection studies. Moreover, substantiating this bias establishes the foundation for analyzing evaluation metrics of community detection algorithms that incorporate a logarithmic function[38]. It also opens the door to crafting new metrics while considering issues rooted in logarithmic functions.

The remainder of this paper is structured as follows: The subsequent section reviews the relevant literature in this domain. Then "Preliminaries and notations" are introduced. "Problem statement" delves into a detailed description of the NMI drawback. The proof of the NMI problem is presented in "Proof of biased behavior of NMI". In "Case study", we present a case study based on a real-world dataset and "Conclusion" concludes the paper with key findings and implications.

Related works

CD algorithms aim to identify groups of nodes characterized by dense interconnections compared to the rest of the network[39,40]. Girvan and Newman[39] introduced the modularity metric to evaluate the accuracy of communities detected by their algorithm, sparking the development of numerous algorithms based on this metric. However, Fortunato[26] highlighted a resolution limit in the modularity metric, indicating its inability to detect small-sized communities. Cai et al.[31] further demonstrated that maximizing modularity is an NP-hard problem and that a random network without any communities can achieve a high Q value. Chen, Nguyen, and Szymanski[41] underscored the inconsistencies of the modularity metric, noting its tendency to favor either small or large communities in different scenarios. They proposed a new measure, modularity density, which combines modularity with split penalty and community density to circumvent the dual problems inherent in modularity.

The NMI was first considered a precise metric by Danon et al.[36], who reported its sensitivity to errors in the community detection procedure. They consider Z_{out} as the average number of links a node has to members of any other community, by increasing Z_{out} NMI tends to be zero. Subsequent research has addressed the limitations of the NMI measure, with Romano et al.[42] emphasizing the role of the number of clusters in the evaluation metrics. Amelio and Pizzuti[30] argued that the NMI is not fair, as solutions with a high number of clusters receive disproportionately high NMI. Zhang[34] demonstrated that the NMI is significantly affected by systematic errors due to finite network sizes and proposed the relative normalized mutual information (rNMI). Lai and Nardini[32] introduced the corrected normalized mutual information (cNMI) to address the reverse finite size problem of the rNMI. Liu, Cheng, and Zhang[33] highlighted the drawbacks of NMI and its improved versions, such as rNMI and cNMI, noting that these measures often overlook the importance of small communities. Rossetti, Pappalardo, and Rinzivillo[43] introduced community precision and community recall to evaluate CD algorithms, addressing the high computational complexity of the NMI. Arab and Hasheminezhad[44] also reported scalability problems with the NMI in large-scale data.

Other researchers have proposed alternative measures for evaluating community detection and clustering algorithms. Meilă[45] introduced the variation of information (VI) metric, an entropy-based measure that operates based on mutual information. Wagner and Wagner[46] categorized measures based on counting pairs, set overlaps, and mutual information and concluded that information theoretical measures outperform counting pairs and set overlaps measures. Santos and Embrechts[47] utilized the ARI for cluster validation and feature selection. Yang and Leskovec[48]compared 13 measures for evaluating community detection algorithms, categorizing them into four groups and concluding that conductance and triad-participation-ratio have the best performance in identifying communities. Saltz, Prat-Pérez, and Dominguez-Sal[49] introduced a new metric for the CD problem, weighted community clustering (WCC), which operates based on the distribution of triangles in the graph.

Preliminaries and notations

Normalized Mutual Information (NMI)

The NMI serves as a metric for assessing the performance of community detection algorithms. The NMI facilitates comparisons between two clusters or communities, yielding a value that ranges from 0 to 1. A higher value indicates a greater degree of similarity between two partitions or communities. As an external metric, the NMI necessitates the availability of class labels for computations, implying that the ground truth is required when employing this metric. The calculation of NMI is executed according to Eq. (1).

$$NMI(A,B) = \frac{2 * I(A,B)}{[H(A) + H(B)]} \tag{1}$$

where $I(A,B)$ is mutual information and H is the entropy as shown in Eqs. (2) and (3).

$$I(A,B) = \sum_{i=1}^{S}\sum_{j=1}^{R} p(community_j \cap class_i) log \frac{p(community_j \cap class_i)}{p(community_j)p(class_i)} \tag{2}$$

$$= \sum_{i=1}^{S}\sum_{j=1}^{R} \frac{|community_j \cap class_i|}{N} log \frac{|community_j \cap class_i|N}{\left|community_j\right| |class_i|}$$

$$H(A) = -\sum_{j=1}^{R} p\left(community_j\right) \log p\left(community_j\right) = -\sum_{j=1}^{R} \frac{|community_j|}{N} log \frac{|community_j|}{N} \tag{3}$$

The expansion of Eq. (1) with respect to (2) and (3) is Eq. (4)

$$NMI(A,B) = \frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij} \log \frac{C_{ij}N}{C_{i.}C_{.j}}}{\sum_{i=1}^{S} C_{i.} \log \frac{C_{i.}}{N} + \sum_{j=1}^{R} C_{.j} \log \frac{C_{.j}}{N}} \tag{4}$$

Suppose there are two networks denoted as Net_1 and Net_2 each consisting of sets of vertices (V) and edges (E). Net_1 consists of R communities denoted as $A = \{A_1, A_2, \ldots., A_R\}$, while Net_2 consists of S communities denoted as $B = \{B_1, B_2, \ldots., B_S\}$. C_{ij} denotes the number of nodes that clusters (communities) A_i and B_j share. If $A = B$, then $NMI(A,B) = 1$; if A and B are completely different, then $NMI(A,B) = 0$.

In addition to the NMI, some other measures can be used to evaluate the accuracy of CD and clustering algorithms. Table 1 lists well-known measures in this domain. We listed these measures here to highlight the differences between the NMI and other measures in practice.

Essentially, to compute the measures listed in Table 1, a contingency table (CT) is employed. This table is created based on the joint members between communities detected by a certain algorithm and the ground truth. The contingency table used for computing the NMI is presented in Table 2.

Table 3 presents important notations.

Problem statement

In this section, we present the main drawback of the NMI. We illustrate this problem through an example. Example 1. Suppose we have 40 members and eight gold standard communities (ground truth) as follows:

$\varnothing_1^S = \{a_1, ..., a_5\}$, $\varnothing_2^S = \{a_6, ..., a_{10}\}$, $\varnothing_3^S = \{a_{11}, ..., a_{15}\}$, $\varnothing_4^S = \{a_{16}, ..., a_{20}\}$, $\varnothing_5^S = \{a_{21}, ..., a_{25}\}$, $\varnothing_6^S = \{a_{26}, ..., a_{30}\}$, $\varnothing_7^S = \{a_{31}, ..., a_{35}\}$, $\varnothing_8^S = \{a_{36}, ..., a_{40}\}$

We analyze all possible states based on the number of communities. Table 4 and Fig. 1 present the results of NMI values for different states compared to those of the ARI, PWF, Fowlkes Mallows, Hubert statistics, and Jaccard.

The number of communities in the ground truth is 8, and it should remain constant across all the experiments. The second column shows the number of communities detected by a specific algorithm, while columns 3–8 display the values for each measure. The last column represents the number of ground truth communities

Metric	Formula	Symbol
Normalized Mutual Information (NMI)	$\frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij} \log \frac{C_{ij}N}{C_{i.}C_{.j}}}{\sum_{i=1}^{S} C_{i.} \log \frac{C_{i.}}{N} + \sum_{j=1}^{R} C_{.j} \log \frac{C_{.j}}{N}}$	*NMI*
Adjusted Rand Index	$\frac{\sum_{ij}\binom{C_{ij}}{2} - \sum_i \binom{C_{i.}}{2} \sum_j \binom{C_{.j}}{2} / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{C_{i.}}{2} + \sum_j \binom{C_{.j}}{2}\right] - \sum_i \binom{C_{i.}}{2} \sum_j \binom{C_{.j}}{2} / \binom{N}{2}}$	*ARI*
Pairwise F measure	$\frac{2 \times precision \times recall}{precision + recall}$	*PWF*
Fowlkes Mallows	$\frac{\sum_{ij}\binom{C_{ij}}{2}}{\sqrt{\sum_i \binom{C_{i.}}{2} \sum_j \binom{C_{.j}}{2}}}$	*FM*
Hubert statistic	$\frac{\binom{N}{2}\sum_{ij}\binom{C_{ij}}{2} - \sum_i \binom{C_{i.}}{2} \sum_j \binom{C_{.j}}{2}}{\sqrt{\sum_i \binom{C_{i.}}{2} \sum_j \binom{C_{.j}}{2} \left[\binom{N}{2} - \sum_i \binom{C_{i.}}{2}\right] \left[\binom{N}{2} - \sum_j \binom{C_{.j}}{2}\right]}}$	Hubert
Jaccard	$\frac{\sum_{ij}\binom{C_{ij}}{2}}{\sum_i \binom{C_{i.}}{2} + \sum_j \binom{C_{.j}}{2} - \sum_{ij}\binom{C_{ij}}{2}}$	Jaccard

Table 1. Well-known metrics for evaluating community detection algorithms.

