A General Clustering Agreement Index: for Comparing Disjoint and Overlapping Clusters

Reihaneh Rabbany and **Osmar R. Zaïane**
Department of Computing Science, University of Alberta
Edmonton, AB, Canada
{*rabbanyk, zaiane*} *@ualberta.ca*

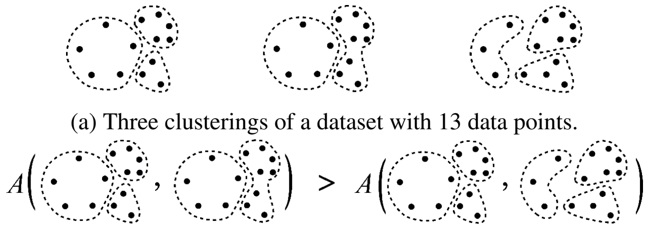
Abstract

A clustering agreement index quantifies the similarity between two given clusterings. It is most commonly used to compare the results obtained from different clustering algorithms against the ground-truth clustering in the benchmark datasets. In this paper, we present a general Clustering Agreement Index (CAI) for comparing disjoint and overlapping clusterings. CAI is generic and introduces a family of clustering agreement indexes. In particular, the two widely used indexes of Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), are special cases of the CAI. Our index, therefore, provides overlapping extensions for both these commonly used indexes, whereas their original formulations are only defined for disjoint cases. Lastly, unlike previous indexes, CAI is flexible and can be adapted to incorporate the structure of the data, which is important when comparing clusters in networks, a.k.a communities.

Introduction

Clustering agreement indexes are well-studied and widely-used in cluster validation (Aggarwal and Reddy 2014), where they measure the agreement between clustering results and the ground-truth clustering, available in the benchmark datasets. More generally, clustering agreement indexes quantify the similarity between two clusterings of the same dataset (see Fig. 1 for a visualized for a toy example), and are utilized to compare and combine multiple clusterings, usually obtained from different objectives, different algorithms, or different parameters of an algorithm; notable cases are consensus clustering (Fred and Jain 2005), ensemble clustering (Strehl and Ghosh 2003), and multi-view clustering (Cui, Fern, and Dy 2007). The *stability and robustness* of clustering algorithms can also be assessed based on the similarity of their results when introducing noise/variations/sampling, and/or changing the order of the data (Fred and Jain 2003; Wagner and Wagner 2007).

More recently, clustering agreement indexes have been applied extensively in the comparison of community mining algorithms (Danon et al. 2005; Lancichinetti and Fortunato 2009; Gustafsson, Hörnquist, and Lombardi 2006); which cluster nodes in a given network, based on the relationships between them. Clusters in networks, a.k.a communities, are shown to be *highly overlapping* (Gregory 2010; Leskovec, Lang, and Mahoney 2010). However, the current extensions of the clustering agreement indexes for overlapping cases (Lancichinetti, Fortunato, and Kertesz 2008; McDaid, Greene, and Hurley 2011; Yang and Leskovec 2013), which are used in the evaluation of overlapping community detection methods, are either inaccurate or inefficient, as discussed in detail in the related works.



(a) Three clusterings of a dataset with 13 data points.

(b) Comparing the pair-wise similarity of clusterings.

Figure 1: A clustering agreement measure compares the pair-wise similarity of clusterings.

In this paper, we present a general clustering agreement formula which naturally extends to overlapping clusters. The main contributions of this paper are five-fold:

1. The introduction of a novel, generalized Clustering Agreement Index (CAI);
2. Proofs that CAI generalizes a number of well-known existing agreement measures;
3. The derivation of a novel NMI measure, CMI, for overlapping clusterings, which reduces to the original NMI if clusterings are disjoint, and is both faster to compute and performs better compared to the current available overlapping extensions.
4. The derivation of a ARI formulation, CRI, for overlapping clusterings, which reduces to the original ARI if clusterings are disjoint, and is faster to compute compared to the current available extension, *i.e.*, computational order is reduced from $O(n^3)$ to $O(n)$;
5. Empirical comparison of four well-known overlapping community detection methods, obtained based on the previous and the newly proposed agreement indexes.

