Improved mutual information measure for clustering, classification, and community detection

M. E. J. Newman,[1,2] George T. Cantwell,[1] and Jean-Gabriel Young[2]
[1]*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2]*Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA*

(Received 7 August 2019; revised manuscript received 13 February 2020; accepted 25 March 2020; published 23 April 2020)

The information theoretic measure known as mutual information is widely used as a way to quantify the similarity of two different labelings or divisions of the same set of objects, such as arises, for instance, in clustering and classification problems in machine learning or community detection problems in network science. Here we argue that the standard mutual information, as commonly defined, omits a crucial term which can become large under real-world conditions, producing results that can be substantially in error. We derive an expression for this missing term and hence write a corrected mutual information that gives accurate results even in cases where the standard measure fails. We discuss practical implementation of the new measure and give example applications.

DOI: 10.1103/PhysRevE.101.042304

I. INTRODUCTION

Mutual information is widely used in physics, statistics, and machine learning as a tool for comparing different labelings of a set of objects [1]. For instance, within physics it is used in statistical mechanics for comparing states of spin models [2] and particularly in network science for comparing partitions of networks into communities, mutual information being perhaps the standard measure for quantifying the performance of community detection algorithms [3,4]: it tells us the extent to which the set of communities found by an algorithm agree with a given set of ground-truth communities. In machine learning and statistics, mutual information is similarly used in classification (supervised learning) and clustering (unsupervised learning) to quantify the similarity of different labelings of sets of objects [5]. For instance, we might attempt to deduce characteristics of a set of users of an online service, such as their age group or gender, and then calibrate our algorithm by using mutual information to compare our efforts against known characteristics of a test set of users. Mutual information also finds use in purely empirical settings for quantifying correlations between any two categorical variables describing the same set of objects. Does your name predict your nationality, for example?

Imagine then that we have some set of individuals or objects, such as people, documents, email messages, or aerial photographs, among many other possibilities. Each object can be classified or labeled as belonging to one of several types, groups, or communities. In supervised learning or classification problems the labels usually correspond to some known characteristic: people could be labeled by sex, race, or blood type, for instance; documents by topic; aerial photographs by type of terrain, and so forth. In unsupervised learning or clustering problems the labels may have no such explicit meaning, or the meaning maybe unknown, but the grouping of the objects is nonetheless of interest—community detection problems are a good example.

Now imagine we have two different sets of labels for our objects. One set might be inferred by some algorithm, for instance, and the other assigned by human experts. The mutual information of the two labelings represents the amount of information that the first labeling gives us about the second—in effect, how good the algorithm is at mimicking the human experts. If the two labelings are exactly the same, then the first tells us everything about about the second and the mutual information is maximal. If the two labelings are completely unrelated, then the first tells us nothing about the second and the mutual information is zero. A nice feature of mutual information is that it is invariant under permutations of the labels of one or both labelings, which is useful in clustering or community detection applications where one is interested primarily in the division of objects into groups and not in the labels themselves.

Despite its advantages, however, mutual information is not perfect. In this paper we show that, as traditionally defined, the mutual information can give inaccurate answers—indeed maximally inaccurate—under certain conditions and particularly when the number of groups or communities differs between the two labelings. In some applications these conditions arise frequently, in which case the standard measure can fail badly. We show how to properly correct for this failure, creating a mutual information measure that gives useful answers even when the traditional measure fails. These findings have potentially broad repercussions. Given the widespread use of mutual information in machine learning, network science, and other areas, the number of studies whose results could be changed, either modestly or in some cases greatly, by switching to the corrected measure, is substantial.

Various shortcomings of the mutual information have been discussed previously and several alternatives or variants have been proposed. Best known is the work of Danon *et al.* [3], who point out that the range of values of the mutual information is somewhat arbitrary and suggest a modified measure, the *normalized mutual information*, whose values run from zero to one. This measure is useful, for instance, for comparing results across data sets of different sizes and it has become widely used, particularly in network science. We discuss the normalized mutual information in more detail in Sec. II and define a normalized version of our own measure in Sec. III C.

Several previous authors have raised objections to the mutual information similar to those discussed in this paper [6–8] and suggested potential solutions [7–11]. Best known of these solutions is the *variation of information*, proposed by Meilă [9], a metric information distance that sees some use with clustering and community detection algorithms. Though formally elegant, however, the variation of information is difficult to interpret, bearing no simple relationship to the labelings it describes. Rosenberg and Hirschberg [10] have proposed *V-measure*, a heuristic one-parameter family of measures that emphasizes a balance between two desirable properties, homogeneity and completeness, while Gates *et al.* [8] have proposed an "element-centric clustering similarity" calculated via a multistep process involving graph- and matrix-based concepts. Perhaps closest to our own work is that of Dom [7], who proposes a measure that shares some features with ours but which, as we will argue, has problems of its own. A useful survey of measures for comparing labelings has been given by Amigó *et al.* [11].

A different shortcoming of the mutual information has been highlighted by Vinh *et al.* [12], who argue that mutual information should by rights be zero for unrelated random labelings, but in practice this is only true in the limit of infinitely large sets of objects. For finite sets, and hence for all practical applications, the expected mutual information of two random labelings is nonzero, which is undesirable. Vinh *et al.* propose a measure they call the *adjusted mutual information*, which corrects for this problem by subtracting the average mutual information of a random pair of labelings. We will show that our own proposed measure also corrects for this shortcoming of the mutual information, but we argue that it does so in a more correct manner. A detailed discussion and comparison of the various measures is given in Sec. V.

II. MUTUAL INFORMATION AND CONDITIONAL ENTROPY

Consider a set of n objects, numbered 1 to n, and consider two divisions of those objects into groups or communities. The two divisions need not have the same number of groups but the groups must be nonoverlapping within each division and they must not be empty. Let us suppose that the first division has R groups labeled by integers $r = 1 \dots R$ and the second division has S groups labeled $s = 1 \dots S$. (Throughout this paper we use the language of general clustering or classification problems—"objects" and "labels"—but all of the developments described apply equally, without modification, in other realms such as network community detection.)