S	R				
	r_1	r_2	…	r_r	Sum
s_1	c_{11}	c_{12}	…	c_{1r}	$c_{1.}$
s_2	c_{21}	c_{22}	…	c_{2r}	$c_{2.}$
…	…	…		…	…
s_s	c_{s1}	c_{s2}	…	c_{sr}	$c_{s.}$
Sum	$c_{.1}$	$c_{.2}$	…	$c_{.r}$	

Table 2. Contingency table.

Notion	Description
$C_{i.}$	Refers to summation of cells in row i of Contingency table
$C_{.j}$	Refers to summation of cells in column j of Contingency table
C_{ij}	Refers to joint members in communities i and j
R	Refers to No. community which detected by certain algorithm
S	Refers to No. community in gold standard (ground truth)
ω	No. Ground truth with full same common members
$\varnothing^S$	Community in S

Table 3. Key notations.

that share common members with the detected communities. For example, if a certain algorithm detects two communities as follows:

$$\varnothing_1^R = \{a_1, ..., a_5\}, \varnothing_1^R = \{a_6, ..., a_{40}\}$$

The ω value is 8 since all members of the 8 ground truth communities share common members with the detected communities.

In this study, we assume the best-case scenario in which the detected communities of a certain algorithm lead to the highest NMI value. Thus, we propose:

Axiom 1: *The highest NMI value is obtained if and only if the detected communities have the highest possible number of common members with the ground truth and if the highest value of* ω *is maintained.*

Now, as shown in Table 4 and Fig. 1, everything appears to be fine when the number of R is less than S. However, the situation changes when the number of R increases and surpasses S. At first glance, it may not seem that there is a strong argument to support the claim that the value of metrics declines as the number of members decreases. However, upon closer examination, it becomes evident that the decrease in the NMI value is less steep compared to the other metrics. This raises the question of what is amiss when comparing all algorithms to a specific metric such as the NMI.

The problem arises when N is equal to the number of communities, indicating that each community includes only one member, which is the worst-case scenario. Surprisingly, even in this unfavorable situation, the NMI value still indicated a high level of efficiency. On the other hand, when R equals one or two, the NMI value returned is very low. For instance, Table 4 demonstrates that when R is 2, the NMI is 0.5, whereas in the worst-case scenario with R being 40, the NMI is 0.72. This finding implies that if Alg1 and Alg2 return 2 and 40 communities respectively, the NMI suggests that the accuracies of Alg1 and Alg2 are 0.5 and 0.72 respectively. Similarly, when R is 4, the NMI is 0.8, whereas when R is 27 the NMI is also 0.8. In both examples, the NMI suggests that a certain algorithm that detects more communities is better. However, in reality, a certain algorithm that detects 4 communities may be better than another algorithm that detects 27 communities with respect to the number of ground truth communities.

Proof of biased behavior of NMI

As discussed in the previous section, when the number of communities (R) increases and exceeds the number of communities in the ground truth (S), the NMI exhibits a biased behavior. Therefore, in this section, we aim to analyze the effect of quantity of communities on this measure. Firstly, we decompose the NMI formula, allowing us to examine how NMI values change from the minimum to the maximum. Subsequently, we present an explanation as to why NMI changes are more pronounced when $R < S$ rather than $R > S$. Before we begin our discussion, it should be noted that for all Lammas and the relevant proof, Axiom 1 should be maintained.

Decomposition

For simplicity, we have Formula 4 (NMI) here.

S	R	ARI	NMI	PWF	FM	Hubert	Jaccard	ω
8	1	0	0	0.222	0.32	0	0.103	8
8	2	0.215	0.5	0.4	0.459	0.347	0.211	8
8	3	0.381	0.665	0.542	0.56	0.485	0.314	8
8	4	0.552	0.8	0.667	0.667	0.617	0.444	8
8	5	0.631	0.857	0.75	0.718	0.679	0.516	8
8	6	0.727	0.909	0.833	0.784	0.756	0.615	8
8	7	0.847	0.957	0.917	0.873	0.857	0.762	8
8	8	1	1	1	1	1	1	8
8	9	0.957	0.98	0.925	0.962	0.958	0.925	7
8	10	0.942	0.969	0.848	0.949	0.943	0.9	7
8	11	0.934	0.962	0.779	0.942	0.936	0.888	7
8	12	0.926	0.954	0.722	0.935	0.929	0.875	7
8	13	0.894	0.94	0.684	0.908	0.899	0.825	6
8	14	0.869	0.929	0.649	0.887	0.877	0.788	6
8	15	0.852	0.919	0.616	0.873	0.862	0.762	6
8	16	0.843	0.912	0.583	0.866	0.854	0.75	6
8	17	0.807	0.899	0.562	0.837	0.823	0.7	5
8	18	0.779	0.889	0.542	0.814	0.799	0.662	5
8	19	0.759	0.88	0.521	0.798	0.782	0.638	5
8	20	0.749	0.873	0.5	0.791	0.774	0.625	5
8	21	0.708	0.862	0.487	0.758	0.741	0.575	4
8	22	0.676	0.852	0.473	0.733	0.715	0.538	4
8	23	0.654	0.844	0.46	0.716	0.697	0.512	4
8	24	0.642	0.838	0.444	0.707	0.688	0.5	4
8	25	0.595	0.827	0.436	0.671	0.651	0.45	3
8	26	0.558	0.818	0.426	0.642	0.622	0.412	3
8	27	0.532	0.811	0.416	0.622	0.602	0.388	3
8	28	0.519	0.805	0.405	0.612	0.592	0.375	3
8	29	0.464	0.796	0.398	0.57	0.549	0.325	2
8	30	0.42	0.787	0.392	0.536	0.516	0.288	2
8	31	0.39	0.78	0.384	0.512	0.492	0.262	2
8	32	0.374	0.775	0.375	0.5	0.48	0.25	2
8	33	0.31	0.766	0.37	0.447	0.428	0.2	1
8	34	0.258	0.758	0.365	0.403	0.385	0.162	1
8	35	0.222	0.752	0.359	0.371	0.354	0.138	1
8	36	0.204	0.747	0.352	0.354	0.337	0.125	1
8	37	0.127	0.739	0.348	0.274	0.26	0.075	0
8	38	0.065	0.731	0.344	0.194	0.184	0.038	0
8	39	0.022	0.725	0.339	0.112	0.106	0.012	0
8	40	0	0.721	0.333	0	0	0	0

Table 4. The results of evaluation metrics based on all possible numbers of communities for a sample network with 40 nodes.

$$NMI(A,B) = \frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log\frac{C_{ij}N}{C_{i.}C_{.j}}}{\sum_{i=1}^{S} C_{i.}\log\frac{C_{i.}}{N} + \sum_{j=1}^{R} C_{.j}\log\frac{C_{.j}}{N}}$$

We decompose the above equation to examine the behavior and performance of each component. The NMI consists of three main components: the common members (C_{ij}), the sum of common members for each community detected by a particular algorithm ($C_{.j}$), and the sum of common members for each community in the ground truth ($C_{i.}$). Here, S represents the number of communities in the ground truth, and R represents the number of communities detected by a specific algorithm. We denote the set of common members as Z. S represents the set of the sum of common members in each ground truth community, while R represents the set of common members in each community detected by a specific algorithm.

