Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures

Martijn Gösgens Alexey Tikhonov Liudmila Prokhorenkova

A. Further Related Work

Several attempts to the comparative analysis of cluster similarity indices have been made in the literature, both in machine learning and complex networks communities. In particular, the problem of indices favoring clusterings with smaller or larger clusters has been identified (Albatineh et al., 2006; Vinh et al., 2009; 2010; Lei et al., 2017). The most popular approach to resolving the bias of an index is to subtract its expected value and normalize the resulting quantity to obtain an index that satisfies the maximum agreement property. This approach has let to 'adjusted' indices such as AR (Hubert & Arabie, 1985) and AMI (Vinh et al., 2009). In Albatineh et al. (2006), the family of pair-counting indices $\mathcal{L}$ is introduced for which adjusted forms can be computed easily. This family corresponds to the set of all pair-counting indices that are linear functions of N_{11} for fixed $N_{11}+N_{10}, N_{11}+N_{01}$. In (Romano et al., 2016), a generalization of information-theoretic indices by the Tsallis q-entropy is given and this is shown to correspond to pair-counting indices for $q=2$. Formulas are provided for adjusting these generalized indices for chance.

A disadvantage of this adjustment scheme is that an index can be normalized in many ways, while it is difficult to grasp the differences between these normalizations intuitively. For example, three variants of AMI have been introduced (Vinh et al., 2009), and we show that normalization by the maximum entropies results in an index that fails monotonicity. Romano et al. (2014) go one step further by standardizing mutual information, while Amelio & Pizzuti (2015) multiply NMI with a penalty factor that decreases with the difference in the number of clusters.

In summary, all these works take a popular biased index and 'patch' it to get rid of this bias. This approach has two disadvantages: firstly, these patches often introduce new problems (e.g., FNMI and SMI fail monotonicity), and secondly, the resulting index is usually less interpretable than the original. We have taken a different approach in our work: instead of patching existing indices, we analyze previously introduced indices to see whether they satisfy more properties. Our analysis shows that AR is dominated by Pearson correlation, which was introduced more than 100 years before AR. Therefore, there was no need to construct AR from Rand in the first place.

In Lei et al. (2017), the biases of pair-counting indices are characterized. They define these biases as a preference towards either few or many clusters. They prove that the direction of Rand's bias depends on the Havrda-Charvat entropy of the reference clustering. In the present work, we show that the number of clusters is not an adequate quantity for expressing these biases. We introduce methods to easily analyze the bias of any pair-counting index and simplify the condition for the direction of Rand's bias to $m_A < N/2$.

A paper closely related to the current research (Amigó et al., 2009) formulates several constraints (axioms) for cluster similarity indices. Their *cluster homogeneity* is a weaker analog of our monotonicity w.r.t. perfect splits while their *cluster equivalence* is equivalent to our monotonicity w.r.t. perfect merges. The third *rag bag* constraint is motivated by a subjective claim that "introducing disorder into a disordered cluster is less harmful than introducing disorder into a clean cluster". While this is important for their particular application (text clustering), we found no other work that deemed this constraint necessary; hence, we disregarded this constraint in the current research. The last constraint by Amigó et al. (2009) concerns the balance between making errors in large and small clusters. Though this is an interesting aspect that has not received much attention in our research, this constraint poses a particular balance while we believe that the desired balance may differ per application. Hence, this property seems to be non-binary and we are not aware of a proper formalization of this "level of balance" in a general form. Hence, we do not include this in our list of formal properties. The most principal difference of our work compared to Amigó et al. (2009) is the constant baseline which was not analyzed in their work. We find this property extremely important while it is failed by most of the widely used indices including their BCubed. To conclude, our research gives a more comprehensive list of constraints and focuses on those that are desirable in a wide range of applications. We also cover all similarity indices often used in the literature and give formal proofs for all index-property

combinations.

A property similar to our monotonicity property is also given in Meilă (2007), where the similarity between clusterings A and B is upper-bounded by the similarity between A and $A \otimes B$ (as defined in Section C.4). One can show that this property is implied by our monotonicity but not vice versa, i.e., the variant proposed by Meilă (2007) is weaker. Our analysis of monotonicity generalizes and unifies previous approaches to this problem, see Theorem 2 of the main text, which relates consistent improvements to perfect splits and merges.

While we focus on *external* cluster similarity indices that compare a candidate partition with a reference one, there are also *internal* similarity measures that estimate the quality of partitions with respect to internal structure of data (e.g., Silhouette, Hubert-Gamma, Dunn, and many other indices). Kleinberg (2002) used an axiomatic approach for internal measures and proved an impossibility theorem: there are three simple and natural constraints such that no internal clustering measure can satisfy all of them. More work in this direction can be found in, e.g., Ben-David & Ackerman (2008). In network analysis, internal measures compare a candidate partition with the underlying graph structure. They quantify how well a community structure (given by a partition) fits the graph and are often referred to as goodness or quality measures. The most well-known example is *modularity* (Newman & Girvan, 2004). Axioms that these measures ought to satisfy are given in (Ben-David & Ackerman, 2008; Van Laarhoven & Marchiori, 2014). Note that all pair-counting indices discussed in this paper can also be used for graph-partition similarity, as we discuss in Section B.3.

B. Cluster Similarity Indices

B.1. General Indices

Here we give the definitions of the indices listed in Table 3. We define the contingency variables as $n_{ij} = |A_i \cap B_j|$. We note that all indices discussed in this paper can be expressed as functions of these contingency variables.

The *F-Measure* is defined as the harmonic mean of recall and precision. Recall is defined as

$$r(A, B) = \frac{1}{n} \sum_{i=1}^{k_A} \max_{j \in [k_B]} \{n_{ij}\},$$

and precision is its symmetric counterpart $r(B, A)$.

In (Amigó et al., 2009), recall is redefined as

$$r'(A, B) = \frac{1}{n} \sum_{i=1}^{k_A} \frac{1}{|A_i|} \sum_{j=1}^{k_B} n_{ij}^2,$$

and *BCubed* is defined as the harmonic mean of $r'(A, B)$ and $r'(B, A)$.

The remainder of the indices are information-theoretic and require some additional definitions. Let $p_1, \ldots, p_\ell$ be a discrete distribution (i.e., all values are nonnegative and sum to 1). The Shannon entropy is then defined as

$$H(p_1, \ldots, p_\ell) := -\sum_{i=1}^{\ell} p_i \log(p_i).$$

The entropy of a clustering is defined as the entropy of the cluster-label distribution of a random item, i.e.,

$$H(A) := H(|A_1|/n, \ldots, |A_{k_A}|/n),$$

and similarly for $H(B)$. The joint entropy $H(A, B)$ is then defined as the entropy of the distribution with probabilities $(p_{ij})_{i \in [k_A], j \in [k_B]}$, where $p_{ij} = n_{ij}/n$.

Variation of Information (Meilă, 2007) is defined as

$$\mathrm{VI}(A, B) = 2H(A, B) - H(A) - H(B).$$

Mutual information is defined as

$$M(A, B) = H(A) + H(B) - H(A, B).$$

The mutual information between A and B is upper-bounded by $H(A)$ and $H(B)$, which gives multiple possibilities to normalize the mutual information. In this paper, we discuss two normalizations: normalization by the average of the entropies $\frac{1}{2}(H(A)+H(B))$, and normalization by the maximum of entropies $\max\{H(A),H(B)\}$. We will refer to the corresponding indices as NMI and $\mathrm{NMI}_{\max}$, respectively:

$$\mathrm{NMI}(A,B) = \frac{M(A,B)}{(H(A)+H(B))/2},$$
$$\mathrm{NMI}_{\max}(A,B) = \frac{M(A,B)}{\max\{H(A),H(B)\}}.$$

Fair NMI is a variant of NMI that includes a factor that penalizes large differences in the number of clusters (Amelio & Pizzuti, 2015). It is given by

$$\mathrm{FNMI}(A,B) = e^{-|k_A-k_B|/k_A}\mathrm{NMI}(A,B).$$

In this definition, NMI may be normalized in various ways. We note that a different normalization would not result in more properties being satisfied.

Adjusted Mutual Information addresses for the bias of NMI by subtracting the expected mutual information (Vinh et al., 2009). It is given by

$$\mathrm{AMI}(A,B) = \frac{M(A,B) - \mathbf{E}_{B'\sim\mathcal{C}(S(B))}[M(A,B')]}{\sqrt{H(A)\cdot H(B)} - \mathbf{E}_{B'\sim\mathcal{C}(S(B))}[M(A,B')]}.$$

Here, a normalization by the geometric mean of the entropies is used, while other normalizations are also used (Vinh et al., 2009).

Standardized Mutual Information standardizes the mutual information w.r.t. random permutations of the items (Romano et al., 2014), i.e.,

$$\mathrm{SMI}(A,B) = \frac{M(A,B) - \mathbf{E}_{B'\sim\mathcal{C}(S(B))}(M(A,B'))}{\sigma_{B'\sim\mathcal{C}(S(B))}(M(A,B'))},$$

where σ denotes the standard deviation. Calculating the expected value and standard deviation of the mutual information is nontrivial and requires significantly more computation power than other indices. For this, we refer to the original paper (Romano et al., 2014). Note that this index is symmetric since it does not matter whether we keep A constant while randomly permuting B or keep B constant while randomly permuting A.

B.2. Pair-counting Indices and Their Equivalences

Pair-counting similarity indices are defined in Table 1. Table 2 lists linearly equivalent indices (see Definition 2). Note that our linear equivalence differs from the less restrictive monotonous equivalence given in (Batagelj & Bren, 1995). In the current work, we have to restrict to linear equivalence as the constant baseline property is not invariant to non-linear transformations.