Let a_r be the number of objects with label r in the first labeling, let b_s be the number with label s in the second, and let c_{rs} be the number labeled r in the first labeling and s in the second. Note that

$$a_r = \sum_{s=1}^{S} c_{rs}, \qquad b_s = \sum_{r=1}^{R} c_{rs}, \tag{1}$$

$$\sum_{r=1}^{R} a_r = \sum_{s=1}^{S} b_s = \sum_{rs} c_{rs} = n, \tag{2}$$

and that $a_r > 0$ and $b_s > 0$ for all r and s since the groups are nonempty. With these definitions we can compute the probabilities $P(r)$ and $P(s)$ that an object chosen uniformly at random has group label r or s, or the probability $P(r, s)$ that it has both labels r and s, from

$$P(r) = \frac{a_r}{n}, \qquad P(s) = \frac{b_s}{n}, \qquad P(r, s) = \frac{c_{rs}}{n}. \tag{3}$$

Now we ask the following question: If we are told the label r of a particular object, then how much additional information is need to specify the other label s of the same object, on average? Information theory tells us that the answer is given by the entropy of the conditional probability distribution $P(s|r) = P(r, s)/P(r)$:

$$\text{Entropy} = -\sum_{s} P(s|r) \log P(s|r). \tag{4}$$

The average of this entropy over all objects, i.e., over the complete distribution of r, is the quantity we call the *conditional entropy* of s given r:

$$\begin{aligned} H(s|r) &= -\sum_{r} P(r) \sum_{s} P(s|r) \log P(s|r) \\ &= -\sum_{rs} P(r, s) \log \frac{P(r, s)}{P(r)}. \end{aligned} \tag{5}$$

The conditional entropy is the average amount of additional information we would need to supply, on top of the value r, to specify the value s. (Traditionally the logarithms would be taken base 2, giving entropy in units of bits, but other choices are possible and produce only an overall multiplier in the entropy. None of our results will depend on what base is used. Note also that for the expression above and subsequent expressions to give correct answers we must adopt the convention that $0 \log 0 = 0$.)

The maximum of the conditional entropy occurs when r and s are independent, so that r tells us nothing about s and $P(r, s) = P(r)P(s)$, which gives $H(s|r) = H(s)$, where

$$H(s) = -\sum_{s} P(s) \log P(s) \tag{6}$$

is the (unconditional) entropy of s. Conversely, the minimum value of the conditional entropy is zero, which occurs when r and s are in perfect agreement.

Arguably the conditional entropy is a little counterintuitive since it is minimized, not maximized, when labelings are identical. One can reverse the scale and create a measure that is maximized for maximum similarity by subtracting $H(s|r)$ from its maximum value $H(s)$, which gives the quantity known as the *mutual information*:

$$\begin{aligned} I(r;s) &= H(s) - H(s|r) \\ &= -\sum_s P(s)\log P(s) + \sum_{rs} P(r,s)\log\frac{P(r,s)}{P(r)} \\ &= \sum_{rs} P(r,s)\log\frac{P(r,s)}{P(r)P(s)}. \end{aligned} \tag{7}$$

This measure is now zero for uncorrelated labels and equal to $H(r) = H(s)$ for identical ones. The mutual information is often described as the amount of information that the known labeling r tells us about the unknown one s. For our purposes, however, it is perhaps more useful to think of it as the amount of information we save if we already know r when specifying s, as compared to if we do not.

In addition to increasing with the similarity of the labelings, the mutual information has the nice property of being symmetric in r and s. In some circumstances, it is convenient to normalize it so as to create a measure whose value runs between zero and one [3]. There are several ways to perform the normalization [13], but the most widely used normalizes by the mean of the entropies $H(r)$ and $H(s)$, which preserves the symmetry with respect to r and s:

$$\text{Normalized mutual information} = \frac{I(r;s)}{\frac{1}{2}[H(r)+H(s)]}. \tag{8}$$

A. Shortcomings of the mutual information

As we have said, the mutual information is not perfect. The particular shortcoming that we focus on here is illustrated by the following two simple examples suggested by Gates *et al.* [8]. First, consider a case where the labeling r consists of just a single group or community containing all n objects. At the same time let us assume that the labeling s is nontrivial, with $S > 1$. It is clear in this case that labeling r communicates no information about labeling s, and hence the mutual information $I(r;s)$ should be zero. And indeed it is, since $P(r) = 1$ and $P(r,s) = P(s)$ and hence

$$I(r;s) = \sum_s P(s)\log\frac{P(s)}{P(s)} = 0. \tag{9}$$

Now consider a second case where the labeling r consists of n groups, each containing a single object. It is again clearly true that r communicates no information about s in this situation and hence the mutual information should be zero, but if we perform the calculation this is no longer what we find. In this case c_{rs} takes only the values zero and one and $a_r = 1$ for all r. Combining Eqs. (1), (3), (6), and (7), we then get

$$\begin{aligned} I(r;s) &= \frac{1}{n}\sum_{rs} c_{rs}\log\frac{n\,c_{rs}}{a_r b_s} \\ &= \frac{1}{n}\sum_{rs}\left[c_{rs}\log c_{rs} - c_{rs}\log a_r - c_{rs}\log\frac{b_s}{n}\right] \\ &= -\sum_s \frac{b_s}{n}\log\frac{b_s}{n} = -\sum_s P(s)\log P(s) \\ &= H(s) \end{aligned} \tag{10}$$

(where, as previously, we are employing the convention that $0\log 0 = 0$).

Equation (10) is the complete opposite of what we expect. The mutual information should be zero; instead it takes its maximal value of $H(s)$. The measure has failed completely.

III. AN IMPROVED MEASURE OF SIMILARITY

In this paper we propose a new measure, the *reduced mutual information*, or RMI for short, which corrects for the failure illustrated in the previous section. Our measure is simple to state—it is equal to the standard mutual information minus a single correction term, thus,

$$\text{RMI} = I(r;s) - \frac{1}{n}\log\Omega(a,b), \tag{11}$$

where $\Omega(a,b)$ is a (usually large) integer equal to the number of $R \times S$ nonnegative integer matrices with row sums $a = \{a_r\}$ and column sums $b = \{b_s\}$. We give a (previously published) formula for calculating this number, to a good approximation, in Sec. III E.

The origin of the correction term in Eq. (11) lies in the fact that even when our two labelings r and s agree perfectly, r does not tell us everything about s. To deduce s from r one also needs to know how the values of the labels used in r map to those used in s. One can encapsulate this mapping in a matrix and it is the information content of this matrix that gives rise to the term in $\Omega(a,b)$. To understand the argument in detail, let us look more closely at the conditional entropy and mutual information, paying careful attention to their derivation, including several terms that are commonly neglected.

A. Information content of a labeling

The central question we want to answer is this: if we know the labels in labeling r, how much additional information is needed to specify the other labeling s? One can think in terms of two individuals, traditionally called Alice and Bob, who are in communication with one another. Alice knows both labelings of the objects, r and s, but Bob only knows r. How much information, in the traditional Shannon sense, must Alice transmit to communicate s to Bob?