$$Z = \{C_{11}, C_{12}, \ldots, C_{21}, C_{22}, \ldots C_{RS}\}$$

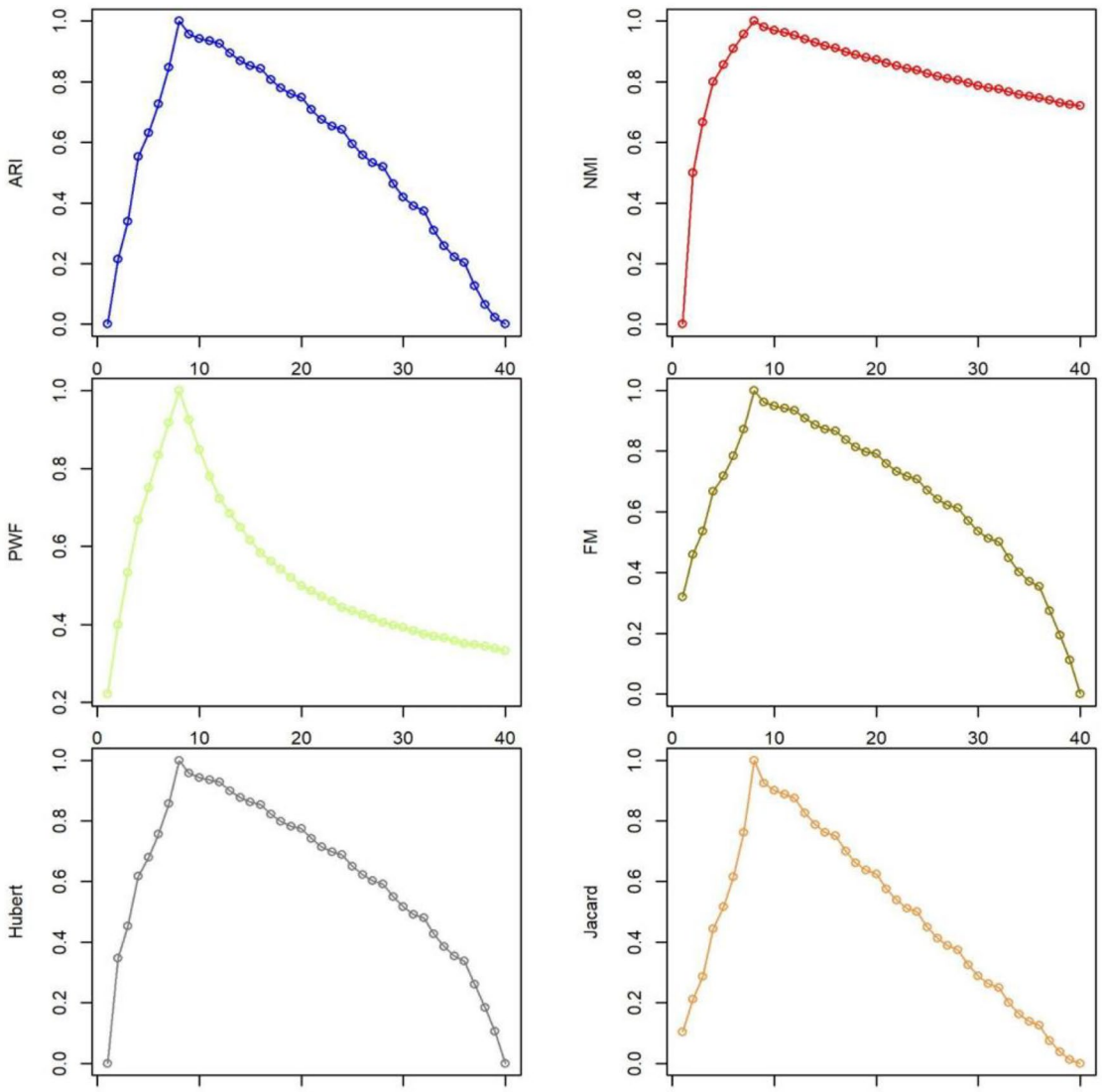

Figure 1. The response of different metrics based on varying numbers of communities.

$S = \{s_1, s_2, \ldots, s_s\}$ is a set representing the number of members in each community of the ground truth, where s_i denotes the number of members in community i. It is important to note that there exists an inverse relationship between each pair of elements in S.

$R = \{r_1, r_2, \ldots, r_r\}$ is a set representing the number of members in each community detected by a certain algorithm, where r_j represents the number of members in community j. Again, there exists an inverse relationship between each pair of elements in R.

Suppose that:

$$K = \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log\frac{C_{ij}N}{C_{i.}C_{.j}}, L = \sum_{i=1}^{S} C_{i.}\log\frac{C_{i.}}{N}, M = \sum_{j=1}^{R} C_{.j}\log\frac{C_{.j}}{N} \tag{5}$$

$$\therefore NMI = \frac{-2K}{L+M} \tag{6}$$

The L and M values are always negative, $\because 0 < C_{i.} < N \therefore \log\frac{C_{i.}}{N} < 0 \Rightarrow C_{i.}\log\frac{C_{i.}}{N} < 0$
In addition, K is positive, therefore, the − 2 multiplied K is also negative and the NMI value is positive.
and:

$$s_i = \sum_{k=1}^{S} C_{ik} = C_i. And\ r_j = \sum_{k=1}^{R} C_{kj} = C_{.j} \tag{7}$$

Lemma 1 $N = \sum_{j=1}^{R} r_j = \sum_{i=1}^{S} s_i$

Proof In simple terms, a certain algorithm detects communities with N members, where these N members are distributed among R communities. Similarly, this term holds true for the ground truth, which consists of S communities. Below, we provide a formal proof:

$$\forall i \exists s_i : s_i = \sum_{K=1}^{S} C_{ik}, \& \forall S \textit{ consist of } |S| \textit{ communities} : S_i \textit{ consist of elements in community } i \ \therefore N = \sum S_i$$

$$\forall j \exists r_j : r_j = \sum_{k=1}^{R} C_{kj}, \& \forall R \textit{ consist of } |R| \textit{ communities} : R_j \textit{ consist of elements in community } i \ \therefore N = \sum R_j$$

From minimum to maximum

Lemma 2 *The minimum value of NMI is obtained if* $R = 1 and S > 1 OR S = 1 and R > 1$

Proof If $R = 1$, then a certain algorithm considers all members in one community, on the other hand $S > 1$ demonstrates that there is more than one community in ground truth. The conclusion is that $C_{ij} = C_{i.}$ and $N = C_{.j}$ therefore $K = 0$. In a similar vein when $S = 1$ and $R > 1$ NMI is zero. Formally this can be shown as:

In both cases K is zero, $\because C_{ij}N = C_{i.}C_{.j} \Rightarrow \log 1 = 0, \therefore NMI = 0$.

$R = S$

Lemma 3 *The maximum value of NMI is 1, if* $R = S \& \forall i, j, i = j \Rightarrow C_{ij} \neq 0 \& s.ti \neq j \Rightarrow C_{ij} = 0$ *and the CT is a square and diagonal matrix.*

Proof When $R = S$ only diagonal elements of CT is non zero, to have a maximum NMI value of 1, the members of each community in the ground truth first should be the same in terms of quantity and second in terms of member similarity with the communities detected by a certain algorithm. It should be noted that this does not imply that the number of members in all communities should be the same. Therefore, when R equals S, a square matrix is formed, ensuring that there are similar members in each pair. This results in members not being distributed across two communities, thus resulting in a diagonal matrix for the contingency table. Having only one non-zero cell leads to:

$$\forall i, j C_{ij} = C_{i.} = C_{.j}$$

$$\frac{C_{ij}N}{C_{i.}C_{j.}} = \frac{C_{i.}N}{C_{i.}C_{i.}} = \frac{N}{C_{i.}} = \left(\frac{C_{i.}}{N}\right)^{-1}$$

$$NMI = \frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{i.}\log\left(\frac{N}{C_{i.}}\right)}{\sum_{i=1}^{S} C_{i.}\log\left(\frac{C_{i.}}{N}\right) + \sum_{i=1}^{S} C_{i.}\log\left(\frac{C_{i.}}{N}\right)}$$