B.3. Defining the Subclass of Pair-counting Indices

From Definition 1 of the main text, it follows that a pair-counting index is a function of two binary vectors $\vec{A},\vec{B}$ of length N. Note that this binary-vector representation has some redundancy: whenever u,v and v,w form intra-cluster pairs, we know that u,w must also be an intra-cluster pair. Hence, not every binary vector of length N represents a clustering. The class of N-dimensional binary vectors is, however, isomorphic to the class of undirected graphs on n vertices. Therefore, pair-counting indices are also able to measure the similarity between graphs. For example, for an undirected graph $G=(V,E)$, one can consider its incidence vector $\vec{G} = (\mathbf{1}\{\{v,w\}\in E\})_{v,w\in V}$. Hence, pair-counting indices can be used to measure the similarity between two graphs or between a graph and a clustering. So, one may see a connection between graph and cluster similarity indices. For example, the Mirkin metric is a pair-counting index that coincides with the Hamming distance between the edge-sets of two graphs (Donnat & Holmes, 2018). Another example is the Jaccard graph distance, which turns out to be more appropriate for comparing sparse graphs (Donnat & Holmes, 2018). Thus, all pair-counting indices and their properties discussed in the current paper can also be applied to graph-graph and graph-partition similarities.

Table 1. A selection of pair-counting indices. Most of these indices are taken from (Lei et al., 2017).

Index (Abbreviation)	**Expression**
Rand (R)	$\frac{N_{11}+N_{00}}{N_{11}+N_{10}+N_{01}+N_{00}}$
Adjusted Rand (AR)	$\frac{N_{11}-\frac{(N_{11}+N_{10})(N_{11}+N_{01})}{N_{11}+N_{10}+N_{01}+N_{00}}}{\frac{(N_{11}+N_{10})+(N_{11}+N_{01})}{2}-\frac{(N_{11}+N_{10})(N_{11}+N_{01})}{N_{11}+N_{10}+N_{01}+N_{00}}}$
Jaccard (J)	$\frac{N_{11}}{N_{11}+N_{10}+N_{01}}$
Jaccard Distance (JD)	$\frac{N_{10}+N_{01}}{N_{11}+N_{10}+N_{01}}$
Wallace1 (W)	$\frac{N_{11}}{N_{11}+N_{10}}$
Wallace2	$\frac{N_{11}}{N_{11}+N_{01}}$
Dice	$\frac{2N_{11}}{2N_{11}+N_{10}+N_{01}}$
Correlation Coefficient (CC)	$\frac{N_{11}N_{00}-N_{10}N_{01}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{00}+N_{10})(N_{00}+N_{01})}}$
Correlation Distance (CD)	$\frac{1}{\pi}\arccos\left(\frac{N_{11}N_{00}-N_{10}N_{01}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{00}+N_{10})(N_{00}+N_{01})}}\right)$
Sokal&Sneath-I ($S\&S_1$)	$\frac{1}{4}\left(\frac{N_{11}}{N_{11}+N_{10}}+\frac{N_{11}}{N_{11}+N_{01}}+\frac{N_{00}}{N_{00}+N_{10}}+\frac{N_{00}}{N_{00}+N_{01}}\right)$
Minkowski	$\sqrt{\frac{N_{10}+N_{01}}{N_{11}+N_{10}}}$
Hubert (H)	$\frac{N_{11}+N_{00}-N_{10}-N_{01}}{N_{11}+N_{10}+N_{01}+N_{00}}$
Fowlkes&Mallow	$\frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}}$
Sokal&Sneath-II	$\frac{\frac{1}{2}N_{11}}{\frac{1}{2}N_{11}+N_{10}+N_{01}}$
Normalized Mirkin[1]	$\frac{N_{10}+N_{01}}{N_{11}+N_{10}+N_{01}+N_{00}}$
Kulczynski	$\frac{1}{2}\left(\frac{N_{11}}{N_{11}+N_{10}}+\frac{N_{11}}{N_{11}+N_{01}}\right)$
McConnaughey	$\frac{N_{11}^2-N_{10}N_{01}}{(N_{11}+N_{10})(N_{11}+N_{01})}$
Yule	$\frac{N_{11}N_{00}-N_{10}N_{01}}{N_{11}N_{10}+N_{01}N_{00}}$
Baulieu-I	$\frac{(N_{11}+N_{10}+N_{01}+N_{00})(N_{11}+N_{00})+(N_{10}-N_{01})^2}{(N_{11}+N_{10}+N_{01}+N_{00})^2}$
Russell&Rao	$\frac{N_{11}}{N_{11}+N_{10}+N_{01}+N_{00}}$
Fager&McGowan	$\frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}}-\frac{1}{2\sqrt{N_{11}+N_{10}}}$
Peirce	$\frac{N_{11}N_{00}-N_{10}N_{01}}{(N_{11}+N_{01})(N_{00}+N_{10})}$
Baulieu-II	$\frac{N_{11}N_{00}-N_{10}N_{01}}{(N_{11}+N_{10}+N_{01}+N_{00})^2}$
Sokal&Sneath-III	$\frac{N_{11}N_{00}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{00}+N_{10})(N_{00}+N_{01})}}$
Gower&Legendre	$\frac{N_{11}+N_{00}}{N_{11}+\frac{1}{2}(N_{10}+N_{01})+N_{00}}$
Rogers&Tanimoto	$\frac{N_{11}+N_{00}}{N_{11}+2(N_{10}+N_{01})+N_{00}}$
Goodman&Kruskal	$\frac{N_{11}N_{00}-N_{10}N_{01}}{N_{11}N_{00}+N_{10}N_{01}}$

In this section, we show that the subclass of pair-counting similarity indices can be uniquely defined by the property of being pair-symmetric.

For two graphs G_1 and G_2 let $M_{G_1G_2}$ denote the $N \times 2$ matrix that is obtained by concatenating their adjacency vectors. Let

[1]Throughout the literature, the Mirkin metric is defined as $2(N_{10}+N_{01})$, but we use this variant as it satisfies the scale-invariance.

Table 2. Equivalent pair-counting indices

Representative Index	Equivalent indices
Rand	Normalized Mirkin Metric, Hubert
Jaccard	Jaccard Distance
Wallace1	Wallace2
Kulczynski	McConnaughey

us write $V_M^{(G)}(M_{G_1G_2})$ for the similarity between two graphs G_1, G_2 according to some graph similarity index $V^{(G)}$. We will now characterize all pair-counting similarity indices as a subclass of the class of similarity indices between undirected graphs.

Definition 1. *We define a graph similarity index $V_M^{(G)}(M_{G_1G_2})$ to be* pair-symmetric *if interchanging two rows of M_{G_1,G_2} leaves the index unchanged.*

We give the following result.

Lemma 1. *The class of pair-symmetric graph similarity indices coincides with the class of pair-counting cluster similarity indices.*

Proof. A matrix is an ordered list of its rows. An unordered list is a multiset. Hence, when we disregard the ordering of the matrix M_{AB}, we get a multiset of the rows. This multiset contains at most four distinct elements with multiplicities corresponding to the four pair-counts. Therefore, each $V_M^{(G)}(M_{AB})$ that is symmetric w.r.t. interchanging rows is equivalently a function of the pair-counts of A and B. □

C. Checking Properties for Indices

In this section, we check all non-trivial properties for all indices. The properties of symmetry, maximal/minimal agreement and asymptotic constant baseline can trivially be tested by simply checking $V(B, A) = V(A, B)$, $V(A, A) = c_{\max}$, $V(0, N_{10}, N_{01}, 0) = c_{\min}$ and $V\left(\overline{N_{11}}, \overline{N_{10}}, \overline{N_{01}}, \overline{N_{00}}\right) = c_{\text{base}}$ respectively. For pair-counting indices, we will frequently use the notation $p_{AB} = N_{11}/N, p_A = (N_{11} + N_{10})/N, p_B = (N_{11} + N_{01})/N$ and write $V^{(p)}(p_{AB}, p_A, p_B)$ instead of $V(N_{11}, N_{10}, N_{01}, N_{00})$.

C.1. Distance

C.1.1. Positive cases

NMI and VI. In (Vinh et al., 2010) it is proven that for max-normalization $1 - \text{NMI}$ is a distance, while in (Meilă, 2007) it is proven that VI is a distance.

Rand. The Mirkin metric $1 - R$ corresponds to a rescaled version of the size of the symmetric difference between the sets of intra-cluster pairs. The symmetric difference is known to be a distance metric.

Jaccard. In (Kosub, 2019), it is proven that the Jaccard distance $1 - J$ is indeed a distance.

Correlation Distance. In Theorem 1 of the main text it is proven that Correlation Distance is indeed a distance.

C.1.2. Negative cases

To prove that an index that satisfies symmetry and maximal agreement is not linearly transformable to a distance metric, we only need to disprove the triangle inequality for one instance of its equivalence class that is nonnegative and equals zero for maximal agreement.

FNMI and Wallace. These indices cannot be transformed to distances as they are not symmetric.

SMI. SMI does not satisfy the maximal agreement property (Romano et al., 2014), so it cannot be transformed to a metric.

FMeasure and BCubed. We will use a simple counter-example, where $|V| = 3, k_A = 1, k_B = 2, k_C = 3$. Let us denote the FMeasure and BCubed by FM, BC respectively. We get

$$1 - \mathrm{FM}(A, C) = 1 - 0.5 > (1 - 0.8) + (1 - 0.8) = (1 - \mathrm{FM}(A, B)) + (1 - \mathrm{FM}(B, C))$$

and

$$1 - \mathrm{BC}(A, C) = 1 - 0.5 > (1 - 0.71) + (1 - 0.8) \approx (1 - \mathrm{BC}(A, B)) + (1 - \mathrm{BC}(B, C)),$$

so that both indices violate the triangle inequality in this case.