Background and Related Works

Clustering agreement indexes are often classified into three main families of set-matching, pair counting, and information theoretic indexes. The former class of indexes are less favoured as they suffer from the "problem of matching" (Meilă 2007); whereas the representatives of the latter two classes: Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and Normalized Mutual Information (NMI) (Vinh, Epps, and Bailey 2010), respectively, are the most commonly used indexes for comparing clusterings. Although classified differently, we have previously shown that the ARI and NMI indexes are derived from a generalized clustering distance formula (Rabbany and Zaïane 2015), which measures the average dispersion in the confusion matrix of the two given clusterings, *i.e.*, their pair-wise cluster overlaps. The above measures are all only defined when clusterings are disjoint.

In more detail, let U denote a clustering of dataset D with n datapoints into k mutually disjoint subsets/clusters, *i.e.*, $U = \{u_1, u_2 \ldots u_k\}$; where $u_i \cap u_j = \emptyset,\ \forall i \neq j$ and $D = \cup_{i=1}^{k} u_i$. The agreement between U and another clusterings of D, say V with r clusters, is computed based on the confusion table (a.k.a contingency table) of U and V, whose $(i,j)^{th}$ element denotes the size of overlap between u_i and v_j, $n_{ij} = |u_i \cap v_j|$, *i.e.*,

	v_1	v_2	$\ldots$	v_r	marg. sums
u_1	n_{11}	n_{12}	$\ldots$	n_{1r}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2r}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_k	n_{k1}	n_{k2}	$\ldots$	n_{kr}	$n_{k.}$
marg. sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.r}$	n

The last row and column are the marginal sums, *i.e.*, $n_{i.} = \sum_j n_{ij}$, and $n_{.j} = \sum_i n_{ij}$. ARI and NMI indexes are both measuring the (normalized) sum of the divergences in rows (and columns) of this table, as shown by our generalization in (Rabbany and Zaïane 2015). Several other clustering agreement indexes are defined based on this confusion matrix, including Jaccard, Rand Index, F-measure, Wallace (1983), Fowlkes and Mallows (1983), Mirkin Index (Wu, Xiong, and Chen 2009), and Variation of Information (Meilă 2007). All these measures are only defined when clusters are disjoint. In the literature, there is also a line of work on extending the agreement indexes for fuzzy clusters with soft memberships (Brouwer 2008; Quere et al. 2010; Campello 2010; Anderson et al. 2010; Hullermeier et al. 2012). The fuzzy measures are not applicable to cases where a data-point could fully belong to more than one cluster, *i.e.*, crisp overlappings; which are common in the clusters of nodes in networks, a.k.a communities.

When comparing overlapping communities, the two overlapping extensions for NMI by Lancichinetti, Fortunato, and Kertesz (2008) and McDaid, Greene, and Hurley (2011), are commonly used. These indexes are both defined based on a matching between the clusters of the two clusterings, and hence fail to measure the agreement accurately when this matching is not perfect. For instance, in Fig. 1, the original ARI and NMI indexes (which are defined only for disjoint cases) both agree with A, however the set-matching measures, including these two overlapping NMI extensions, suggest the opposite. This "problem of matching", coined by (Meilă 2007), is present in all the matching based agreement indexes, including the Balanced Error Rate with alignment introduced by Yang and Leskovec (2013), and the average $F1$ score, and Recall measures used by Mcauley and Leskovec (2014).

The reason for resorting back to the set-matching formulations is the inherent difficulty of extending any formula based on the confusion matrix for overlapping clusters; since this matrix can not differentiate between the natural overlaps in the data and the cluster overlaps used for measuring the (dis)agreement. To tackle this issue, the agreements between two given clusterings can be measured directly based on the co-memberships of data-points in their clusters, instead of the overlaps between their clusters; as implemented by our overlapping ARI extension proposed previously in (Rabbany and Zaïane 2015). Although this extension accurately extends the ARI for overlapping cases, it is computationally expensive and not scalable to most cases. In the next section, we show that this previous extension is a special case of our new CAI index, whereas the new formulation is more general and significantly more efficient.

Here, we introduce a novel generalized formula which measures the overlaps within the clusterings themselves, and factors them from their confusion matrix; hence naturally extends to overlapping clusterings. Our formulation reduces to the original formula if clusterings are disjoint, while it does not restrict any condition on the clusterings, and works for disjoint, fuzzy, or crisp overlapping clusters.