Since the CT is diagonal, only one cell in each row or column is considered to be non-zero $\therefore$

$$\sum_{i=1}^{S}\sum_{j=1}^{R} C_{i.}\log\left(\frac{N}{C_{i.}}\right) = \sum_{i=1}^{S} C_{i.}\log\left(\frac{N}{C_{i.}}\right) \tag{8}$$

In addition:

$$\sum_{i=1}^{S} C_{i.}\log\frac{C_{i.}}{N} = \sum_{j=1}^{R} C_{.j}\log\frac{C_{.j}}{N} \tag{9}$$

$\therefore$

$$\therefore (8)(9) \Rightarrow \frac{-2\sum_{i=1}^{S} C_{i.}\log\left(\frac{N}{C_{i.}}\right)}{2\sum_{i=1}^{S} C_{i.}\log\frac{C_{i.}}{N}} = \frac{-2\sum_{i=1}^{S} C_{i.}\log\frac{N}{C_{i.}}}{-2\sum_{i=1}^{S} C_{i.}\log\frac{N}{C_{i.}}} = 1$$

Furthermore, this can be demonstrated through *contradiction* by considering the case when L and M are not equal, therefore $R \neq S$. If $R > S$, it implies that some members are distributed across different communities. This distribution decreases the number of common members, consequently increasing the absolute value of the denominator in the fraction. Consequently, the NMI decreases, as shown in Eqs. (10) and (11):

$$\text{Where } a = b + c \; then |\log\frac{a}{N}| < |\log\frac{b}{N}| + |\log\frac{c}{N}| \tag{10}$$

Now, if $S > R$, the absolute value of the numerator decreases. This occurs because the logarithmic reduction has a smaller slope compared to the effect of the coefficient.

$$a|\log\frac{a}{N}| < b|\log\frac{b}{N}| + c|\log\frac{c}{N}| \text{ Where } a > N \tag{11}$$

In the given example, the roles of K and M are crucial in the NMI formula, where L is constant. Let us consider the two algorithms, Alg 1 and Alg 2, along with their respective communities.

Example 2 Suppose the ground truth consists of 4 communities: $\{a_1, a_2\}, \{a_3, a_4\}, \{a_5, a_6\}, \{a_7, a_8\}$

Algorithm Alg 1 consists of two communities: $\{a_1, a_2, a_3, a_4\}, \{a_5, a_6, a_7, a_8\}$. In this case, the values for *K, L, M*, and NMI for Alg 1 are as follows:

$$K = 2.76, L = -11,09, M = -5,54 \; and \; NMI = 0.66$$

Now, consider Algorithm Alg 2, which consists of six communities:$\{a_1, a_2\}, \{a_3, a_4\}, \{a_5\}, \{a_6\}, \{a_7\}, \{a_8\}$. In this case, the values for *K, L, M*, and NMI for Alg 2 are as follows:

$$K = 11,09, L = -11,09, M = -13,86 \; and \; NMI = 0.88$$

These values demonstrate the roles of K and M in the NMI formula, and how they contribute to the calculation of the NMI value for Alg 1 and Alg 2, respectively.

$\boldsymbol{R \neq S}$

Now, let us discuss the scenarios where the number of communities is greater or less than the number of communities in the ground truth.

$\boldsymbol{R > S}$

It is clear that $|\boldsymbol{Z}| = \boldsymbol{R} \times \boldsymbol{S}$
For this state:

Lemma 4 *If $R > S$, then the maximum value of NMI is obtained if and only if the cardinality of the set of common members (Z) is equal to $\boldsymbol{R} \times (\boldsymbol{R} - 1)$, and if the number of pairwise common members (C_{RS}) that are not equal to zero is equal to or greater than the cardinality of $\boldsymbol{R}$.*

First it is important to note that values smaller than |R|, indicating non-zero values in the cells of the contingency table, are not possible. This is because having a community without any common members violates the *pigeonhole principle* in mathematics. According to this principle, if n items are placed into m containers, where $n > m$, then at least one container must contain more than one item.

In this scenario, at least one detected community should have at least one common member, resulting in at least $|R|$ being the non-zero member in Z. This can be illustrated with an example.

Example 3 Let us suppose that the ground truth has 2 communities, and a certain algorithm detects 3 communities as follows:

$$\varnothing_1^S = \{a_1, a_2, a_3, a_4\}, \varnothing_2^S = \{a_5, a_{6,} a_7, a_8\}$$

$$\varnothing_1^R = \{a_1, a_2, a_3, a_4\}, \varnothing_2^R = \{a_5, a_6\}, \varnothing_3^R = \{a_7, a_8\}$$

Table 5 displays the number of common members (Z). The number of nonzero elements in Z is 3, which is equal to or greater than $|R|$.

Proof (lemma 4) The greater the difference between R and S, the greater the number of communities detected. This results in fewer common members between the communities in R and S, subsequently leading to a decrease

	r_1	r_2	r_3
s_1	4	0	0
s_2	0	2	2

Table 5. Contingency table of the sample community detection algorithm.

in NMI. Additionally, if the number of non-zero members in Z is greater than R, it indicates that the number of common members between S and R is less, further contributing to the reduction in NMI.

$$\boldsymbol{R} < \boldsymbol{S}$$

For this state:

Lemma 5 *If* $\boldsymbol{R} < \boldsymbol{S}$, *then the maximum value of NMI is obtained if and only if the cardinality of the set of common members (Z) is equal to* $\boldsymbol{S} \times (\boldsymbol{S} - 1)$, *and if the number of pairwise common members* (C_{RS}) *that are not equal to zero is equal to the cardinality of S.*

Like Lemma 4, Lemma 5 can be proven.

Slope of change in the NMI

Lemma 6 *When the difference between R and S increases, in cases where* $R < S$**,** *the slope of changes of the NMI values are greater than when* $R > S$**.**

Proof By applying Lemma 4 and Lemma 5, we can demonstrate that if there is a difference between R and S, the optimal NMI value is obtained when the difference between R and S is 1. In other words, as the difference between R and S increases, the NMI decreases.

Equation (6) is equal to function 12 where we consider NMI as $f(x, y, z)$, $K = g(x, y, z)$, $M = h(z)$:
s.t. $x = C_{ij}, y = C_{i.}, z = C_{.j}$

$$f(x, y, z) = NMI = \frac{-2g(x, y, z)}{L + h(z)} \tag{12}$$

$$L = \sum_{i=1}^{S} C_{i.}\log(\frac{C_{i.}}{N}) = \sum_{i=1}^{S} C_{i.}\log(C_{i.}) - \sum_{i=1}^{S} C_{i.}\log N = \sum_{i=1}^{S} C_{i.}\log(C_{i.}) - N\log N$$

$$= \left(C_{1.}\log(C_{1.}) + C_{2.}\log(C_{2.}) + \cdots + C_{S.}\log(C_{S.})\right) - N\log N$$

where $O = (C_{1.}\log(C_{1.}) + C_{2.}\log(C_{2.}) + \cdots + C_{S.}\log(C_{S.}))$

$$L = O - N\log N \tag{13}$$

$$h(z) = \sum_{j=1}^{R} C_{.j}\log(\frac{C_{.j}}{N}) = \sum_{j=1}^{R} C_{.j}\log(C_{.j}) - \sum_{j=1}^{R} C_{.j}\log N = \sum_{j=1}^{R} C_{.j}\log(C_{.j}) - N\log N$$

$$= \left(C_{.1}\log(C_{.1}) + C_{.2}\log(C_{.2}) + \cdots + C_{.R}\log(C_{.R})\right) - N\log N$$

$$\text{where } P = (C_{.1}\log(C_{.1}) + C_{.2}\log(C_{.2}) + ... + C_{.R}\log(C_{.R})) \tag{14}$$

$$h(z) = P - N\log N \tag{15}$$

For all cases, $N\log N$ is a constant, therefore $h(z)$ is completely dependent on P, s.t $P < N\log N$

When $S > R$, increasing R results in a decrease in P, causing $|h(z)|$ to increase. This might suggest that NMI decreases. However, $g(x, y, z)$ increases, leading to an increase in NMI. Let us examine $g(x, y, z)$ to further understand this.