Adjusted Rand, Dice, Correlation Coefficient, Sokal&Sneath and AMI. For these indices, we use the following counter-example: Let $A = \{\{0, 1\}, \{2\}, \{3\}\}, B = \{\{0, 1\}, \{2, 3\}\}, C = \{\{0\}, \{1\}, \{2, 3\}\}$. Then $p_{AB} = p_{BC} = 1/6$ and $p_{AC} = 0$ while $p_A = p_C = 1/6$ and $p_B = 1/3$. By substituting these variables, one can see that

$$1 - V^{(p)}(p_{AC}, p_A, p_C) > (1 - V^{(p)}(p_{AB}, p_A, p_B)) + (1 - V^{(p)}(p_{BC}, p_B, p_C)),$$

holds for each of these indices, contradicting the triangle inequality. The same A, B and C also form a counter-example for AMI.

C.2. Linear Complexity

We will frequently make use of the following lemma:

Lemma 2. *The nonzero values of n_{ij} can be computed in $O(n)$.*

Proof. We will store these nonzero values in a hash-table that maps the pairs (i, j) to their value n_{ij}. These values are obtained by iterating through all n elements and incrementing the corresponding value of n_{ij}. For hash-tables, searches and insertions are known to have amortized complexity complexity $O(1)$, meaning that any sequence of n such actions has worst-case running time of $O(n)$, from which the result follows. □

C.2.1. Positive cases

NMI, FNMI and VI. Given the positive values of n_{ij}, it is clear that the joint and marginal entropy values can be computed in $O(n)$. From these values, the indices can be computed in constant time, leading to a worst-case running time of $O(n)$.

FMeasure and BCubed. Note that in the expressions of recall and precision as defined by these indices, only the positive values of n_{ij} contribute. Furthermore, all of the variables a_i, b_j and n_{ij} appear at most once, so that these can indeed be computed in $O(n)$.

Pair-counting indices. Note that $N_{11} = \sum_{n_{ij}>1} \binom{n_{ij}}{2}$ can obviously be computed in $O(n)$. Similarly, $m_A = \sum_{i=1}^{k_A} \binom{a_i}{2}$ and m_B can be computed in $O(k_A), O(k_B)$ respectively. The other pair-counts are then obtained by $N_{10} = m_A - N_{11}$, $N_{01} = m_B - N_{11}$ and $N_{00} = N - m_A - m_B + N_{11}$.

C.2.2. Negative cases: AMI and SMI.

Both of these require the computation of the expected mutual information. It has been known (Romano et al., 2016) that this has a worst-case running time of $O(n \cdot \max\{k_A, k_B\})$ while $\max\{k_A, k_A\}$ can be $O(n)$.

C.3. Strong Monotonicity

C.3.1. Positive cases

Correlation Coefficient. This index has the property that inverting one of the binary vectors results in the index flipping sign. Furthermore, the index is symmetric. Therefore, we only need to prove that this index is increasing in N_{11}. We take

the derivative and omit the constant factor $((N_{00}+N_{10})(N_{00}+N_{01}))^{-\frac{1}{2}}$:

$$\frac{N_{00}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}} - \frac{(N_{11}N_{00}-N_{10}N_{01})\cdot\frac{1}{2}(2N_{11}+N_{10}+N_{01})}{[(N_{11}+N_{10})(N_{11}+N_{01})]^{1.5}}$$
$$=\frac{\frac{1}{2}N_{11}N_{00}(N_{10}+N_{01})+N_{00}N_{10}N_{01}}{[(N_{11}+N_{10})(N_{11}+N_{01})]^{1.5}} + \frac{\frac{1}{2}N_{10}N_{01}(2N_{11}+N_{10}+N_{01})}{[(N_{11}+N_{10})(N_{11}+N_{01})]^{1.5}} > 0.$$

Correlation Distance. The correlation distance satisfies strong monotonicity as it is a monotone transformation of the correlation coefficient, which meets the property.

Sokal&Sneath. All four fractions are nondecreasing in N_{11}, N_{00} and nonincreasing in N_{10}, N_{01} while for each of the variables there is one fraction that satisfies the monotonicity strictly so that the index is strongly monotonous.

Rand Index. For the Rand index, it can be easily seen from the form of the index that it is increasing in N_{11}, N_{00} and decreasing in N_{10}, N_{01} so that it meets the property.

C.3.2. NEGATIVE CASES

Jaccard, Wallace, Dice. All these three indices are constant w.r.t. N_{00}. Therefore, these indices do not satisfy strong monotonicity.

Adjusted Rand. It holds that

$$AR(1,2,1,0) < AR(1,3,1,0),$$

so that the index does not meet the strong monotonicity property.

C.4. Monotonicity

C.4.1. POSITIVE CASES

Rand, Correlation Coefficient, Sokal&Sneath, Correlation Distance. Strong monotonicity implies monotonicity. Therefore, these pair-counting indices satisfy the monotonicity property.

Jaccard and Dice. It can be easily seen that these indices are increasing in N_{11} while decreasing in N_{10}, N_{01}. For N_{00}, we note that whenever N_{00} gets increased, either N_{10} or N_{01} must decrease, resulting in an increase of the index. Therefore, these indices satisfy monotonicity.

Adjusted Rand. Note that for $b, b+d > 0$, it holds that

$$\frac{a+c}{b+d} > \frac{a}{b} \Leftrightarrow c > \frac{ad}{b}. \tag{1}$$

We will let a, b denote the numerator and denomenator of Adjusted Rand while c, d will denote their change when incrementing N_{11} or N_{00} while decrementing N_{10} or N_{01}. For Adjusted Rand, we have

$$a = N_{11} - \frac{1}{N}(N_{11}+N_{10})(N_{11}+N_{01}), \quad b = a + \frac{1}{2}(N_{10}+N_{01}).$$

Because of this, when we increment either N_{11} or N_{00} while decrementing either N_{10} or N_{01}, we get $d = c - \frac{1}{2}$. Hence, we need to prove $c > a(c-\frac{1}{2})/b$, or, equivalently

$$c > -\frac{a}{2(b-a)} = \frac{\frac{1}{N}(N_{11}+N_{10})(N_{11}+N_{01}) - N_{11}}{N_{10}+N_{01}}.$$

For simplicity we rewrite this to

$$c + \frac{p_{AB} - p_A p_B}{p_A + p_B - 2p_{AB}} > 0,$$

where $p_{AB} = \frac{N_{11}}{N}$, $p_A = \frac{1}{N}(N_{11} + N_{10})$ and $p_B = \frac{1}{N}(N_{11} + N_{01})$. If we increment N_{00} while decrementing either N_{10} or N_{01}, then $c \in \{p_A, p_B\}$. The symmetry of AR allows us to w.l.o.g. assume that $c = p_A$. We write

$$p_A + \frac{p_{AB} - p_A p_B}{p_A + p_B - 2p_{AB}} = \frac{p_A^2 + (1 - 2p_A)p_{AB}}{p_A + p_B - 2p_{AB}}.$$

When $p_A \leq \frac{1}{2}$, then this is clearly positive. For the case $p_A > \frac{1}{2}$, we bound $p_{AB} \leq p_A$ and bound the numerator by

$$p_A^2 + (1 - 2p_A)p_A = (1 - p_A)p_A > 0.$$

This proves the monotonicity for increasing N_{00}. When incrementing N_{11} while decrementing either N_{10} or N_{01}, we get $c \in \{1 - p_A, 1 - p_B\}$. Again, we assume w.l.o.g. that $c = 1 - p_A$ and write

$$1 - p_A + \frac{p_{AB} - p_A p_B}{p_A + p_B - 2p_{AB}} = \frac{p_A(1 - p_A) + (1 - 2p_A)(p_B - p_{AB})}{p_A + p_B - 2p_{AB}}.$$

This is clearly positive whenever $p_A \leq \frac{1}{2}$. When $p_A > \frac{1}{2}$, we bound $p_{AB} \geq p_A + p_B - 1$ and rewrite the numerator as

$$p_A(1 - p_A) + (1 - 2p_A)(p_A - 1) = (1 - p_A)(3p_A - 1) > 0.$$

This proves monotonicity for increasing N_{11}. Hence, the monotonicity property is met.

NMI and VI. Let B' be obtained by a perfect split of a cluster B_1 into B'_1, B'_2. Note that this increases the entropy of the candidate while keeping the joint entropy constant. Let us denote this increase in the candidate entropy by the conditional entropy $H(B'|B) = H(B') - H(B) > 0$. Now, for NMI, the numerator increases by $H(B'|B)$ while the denominator increases by at most $H(B'|B)$ (dependent on $H(A)$ and the specific normalization that is used). Therefore, NMI increases. Similarly, VI decreases by $H(B'|B)$. Concluding, both NMI and VI are monotonous w.r.t. perfect splits. Now let B'' be obtained by a perfect merge of B_1, B_2 into B''_1. This results in a difference of the entropy of the candidate $H(B'') - H(B) = -H(B|B'') < 0$. The joint entropy decreases by the same amount, so that the mutual information remains unchanged. Therefore, the numerator of NMI remains unchanged while the denominator may or may not change, depending on the normalization. For min- or max-normalization, it may remain unchanged while for any other average it increases. Hence, NMI does not satisfy monotonicity w.r.t. perfect merges for min- and max-normalization but does satisfy this for average-normalization. For VI, the distance will decrease by $H(B|B'')$ so that it indeed satisfies monotonicity w.r.t. perfect merges.