Clustering Agreement Index (CAI)

Consider clustering U which clusters dataset D with n datapoints. For each data point $i \in D$, and cluster $u \in U$, let $u_{\leftarrow i}$ denote the membership strength of data-point i in cluster u. The restrictive assumption on this definition is $\sum_u u_{\leftarrow i} = 1,\ \forall\ i \in D$, for disjoint and fuzzy cases, whereas in the former we also have $u_{\leftarrow i} \in \{0,1\}$, and in the latter $u_{\leftarrow i} \in [0,1]$. Here, we do not put any assumption on the form of the clusterings, and define a general agreement measure for two given clusterings of a same data-set. More formally, we assume $i \in u$ iff $u_{\leftarrow i} > 0$ and obtain the size of each cluster $u \in U$ as:

$$o_u = \sum_{i \in u} u_{\leftarrow i} \tag{1}$$

Similarly, we consider pairwise overlap between each pair of clusters u and v, where $i \in u \cap v$ iff $u_{\leftarrow i} > 0 \wedge v_{\leftarrow i} > 0$, and compute the size of this overlap as:

$$o_{uv} = \sum_{i \in u \cap v} u_{\leftarrow i} \times v_{\leftarrow i} \tag{2}$$

If we consider two clusterings U and V of the (same) dataset D, then $\{\{o_{uv}\ \forall u \in U\}\ \forall v \in V\}$ constitutes the confusion matrix of U and V. In the same manner, we can represent the intrinsic overlaps of the clustering themselves as $\{\{o_{uu'}\ \forall u \in U\}\ \forall u' \in U\}$ and $\{\{o_{vv'}\ \forall v \in V\}\ \forall v' \in V\}$. We quantify and sum up these overlaps using a generic function, $\varphi : \mathbb{R}^{\geq 0} \mapsto \mathbb{R}$; and also consider an expectation form, $\mathcal{E}$, to make the parallel with the previous formulations. More specifically, we consider:

$$\mathcal{O}_{UU} = \sum_{u \in U} \sum_{u' \in U} \varphi(o_{uu'})\ , \quad \mathcal{O}_{VV} = \sum_{v \in V} \sum_{v' \in V} \varphi(o_{vv'})$$

$$\mathcal{O}_{UV} = \sum_{u \in U} \sum_{v \in V} \varphi(o_{uv})\ , \quad \mathcal{E}_{UV} = \sum_{u \in U} \sum_{v \in V} \varphi(\frac{o_u o_v}{n})$$

Based on these we then define the Clustering Agreement Index (CAI) as:

$$CAI(U,V) = \frac{\mathcal{O}_{UV} - \mathcal{E}_{UV}}{\frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV}) - \mathcal{E}_{UV}} \tag{3}$$

We can derive different agreement measures using different $\varphi(.)$ functions. In particular, we introduce two agreement indexes derived with $\varphi(x) = x^2$ and $\varphi(x) = x\log(x)$; which are respectively called CRI and CMI. In the following, we prove that these two derivations reduce respectively to the original ARI and NMI when clusterings are disjoint. Hence, they provide overlapping extensions for these two commonly used indexes.

Theorem 1. *CRI, derived from CAI using $\varphi(x) = x^2$, reduces to the ARI for disjoint clusters.*

Proof. CRI is derived from CAI with $\varphi(x) = x^2$ as:

$$\frac{\sum\limits_{u \in U} \sum\limits_{v \in V} o_{uv}^2 - \sum\limits_{u \in U} \sum\limits_{v \in V} (\frac{o_u o_v}{n})^2}{\frac{1}{2}\left[\sum\limits_{u,u' \in U} o_{uu'}^2 + \sum\limits_{v,v' \in V} o_{vv'}^2\right] - \sum\limits_{u \in U} \sum\limits_{v \in V} (\frac{o_u o_v}{n})^2} \tag{4}$$