The exact answer depends on how Alice encodes her communication with Bob. There are more and less efficient ways to transmit the labeling s that will accomplish the desired outcome with shorter or longer communications. At the simplest level, Alice could entirely ignore r and just send a complete record of s to Bob as a single message. Since there are S possible labels for each object, there are S^n possible such messages Alice might need to send. If the message is sent in binary, then k bits will suffice to encode uniquely each of these S^n possibilities provided $2^k \geqslant S^n$, or $k \geqslant \log_2 S^n = n\log_2 S$. Since k is an integer, the number of bits needed to communicate s in this manner is thus given by $\lceil n\log S\rceil$, the smallest integer not less than the log of the number of different messages one might send. Then the amount of information *per object* is

$$H_1 = \frac{1}{n}\lceil n\log S\rceil = \log S + \mathrm{O}(1/n). \tag{12}$$

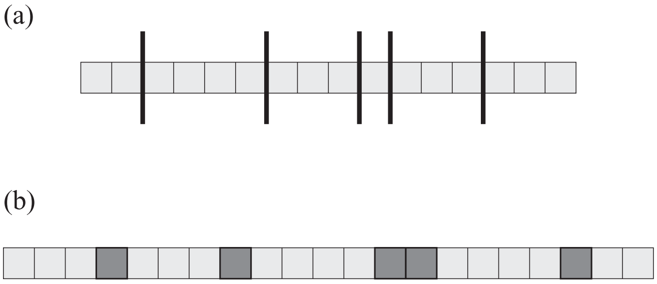

FIG. 1. (a) The number of ways of choosing S strictly positive integers b_s such that they sum to n is equal to the number of ways of placing $S-1$ dividers (the vertical bars) between n objects (represented by the line of gray squares), with no two dividers allowed to fall in the same place. The number of squares between each divider and the next represents the value of one of the b_s, and trivially they sum to n. Since there are $n-1$ places the dividers can go, the total number of ways of choosing the b_s is $\binom{n-1}{S-1}$. (b) The number of ways of choosing S integers b_s such that they sum to n, where individual integers are now allowed to be zero (but not negative), is equal to the number of ways of placing an additional $S-1$ divider objects (the dark squares) among the n lighter squares, for a total of $n+S-1$ squares overall. The values of the b_s are again the numbers of squares between the dividers, which again necessarily sum to n. Since two divider squares can now be adjacent, however, we can have $b_s=0$, and the total number of divisions is $\binom{n+S-1}{S-1}$.

In the common case where the number n of objects is large we neglect the term of order $1/n$ and just say that the amount of information is $\log S$. (As before, the logarithm would traditionally be taken base 2, but the choice of base only affects the overall multiplier of the information and is not important for our purposes. Henceforth, we will for convenience use natural logarithms in this paper, converting to units of bits only where necessary.)

B. A more efficient encoding

Equation (12) provides an upper bound on the amount of information Alice needs to send per object. It is always possible to send a simple message like this and achieve the desired communication. It is, however, often also possible to do better. For example, suppose that not all labels occur with equal frequency, i.e., that b_s varies with s. Then in most cases Alice can transmit s more compactly by first sending Bob the values of the b_s, followed by the particular value of s. This approach works as follows.

The number of possible choices of b_s is equal to the number of different sets of S nonnegative integers that sum to n, which is $\binom{n-1}{S-1}$, and the amount of information needed to communicate one of these sets is, once again, given by the log of this number. (See Fig. 1 for a simple visual explanation of why this is the correct combinatoric factor.) Once the values of the b_s are specified, Alice need then only tell Bob which labeling s is the correct one out of the set of all choices compatible with those values. Normally this set of compatible choices is much smaller than the total number of possible labelings, which is why this approach is more economical.

The number of compatible choices of the labeling s is given by the multinomial coefficient $n!/\prod_s b_s!$ and the information needed to specify a particular choice is the logarithm of this number. So the total information needed to communicate the value of s is $\log\binom{n-1}{S-1}+\log(n!/\prod_s b_s!)$, and the information per object is

$$H_2=\frac{1}{n}\left[\log\binom{n-1}{S-1}+\log\frac{n!}{\prod_s b_s!}\right]. \tag{13}$$

This expression is exact and easily computable. We can, however, gain further insight by approximating the second term using Stirling's formula in the form

$$\log n! = n\log n - n + \tfrac{1}{2}\log n + \mathrm{O}(1), \tag{14}$$

which gives us

$$\begin{aligned}\frac{1}{n}\log\frac{n!}{\prod_s b_s!} &= \sum_s \frac{b_s}{n}\log\frac{n}{b_s}+\mathrm{O}\left(\frac{\log n}{n}\right)\\ &= -\sum_s P(s)\log P(s)+\mathrm{O}\left(\frac{\log n}{n}\right),\end{aligned} \tag{15}$$

where as previously $P(s)=b_s/n$ is the probability that a randomly chosen object is in group s.

The sum in Eq. (15) is the Shannon entropy of s, Eq. (6). Note that this entropy represents only the second term in Eq. (13). It omits the first term, which measures the information needed to transmit the values of the b_s. Applying Stirling's approximation to this term too, we find that

$$\frac{1}{n}\log\binom{n-1}{S-1}=(S-1)\frac{\log n}{n}+\mathrm{O}(1/n), \tag{16}$$

for S constant as n grows. Thus, this term is of the same order as the terms we have neglected in Eq. (15) and it is often legitimate to neglect it too, in which case our information is given approximately by the conventional entropy, Eq. (15).

Assuming this to be the case, we can apply Jensen's inequality, which says that $\sum_s q_s\log x_s\leqslant\log\sum_s q_s x_s$ for any set of nonnegative numbers x_s and weights q_s with $\sum_s q_s=1$, with the exact equality applying only when all x_s are equal. Setting $q_s=b_s/n$ and $x_s=n/b_s$, we then get

$$H_2=\sum_s\frac{b_s}{n}\log\frac{n}{b_s}<\log\sum_s 1=\log S=H_1, \tag{17}$$

except when all b_s are equal. Thus, $H_2<H_1$ and this new scheme for transmitting the labeling s requires less information than simply sending the complete labeling. In the language of information theory, we have found a more efficient encoding of the information by breaking it up and sending the values of b_s and s separately.

C. Employing shared information

So far Alice has made no use of the fact that Bob already knows the other labeling r of our n objects. If r is correlated with s, however, then it may be possible to use that fact to convey the value of s more succinctly. Again there is more than one way to do this, corresponding to different encoding schemes, but we can illustrate the point with the following simple approach. Alice first communicates to Bob the complete set of values c_{rs}, defined in Sec. II, which are the numbers of objects with each combination of labels r, s. The matrix of values c_{rs} is known in this context as a *contingency table*. After sending the contingency table, Alice then follows up by sending the particular value of s. As in the previous section, knowing the contingency table narrows down the number of possible choices for s and hence (in most cases) reduces the total amount of information Alice needs to send to specify the labeling s. The accounting works as follows.