AMI. Let B' be obtained by splitting a cluster B_1 into B'_1, B'_2. This split increases the mutual information by $H(B'|B) - H(A \otimes B'|A \otimes B)$. Recall the definition of the meet $A \otimes B$ from C.4 and note that the joint entropy equals $H(A \otimes B)$. For a perfect split we have $H(A \otimes B'|A \otimes B) = 0$. The expected mutual information changes with

$$\mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B') - M(A', B)] = H(B'|B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[H(A' \otimes B') - H(A' \otimes B)],$$

where we choose to randomize A instead of B' and B for simplicity. Note that for all A',

$$H(A' \otimes B) - H(A' \otimes B') = H(A' \otimes B'|A' \otimes B) \geq 0,$$

with equality if and only if the split is a perfect split w.r.t. A'. Unless A consists exclusively of singleton clusters, there is a positive probability that this split is not perfect, so that the expected value is positive. Furthermore, for the normalization term, we have $\sqrt{H(A)H(B')} < \sqrt{H(A)H(B)} + H(B'|B)$. Combining this, we get

$$\begin{aligned}
&\mathrm{AMI}(A, B') \\
=&\frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)] + \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[H(A' \otimes B'|A' \otimes B)]}{\sqrt{H(A)H(B')} - H(B'|B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)] + \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[H(A' \otimes B'|A' \otimes B)]} \\
>&\frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)] + \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[H(A' \otimes B'|A' \otimes B)]}{\sqrt{H(A)H(B)} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)] + \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[H(A' \otimes B'|A' \otimes B)]} \\
>&\frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]}{\sqrt{H(A)H(B)} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]} = \mathrm{AMI}(A, B).
\end{aligned}$$

This proves that AMI satisfies monotonicity w.r.t. perfect splits.

Now let B'' be obtained by a perfect merge of B_1, B_2 into B_1''. Again, we have $H(B'') - H(B) = -H(B|B'' < 0)$ and $M(A, B'') = M(A, B)$. Let $A' \sim \mathcal{C}(S(A))$ (again, randomizing A instead of B and B'' for simplicity), then $H(A' \otimes B'') \geq H(A' \otimes B) - H(B|B'')$ with equality if and only if B'' is a perfect merge w.r.t. A' which happens with probability strictly less than 1 (unless A consists of a single cluster). Therefore, as long as $k_A > 1$, the expected mutual information decreases. For the normalization, we have $\sqrt{H(A)H(B'')} < \sqrt{H(A)H(B)}$. Hence,

$$\begin{aligned}
\mathrm{AMI}(A, B'') &= \frac{M(A, B'') - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B'')]}{\sqrt{H(A)H(B'')} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B'')]} \\
&= \frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B'')]}{\sqrt{H(A)H(B'')} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B'')]} \\
&> \frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]}{\sqrt{H(A)H(B'')} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]} \\
&> \frac{M(A, B) - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]}{\sqrt{H(A)H(B)} - \mathbf{E}_{A' \sim \mathcal{C}(S(A))}[M(A', B)]} \\
&= \mathrm{AMI}(A, B).
\end{aligned}$$

BCubed. Note that a perfect merge increases BCubed recall while leaving BCubed precision unchanged and that a perfect split increases precision while leaving recall unchanged. Hence, the harmonic mean increases.

C.4.2. NEGATIVE CASES

FMeasure. We give a numerical counter-example: consider $A = \{\{0, \ldots, 6\}\}, B = \{\{0, 1, 2, 3\}, \{4, 5\}, \{6\}\}$ and merge the last two clusters to obtain $B' = \{\{0, 1, 2, 3\}, \{4, 5, 6\}\}$. Then, the FMeasure remains unchanged and equal to 0.73, violating monotonicity w.r.t. perfect merges.

FNMI We will give the following numerical counter-example: Consider $A = \{\{0, 1\}, \{2\}, \{3\}\}, B = \{\{0\}, \{1\}, \{2, 3\}\}$ and merge the first two clusters to obtain $B' = \{\{0, 1\}, \{2, 3\}\}$. This results in

$$\mathrm{FNMI}(A, B) \approx 0.67 > 0.57 \approx \mathrm{FNMI}(A, B').$$

This non-monotonicity is caused by the penalty factor that equals 1 for the pair A, B and equals $\exp(-1/3) \approx 0.72$ for A, B'.

SMI. For this numerical counter-example we rely on the Matlab-implementation of the index by its original authors (Romano et al., 2014). Let $A = \{\{0, \ldots, 4\}, \{5\}\}, B = \{\{0, 1\}, \{2, 3\}, \{4\}, \{5\}\}$ and consider merging the two clusters resulting in $B' = \{\{0, 1, 2, 3\}, \{4\}, \{5\}\}$. The index remains unchanged and equals 2 before and after the merge.

Wallace. Let $k_A = 1$ and let $k_B > 1$. Then any merge of B is a perfect merge, but no increase occurs since $W_1(A, B) = 1$.

C.5. Constant Baseline

C.5.1. POSITIVE CASES

AMI and SMI. Both of these indices satisfy the constant baseline by construction since the expected mutual information is subtracted from the actual mutual information in the numerator.

Adjusted Rand, Correlation Coefficient and Sokal&Sneath. These indices all satisfy ACB while being linear in p_{AB}-linear for fixed p_A, p_B. Thus, by linearity of expectation, the expected value equals the asymptotic constant.

C.5.2. NEGATIVE CASES

For all the following indices, we will analyse the counter-example given by $k_A = k_B = n - 1$. For each index, we will compute the expected value and show that it is not constant. All of these indices satisfy the maximal agreement property

and maximal agreement is achieved with probability $1/N$ (the probability that the single intra-pair of A coincides with the single intra-pair of B). Furthermore, each case where the intra-pairs do not coincide will result in the same contingency variables and hence the same value of the index. We will refer to this value as $c_n(V)$. Therefore, the expected value will only have to be taken over two values and will be given by

$$\mathbf{E}[V(A,B)] = \frac{1}{N}c_{\max} + \frac{N-1}{N}c_n(V).$$

For each of these indices we will conclude that this is a non-constant function of n so that the index does not satisfy the constant baseline property.

Jaccard and Dice. For both these indices we have $c_{\max} = 1$ and $c_n(V) = 0$ (as $N_{11} = 0$ whenever the intra-pairs do not coincide). Hence, $\mathbf{E}[V(A,B)] = \frac{1}{N}$, which is not constant.

Rand and Wallace. As both functions are linear in N_{11} for fixed $m_A = N_{11} + N_{10}, m_B = N_{11} + N_{01}$, we can compute the expected value by simply substituting $N_{11} = m_A m_B/N$. This will result in expected values $1 - 2/N + 2/N^2$ and $1/N$ for Rand and Wallace respectively, which are both non-constant.

Correlation distance. Here $c_{\max} = 0$ and

$$c_n(CD) = \frac{1}{\pi}\arccos\left(\frac{0 - 1/N^2}{(N-1)/N^2}\right),$$

so that the expected value will be given by

$$\mathbf{E}[CD(A,B)] = \frac{N-1}{N\pi}\arccos\left(-\frac{1}{N-1}\right).$$

This is non-constant (it evaluates to $0.44, 0.47$ for $n = 3, 4$ respectively). Note that this expected value converges to $\frac{1}{2}$ for $n \to \infty$, which is indeed the asymptotic baseline of the index.

FNMI and NMI. Note that in this case $k_A = k_B$ so that the penalty term of FNMI will equal 1 and FNMI will coincide with NMI. Again $c_{\max} = 1$. For the case where the intra-pairs do not coincide, the joint entropy will equal $H(A,B) = \ln(n)$ while each of the marginal entropies will equal

$$H(A) = H(B) = \frac{n-2}{n}\ln(n) + \frac{2}{n}\ln(n/2) = \ln(n) - \frac{2}{n}\ln(2).$$

This results in

$$c_n(\text{NMI}) = \frac{2H(A) - H(A,B)}{H(A)} = 1 - \frac{2\ln(n)}{n\ln(n) - 2\ln(2)},$$

and the expected value will be given by the non-constant

$$\mathbf{E}[\text{NMI}(A,B)] = 1 - \frac{N-1}{N}\frac{2\ln(n)}{n\ln(n) - 2\ln(2)}.$$

Note that as $H(A) = H(B)$, all normalizations of MI will be equal so that this counter-example proves that none of the variants of (F)NMI satisfy the constant baseline property.

Variation of Information. In this case $c_{\max} = 0$. We will use the entropies from the NMI-computations to conclude that

$$\mathbf{E}[\text{VI}(A,B)] = \frac{N-1}{N}(2H(A,B) - H(A) - H(B)) = \frac{N-1}{N}\frac{4}{n}\ln(2),$$

which is again non-constant.

F-measure. Here $c_{\max} = 1$. In the case where the intra-pairs do not coincide, all contingency variables will be either one or zero so that both recall and precision will equal $1 - 1/n$ so that $c_n(\text{FM}) = 1 - 1/n$. This results in the following non-constant expected value

$$\mathbf{E}[\text{FM}(A, B)] = 1 - \frac{N-1}{N}\frac{1}{n}.$$

Note that because recall equals precision in both cases, this counter-example also works for other averages than the harmonic average.

BCubed. Again $c_{\max} = 1$. In the other case, the recall and precision will again be equal. Because for BCubed, the contribution of cluster i is given by $\frac{1}{n}\max\{n_{ij}^2\}/|A_i|$, the contributions of the one- and two-clusters will be given by $\frac{1}{n}$, $\frac{1}{2n}$ respectively. Hence, $c_n(\text{BC}) = \frac{n-2}{n} + \frac{1}{2n} = 1 - \frac{3}{2n}$ and we get the non-constant

$$\mathbf{E}[BC(A, B)] = 1 - \frac{N-1}{N} \cdot \frac{3}{2n}.$$

We note that again, this counter-example can be extended to non-harmonic averages of the BCubed recall and precision.