ARI, on the other hand, is defined by Hubert and Arabie (1985) based on the pair counts, *i.e.*, the number of pairs of data points clustered in the same and different clusters in the two given clusterings. These pair counts can be calculated based on the confusion matrix, whereas, the ARI for two clusterings with k and r disjoint clusters is often formulated as (Albatineh, Niewiadomska-Bugaj, and Mihalko 2006):

$$ARI = \frac{\sum\limits_{p=1}^{k} \sum\limits_{q=1}^{r} n_{pq}^2 - (\sum\limits_{p=1}^{k} n_{p.}^2)(\sum\limits_{q=1}^{r} n_{.q}^2)/n^2}{\frac{1}{2}[\sum\limits_{p=1}^{k} n_{p.}^2 + \sum\limits_{q=1}^{r} n_{.q}^2] - (\sum\limits_{p=1}^{k} n_{p.}^2)(\sum\limits_{q=1}^{r} n_{.q}^2)/n^2} \tag{5}$$

where n_{pq} measures the overlap between the p^{th} cluster in the first clustering and the q^{th} cluster in the second clustering, which we respectively denote by u and v, for short. Hence, we have $n_{pq} = |u \cap v| = \sum_{i \in u \cap v} 1$. When clusterings are disjoint, we have $u_{\leftarrow i} = 1$ iff $i \in u$ and zero otherwise; and we can write: $n_{pq} = \sum_{i \in u \cap v} u_{\leftarrow i} \times v_{\leftarrow i} = o_{uv}$. Therefore, for disjoint clusters (when $\varphi(x) = x^2$) we have:

$$\sum_{p=1}^{k} \sum_{q=1}^{r} n_{pq}^2 = \sum_{u \in U} \sum_{v \in V} o_{uv}^2 = \mathcal{O}_{UV}$$

Similarly, $n_{p.} = \sum_{q=1}^{r} n_{pq}$, which represent the size of cluster p. In case of disjoint covering clusters, we then have $n_{p.} = \sum_{i \in u} u_{\leftarrow i} = o_u$. Hence, with $\varphi(x) = x^2$, we have:

$$(\sum_{p=1}^{k} n_{p.}^2)(\sum_{q=1}^{r} n_{.q}^2)/n^2 = \sum_{p=1}^{k} \sum_{q=1}^{r} (n_{p.}n_{.q}/n)^2 = \mathcal{E}_{UV}$$

Furthermore, since clusters are disjoint, $o_{uu'} = o_u$ iff $u = u'$ and zero otherwise, *i.e.*, disjoint clusters only have overlap with themselves which is equal to their size, we further have:

$$\sum_{p=1}^{k} n_{p.}^2 = \sum_{u \in U} o_u^2 = \sum_{u \in U} \sum_{u' \in U} o_{uu'}^2 = \mathcal{O}_{UU}$$

Substituting these terms in the ARI of Equation 5 results in the CRI in Equation 4, which concludes the proof. □

Theorem 2. *CMI, derived from CAI using $\varphi(x) = x\log(x)$, reduces to the NMI_{sum} for disjoint clusters.*

Proof. NMI_{sum} is a common normalization for the mutual information between two given clusterings. It is formally defined for disjoint clusters as (Vinh, Epps, and Bailey 2009):

$$\begin{aligned} NMI_{sum}(U,V) &= \frac{I(U,V)}{\frac{1}{2}\left[H(U) + H(V)\right]} \\ &= \frac{H(U,V) - H(V) - H(U)}{\frac{1}{2}\left[H(U) + H(V)\right] - H(V) - H(U)} \end{aligned} \tag{6}$$

where $H(U)$ denotes the entropy of clustering U; $I(U,V)$ denotes the mutual information between two clustering U and V, and $H(U,V)$ denotes their joint entropy; which are all calculated based on the confusion matrix, assuming this matrix shows the joint probability distribution of, memberships of data points in, clustering U and V. More precisely:

$$H(U,V) = -\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{n_{ij}}{n} \log(\frac{n_{ij}}{n})$$

Similar to the ARI, n_{ij} denotes the overlap between the i^{th} cluster in U and the j^{th} cluster in V, which we call u and v for short, and then we write:

$$\begin{aligned} H(U,V) &= -\sum_{u \in U} \sum_{v \in V} \frac{o_{uv}}{n} \log(\frac{o_{uv}}{n}) \\ &= -\frac{1}{n} \sum_{u \in U} \sum_{v \in V} o_{uv}(\log(o_{uv}) - \log(n)) \\ &= -\frac{1}{n} \sum_{u \in U} \sum_{v \in V} o_{uv} \log(o_{uv}) + \frac{\log(n)}{n} \sum_{u \in U} \sum_{v \in V} o_{uv} \end{aligned}$$

Since clusters are covering and disjoint we have $\sum\limits_{u \in U} \sum\limits_{v \in V} o_{uv} = n$, hence we get:

$$H(U,V) = -\frac{1}{n} \sum_{u \in U} \sum_{v \in V} o_{uv} \log(o_{uv}) + \log(n)$$