Since both Alice and Bob already know r they also know a_r, which are the row sums of the contingency table, so, even before Alice sends the table, Bob knows that it must have these row sums. The number of choices for a row with S elements which sum to a_r (and which are allowed to be zero) is $\binom{a_r+S-1}{S-1}$ (see Fig. 1 again for a visual explanation), and hence the number of tables with row sums a_r is $\prod_r \binom{a_r+S-1}{S-1}$ and the information needed to transmit the table is the log of this number. At the same time, the number of possible choices of s that are compatible with a particular contingency table is given by a product of multinomial coefficients $\prod_r [a_r!/\prod_s c_{rs}!]$ and again the information is the logarithm. Hence the total information, per object, needed to communicate s to Bob is

$$H_3 = \frac{1}{n}\sum_r \left[\log \binom{a_r+S-1}{S-1} + \log \frac{a_r!}{\prod_s c_{rs}!}\right]. \tag{18}$$

Focusing on the last term in this expression and again applying Stirling's approximation, Eq. (14), we find that

$$\begin{aligned}\frac{1}{n}\sum_r \log \frac{a_r!}{\prod_s c_{rs}!} &= \sum_{rs} \frac{c_{rs}}{n} \log \frac{a_r}{c_{rs}} + \mathrm{O}\left(\frac{\log n}{n}\right)\\ &= -\sum_{rs} P(r,s) \log \frac{P(r,s)}{P(r)} + \mathrm{O}\left(\frac{\log n}{n}\right),\end{aligned} \tag{19}$$

where $P(r,s) = c_{rs}/n$ is, as previously, the probability that a randomly chosen object has labels r, s. The sum in the second line of Eq. (19) is precisely the conditional entropy $H(s|r)$ of Eq. (5) and thus we can to a good approximation write

$$H_3 \simeq H(s|r) + \frac{1}{n}\sum_r \log \binom{a_r+S-1}{S-1}. \tag{20}$$

Thus, we see that the information needed to communicate s is not exactly equal to the standard conditional entropy, but contains an additional term.

Equation (20) was proposed previously as a measure of the similarity of labelings by Dom [7], who argued for its use on the grounds that the second term compensates to some extent for the shortcomings of the conditional entropy discussed in Sec. II A.

The measure has some issues however. For instance, it does not equal $H(s)$—as we would expect—in the case discussed in Sec. II A where the labeling r places each object in a separate group on its own. In that case, $H(s|r) = 0$ and, putting $a_r = 1$ for all r, we get $H_3 = \log S$, which can be in error by a wide margin depending on the distribution of the b_s.

Instead therefore we here adopt a different approach as follows. In Sec. III B we found that specifying the values of the b_s reduced the total amount of information Alice needed to send. The same is true here. Alice can use a three-part encoding, in which she first sends the values of the b_s, then the contingency table, then s itself. Each step reduces the number of possibilities on the next and hence saves information. The accounting is as follows.

First Alice sends the values of the b_s. As we have said, this takes an amount of information equal to $\log \binom{n-1}{S-1}$. Second, she sends the contingency table. Since she and Bob now know both a_r and b_s for all r and s, both the row sums and the column sums of the table are fixed, and hence we need choose only among the set of tables that satisfy these constraints. Suppose that the number of such tables is $\Omega(a,b)$. Then the information needed to transmit the contingency table is $\log \Omega(a,b)$. Third, Alice sends s itself, choosing among only those values that are compatible with the contingency table. As before, this requires an amount of information equal to $\sum_r \log(a_r!/\prod_s c_{rs}!)$. Thus, the total information, per object, is

$$H_4 = \frac{1}{n}\left[\log \binom{n-1}{S-1} + \log \Omega(a,b) + \sum_r \log \frac{a_r!}{\prod_s c_{rs}!}\right]. \tag{21}$$

If, as previously, the first term in this expression is negligible [since it vanishes for large n—see Eq. (16)], and given that the number $\Omega(a,b)$ of contingency tables with row and column constraints can never exceed the number with row constraints alone, we have $H_4 \leqslant H_3$, making Eq. (21) the most efficient of the encoding schemes we have considered. Technically, like all of the measures we have given, this one is still just an upper bound on the amount of information: in any particular situation there could exist an encoding that transmits s more succinctly. However, as we argue in Sec. VI, Eq. (21) is the best bound we can establish without making use of domain-specific knowledge about the particular system of study. If we wish to define a general measure that does not require additional tuning for each individual application, then Eq. (21) is the best we can do.

As discussed in Sec. II, it is in fact conventional to invert the scale on which information is measured by subtracting the information needed to communicate s, given in this case by Eq. (21), from the amount needed when the receiver does not know r, given by Eq. (13). This gives us a mutual-information-style measure, thus

$$H_2 - H_4 = \frac{1}{n}\left[\log \frac{n!}{\prod_s b_s!} - \sum_r \log \frac{a_r!}{\prod_s c_{rs}!} - \log \Omega(a,b)\right]. \tag{22}$$

This is the quantity we call the *reduced mutual information* or RMI. The reduced mutual information is equal to the amount of information that Alice saves by making use of the fact that Bob already knows r. Note that the first term in Eq. (21), which measures the amount of information needed to transmit the values of the b_s, has canceled out, so in fact we need not assume (as we previously did) that this term is negligible.

Equation (22) can be rearranged to give a manifestly symmetric expression for the reduced mutual information, thus

$$\mathrm{RMI} = \frac{1}{n}\left[\log \frac{n!\prod_{rs} c_{rs}!}{\prod_r a_r! \prod_s b_s!} - \log \Omega(a,b)\right], \tag{23}$$

which is the form we use henceforth. Applying Stirling's approximation to the first term, we find that

$$\frac{1}{n}\log\frac{n!\prod_{rs}c_{rs}!}{\prod_r a_r!\prod_s b_s!} \simeq \sum_{rs} P(r,s)\log\frac{P(r,s)}{P(r)P(s)}, \tag{24}$$

which is the standard mutual information $I(r;s)$ of Eq. (7). So, as claimed at the start of Sec. III, the reduced mutual information is, to a good approximation, equal to the standard mutual information minus the information needed to specify the contingency table: $I(r;s) - (1/n)\log\Omega(a,b)$. If one computes only the first term in this expression, as one normally does when comparing labelings, then one is neglecting the contingency table. This is the root cause of the problem described in Sec. II A. For correct answers in all cases, one should compare labelings using the full expression, Eq. (23), with both terms.

The reduced mutual information will normally be positive but it is possible for it to take negative values. Referring to Eq. (22), we see that this happens whenever $H_2 < H_4$—in other words when it is more efficient for Alice to just send s directly to Bob than to exploit correlations between r and s. This tells us that in fact there *are* no correlations, a least at the level where they give us pertinent information about the labeling, a result that could be useful in some situations. The standard mutual information, by contrast, is never negative and hence has no natural threshold value that indicates absence of correlation between two variables. The adjusted mutual information discussed in the Introduction [12] does have such a threshold, created by subtracting a small offset value from the mutual information, but, as we argue in Sec. V, Eq. (23) is more correct in most situations.