D. Further Analysis of Constant Baseline Property

D.1. Analysis of Exact Constant Baseline Property

In this section we will prove equivalence between Definition 8 of the main text and another formulation. Let $S(B)$ denote the specification of the cluster sizes of the clustering B, i.e., $S(B) := [|B_1|, \ldots, |B_{k_B}|]$, where $[\ldots]$ denotes a multiset. For a cluster sizes specification s, let $\mathcal{C}(s)$ be the uniform distribution over clusterings B with $S(B) = s$. We prove the following result:

Lemma 3. *An index V has a constant baseline if and only if there exists a constant c_{base} so that, for any clustering A with $1 < k_A < n$ and cluster sizes specification s, it holds that $\mathbf{E}_{B\sim\mathcal{C}(s)}[V(A, B)] = c_{base}$.*

Proof. One direction follows readily from the fact that $\mathcal{C}(s)$ is an element-symmetric distribution for every s. For the other direction, we write

$$\begin{aligned}
\mathbf{E}_{B\sim\mathcal{B}}[V(A, B)] &= \sum_s \mathbf{P}_{B\sim\mathcal{B}}(S(B) = s)\, \mathbf{E}_{B\sim\mathcal{B}}[V(A, B)|S(B) = s] \\
&= \sum_s \mathbf{P}_{B\sim\mathcal{B}}(S(B) = s)\, \mathbf{E}_{B\sim\mathcal{C}(s)}[V(A, B)] \\
&= \sum_s \mathbf{P}_{B\sim\mathcal{B}}(S(B) = s)\, c_{\text{base}} = c_{\text{base}},
\end{aligned}$$

where the sum ranges over cluster-sizes of n elements. □

Symmetry of constant baseline Note that drawing $B' \sim \mathcal{C}(S(B))$ is equivalent to obtaining B' by randomly permuting the cluster-assignments of B. Note that for the expectation $\mathbf{E}_{B'\sim\mathcal{C}(S(B))}[V(A, B')]$, it does not matter whether we randomly permute the labels of B or A, i.e.

$$\mathbf{E}_{B'\sim\mathcal{C}(S(B))}[V(A, B')] = \mathbf{E}_{A'\sim\mathcal{C}(S(A))}[V(A', B)].$$

This shows that the definition of constant baseline is indeed symmetric.

D.2. Analysis of Asymptotic Constant Baseline Property

Definition 2. *An index V is said to be* scale-invariant*, if it can be expressed as a continuous function of the three variables $p_A := m_A/N, p_B := m_B/N$ and $p_{AB} := N_{11}/N$.*

All indices in Table 4 are scale-invariant. For such indices, we will write $V^{(p)}(p_{AB}, p_A, p_B)$. Note that when $B \sim \mathcal{C}(s)$ for some s, the values p_A, p_B are constants while p_{AB} is a random variable. Therefore, we further write P_{AB} to stress that this is a random variable.

Theorem 1. *Let V be a scale-invariant pair-counting index, and consider a sequence of clusterings $A^{(n)}$ and cluster-size specifications $s^{(n)}$. Let $N_{11}^{(n)}, N_{10}^{(n)}, N_{01}^{(n)}, N_{00}^{(n)}$ be the corresponding pair-counts. Then, for any $\varepsilon > 0$, as $n \to \infty$,*

$$\mathbf{P}\left(\left|V\left(N_{11}^{(n)}, N_{10}^{(n)}, N_{01}^{(n)}, N_{00}^{(n)}\right) - V\left(\overline{N_{11}^{(n)}}, \overline{N_{10}^{(n)}}, \overline{N_{01}^{(n)}}, \overline{N_{00}^{(n)}}\right)\right| > \varepsilon\right) \to 0.$$

Proof. We prove the equivalent statement

$$V^{(p)}\left(P_{AB}^{(n)}, p_A^{(n)}, p_B^{(n)}\right) - V^{(p)}\left(p_A^{(n)} p_B^{(n)}, p_A^{(n)}, p_B^{(n)}\right) \xrightarrow{P} 0\,.$$

We first prove that $P_{AB}^{(n)} - p_A^{(n)} p_B^{(n)} \xrightarrow{P} 0$ so that the above follows from the continuous mapping theorem. Chebychev's inequality gives

$$\mathbf{P}\left(\left|P_{AB}^{(n)} - p_A^{(n)} p_B^{(n)}\right| > \varepsilon\right) \leq \frac{1}{\binom{n}{2}^2 \varepsilon^2} \mathrm{Var}\left(N_{11}^{(n)}\right) \to 0.$$

The last step follows from the fact that $\mathrm{Var}(N_{11}) = o(n^4)$, as we will prove in the remainder of this section. Even though in the definition, A is fixed while B is randomly permuted, it is convenient to equivalently consider both clusterings are randomly permuted for this proof.

We will show that $\mathrm{Var}(N_{11}) = o(n^4)$. To compute the variance, we first inspect the second moment. Let $A(S)$ denote the indicator function of the event that all elements of $S \subset \{1, \ldots, n\}$ are in the same cluster in A. Define $B(S)$ similarly and let $AB(S) = A(S)B(S)$. Let e, e_1, e_2 range over subsets of $\{1, \ldots, n\}$ of size 2. We write

$$\begin{aligned}
N_{11}^2 &= \left(\sum_e AB(e)\right)^2 \\
&= \sum_{e_1, e_2} AB(e_1)AB(e_2) \\
&= \sum_{|e_1 \cap e_2| = 2} AB(e_1)AB(e_2) + \sum_{|e_1 \cap e_2| = 1} AB(e_1)AB(e_2) + \sum_{|e_1 \cap e_2| = 0} AB(e_1)AB(e_2) \\
&= N_{11} + \sum_{|e_1 \cap e_2| = 1} AB(e_1 \cup e_2) + \sum_{e_1 \cap e_2 = \emptyset} AB(e_1)AB(e_2).
\end{aligned}$$

We take the expectation

$$\mathbf{E}[N_{11}^2] = \mathbf{E}[N_{11}] + 6\binom{n}{3}\mathbf{E}[AB(\{v_1, v_2, v_3\})] + \binom{n}{2}\binom{n-2}{2}\mathbf{E}[AB(e_1)AB(e_2)],$$

where $v_1, v_2, v_3 \in V$ distinct and $e_1 \cap e_2 = \emptyset$. The first two terms are obviously $o(n^4)$. We inspect the last term

$$\binom{n}{2}\binom{n-2}{2}\mathbf{E}[AB(e_1)AB(e_2)] = \binom{n}{2}\sum_{i,j}\mathbf{P}(e_1 \subset A_i \cap B_j) \times \binom{n-2}{2}\mathbf{E}[AB(e_2)|e_1 \subset A_i \cap B_j]\,. \tag{2}$$

Now we rewrite $\mathbf{E}[N_{11}]^2$ to

$$\mathbf{E}[N_{11}]^2 = \binom{n}{2}\sum_{i,j}\mathbf{P}(e_1 \subset A_i \cap B_j)\binom{n}{2}\mathbf{E}[AB(e_2)].$$

Note that $\binom{n}{2}\mathbf{E}[AB(e_2)] > \binom{n-2}{2}\mathbf{E}[AB(e_2)]$ so that the difference between (2) and $\mathbf{E}[N_{11}]^2$ can be bounded by

$$\binom{n}{2}\binom{n-2}{2}\sum_{i,j}\mathbf{P}(e_1 \subset A_i \cap B_j) \cdot (\mathbf{E}[AB(e_2)|e_1 \subset A_i \cap B_j] - \mathbf{E}[AB(e_2)]).$$

As $\binom{n}{2}\binom{n-2}{2} = O(n^4)$, what remains to be proven is

$$\sum_{i,j}\mathbf{P}(e_1 \subset A_i \cap B_j) \cdot (\mathbf{E}[AB(e_2)|e_1 \subset A_i \cap B_j] - \mathbf{E}[AB(e_2)]) = o(1).$$

Note that it is sufficient to prove that

$$\mathbf{E}[AB(e_2)|e_1 \subset A_i \cap B_j] - \mathbf{E}[AB(e_2)] = o(1),$$

for all i, j. Note that $\mathbf{E}[AB(e_2)] = m_A m_B / N^2$, while

$$\mathbf{E}[AB(e_2)|e_1 \subset A_i \cap B_j] = \frac{(m_A - (2a_i - 3))(m_B - (2b_j - 3))}{(N - (2n-3))^2}.$$

Hence, the difference will be given by

$$\begin{aligned}
&\frac{(m_A - (2a_i - 3))(m_B - (2b_j - 3))}{(N - (2n-3))^2} - \frac{m_A m_B}{N^2} \\
=&\frac{N^2(m_A - (2a_i - 3))(m_B - (2b_j - 3))}{N^2(N - (2n-3))^2} - \frac{(N - (2n-3))^2 m_A m_B}{N^2(N - (2n-3))^2} \\
=&\frac{N^2((2a_i - 3)(2b_j - 3) - m_A(2b_j - 3) - m_B(2a_i - 3))}{N^2(N - (2n-3))^2} + \frac{m_A m_B (2N(2n-3) - (2n-3)^2)}{N^2(N - (2n-3))^2} \\
=&\frac{((2a_i - 3)(2b_j - 3) - m_A(2b_j - 3) - m_B(2a_i - 3))}{(N - (2n-3))^2} + \frac{m_A m_B}{N^2} \frac{(2N(2n-3) - (2n-3)^2)}{(N - (2n-3))^2} \\
=&\frac{O(n^3)}{(N - (2n-3))^2} + \frac{m_A m_B}{N^2} \frac{O(n^3)}{N^2(N - (2n-3))^2} \\
=&o(1),
\end{aligned}$$

as required.

□

D.3. Statistical Tests for Constant Baseline

In this section, we provide two statistical tests: one test to check whether an index V satisfies the constant baseline property and another to check whether V has a selection bias towards certain cluster sizes.

Checking constant baseline. Given a reference clustering A and a number of cluster sizes specifications $s_1, \ldots, s_k$, we test the null hypothesis that

$$\mathbf{E}_{B \sim \mathcal{C}(s_i)}[V(A, B)]$$

is constant in $i = 1, \ldots, k$. We do so by using one-way Analysis Of Variance (ANOVA). For each cluster sizes specification, we generate r clusterings. Although ANOVA assumes the data to be normally distributed, it is known to be robust for sufficiently large groups (i.e., large r).