$$g(x, y, z) = \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(\frac{C_{ij}N}{C_{i.}C_{.j}}) = \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{ij}N) - \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{i.}C_{.j})$$

$$= \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{ij}) + \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(N) - \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{i.}) - \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{.j})$$

$$= \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{ij}) + N\log N - \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{i.}) - \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{.j}) \tag{16}$$

$$\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log(C_{i.}) = \sum_{i=1}^{S} C_{i.}\log(C_{i.}) = O \tag{17}$$

$$\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log\left(C_{.j}\right) = \sum_{j=1}^{R} C_{.j}\log(C_{.j}) = P \tag{18}$$

$$(16), (17), and\ (18) \Rightarrow \mathrm{g}(\mathrm{x},\mathrm{y},\mathrm{z}) = \sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log C_{ij} + N\log N - P - O \tag{19}$$

$$(12), (13), (15) \text{ and } (19) \Rightarrow \mathrm{NMI} = \frac{-2(\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log C_{ij} + N\log N - P - O)}{O + P - 2(N\log N)}$$

$$= \frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log C_{ij} - 2(N\log N - P - O)}{O + P - 2(N\log N)} = \frac{-2\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log C_{ij} + 2(O + P - N\log N)}{O + P - 2(N\log N)}$$

suppose $\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij}\log C_{ij} = T$ then

$$= \frac{2(O+P) - 2N\log N - 2T}{O + P - 2N\log N} \tag{20}$$

O and $N\log N$ are constant for all scenarios. Let us suppose $O = C_1$ and $2N\log N = C_2$:

$$(20) \Rightarrow NMI = \frac{2(C_1 + P) - C_2 - 2T}{C_1 + P - C_2} \tag{21}$$

Now, let us consider Eq. (21) for two possible scenarios:

1. $R < S$ (the number of detected communities is lower than that of the ground truth):

In this scenario, let us suppose that Algorithm 1 returns N_1 communities in one run and N_2 communities in another run, where $N_1 < N_2$.. According to *Axiom 1*, as the number of communities increases, the possibility of member distribution also increases, which leads to a reduction in the P value according to (22) (14) and (13). A decrease in P results in a larger numerator in Eq. (21). This is because it increases the difference with C_2, where C_2, is a constant value for all the scenarios.

$\because$

$$Na = \frac{Na}{2} + \frac{Na}{2}$$

$$N\log a > \frac{N}{2}\log\frac{a}{2} + \frac{N}{2}\log\frac{a}{2}$$

$$\because N\log a > \frac{N}{2}\log a - \frac{N}{2}\log 2 + \frac{N}{2}\log a - \frac{N}{2}\log 2$$

$$N\log a > N\log a - N\log 2 \tag{22}$$

For Alg 1 $R_1 = \{r_{1_{.1}}, r_{1_{.2}}, ..., r_{1_{.N_1}}\}$
For Alg 2, $R_2 = \{r_{2_{.1}}, r_{2_{.2}}, ..., r_{2_{.N_2}}\}$
$N_2 > N_1 : \frac{N}{N_2} < \frac{N}{N_1} \therefore \forall i r_{2_{.i}} \leq r_{1_{.i}} \& \nexists j r_{2_{.j}} > r_{1_{.j}} \& \exists k r_{2_{.k}} < r_{1_{.k}}$
$\therefore$ According to (14) and (22) $P1 > P2$
To further illustrate this, let us consider the following example:

Example 4 Suppose the ground truth consists of 4 communities, Algorithm 1 returns 2 communities, Algorithm 2 returns 3 communities (in both cases, $R < S$), and $N = 8$.

$$\varnothing_1^S = \{a_1, a_2\}, \varnothing_2^S = \{a_3, a_4\}, \varnothing_3^S = \{a_5, a_6\}, \varnothing_4^S = \{a_7, a_8\}$$

$$\varnothing_1^{Alg1} = \{a_1, a_2, a_3, a_4\}, \varnothing_2^{Alg1} = \{a_5, a_6, a_7, a_8\}$$

S	R		
	r_1	r_2	$C_{i.}$
s_1	2	0	2
s_2	2	0	2
s_3	0	2	2
s_4	0	2	2
$C_{.j}$	4	4	

Table 6. Contingency table of Alg1.

$$\varnothing_1^{Alg2} = \{a_1, a_2\}, \varnothing_2^{Alg2} = \{a_3, a_4\}, \varnothing_3^{Alg2} = \{a_5, a_6, a_7, a_8\}$$

Table 6 displays the contingency table of Alg1.

$$P_{Alg1} = 4log(4) + 4log(4) \approx 11.09$$

$$O_{Alg1} = 2log(2) + 2log(2) + 2log(2) + 2log(2) \approx 5.54$$

$$T_{Alg1} = 2log(2) + 2log(2) + 2log(2) + 2log(2) \approx 5.54$$

$$NMI_{Alg1} \approx \frac{33.26 - 2(5.54) - C_2}{16.63 - C_2} \approx \frac{22.18 - C_2}{16.63 - C_2} \approx \frac{22.18 - 33.27}{16.63 - 33.27} \approx \frac{-11.09}{-16.64} \approx 0.66$$

Table 7 presents the contingency table of Alg 2.

$$P_{Alg2} = 2log(2) + 2log(2) + 4log(4) \approx 8.31$$

$$O_{Alg2} = 2log(2) + 2log(2) + 2log(2) + 2log(2) \approx 5.54$$

$$T_{Alg2} = 2log(2) + 2log(2) + 2log(2) + 2log(2) \approx 5.54$$

$$NMI_{Alg2} \approx \frac{27.7 - 2(5.54) - C_2}{13.86 - C_2} \approx \frac{16.61 - 33.27}{13.86 - 33.27} \approx \frac{16.66}{19.41} \approx 0.85$$

Therefore, the NMIs of different algorithms depend on P and T. Let us consider Alg1 and Alg2 with P_1, T_1, P_2, and T_2 respectively . Suppose Alg1 has a lower number of communities than Alg2. In this case, it can be proven that Eq. (23) holds.

$$\frac{2(C_1 + P) - C_2 - 2T}{C_1 + P - C_2} < \frac{2(C_1 + P) - C_2 - 2T}{C_1 + P - C_2} \tag{23}$$

1. R > S (the number of detected communities is greater than that of the ground truth)

To analyze the scenario where $R > S$, we start by incrementing R by one unit. This implies that, after the case where $R = S$, in the next state, a certain algorithm returns $R + 1$ communities. As a reminder, S is constant. In

S	R			
	r_1	r_2	r_3	$C_{i.}$
s_1	2	0	0	2
s_2	0	2	0	2
s_3	0	0	2	2
s_4	0	0	2	2
$C_{.j}$	2	2	4	

Table 7. Contingency table of Alg 2.

this case, one of the community members is distributed across two communities to maintain Axiom 1. This can be expressed as:

$$C_{SR-new} = [\frac{C_{SR-old}}{2}], C_{S,R+1} = C_{SR-old} - [\frac{C_{SR-old}}{2}]$$

By further increasing R, this process is repeated until the new community consists of only one member. The distribution of shared members from one community to new communities results in a lower value of P compared to previous states according to (22) and (14). Additionally, $P = T$ in this case ($R > S$), the numerator remains constant for all scenarios.

The key issue here is that increasing R results in a non-linear (approximately logarithmic) decrease in P. The maximum value of P with respect to R is achieved if:

$$R = 1 \Rightarrow P_1 \approx N\log N \tag{24}$$

$$R = 2 \Rightarrow P_2 \approx \frac{N}{2}\log\frac{N}{2} + \frac{N}{2}\log\frac{N}{2} \tag{25}$$

....

$$R = S \Rightarrow P_R \approx \sum_{j=1}^{R} \frac{N}{S}\log\frac{N}{S} \tag{26}$$

.....

$$R = N \Rightarrow P_N \approx \sum_{j=1}^{R} \frac{N}{N}\log\frac{N}{N} = 0 \tag{27}$$

First, according to (22)

$$P_1 > P_2 > .... > P_n$$

and

$$\forall i \geq 1, i < S \exists j \geq S \Rightarrow (P_i - P_{i+1}) > (P_j - P_{j+1}) \tag{28}$$

According to Eq. (28), we can conclude that when $R < S$, the slope of the change curve is greater than when $R > S$. This results in a larger P value when $R < S$ compared to states where $R > S$. Additionally, it can be proven that the T value when $R < S$ is constant. Therefore, NMI when $R < S$ changes with respect to P, and a greater change in P when $R < S$ leads to a greater change in NMI.