It is also possible to create a normalized version of Eq. (23), akin to the normalized mutual information of Danon *et al.* [3], Eq. (8), by dividing by the average of the values obtained when both labelings are equal to r and when both are equal to s. This gives a normalized information of

$$\text{Normalized RMI} = \frac{2\log\dfrac{n!\prod_{rs}c_{rs}!}{\prod_r a_r!\prod_s b_s!} - 2\log\Omega(a,b)}{\log\dfrac{n!}{\prod_r a_r!} + \log\dfrac{n!}{\prod_s b_s!} - \log\Omega(a,a) - \log\Omega(b,b)}, \tag{25}$$

which is now equal to 1 when r and s are identical.

D. Simple examples

To illustrate the importance of the second term in Eq. (23) let us revisit our two simple examples from Sec. II A. In the first example, we considered a labeling r in which all objects belong to the same single group. In this case the contingency table has just one row, and S columns with entries $c_{1s} = b_s$. Hence, all entries in the table are fixed and there is only one possible contingency table with the given row and column sums. Thus, the second term in Eq. (23) is zero. At the same time there is only a single a_r with value $a_1 = n$ and hence all factors in the first term cancel making this term zero as well. Thus, we get a reduced mutual information of zero, which is the correct answer, although the standard mutual information also correctly gives zero in this case, as we showed in Sec. II A—see Eq. (9).

But now take our second example, in which the labeling r consists of n groups of one object each. In Sec. II A we found that in this case the standard mutual information gave the wrong answer—it took on the maximal value of $H(s)$ when common sense dictates that it should again be zero. This is precisely because we neglected the information content of the contingency table. In this case we have all $a_r = 1$ and all $c_{rs} = 0$ or 1, so that $a_r! = c_{rs}! = 1$ for all r, s and the first term in Eq. (23) is $\log(n!/\prod_s b_s!)$, which is indeed approximately equal to $H(s)$ by Stirling's formula. The second term, however, is now large. The contingency table has n rows and S columns with a single 1 in each row and all other elements zero. The number of such contingency tables that have the appropriate column sums b_s is equal to the number of ways of placing the n objects in groups of sizes b_s, which is given by the multinomial coefficient $n!/\prod_s b_s!$. Hence, the two terms in Eq. (23) cancel exactly to give zero once again, which is the correct answer.

The intuitive explanation for this result is that in this latter case the contingency table in fact tells us the complete labeling s. Given that Bob knows the unique label r of each object, then once Alice tells him which value of s each r corresponds to he can reconstruct s without any further information. Thus, the contingency table contains as much information as s itself, so Alice necessarily saves no information by sending the table first: She might as well have just sent s directly.

E. Counting contingency tables

To compute the reduced mutual information of Eq. (23) we need to know the number $\Omega(a,b)$ of possible $R \times S$ contingency tables with row and column sums equal to the given values of a_r and b_s. We are aware of no general closed-form expression for this number, although there are expressions for small (two- and three-row) tables [14]. Tables can be counted numerically, for instance by exhaustively enumerating all $R \times S$ matrices with the desired row sums and counting how many have the desired column sums. The number of matrices with row sums a_r is $\prod_r \binom{a_r+S-1}{S-1} \leqslant \binom{n+S-1}{S-1}^R \leqslant (n+S-1)^{RS}$ and hence the running time for this procedure is polynomial in n, but in practice the degree of the polynomial is so large as to make the method unworkable for all but the smallest of matrices. More sophisticated algorithms have been proposed with somewhat better running times [14,15], but they are still practical only for small cases. For larger values of n, R, and S one can calculate approximate answers by Monte Carlo sampling [14,16], but a more convenient solution for our purposes is to make use of analytic approximations such as those given in Refs. [14,17–20]. We are particularly interested in two limits. The first is the sparse limit, typified by our example above in which each object is placed in a group on its own. In cases like this where most elements of the contingency table are zero and the elements that are nonzero are small, Békéssy *et al.* [14,17] have shown that

$$\log\Omega(a,b) \simeq \log\frac{n!}{\prod_r a_r!\prod_s b_s!} + \frac{2}{n^2}\sum_r\binom{a_r}{2}\sum_s\binom{b_s}{2}. \tag{26}$$

Applying Stirling's formula we then have

$$\frac{1}{n}\log\Omega(a,b)\simeq H(s)-\frac{1}{n}\sum_r\log a_r!+\mathrm{O}(1/n). \tag{27}$$

Thus, in general, the contingency table requires an amount of information on the order of $H(s)$ per object for its specification, making it essential that it be included in the calculation if we are to obtain useful results.

Substituting Eq. (26) into Eq. (23) we get

$$\mathrm{RMI}\simeq\frac{1}{n}\sum_{rs}\log c_{rs}!-\frac{2}{n^3}\sum_r\binom{a_r}{2}\sum_s\binom{b_s}{2}, \tag{28}$$

which can be used to calculate the reduced mutual information in this regime.

This, however, is not the regime in which we are most commonly operating. Usually we are interested in cases where the numbers of groups R and S are substantially smaller than n and the contingency table is relatively dense, with many elements much larger than 1. In this regime we use a different approximation, a symmetrized version of a formula due to Diaconis and Efron [21]:

$$\begin{aligned}\log\Omega(a,b)&\simeq(R-1)(S-1)\log\left(n+\frac{1}{2}RS\right)+\frac{1}{2}(R+\nu-2)\\&\quad\times\sum_s\log y_s+\frac{1}{2}(S+\mu-2)\sum_r\log x_r\\&\quad+\frac{1}{2}\log\frac{\Gamma(\mu R)\Gamma(\nu S)}{[\Gamma(\nu)\Gamma(R)]^S[\Gamma(\mu)\Gamma(S)]^R},\end{aligned} \tag{29}$$

where

$$w=\frac{n}{n+\frac{1}{2}RS}, \tag{30}$$

$$x_r=\frac{1-w}{R}+\frac{wa_r}{n},\qquad y_s=\frac{1-w}{S}+\frac{wb_s}{n}, \tag{31}$$

$$\mu=\frac{R+1}{R\sum_r y_r^2}-\frac{1}{R},\qquad \nu=\frac{S+1}{S\sum_r x_r^2}-\frac{1}{S}. \tag{32}$$

The leading term in Eq. (29) is of order $RS\log n$, and hence

$$\frac{1}{n}\log\Omega(a,b)=\mathrm{O}\left(\frac{RS\log n}{n}\right). \tag{33}$$

If R and S are bounded as n becomes large, then this expression is of the same order of magnitude as terms neglected in the conventional definition of the mutual information, Eq. (19), and therefore can itself be neglected for large enough n. This is why the traditional mutual information works well in some circumstances, but there are also common situations in which these conditions do not apply. First, either or both of the number of groups R and S may increase with n. For instance, in community structure problems in networks it is reasonable to suppose that R and/or S would grow: the number of groups of friends in the world, for example, probably grows with world population. In this case Eq. (29) cannot be neglected for large n. Indeed, if R and S are of order $\sqrt{n}$ or larger, which they often are, then Eq. (29) cannot be neglected under any circumstances.