Similarly, for disjoint covering clusters we also have $\sum_{u \in U} o_u = n$, and we can also rewrite $H(U)$ as:

$$\begin{aligned} H(U) &= -\sum_{i=1}^{k} \frac{n_{i.}}{n} \log(\frac{n_{i.}}{n}) \\ &= -\sum_{u \in U} \frac{o_u}{n} \log(\frac{o_u}{n}) = -\frac{1}{n} \sum_{u \in U} o_u \log(o_u) + \log(n) \end{aligned}$$

Since when clusters are disjoint we have $o_{uu'} = o_u$ iff $u = u'$ and zero otherwise, $H(U)$ is further equivalent to (when $0 \log 0 = 0$):

$$
\begin{aligned}
H(U) &= -\frac{1}{n} \sum_{u \in U} o_u \log(o_u) + \log(n) \\
&= -\frac{1}{n} \sum_{u \in U} \sum_{u' \in U} o_{uu'} \log(o_{uu'}) + \log(n)
\end{aligned}
$$

Therefore we can write:

$$
\begin{aligned}
H(U)+&H(V) = 2\log(n) - \\
&\frac{1}{n}\Big(\sum_{u,u' \in U} o_{uu'} \log(o_{uu'}) + \sum_{v,v' \in V} o_{vv'} \log(o_{vv'})\Big)
\end{aligned}
$$

On the other hand, for $H(V) + H(U)$ we also have:

$$
H(U) + H(V) = -\sum_{u \in U} \frac{o_u}{n} \log(\frac{o_u}{n}) - \sum_{v \in V} \frac{o_v}{n} \log(\frac{o_v}{n})
$$

For disjoint covering clusters, we can further show that:

$$
\begin{aligned}
-&\sum_{u \in U} \frac{o_u}{n} \log(\frac{o_u}{n}) - \sum_{v \in V} \frac{o_v}{n} \log(\frac{o_v}{n}) \\
&= -\frac{1}{n} \sum_{u \in U} \sum_{v \in V} \frac{o_u o_v}{n} \log(\frac{o_u o_v}{n}) + \log(n)
\end{aligned}
$$

By substituting these terms in the NMI_{sum} of Equation 6, we get the CMI formula, which concludes the proof. □

Lastly, we show that the CRI provides an alternative (and computationally feasible) formulation for the overlapping ARI proposed in (Rabbany and Zaïane 2015).

Theorem 3. *CRI of Equation 4 is equivalent to the overlapping extension of ARI proposed in (Rabbany and Zaïane 2015).*

Proof. Consider $U_{n \times k}$ denotes k clusters over a dataset with n data-points; *i.e.*, u_{ik} shows the strength of the data-point i's memberships in cluster k. Using this matrix representation, the size of pairwise cluster overlaps in U can be computed by $U^T U$, whereas the number of clusters that each pairs of data-points appeared together in, can be computed by UU^T. The Clustering Co-Membership Difference Matrix (Δ) is then defined as: $\Delta(U,V) = (UU^T - VV^T)_{n \times n}$; a normalization of which gives the generalized ARI, *i.e.*,

$$
\begin{aligned}
&ARI_\delta(U,V) = \\
&\quad 1 - \frac{\|\Delta(U,V)\|_F^2}{\|UU^T\|_F^2 + \|VV^T\|_F^2 - \frac{2}{n^2}|UU^T||VV^T|} \qquad (7)
\end{aligned}
$$

where $|.|$ is the sum of all elements in the matrix, and $\|.\|_F^2$ is the squared Frobenius norm, *i.e.*, sums the squared values of the given matrix. First, since we have:

$$
(UU^T)_{ij} = U_{i.}U_{.j}^T = U_{i.}U_{j.} = \sum_{p=1}^{k} U_{ip}U_{jp}
$$

hence we can write $\|UU^T\|_F^2$ as:

$$
\sum_{i=1}^{n}\sum_{j=1}^{n} \Big((UU^T)_{ij}\Big)^2 = \sum_{p=1}^{k}\sum_{p'=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ip}U_{jp}U_{ip'}U_{jp'}
$$