Figure 2 illustrates a nonlinear (approximately logarithmic) decrease in the P value as the number of communities (R) increases.

Lemma 7 *When $R < S$ and Axiom 1 holds, T is a constant.*

Proof

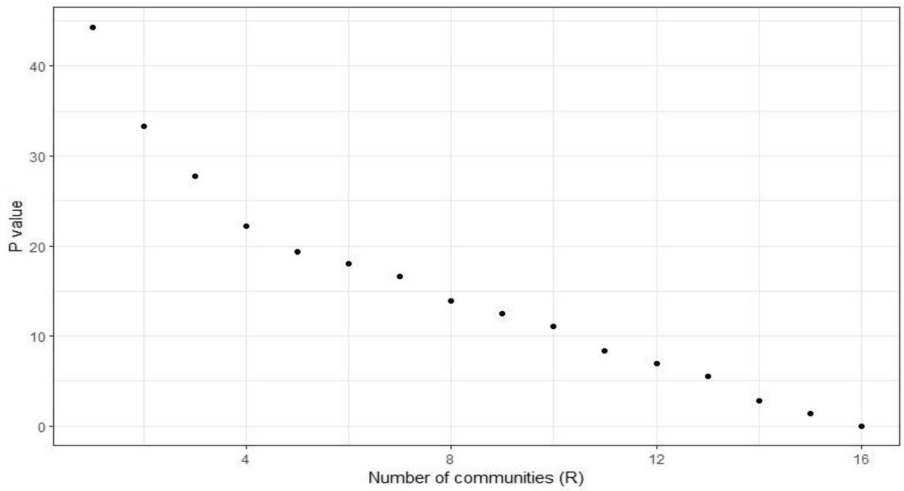

Figure 2. Nonlinear (approximately logarithmic) decreases in the P value.

$$\forall i, \exists j = kC_{ij} \neq 0 \,\&\, \forall k \neq jC_{ik} = 0 \Rightarrow C_{ij} = C_{i.}$$

$$\therefore T = O \Rightarrow T \text{ is constant}$$

As the number of communities (R) becomes larger than the number of communities in the ground truth (S), the changes in P decrease. In this case, the value of T is equal to P, resulting in a constant numerator for the scenario when $R > S$. However, due to a small change in the denominator, the magnitude of change in NMI is smaller when $R > S$ than when $R < S$.

Lemma 8 *When $R > S$ and Axiom 1 holds, then T equals P.*

Proof When $R = S$, the contingency table becomes a scalar matrix.

$$\forall i \exists j, s.t \text{ where } i = j \text{ then } C_{ij} \neq 0, \text{where } i \neq jC_{ij} = 0$$

By increasing each unit in R when $R > S$ then:

$$\exists i \exists j \exists k s.t C_{i.} \cap C_{.j} \neq \varnothing, C_{i.} \cap C_{.k} \neq \varnothing and \forall m s.t m \neq k, jC_{i.} \cap C_{.m} = \varnothing$$

$$\therefore C_{ij} = C_{.i} \Rightarrow T = P$$

Regarding the values of (24) to (27), for the case when $R > S$, the T value is equal to P, resulting in a constant numerator for this scenario.

Overall, when we summarize the NMI formula in Eq. (21), the logarithm function remains the key factor. One of the main characteristics of the logarithm function is its low slope change with respect to its variable. This characteristic is reflected in the descending derivative of the logarithm function. This observation forms the main intuition behind Eq. (28), which explains the biased behavior of NMI. Furthermore, this phenomenon can be generalized to other entropy-based metrics that utilize the logarithm function in their formulas.

Case study

In this section, we provide a case study to demonstrate why and how NMI returns biased results in practice. To do this, we utilized the email-Eu-core network dataset[50], which was generated using email data from a large European research institution. This dataset is provided by Jure Leskovec from Stanford university[51], and according to[50-p.12] " we have anonymized information about all incoming and outgoing email of the research institution". It consists of 1005 nodes (N = 1005) and 25,571 edges. To compare the NMI value and reveal its biased behavior in different scenarios according to R value, seven well-known community detection algorithms were deployed, including multi-level, Louvain, leading eigenvector, infomap, walktrap, edge betweenness (GN), and Leiden. These algorithms are implemented through specific functions developed to detect communities within the 'igraph' package of the R programming language. The functions corresponding to the aforementioned algorithms are as follows: multilevel.community, cluster_louvain, cluster_leading_eigen, cluster_infomap, cluster_walktrap, cluster_edge_betweenness, and cluster_leiden. These functions take as input a network comprising nodes and edges and return the node name along with its corresponding community number. Interestingly, a diverse range of R values obtained by the above-mentioned algorithms helps us clearly show the limitations of NMI. Additionally, a plot of the cumulative distribution function (CDF) of the percentage of common members according to CT is illustrated for each community detection algorithm (see Fig. 3). Each point in this plot is labeled with a number representing the number of common members. For instance, if a point is labeled 1 and the corresponding value on the y-axis is 97, it means that 97% of the cells in CT have 1 or fewer common members. The labels '0' and corresponding CDF values on the y-axis represent the percentages of cells in the CT with a zero value. This indicates a lack of common members between the communities detected by a certain algorithm and the communities in the ground truth. The x-axis illustrates the common members observed in the cells of the CT. For example, in Fig. 3's top plot, the CDF of common members in the contingency table for the GN algorithm reveals the occurrence of values 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 17, 19, and 53. These findings suggest that these values represent the number of common members between the communities detected by the GN algorithm and the ground truth. The CDF essentially shows the frequency of these values in the CT. It is evident that lower values are more commonly observed in the cells of the CT, with the exception of the Leiden algorithm. As we will discuss later, the Leiden algorithm returns a CT in which all cells are filled with the value '1'."

Table 8 shows the NMI values obtained by applying the above-mentioned algorithms to the email-Eu-core network. The abnormal NMI value, which alone can highlight the drawback of the NMI measure, is obtained by the Leiden algorithm, with a value of 0.65. Leiden returns 1005 communities, meaning that each community consists of only one node. In practice, this is a worst-case scenario, as considering each node as a separate community in relation to the ground truth (42 communities) is a noticeable fault.

Furthermore, as shown in Fig. 3, 100% of the cells in CT for the Leiden algorithm have a value of 1, indicating that the communities detected by Leiden have only one common member with the ground truth communities. It is evident that the NMI value of 0.65 is a significantly biased value, whereas the ARI is zero for this algorithm, providing a more realistic measure of common members with the ground truth. The problem arises when an analyst deploys only the NMI without considering information about R, S, or common members, which is a common approach in scientific experiments in community detection studies. In such cases, the analyst may interpret Leiden or infomap as the best algorithm, while Leiden is actually the worst.