Second, in practice n is not infinite and in many cases it is not even particularly large: Values of a few dozen to a few hundred are common in practical situations. In this regime, Eq. (29) may well be comparable in magnitude with the standard mutual information and again must be included to get accurate results.

IV. EXAMPLE APPLICATIONS

As a simple example of the use of our measure, consider Fig. 2, which depicts community structure in a well-known test network, the "karate club" network of Zachary [22]. This network, which depicts friendships among a group of university students, is widely agreed to divide into two communities, represented by the shaded areas in the figure. Figures 2(a) and 2(b) show two different possible divisions of the network found using popular community detection methods, the first by a statistical inference technique [23] and the second by the method of maximum modularity [24]. The two divisions are quite different. The first is closely similar to the accepted ground truth, but differs from it for one node in the center of the figure. The second, by contrast, divides the network into four smaller communities, which arguably align quite well with the ground truth, although again one node is clearly wrong. Which of these two divisions is more similar to the ground truth? Most observers would probably say the two-group division on the left. Both get a single node wrong, but the two-group division is closer in overall form to the ground truth.

Traditional mutual information, however, says the reverse. Calculating the mutual information between each division and the ground truth we find 0.788 bits per node for the two-group division and 0.807 for the four-group division, indicating that the four-group division is slightly superior. If, however, one includes the information in the contingency table, then the outcome is reversed. For this small example we have no need of the approximation in Eq. (29) to calculate the number of contingency tables—the tables can be counted exactly by exhaustive enumeration, giving $\Omega(a,b)=16$ in the first case and 428 in the second. Substituting these figures into Eq. (23) we find a reduced mutual information of 0.670 bits per node for the two-group division and 0.550 for the four-group one, making the two-group division now the favored choice.

A more complex example is depicted in Fig. 3, which shows results for a widely studied data set from the machine learning literature representing 178 different wines, each produced in the same region of Italy and each made from one of three different grapes [25]. Chemical and photometric analysis of the wines produces 13 different measurements for each one, for attributes such as alcohol content, flavinoids, and color intensity. One of these attributes we discard based on standard feature selection criteria [26]. The remaining 12 can be used to place each wine at a point in a 12-dimensional data space, and previous studies have shown that it is possible, by clustering these points, to infer at least approximately which wine was made from which grape. Figure 3 shows the results of clustering using what is perhaps the simplest of methods, the k-means algorithm. Figures 3(a) and 3(b) show two such clusterings, one into three groups (which is the "correct" number according to the ground truth) and the other into six. In both panels the calculated clustering is indicated by the colors of the data points while their shapes indicate the ground truth, and most observers would probably say that

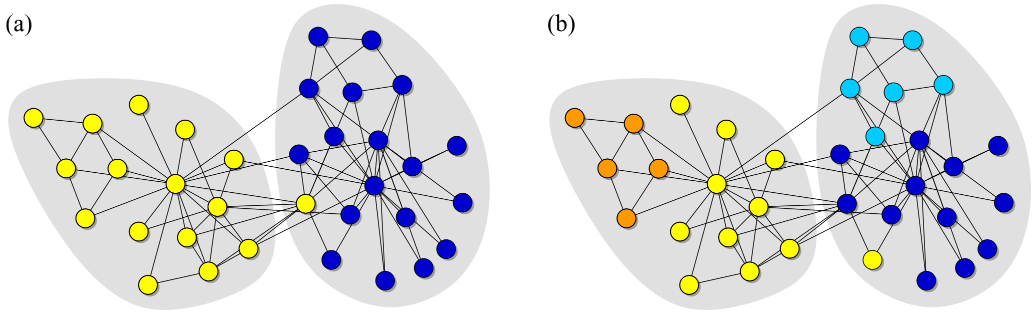

FIG. 2. Two divisions into communities of the “karate club” network of Zachary [21]. (a) Division found by statistical inference, maximizing the profile likelihood for the degree-corrected stochastic block model [22]. (b) Division found by maximizing modularity [23]. The accepted ground-truth division is indicated by the shaded regions. According to the standard mutual information, division (b) has the stronger similarity to the ground truth, but division (a) is more similar according to the reduced mutual information of Eq. (23).

the division into three groups in Fig. 3(a) is closer to the ground truth—remarkably it classifies all but five of the wines correctly.

Once again, however, traditional mutual information reaches the opposite conclusion, saying that the division into six groups is better: the mutual information between the three-group division and the ground truth is 1.380 bits per data point, while for the six-group division we find a slightly better score of 1.409. If one includes the correction due to the information content of the contingency table, however, then the results are reversed. In this case it is impractical to count the number of contingency tables exhaustively: for the six-group division there are estimated to be over 10^{11} tables. Instead, therefore, we make use of the approximate formula in Eq. (29). Combining this formula with Eq. (23) we find a reduced mutual information of 1.266 bits per data point for the three-group division and a lower 1.202 for the six-group one, so that the three-group division is now favored. Figure 3(c) shows results for both the standard and reduced mutual information for k-means clusterings of the data points into 2, 3, 4, 5, and 6 groups. As the figure shows, the standard measure increases with the number of groups while the modified measure peaks at the correct value of three then falls off. Indeed it seems likely that the standard measure will continue to increase with the number of groups as far as we are willing to take the calculation. Such an increase has been observed previously in numerical studies of mutual information, for instance by White and Liu [6].

V. COMPARISON WITH OTHER MEASURES

As discussed in the Introduction, the measure we give for comparing clusterings is not the only one that has been proposed for this task. In this section we review some of the previously proposed measures and their relationship to our reduced mutual information.

In addition to the standard mutual information, whose disadvantages we have already discussed in depth, the measures that are most directly related to ours are the measure proposed by Dom in Ref. [7], the adjusted mutual information proposed