Checking selection bias. In (Romano et al., 2014) it is observed that some indices with a constant baseline do have a *selection bias*; when we have a pool of random clusterings of various sizes and select the one that has the highest score w.r.t. a reference clustering, there is a bias of selecting certain cluster sizes. We test this bias in the following way: given a reference clustering A and cluster sizes specifications $s_1, \ldots, s_k$, we repeatedly generate $B_1 \sim \mathcal{C}(s_1), \ldots, B_k \sim \mathcal{C}(s_k)$. The null-hypothesis will be that each of these clusterings B_i has an equal chance of maximizing $V(A, B_i)$. We test this hypothesis by generating r pools and using the Chi-squared test.

We emphasize that these statistical tests cannot prove whether an index satisfies the property or has a bias. Both will return a confidence level p with which the null hypothesis can be rejected. Furthermore, for an index to not have these biases, the null hypothesis should be true for all choices of $A, s_1, \ldots, s_k$, which is impossible to verify statistically.

The statistical tests have been implemented in Python and the code is available at `https://github.com/MartijnGosgens/validation_indices`. We applied the tests to the indices of Tables 3 and 4. We chose $n = 50, 100, 150, \ldots, 1000$ and $r = 500$. For the cluster sizes, we define the *balanced cluster sizes* $BS(n, k)$ to be the cluster-size specification for k clusters of which $n - k * \lfloor n/k \rfloor$ clusters have size $\lceil n/k \rceil$ while the remainder have size $\lfloor n/k \rfloor$. Then we choose $A^{(n)}$ to be a clustering with sizes $BS(n, \lfloor n^{0.5} \rfloor)$ and consider candidates with sizes

$s_1^{(n)} = BS(n, \lfloor n^{0.25} \rfloor), s_2^{(n)} = BS(n, \lfloor n^{0.5} \rfloor), s_3^{(n)} = BS(n, \lfloor n^{0.75} \rfloor)$. For each n, the statistical test returns a p-value. We use Fisher's method to combine these p-values into one single p-value and then reject the constant baseline if $p < 0.05$. The obtained results agree with Tables 3 and 4 except for Correlation Distance, which is so close to having a constant baseline that the tests are unable to detect it.

D.4. Illustrating Significance of Constant Baseline

In this section, we conduct two experiments illustrating the biases of various indices. We perform two experiments that allow us to identify the direction of the bias in different situations. Our reference clustering corresponds to the expert-annotated clustering of the production experiment described in Section 3 of the main text and Appendix F.3, where $n = 924$ items are grouped into $k_A = 431$ clusters (305 of them consist of a single element).

In the first experiment, we randomly cluster the items into k approximately equally sized clusters for various k. Figure 1 shows the averages and 90% confidence bands for each index. It can be seen that some indices (e.g., NMI and Rand) have a clear increasing baseline while others (e.g., Jaccard and VI) have a decreasing baseline. In contrast, all unbiased indices have a constant baseline.

In Section 4.6 we argued that these biases could not be described in terms of the number of clusters alone. Our second experiment illustrates that the bias also heavily depends on the sizes of the clusters. In this case, items are randomly clustered into 32 clusters, 31 of which are "small" clusters of size s while one cluster has size $n - 31 \cdot s$, where s is varied between 1 and 28. In Figure 2, that the biases are clearly visible. This shows that, even when fixing the number of clusters, biased indices may heavily distort an experiment's outcome.

Finally, recall that we have proven that the baseline of CD is only asymptotically constant. Figures 1 and 2 show that for practical purposes its baseline can be considered constant.

E. Additional Results

E.1. Proof of Theorem 2 in the Main Text

Let B' be an A-consistent improvement of B. We define

$$B \otimes B' = \{B_j \cap B'_{j'} | B_j \in B, B'_{j'} \in B', B_j \cap B'_{j'} \neq \emptyset\}$$

and show that $B \otimes B'$ can be obtained from B by a sequence of perfect splits, while B' can be obtained from $B \otimes B'$ by a sequence of perfect merges. Indeed, the assumption that B' does not introduce new disagreeing pairs guarantees that any $B_j \in B$ can be split into $B_j \cap B'_1, \ldots, B_j \cap B'_{k_{B'}}$ without splitting over any intra-cluster pairs of A. Let us prove that B' can be obtained from $B \otimes B'$ by perfect merges. Suppose there are two $B''_1, B''_2 \in B \otimes B'$ such that both are subsets of some $B'_{j'}$. Assume that this merge is not perfect, then there must be $v \in B''_1, w \in B''_2$ such that v, w are in different clusters of A. As v, w are in the same cluster of B', it follows from the definition of $B \otimes B'$ that v, w must be in different clusters of B. Hence, v, w is an inter-cluster pair in both A and B, while it is an intra-cluster pair of B', contradicting the assumption that B' is an A-consistent improvement of B. This concludes the proof.

E.2. Deviation of CD from Constant Baseline

Theorem. *Given ground truth A with a number of clusters $1 < k_A < n$, a cluster-size specification s and a random partition $B \sim \mathcal{C}(s)$, the expected difference between Correlation Distance and its baseline is given by*

$$\mathbf{E}_{B \sim \mathcal{C}(s)}[\mathrm{CD}(A, B)] - \frac{1}{2} = -\frac{1}{\pi} \sum_{k=1}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2} \frac{\mathbf{E}_{B \sim \mathcal{C}(s)}[\mathrm{CC}(A, B)^{2k+1}]}{2k+1}.$$

Proof. We take the Taylor expansion of the arccosine around $\mathrm{CC}(A, B) = 0$ and get

$$\mathrm{CD}(A, B) = \frac{1}{2} - \frac{1}{\pi} \sum_{k=0}^{\infty} \frac{(2k)!}{2^{2k}(k!)^2} \frac{\mathrm{CC}(A, B)^{2k+1}}{2k+1}.$$

We take the expectation of both sides and note that the first moment of CC equals zero, so the starting index is $k = 1$. □

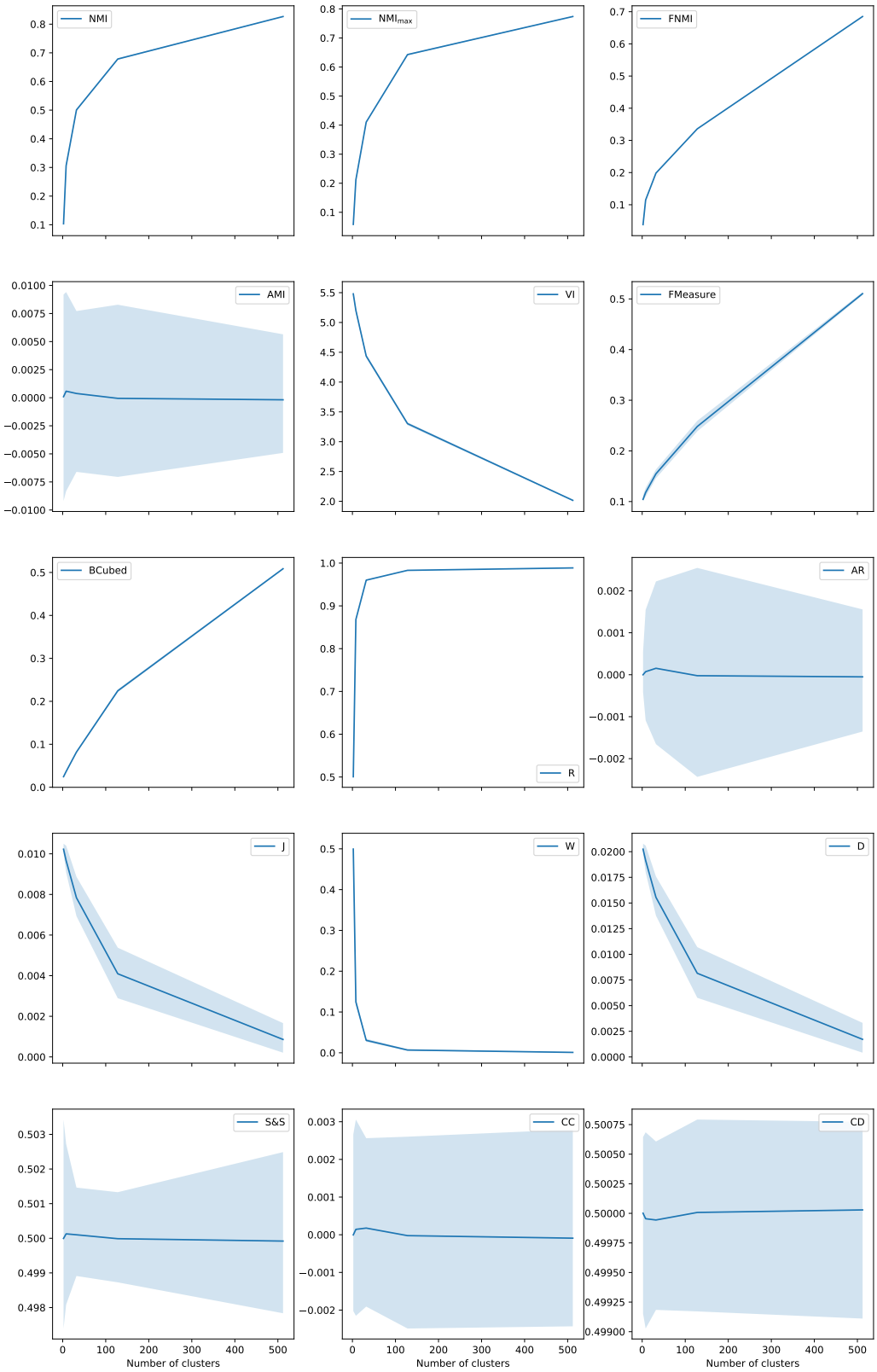


Figure 1. The reference clustering of Appendix F.3 ($n = 924$ and $k_A = 431$) is compared to random clusterings. Each clustering consists of k approximately equally-sized clusters, where k is varied between 2 and 512. For each k, 200 random clusterings are generated. For each index, we plot the average score, along with a 90% confidence band.