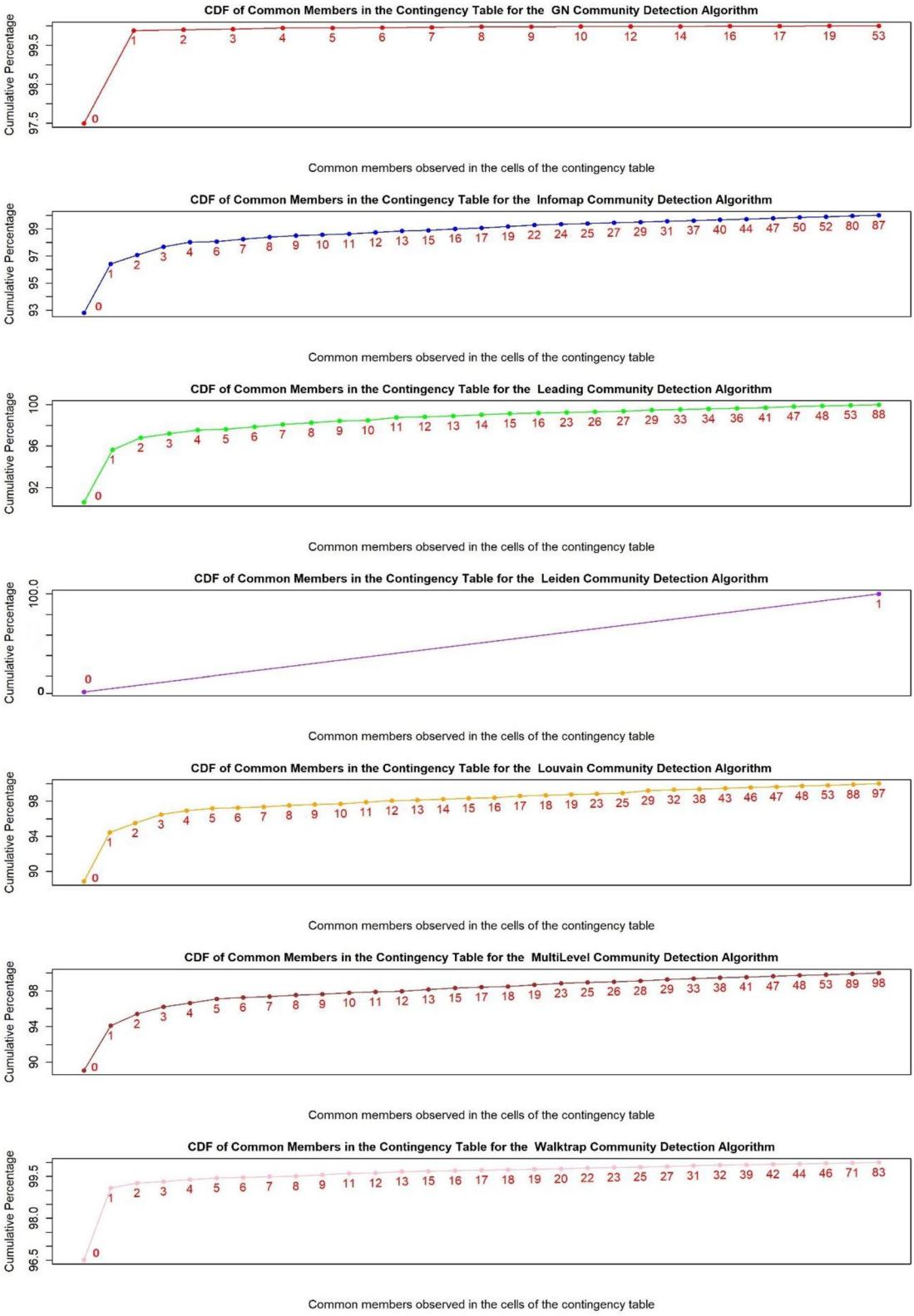

Figure 3. Cumulative Distribution Function (CDF) of common members based on CT.

Now, let us consider the second-highest value of R in Table 8, which is the result of the edge betweenness (GN) algorithm. Its R value is 731, which is significantly greater than S, leading to the distribution of members in numerous communities. Consequently, it is less likely that the communities detected by GN align with the

Algorithm	R	S	P	O	NMI	ARI
Louvain	27	42	4840	3613	0.59	0.34
multi-level	27	42	4847	3613	0.58	0.32
Leading eigenvector	41	42	4718	3613	0.52	0.26
infomap	43	42	4500	3613	0.65	0.30
WalkTrap	145	42	4317	3613	0.58	0.20
Edge betweenness	731	42	1305	3613	0.60	0.06
Leiden	1005	42	0	3613	0.65	0

Table 8. NMI values based on seven well-known community detection algorithms applied to the email-Eu-core network dataset.

ground truth. However, the NMI for this algorithm is 0.6, which is a high value compared to that of the other algorithms in this case study and an inaccurate result. According to Fig. 3, approximately 97.5% of the CT cells are zero, more than 99% are 1 or 0, and only 0.5% of the cells are greater than 1.

Comparing Leiden and GN to Louvain highlights the incompetence of NMI. There are 27 communities returned by Louvain, suggesting a greater likelihood of having more common members with the ground truth. However, this is not guaranteed, as meeting Axiom 1 in a real-world scenario is not assured. Investigating common members based on the CDF plot of the number of cells depicted in Fig. 3 demonstrates that the Louvain algorithm leads to more similar communities with ground truth communities. Despite this fact, NMI returned higher values for GN and Leiden than for Louvain. It can be concluded that when $R > S$, NMI exhibits biased behavior. A comparison with other algorithms also highlights this problem.

Another interesting issue we find from this case study is how Eq. (28) behaves when $R > S$ and $R < S$. As expressed in the explanation of Eq. (28), the changes in each pair of P when $R < S$ are greater than when $R > S$. For instance, when R is 27, P is 4847, when R is 41, P is 4718, and the difference in P is 129. On the other hand, the difference in P when R is 43 and 145 is 183. Therefore, the changes in the case of $R < S$ per unit are $129/(41-27) = 9.21$, which is greater than when $R > S$, where the difference per unit is $183/(145-43) = 1.79$. Therefore, it leads to a sharper change per unit in the NMI value when $R < S$, which is $(0.58-0.52)/(41-27) \times 100 = 0.43$, compared to $R > S$, which is $(0.65-0.58)/102 \times 100 = 0.07$. This computation can be extended to other pairs of R values (e.g., 145, 731, and 1005) with respect to their P value. However, a meaningful comparison will be achieved if we have all possible R values, but in the real world, it is not an easy task to achieve this due to the limited number of community detection algorithms and the large number of nodes. Therefore, by considering a percentage and simulating it per unit, we provide an estimation of this problem.

Ethical approval

The authors declare that there is no any human or live human cells involved in the study.

Conclusion

In this study, we conducted an in-depth analysis of one of the most recognized measures employed in the evaluation of clustering and community detection algorithms. Utilizing a mathematical approach, we demonstrated the inherent bias of this measure (NMI), a bias that becomes particularly pronounced when the number of communities detected by a given algorithm surpasses the number of communities present in the ground truth. Our findings underscore the significant impact that the number of detected communities can have on each evaluation metric, an effect that is especially notable in logarithmic entropy-based metrics such as NMI. This observation is critical because it highlights the potential for skewed results and misinterpretations when using this metric in different applications.

The findings of this study can be generalized and applied in various contexts that utilize community detection evaluation metrics, ranging from friendship networks to critical applications such as identifying certain cancer diseases in protein–protein interaction (PPI) networks. Our study has highlighted and formally proven that the NMI is a biased metric for evaluating the results of community detection algorithms in any application. For instance, an algorithm that fails to detect the protein structure in PPI networks may be considered successful based on the NMI, potentially leading to the misdiagnosis of cancer. Therefore, our findings underscore the need for careful consideration of the characteristics and limitations of NMI, particularly in scenarios where the number of detected communities is high.Therefore, this study formally indicates that in the future, any field of study intending to base decisions on community detection algorithms should exercise caution in selecting the appropriate metric for evaluating these algorithms. This applies across various domains, such as marketing, where accurately targeting communities is crucial, or in information diffusion methods, where identifying dense communities is vital for activating more nodes.

However, the primary contribution of this study lies in revealing the root cause of this biased behavior, originating from the logarithmic function and its corresponding derivative. This insight holds significant value for future studies focused on designing new and equitable metrics within this domain. Understanding the mathematical behavior of this logarithmic metric can substantially aid in the creation of more precise evaluation metrics.

Data availability

The data and codes used and/or analyzed during the current study available from the corresponding author on request.