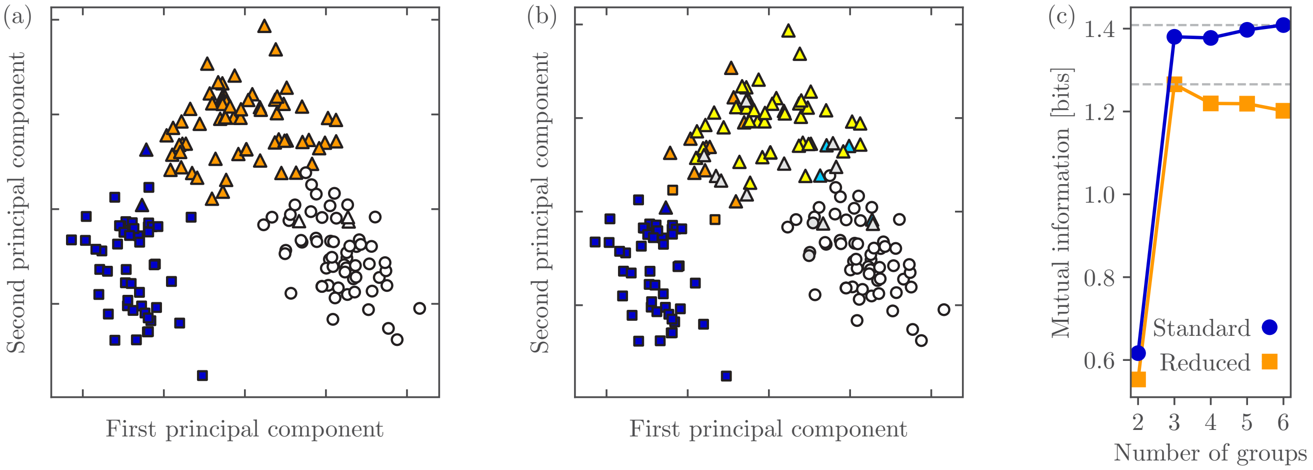

FIG. 3. Clustering analysis of the wine data set described in the text [25], which represents 12 measured attributes of 178 different Italian wines. (a) A three-group clustering of the data found using k-means applied to a normalized version of the data. The data points are projected onto the data set’s two most significant principal components, with colors representing the calculated clustering and shapes representing the ground truth. (b) A six-group clustering of the same data, again computed using k-means. (c) The standard mutual information and the reduced mutual information of Eq. (23) for divisions into 2, 3, 4, 5, and 6 groups using k-means. While the standard measure increases with increasing number of groups, the modified measure peaks at the correct ground-truth value of three.

by Vinh *et al.* [12], and the variation of information proposed by Meilă [9]. Let us consider each of these in turn.

The measure proposed by Dom [7] is equal to our measure H_3, Eq. (18), when the conditional entropy term is approximated using Stirling's formula as in Eq. (20). This measure does compensate to some extent for the shortcomings of the standard conditional entropy but, as discussed Sec. III C, it also has some undesirable features. For instance, it does not take the expected value of $H(s)$ when labeling r places every node in a separate group of size one, and it is moreover normally larger than our measure H_4 and hence necessarily a poorer approximation to the amount of information needed to specify s—both measures by definition give upper bounds on the information, but H_4 gives a lower, and hence better, bound. In that the stated goal of both measures is to calculate the information, it is therefore normally better to use H_4 or, equivalently, the reduced mutual information.

A more subtle case is presented by the adjusted mutual information [12] (and variants such as the relative mutual information of Zhang [27]). Like our measure, the adjusted mutual information subtracts a correction term from the standard mutual information $I(r;s)$. For the adjusted mutual information this term is equal to the average of $I(r;s)$ over all possible choices of the labels s, consistent with the values of the b_s. It is helpful in comparing this measure to our own to re-express this adjustment in terms of contingency tables as follows.

As Vinh *et al.* point out, if all labelings s are equally likely then all contingency tables are not. For a given labeling r, the same table can be generated by many different choices of s. As discussed in Sec. III C, the number of labelings s that correspond to a particular contingency table is $\prod_r a_r!/\prod_{rs} c_{rs}!$, so if all s are equally likely then the probability Q_T of generating a particular table T is equal to this number divided by the total number of labelings s, which is $n!/\prod_s b_s!$. Thus,

$$Q_T = \frac{\prod_r a_r!/\prod_{rs} c_{rs}!}{n!/\prod_s b_s!} = \frac{\prod_r a_r!\prod_s b_s!}{n!\prod_{rs} c_{rs}!}. \tag{34}$$

This probability distribution over contingency tables is known as the multivariate hypergeometric distribution.

Given this distribution, information theory tells us that the amount of information needed to transmit a particular table T is

$$-\log Q_T = \log \frac{n!\prod_{rs} c_{rs}!}{\prod_r a_r!\prod_s b_s!} \simeq nI(r;s), \tag{35}$$

where we have used Eq. (24). Averaging over all tables T, we then find that the average information per object needed to transmit the contingency table is

$$-\frac{1}{n}\sum_T Q_T \log Q_T = \sum_T Q_T I(r;s) = \langle I(r;s)\rangle, \tag{36}$$

where $\langle\ldots\rangle$ denotes the average over the hypergeometric distribution. This is precisely the average mutual information that Vinh *et al.* subtract to create their adjusted mutual information.

In other words, though this is not how it was originally motivated, the adjusted mutual information can be interpreted in a manner analogous to our own measure, as the standard mutual information minus the amount of information needed to specify the contingency table. The crucial difference, however, is that in calculating the correction Vinh *et al.* assume that all *labelings s* are equally likely. This contrasts with our own measure, Eq. (23), which assumes that all *contingency tables* are equally likely.

We argue that the former is not a good assumption under typical real-world conditions. This is because the hypergeometric distribution over contingency tables, Eq. (34), is strongly peaked around tables whose elements are uniform, i.e., those for which r and s are uncorrelated. If r and s are uncorrelated, then r provides no information about s, so the assumptions of the adjusted mutual information are equivalent to assuming that r tells us nothing about s. But this is not a good description of the conditions under which mutual information is used. Under normal use conditions, r will sometimes contain information about s and sometimes not, and we use mutual information precisely to tell us when it does and to what extent. If it were always true that r contains no information about s, as the adjusted mutual information effectively assumes, then there would be no need to use mutual information at all.

In other words, in calculating the average information in the contingency table, adjusted mutual information uses an unrealistic distribution over tables. If we are lucky it is possible that the average calculated from this distribution in Eq. (36) could still be approximately correct, but it could also be in error by a substantial factor and we would have no way of knowing. Our reduced mutual information, by contrast, assumes a uniform distribution over contingency tables, making it unbiased with respect to the form of the table, and, in particular, giving equal weight to uniform and nonuniform tables alike. On a purely practical level, the adjusted mutual information is also quite time-consuming to calculate, whereas our own measure can be calculated in time $O(RS)$ when the contingency table is known or $O(n)$ when it is not, which is the same time complexity as the standard mutual information.

Third, let us consider the variation of information [9], which is a somewhat different animal. It is defined as

$$\text{Variation of information} = H(s|r) + H(r|s). \tag{37}$$

Unlike the mutual information (and similar measures) the variation of information is a *dissimilarity* measure, taking larger values for unlike labelings. The variation of information is satisfactory in that it does not suffer from the failings of the simple mutual information. When labeling r places every object in a group of its own, for instance, it gives an answer that diverges as n becomes large. Arguably this is not exactly the answer we are looking for, but it is definitely nonzero, indicating that the two labelings are very different from one another. Moreover, the variation of information has a number of attractive formal properties. In particular, unlike any of the other measures we have considered (including our own), it is a true metric distance measure: It is symmetric, nonnegative, and satisfies the triangle inequality (meaning that the distance, in terms of variation of information, from r to t is never greater than the sum of the distance from r to s and from s to t). This means one can unambiguously say when one pair of labelings are closer together than another and hence define a topology on the space of all labelings.