The expression $U_{ip}U_{jp}U_{ip'}U_{jp'}$ is zero if either i or j does not belong to cluster p or p', *i.e.*, one of the terms becomes zero, hence we can further rewrite $\|UU^T\|_F^2$ as:

$$
\sum_{p=1}^{k}\sum_{p'=1}^{k}\sum_{i,j \in p \cap p'} U_{ip}U_{jp}U_{ip'}U_{jp'} = \sum_{p=1}^{k}\sum_{p'=1}^{k}\left(\sum_{i \in p \cap p'} U_{ip}U_{ip'}\right)^2
$$

For clarity, we use p to denote both the index of the cluster and the cluster itself, *i.e.*, we simply write $p \cap p'$ instead of $U_{.p} \cap U_{.p'}$. Now, we can rewrite $\|UU^T\|_F^2$ as:

$$
\sum_{u \in U}\sum_{u' \in U}\left(\sum_{i \in u \cap u'} u_{\leftarrow i} \times u'_{\leftarrow i}\right)^2 = \sum_{u \in U}\sum_{u' \in U} o_{uu'}^2 \overset{*}{=} \mathcal{O}_{UU}
$$

whereas $\overset{*}{=}$ denotes that the equivalence holds since we have $\varphi(x) = x^2$. Similarly, we have:

$$
\begin{aligned}
&\|UU^T - VV^T\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\Big((UU^T)_{ij} - (VV^T)_{ij}\Big)^2 \\
&\overset{*}{=} \mathcal{O}_{UU} + \mathcal{O}_{VV} - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\left((\sum_{p=1}^{k} U_{ip}U_{jp})(\sum_{q=1}^{r} V_{iq}V_{jq})\right)
\end{aligned}
$$

where

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\left((\sum_{p=1}^{k} U_{ip}U_{jp})(\sum_{q=1}^{r} V_{iq}V_{jq})\right) \\
&= \sum_{p=1}^{k}\sum_{q=1}^{r}\left(\sum_{i \in p \cap q} U_{ip}V_{iq}\right)^2 = \sum_{p=1}^{k}\sum_{q=1}^{r} o_{pq}^2 \overset{*}{=} \mathcal{O}_{UV}
\end{aligned}
$$

We can further show that $|UU^T||VV^T| = n^2\mathcal{E}_{UV}$, since:

$$
\begin{aligned}
|UU^T||VV^T| &= \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(UU^T)_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(VV^T)_{ij}\right) \\
&= \left(\sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ip}U_{jp}\right)\left(\sum_{q=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{n} V_{iq}V_{jq}\right) \\
&= \left(\sum_{p=1}^{k}\sum_{i,j \in p} U_{ip}U_{jp}\right)\left(\sum_{q=1}^{r}\sum_{i,j \in q} V_{iq}V_{jq}\right) \\
&= \sum_{p=1}^{k}\left(\sum_{i \in p} U_{ip}\right)^2 \sum_{q=1}^{r}\left(\sum_{i \in q} V_{iq}\right)^2 \\
&= \sum_{p=1}^{k}(o_p)^2 \sum_{q=1}^{r}(o_q)^2 = \sum_{p=1}^{k}\sum_{q=1}^{r}(o_p o_q)^2
\end{aligned}
$$

By simple substitution, we now re-formulate Equation 7 as:

$$
ARI_\delta(U,V) \overset{*}{=} 1 - \frac{\mathcal{O}_{uu} + \mathcal{O}_{vv} - 2\mathcal{O}_{uv}}{\mathcal{O}_{uu} + \mathcal{O}_{vv} - 2\mathcal{E}_{UV}}
$$

Which holds when we have $\varphi(x) = x^2$, which concludes the proof. □

Although equivalent, the CRI formulation is more efficient than the ARI_δ, and is in fact on par with the disjoint variation in terms of the efficiency. We discuss this in more details in the next section.