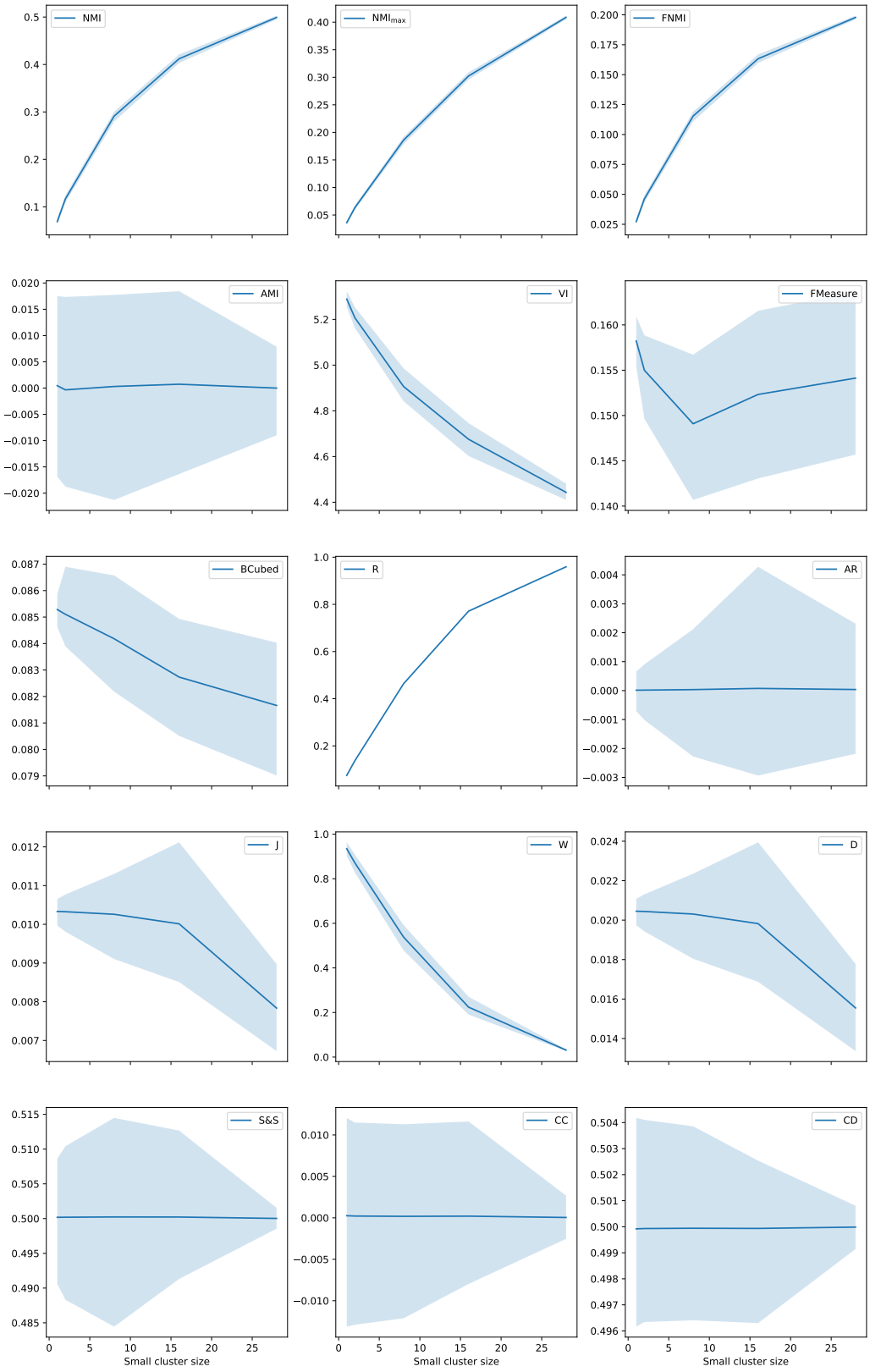


Figure 2. The reference clustering of Appendix F.3 ($n = 924$ and $k_A = 431$) is compared to random clusterings. Each clustering consists of 31 "small" clusters of size s while the last cluster has size $924 - 31 \cdot s$, where s is varied between 1 and 28. For each s, 200 random clusterings are generated. For each index, we plot the average score, along with a 90% confidence band.

For $B \sim \mathcal{C}(s)$ and large n, the value $\mathrm{CC}(A, B)$ will be concentrated around 0. This explains that in practice, the mean tends to be very close to the asymptotic baseline.

E.3. Comparison with Lei et al. (2017)

Lei et al. (2017) describe the following biases for cluster similarity indices: *NCinc* — the average value for a random guess increases monotonically with the Number of Clusters (NC) of the candidate; *NCdec* — the average value for a random guess decreases monotonically with the number of clusters, and *GTbias* — the direction of the monotonicity depends on the specific Ground Truth (GT), i.e., on the reference partition. In particular, the authors conclude from numerical experiments that Jaccard suffers from NCdec and analytically prove that Rand suffers from GTbias, where the direction of the bias depends on the quadratic entropy of the ground truth clustering. Here we argue that these biases are not well defined, suggest replacing them by well-defined analogs, and show how our analysis allows to easily test indices on these biases.

We argue that the quantity of interest should not be the *number of clusters*, but the *number of inter-cluster pairs* of the candidate. Theorem 1 shows that the asymptotic value of the index depends on the number of intra-cluster pairs of both clusterings (or equivalently, the number of inter-cluster pairs). The key insight is that more clusters do not necessarily imply more inter-cluster pairs. For example, let s denote a cluster-sizes specification for 3 clusters each of size $\ell > 2$. Now let s' be the cluster-sizes specification for one cluster of size 2ℓ and ℓ clusters of size 1. Then, any $B \sim \mathcal{C}(s)$ will have 3 clusters and $N - 3\binom{\ell}{2}$ inter-cluster pairs while any $B' \sim \mathcal{C}(s')$ will have $\ell + 1 > 3$ clusters and $N - \binom{2\ell}{2} < N - 3\binom{\ell}{2}$ intra-cluster pairs. For any ground truth A with cluster-sizes s, we have $\mathbf{E}[J(A, B')] > \mathbf{E}[J(A, B)]$ because of a smaller amount of inter-cluster pairs In contrast, Lei et al. (2017) classifies Jaccard as an NCdec index, so that we would expect the inequality to be the other way around, contradicting the definition of NCdec. The PairInc and PairDec biases that are defined in Definition 10 of the main text are sound versions of these NCinc and NCdec biases because they depend on the expected number of agreeing pairs. This allows to analytically determine which bias a given pair-counting index has.

F. Experiment

F.1. Synthetic Experiment

In this experiment, we construct several simple examples to illustrate the inconsistency among the indices. Recall that two indices V_1 and V_2 are inconsistent for a triplet of partitions (A, B_1, B_2) if $V_1(A, B_1) > V_1(A, B_2)$ but $V_2(A, B_1) < V_2(A, B_2)$.

We take all indices from Tables 3 and 4 and construct several triplets of partitions to distinguish them all. Let us note that the pairs Dice vs Jaccard and CC vs CD cannot be inconsistent since they are monotonically transformable to each other. Also, we do not compare with SMI since it is much more computationally complex than all other indices. Thus, we end up with 13 indices and are looking for simple inconsistency examples.

The theoretical minimum of examples needed to find inconsistency for all pairs of 13 indices is 4. We were able to find such four examples, see Figure 3. In this figure, we show four inconsistency triplets. For each triplet, the shapes (triangle, square, etc.) denote the reference partition A. Left and right figures show candidate partitions B_1 and B_2. In the caption, we specify which similarity indices favor this candidate partition over the other one.

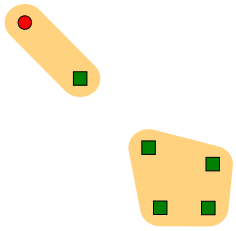
It is easy to see that for each pair of indices, there is a simple example where they disagree. For example, NMI and $\mathrm{NMI}_{\max}$ are inconsistent for triplets 3. Also, we know that Jaccard in general favors larger clusters, while Rand and NMI often prefer smaller ones. Hence, they often disagree in this way (see the triplets 2 and 4).

F.2. Experiments on Real Datasets

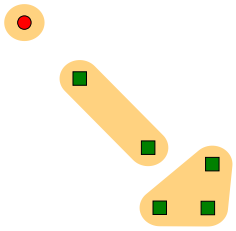
In this section, we test whether the inconsistency affects conclusions obtained in experiments on real data.

For that, we used the following 16 UCI datasets (Dua & Graff, 2017): Arrhythmia, Balance Scale, Ecoli, Heart Statlog, Letter, Segment, Vehicle, WDBC, Wine, Wisc, Cpu, Iono, Iris, Sonar, Thy, Zoo (see GitHub (2020) for datasets and references). The values of the "target class" field were used as a reference partition.