Received: 13 June 2023; Accepted: 7 April 2024
Published online: 19 April 2024

References

1. Manipur, I., Giordano, M., Piccirillo, M., Parashuraman, S. & Maddalena, L. Community detection in protein–protein interaction networks and applications. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 217–237 (2023).
2. Roy, S., Kundu, S., Sarkar, D., Giri, C. & Jana, P. Community detection and design of recommendation system based on criminal incidents. In *Proceedings of International Conference on Frontiers in Computing and Systems* 71–80 (Springer Singapore, 2021).
3. Ferretti, S. On the Modeling and simulation of portfolio allocation schemes: An approach based on network community detection. *Comput. Econ.* https://doi.org/10.1007/s10614-022-10288-w (2022).
4. Wei, S. & Wang, L. Community detection, road importance assessment, and urban function pattern recognition: A big data approach. *J. Spat. Sci.* **68**, 23–43 (2023).
5. Vicario, M. D. *et al.* The spreading of misinformation online. *Proc. Natl. Acad. Sci.* **113**, 554–559 (2016).
6. Mukerjee, S. A systematic comparison of community detection algorithms for measuring selective exposure in co-exposure networks. *Sci. Rep.* **11**, 15218 (2021).
7. Neff, T. *et al.* Vaccine hesitancy in online spaces: A scoping review of the research literature, 2000–2020. *Harvard Kennedy School Misinf. Rev.* https://doi.org/10.37016/mr-2020-82 (2021).
8. Jemielniak, D. & Krempovych, Y. An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter. *Public Health* **200**, 4–6 (2021).
9. Benkler, Y., Faris, R. & Roberts, H. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. (Oxford University Press, 2018).
10. Mosleh, M. & Rand, D. G. Measuring exposure to misinformation from political elites on Twitter. *Nat. Commun.* **13**, 7144 (2022).
11. Górska, A., Kulicka, K. & Jemielniak, D. Men NOT Going Their Own Way: A Thick Big Data Analysis of #MGTOW and #Feminism Tweets. *Feminist Media Studies* (second round of revisions) (2022).
12. Ophir, Y. *et al.* Weaponizing reproductive rights: a mixed-method analysis of White nationalists' discussion of abortions online. *Inf. Commun. Soc.* **26**, 1–26 (2022).
13. Panizo-LLedot, A., Torregrosa, J., Bello-Orgaz, G., Thorburn, J. & Camacho, D. Describing alt-right communities and their discourse on twitter during the 2018 US Mid-term elections. In *Complex Networks and Their Applications VIII* 427–439 (Springer International Publishing, 2020).
14. Okruszek, Ł, Piejka, A., Banasik-Jemielniak, N. & Jemielniak, D. Climate change, vaccines, GMO: The N400 effect as a marker of attitudes toward scientific issues. *PLoS One* **17**, e0273346 (2022).
15. Grusauskaite, K., Carbone, L., Harambam, J. & Aupers, S. Debating (in) echo chambers: How culture shapes communication in conspiracy theory networks on YouTube. *New Media Soc.* 14614448231162585 (2023).
16. Kaiser, J., Rauchfleisch, A. & Córdova, Y. Comparative approaches to mis/disinformation| fighting Zika with honey: An analysis of YouTube's video recommendations on Brazilian YouTube. *Int. J. Commun. Syst.* **15**, 19 (2021).
17. Humprecht, E., Esser, F. & Van Aelst, P. Resilience to online disinformation: A framework for cross-national comparative research. *Int. J. Press/Polit.* **25**, 493–516 (2020).
18. Ahmad, N., Milic, N. & Ibahrine, M. Data and disinformation. *Computer* **54**, 105–110 (2021).
19. Lewandowsky, S., Ecker, U. K. H. & Cook, J. Beyond misinformation: Understanding and coping with the 'Post-Truth' era. *J. Appl. Res. Mem. Cogn.* **6**, 353–369 (2017).
20. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.* **19**, 22–36 (2017).
21. Darius, P. & Stephany, F. How the far-right polarises twitter: 'Hashjacking' as a disinformation strategy in times of COVID-19. In *Complex Networks & Their Applications X* 100–111 (Springer International Publishing, 2022).
22. De Clerck, B. *et al.* Maximum entropy networks applied on twitter disinformation datasets. In *Complex Networks & Their Applications X* 132–143 (Springer International Publishing, 2022).
23. Hasan Ahmed Abdulla, H. H. & Abdulla, H. H. Fake news detection: A graph mining approach. In *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)* 1–5 (2023).
24. Kaur, K. & Gupta, S. Towards dissemination, detection and combating misinformation on social media: a literature review. *J. Bus. Ind. Market.* (2022) (**ahead-of-print**).
25. Ali, M. *et al.* Social media content classification and community detection using deep learning and graph analytics. *Technol. Forecast. Soc. Change* **188**, 122252 (2023).
26. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).
27. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8577–8582 (2006).
28. Meilă, M. Comparing clusterings—An information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
29. Lancichinetti, A. & Fortunato, S. Limits of modularity maximization in community detection. *Phys. Rev. E Stat. Nonlinear Soft. Matter Phys.* **84**, 066122 (2011).
30. Amelio, A. & Pizzuti, C. Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* 1584–1585 (Association for Computing Machinery, 2015).
31. Cai, Q., Ma, L., Gong, M. & Tian, D. A survey on network community detection based on evolutionary computation. *Int. J. Bio-Inspir. Comput.* **8**, 84 (2016).
32. Lai, D. & Nardini, C. A corrected normalized mutual information for performance evaluation of community detection. *J. Stat. Mech.* **2016**, 093403 (2016).
33. Liu, X., Cheng, H.-M. & Zhang, Z.-Y. Evaluation of community detection methods. *IEEE Trans. Knowl. Data Eng.* **32**, 1736–1746 (2020).
34. Zhang, P. Evaluating accuracy of community detection using the relative normalized mutual information. *J. Stat. Mech.* **2015**, P11006 (2015).
35. Yang, Z., Algesheimer, R. & Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* **6**, 30750 (2016).
36. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech.* **2005**, P09008 (2005).
37. Karataş, A. & Şahin, S. Application areas of community detection: A review. In *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)* 65–70 (2018).

38. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).
39. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7821–7826 (2002).
40. Mahmoudi, A., Bakar, A. A., Sookhak, M. & Yaakub, M. R. A temporal user attribute-based algorithm to detect communities in online social networks. *IEEE Access* **8**, 154363–154381 (2020).
41. Chen, M., Nguyen, T. & Szymanski, B. K. A New Metric for Quality of Network Community Structure. *arXiv [cs.SI]* (2015).
42. Romano, S., Bailey, J., Nguyen, V. & Verspoor, K. Standardized mutual information for clustering comparisons: One step further in adjustment for chance. In *Proceedings of the 31st International Conference on Machine Learning* (eds. Xing, E. P. & Jebara, T.) vol. 32 1143–1151 (PMLR, 2014).
43. Rossetti, G., Pappalardo, L. & Rinzivillo, S. A novel approach to evaluate community detection algorithms on ground truth. In *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016* (eds. Cherifi, H., Gonçalves, B., Menezes, R. & Sinatra, R.) 133–144 (Springer International Publishing, 2016).
44. Arab, M. & Hasheminezhad, M. Limitations of quality metrics for community detection and evaluation. In *2017 3th International Conference on Web Research (ICWR)* 7–14 (2017).
45. Meilă, M. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines* 173–187 (Springer Berlin Heidelberg, 2003).
46. Wagner, S. & Wagner, D. *Comparing clusterings—An overview*. https://publikationen.bibliothek.kit.edu/1000011477 (2007) https://doi.org/10.5445/IR/1000011477.
47. Santos, J. M. & Embrechts, M. On the use of the adjusted rand index as a metric for evaluating supervised classification. 175–184 (2009).
48. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics* 1–8 (Association for Computing Machinery, 2012).
49. Saltz, M., Prat-Pérez, A. & Dominguez-Sal, D. Distributed community detection with the WCC metric. In *Proceedings of the 24th International Conference on World Wide Web* 1095–1100 (Association for Computing Machinery, 2015).
50. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 2-es (2007).
51. email-Eu-core network. https://snap.stanford.edu/data/email-Eu-core.html.

Author contributions

A.M. developed the theory and performed the computations. developed the theoretical formalism, performed the analytic calculations and performed the numerical simulations. Prove the hypothesis of the manuscript and provide mathematical proof of this paper. Prepared figures and/or tables, authored and reviewed drafts of the article, and approved the final draft.D.J. investigated and supervised the findings of this work, editing and revising the whole text. Contributed to literature review. Supervised the project. Authored and reviewed drafts of the article, and approved the final draft.

Funding

This work was supported by Narodowe Centrum Nauki (National Science Centre, Poland) under Grant 2020/38/A/HS6/00066.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.M.