The main disadvantage of the variation of information is that it does not have a simple interpretation in terms of information content. It is not, for example, equal to the information needed to transmit the unknown labeling s under any encoding we are aware of. Instead it is equal to the information needed to transmit s given r plus the information needed to transmit r given s, not including the information needed to send the contingency table. This makes it a less natural choice for the type of applications we are considering.

VI. CONCLUSIONS

Mutual information is widely used as a quantitative measure for comparing competing labelings or divisions of a set of objects, such as arise in clustering or community detection problems. In this paper we have argued, however, that the standard mutual information is not always adequate for this purpose, giving incorrect answers—sometimes maximally incorrect—particularly when the labelings have different numbers of groups. We have shown how this shortcoming can be rectified by the inclusion of an additional term in the mutual information equal to minus the logarithm of the number of contingency tables compatible with the sizes of the groups. We have given two example applications of the resulting modified measure, which we call the reduced mutual information, showing how the inclusion of the additional term can make not only quantitative but also qualitative differences to the outcomes of real-world calculations. We have also given a detailed comparison of our measure to other variants of the mutual information and argued that in practical situations it better captures the similarity of pairs of labelings than its competitors. Because of the widespread use of mutual information in many different fields, adoption of the new measure proposed here could change the outcomes and conclusions of many studies, in some cases substantially.

In closing, let us consider whether it would be possible to further improve our measure beyond what is described in this paper. As we have said, all measures of the information needed to describe a labeling are actually upper bounds on the true information. In any individual case it may be possible that a labeling has some special form or structure that allows us to communicate it more succinctly. If one could calculate the true minimum amount of information needed to transmit a labeling in all cases, then by definition this would give a better measure of similarity than any of those discussed in this paper.

There are, however, a number of reasons why this point of view is not a helpful one. First, calculating the true minimum information is normally impossible. It would be equivalent to calculating the so-called Kolmogorov complexity [28], whose computation is provably impossible in general. Second, a measure such as our reduced mutual information, Eq. (23), has the desirable feature of expressing the information as a correction to the conventional mutual information. In effect, it decomposes the information into the standard message that one would send under traditional information theory plus an additional term that has previously been neglected. Thus, it not only gives us a better measure of similarity; it also tells us in a straightforward manner how that measure is related to others we are familiar with.

Even if we focus on measures with this latter property, however, and if we abandon the hope of finding the true minimum information, it is still possible that we could find a better bound. Given that we are fixing the mutual information term, such an improvement would have to come from finding a more efficient scheme for transmitting the contingency table (or some equivalent object). As discussed in Sec. V, we have assumed that all contingency tables are equally likely, which by definition maximizes the information needed to send them. If, in fact, some tables are more likely than others, then we can in principle use that fact to improve our encoding and transmit the table more succinctly. The adjusted mutual information, discussed in Sec. V, offers one way of doing this, but the particular probability distribution over tables that it employs is, we have argued, incorrect. If we could, however, accurately characterize the true distribution of tables in any particular system, then we could in principle find a better encoding for labelings. Unfortunately, this means that any improved information measure would have to make use of domain-specific knowledge about the particular application of interest. In the absence of any such knowledge, the uniform distribution over contingency tables gives the only bound on the information that is correct in all cases and hence we argue that the measure proposed in this paper is the most natural choice for most applications.

ACKNOWLEDGMENTS

This work was funded in part by the U.S. National Science Foundation under Grant No. DMS–1710848 (M.E.J.N.) and by the James S. McDonnell Foundation (J.G.Y.).

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley, New York, 1991).

[2] J. Wilms, M. Troyer, and F. Verstraete, Mutual information in classical spin models, J. Stat. Mech. (2011) P10011.

[3] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas, Comparing community structure identification, J. Stat. Mech. (2005) P09008.

[4] S. Fortunato, Community detection in graphs, Phys. Rep. **486**, 75 (2010).

[5] C. C. Aggarwal and C. K. Reddy (eds.), *Data Clustering: Algorithms and Applications* (CRC Press, Boca Raton, FL, 2014).

[6] A. P. White and W. Z. Liu, Bias in information-based measures in decision tree induction, Mach. Learn. **15**, 321 (1994).

[7] B. E. Dom, An information-theoretic external cluster-validity measure. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, edited by A. Darwiche and N. Friedman (Morgan Kaufmann, San Francisco, CA, 2002), pp. 137–145.

[8] A. J. Gates, I. B. Wood, W. P. Hetrick, and Y.-Y. Ahn, Element-centric clustering comparison unifies overlaps and hierarchy, Sci. Rep. **9**, 8574 (2019).

[9] M. Meilă, Comparing clusterings: An information based distance, J. Multivariate Anal. **98**, 873 (2007).

[10] A. Rosenberg and J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Association of Computational Linguistics, Stroudsburg, PA, 2007), pp. 410–420.

[11] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints, Info. Retrieval **12**, 461 (2009).

[12] N. X. Vinh, J. Epps, and J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, J. Mach. Learn. Res. **11**, 2837 (2010).

[13] A. F. McDaid, D. Greene, and N. Hurley, Normalized mutual information to evaluate overlapping community finding algorithms, arxiv:1110.2515 (2011).

[14] P. Diaconis and A. Gangolli, Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms*, edited by D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele (Springer, New York, 1995), pp. 15–41.

[15] F. Greselin, Counting and enumerating frequency tables with given margins, Statistica & Applicazioni **1**, 87 (2003).

[16] M. Jerrum, L. Valiant, and V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theor. Comput. Sci. **43**, 169 (1986).

[17] A. Békéssy, P. Békéssy, and J. Komlos, Asymptotic enumeration of regular matrices, Stud. Sci. Math. Hungar. **7**, 343 (1972).

[18] E. A. Bender, The asymptotic number of nonnegative integer matrices with given row and column sums, Discrete Math. **10**, 217 (1974).

[19] C. Greenhill and B. D. McKay, Asymptotic enumeration of sparse nonnegative integer matrices with specified row and column sums, Adv. Appl. Math. **41**, 459 (2008).

[20] A. Barvinok, Matrices with prescribed row and column sums, Linear Alg. Appl. **436**, 820 (2012).

[21] P. Diaconis and B. Efron, Testing for independence in a two-way table: New interpretations of the chi-square statistic, Ann. Stat. **13**, 845 (1985).

[22] W. W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. **33**, 452 (1977).

[23] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E **83**, 016107 (2011).

[24] M. E. J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[25] C. L. Blake and C. J. Merz, UCI Machine Learning Repository (1998), http://archive.ics.uci.edu/ml.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[27] P. Zhang, Evaluating accuracy of community detection using the relative normalized mutual information, J. Stat. Mech. (2015) P11006.

[28] C. S. Wallace and D. L. Dowe, Minimum message length and Kolmogorov complexity, Comput. J. **42**, 270 (1999).