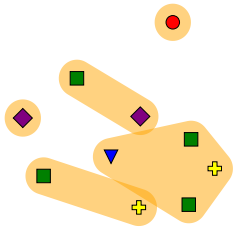
On these datasets, we ran 8 well-known clustering algorithms (Scikit-learn, 2020): KMeans, AffinityPropagation, MeanShift, AgglomerativeClustering, DBSCAN, OPTICS, Birch, GaussianMixture. For AgglomerativeClustering, we used 4 different linkage types ('ward', 'average', 'complete', 'single'). For GaussianMixture, we used 4 different covariance



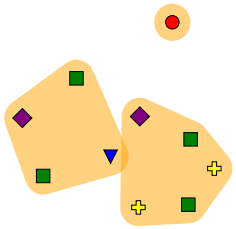
(1a) FNMI, Rand, AdjRand, Jaccard, Dice, Wallace, FMeasure, BCubed



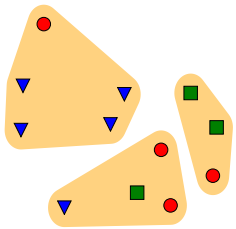
(1b) NMI, NMI_{max}, VI, AMI, S&S, CC, CD



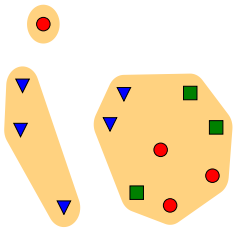
(2a) NMI, NMI_{max}, FNMI, Rand, FMeasure, BCubed



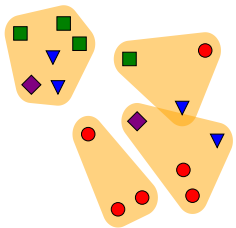
(2b) VI, AMI, AdjRand, Jaccard, Dice, Wallace, S&S, CC, CD



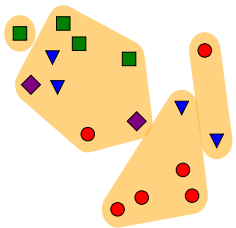
(3a) NMI_{max}, Rand, AdjRand, Jaccard, Dice, S&S, CC, CD, FMeasure



(3b) NMI, VI, FNMI, AMI, Wallace, BCubed



(4a) NMI, NMI_{max}, FNMI, AMI, Rand, AdjRand, CC, CD



(4b) VI, Jaccard, Dice, Wallace, S&S, FMeasure, BCubed

Figure 3. Inconsistency of indices: each row corresponds to a triplet of partitions, shapes denote the reference partitions, the captions indicate which indices favor the corresponding candidate.

Table 3. Inconsistency of indices on real-world clustering datasets, %

	NMI	NMI_{max}	VI	FNMI	AMI	R	AR	J	W	S&S	CC	FMeas	BCub
NMI	–	5.4	40.3	17.3	9.2	13.4	15.7	35.2	68.4	20.1	18.5	31.7	32.0
NMI_{max}		–	41.1	16.5	13.2	12.5	14.1	34.3	68.8	21.1	18.9	30.3	32.4
VI			–	34.7	41.8	45.2	37.6	17.1	28.8	36.0	37.2	18.1	13.6
FNMI				–	23.3	24.0	19.0	29.9	57.0	26.7	23.8	27.5	26.7
AMI					–	21.1	17.3	33.3	61.3	15.1	13.6	35.0	34.4
R						–	15.5	35.6	71.5	21.1	20.7	32.5	35.8
AR							–	23.5	59.4	11.7	8.3	25.3	28.1
J								–	35.9	23.1	23.8	10.7	9.7
W									–	53.5	54.8	40.7	37.4
S&S										–	3.6	26.2	27.8
CC											–	27.0	28.8
FMeas												–	7.7
BCub													–

Table 4. Algorithms preferred by different indices

	NMI	NMI_{max}	VI	FNMI	AMI	R	AR	J	W	S&S1	CC	FMeas	BCub
$k = 2$	2	1	9	4	2	0	4	6	10	3	3	7	7
$k = 2 \cdot \text{ref}$	8	9	1	6	8	10	6	4	0	7	7	3	3

types ('spherical', 'diag', 'tied', 'full'). For methods requiring the number of clusters as a parameter (KMeans, Birch, AgglomerativeClustering, GaussianMixture), we took up to 4 different values (less than 4 if some of them are equal): 2, ref-clusters, max(2,ref-clusters/2), min(items, 2·ref-clusters), where ref-clusters is the number of clusters in the reference partition and items is the number of elements in the dataset. For MeanShift, we used the option $cluster\_all = True$. All other settings were default or taken from examples in the sklearn manual.

For all datasets, we calculated all the partitions for all methods described above. We removed all partitions having only one cluster or which raised any calculation error. Then, we considered all possible triplets A, B_1, B_2, where A is a reference partition and B_1 and B_2 are candidates obtained with two different algorithms. We have 8688 such triplets in total. For each triplet, we check whether the indices are consistent. The inconsistency frequency is shown in Table 3. Note that Wallace is highly asymmetrical and does not satisfy most of the properties, so it is not surprising that it is in general very inconsistent with others. However, the inconsistency rates are significant even for widely used pairs of indices such as, e.g., Variation of Information vs NMI (40.3%, which is an extremely high disagreement). Interestingly, the best agreeing indices are S&S and CC which satisfy most of our properties. This means that conclusions made with these indices are likely to be similar.

Actually, one can show that all indices are inconsistent using only one dataset. This holds for 11 out of 16 datasets: heart-statlog, iris, segment, thy, arrhythmia, vehicle, zoo, ecoli, balance-scale, letter, wine. We do not present statistics for individual datasets since we found the aggregated Table 3 to be more useful.

Finally, to illustrate the biases of indices, we compare two KMeans algorithms with $k = 2$ and $k = 2$·ref-clusters. The comparison is performed on 10 datasets (where both algorithms are successfully completed). The results are shown in Table 4. In this table, biases and inconsistency are clearly seen. We see that NMI and NMI_{max} almost always prefer the larger number of clusters. In contrast, Variation of Information and Rand usually prefer $k = 2$ (Rand prefers $k = 2$ in all cases).

F.3. Production Experiment

To show that the choice of similarity index may have an effect on the final quality of a production algorithm, we conducted an experiment within a major news aggregator system. The system aggregates all news articles to *events* and shows the list of most important events to users. For grouping, a clustering algorithm is used and the quality of this algorithm affects the user experience: merging different clusters may lead to not showing an important event, while too much splitting may cause

[1] The code is available at `https://github.com/MartijnGosgens/validation_indices`.

Table 5. Similarity of candidate partitions to the reference one. In bold are the inconsistently ranked pairs of partitions. For some indices, we flipped the sign of the index, so that larger values correspond to better agreement.

	A_{prod}	A_1	A_2
NMI	0.9326	0.9479	0.9482
NMI$_{\max}$	0.8928	**0.9457**	**0.9298**
FNMI	0.7551	**0.9304**	**0.8722**
AMI	0.6710	**0.7815**	**0.7533**
VI	-0.6996	-0.5662	-0.5503
FMeasure	0.8675	0.8782	0.8852
BCubed	0.8302	0.8431	0.8543
R	0.9827	**0.9915**	**0.9901**
AR	0.4911	0.5999	0.6213
J	0.3320	0.4329	0.4556
W	**0.8323**	**0.6287**	0.8010
D	0.4985	0.6042	0.6260
S&S	0.7926	0.8004	0.8262
CC	0.5376	0.6004	0.6371
CD	-0.3193	-0.2950	-0.2802

the presence of duplicate events.

There is an algorithm $\mathcal{A}_{prod}$ currently used in production and two alternative algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$. To decide which alternative is better for the system, we need to compare them. For that, it is possible to either perform an online experiment or make an offline comparison, which is much cheaper and allows us to compare more alternatives. For the offline comparison, we manually grouped 1K news articles about volleyball, collected during a period of three days, into events. Then, we compared the obtained reference partition with partitions A_{prod}, A_1, and A_2 obtained by $\mathcal{A}_{prod}$, $\mathcal{A}_1$, and $\mathcal{A}_2$, respectively (see Table 5). According to most of the indices, A_2 is closer to the reference partition than A_1, and A_1 is closer than A_{prod}. However, according to some indices, including the well-known NMI$_{\max}$, NMI, and Rand, A_1 better corresponds to the reference partition than A_2. As a result, we see that in practical application *different similarity indices may differently rank the algorithms*.

To further see which algorithm better agrees with user preferences, we launched the following online experiment. During one week we compared $\mathcal{A}_{prod}$ and $\mathcal{A}_1$ and during another — $\mathcal{A}_{prod}$ and $\mathcal{A}_2$ (it is not technically possible to compare $\mathcal{A}_1$ and $\mathcal{A}_2$ simultaneously). In the first experiment, $\mathcal{A}_1$ gave $+0.75\%$ clicks on events shown to users; in the second, $\mathcal{A}_2$ gave $+2.7\%$, which clearly confirms that these algorithms have different effects on user experience and $\mathcal{A}_2$ is a better alternative than $\mathcal{A}_1$. Most similarity indices having nice properties, including CC, CD, and S&S, are in agreement with user preferences. In contrast, AMI ranks $\mathcal{A}_1$ higher than $\mathcal{A}_2$. This can be explained by the fact that AMI gives more weight to small clusters compared to pair-counting indices, which can be undesirable for this particular application, as we discuss in Section 5 of the main text.

References

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

Amelio, A. and 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*, pp. 1584–1585, 2015.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486, 2009.

Batagelj, V. and Bren, M. Comparing resemblance measures. *Journal of classification*, 12(1):73–90, 1995.

Ben-David, S. and Ackerman, M. Measures of clustering quality: A working set of axioms for clustering. *Advances in neural information processing systems*, 21:121–128, 2008.

Donnat, C. and Holmes, S. Tracking network dynamics: A survey of distances and similarity metrics. *arXiv preprint arXiv:1801.07351*, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

GitHub. Clustering datasets. `https://github.com/deric/clustering-benchmark`, 2020.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Kleinberg, J. An impossibility theorem for clustering. *Advances in neural information processing systems*, 15:463–470, 2002.

Kosub, S. A note on the triangle inequality for the jaccard distance. *Pattern Recognition Letters*, 120:36–38, 2019.

Lei, Y., Bezdek, J. C., Romano, S., Vinh, N. X., Chan, J., and Bailey, J. Ground truth bias in external cluster validity indices. *Pattern Recognition*, 65:58–70, 2017.

Meilă, M. Comparing clusteringsan information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Romano, S., Bailey, J., Nguyen, V., and Verspoor, K. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *International Conference on Machine Learning*, pp. 1143–1151, 2014.

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

Scikit-learn. Clustering algorithms. `https://scikit-learn.org/stable/modules/clustering.html`, 2020.

Van Laarhoven, T. and Marchiori, E. Axioms for graph clustering quality functions. *The Journal of Machine Learning Research*, 15(1):193–215, 2014.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009.

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