Complexity of CAI

Let k and r, denote the number of clusters in the two given clusterings, where $k \geq r$. Also let m denote the cardinality of the largest cluster in the two clusterings. The complexity of ARI (Equation 5) and NMI (Equation 6), which are only defined for disjoint clusters, is $O(krm)$. Since they are both formulated based on the $k \times r$ confusion matrix, whose entries are the pairwise cluster overlaps, that are measured by set intersection in $O(m)$. The CAI (Equation 3), is also formulated based on $k \times r$ pairwise cluster overlaps, which are defined in Equation 2, and can be computed in $O(m)$. Moreover, CAI also measures the pairwise cluster overlaps in each of the two clusterings themselves, which costs $O(k^2m)$ and $O(r^2m)$. Hence the total complexity of CAI formula, is $O((kr + k^2 + r^2)m)$, which is in the order of $O(k^2m)$. This is about the same as the cost for the disjoint clusters. This is much more efficient when compared to the overlapping formulation of ARI (Equation 7) presented in (Rabbany and Zaïane 2015), which is based on pair-wise co-memberships and computes in the order of $O(n^3)$, where n denotes the number of data points. Therefore, our new extension can be computed more efficiently in real world applications where n is large.

Properties of CAI

Here, we show basic properties of CAI which are the basic axioms for a 'good' clustering agreement measure (Wagner and Wagner 2007).

Proposition 1. $CAI(U,U) = 1$

Proposition 2. *Symmetry:* $CAI(U,V) = CAI(V,U)$

Proof. We can see that $o_{uv} = \sum_{i \in u \cap v} u_{\leftarrow i} \times v_{\leftarrow i} = o_{vu}$. Based on which simply follows that $\mathcal{O}_{UV} = \mathcal{O}_{VU}$ and $\mathcal{E}_{UV} = \mathcal{E}_{VU}$, hence $CAI(U,V) = CAI(V,U)$. □

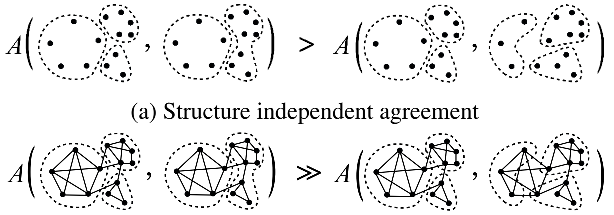
Proposition 3. *Upper bound:* $CAI(U,V) \leq 1$ *when* $\frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV})$ *is the upper bound for* $\mathcal{O}_{UV}$ *and* $\mathcal{E}_{UV}$.

Proof. $CAI(U,V) \leq 1$ iff $\frac{\mathcal{O}_{UV} - \mathcal{E}_{UV}}{\frac{1}{2}(\mathcal{O}_{UU}+\mathcal{O}_{VV}) - \mathcal{E}_{UV}} \leq 1$ iff $\frac{\mathcal{O}_{UV} - \frac{1}{2}(\mathcal{O}_{UU}+\mathcal{O}_{VV})}{\frac{1}{2}(\mathcal{O}_{UU}+\mathcal{O}_{VV}) - \mathcal{E}_{UV}} \leq 0$; which holds if we have $\mathcal{O}_{UV} \leq \frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV})$ and $\mathcal{E}_{UV} \leq \frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV})$, *i.e.*, when $\frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV})$ is the upper bound for $\mathcal{O}_{UV}$ and $\mathcal{E}_{UV}$. □

Proposition 4. *Lower bound:* $0 \leq CAI(U,V)$ *when* $\mathcal{E}_{UV} \leq \mathcal{O}_{UV}$, *and also* $\mathcal{E}_{UV} \leq \frac{1}{2}(\mathcal{O}_{UU} + \mathcal{O}_{VV})$.

Proof. From the definition of CAI, it is easy to see that this proposition holds. □

It is easy to show that the upper bound of Proposition 3 is valid for the CRI (derivation of CAI with $\varphi(x) = x^2$); and the lower bound of Proposition 4 is valid for CMI (derivation of CAI with $\varphi(x) = x\log(x)$). We also experimentally observe that the upper bound of Proposition 3 is also valid for CMI, and the lower bound of Proposition 4 is valid for CRI. However, proofs are not straightforward and are future works. It is interesting to study if there exists a property or constraint for φ which guarantees these bounds, and encompasses both $\varphi(x) = x^2$ (CRI) and $\varphi(x) = x \log x$ (CMI).



(a) Structure independent agreement

(b) Structure dependent agreement

Figure 2: Including the connections between the data-points makes the agreement of the first pair of clusterings stronger, since it is better aligned with the structure of the data.

Flexibility of CAI

CAI does not enforce any assumptions on the form of memberships of data-points. This flexibility is in particular important when adapting the clustering agreement indexes to compare communities, *i.e.*, clusters in networks. More specifically, communities are defined based on the connections between the nodes, whereas all the classic clustering agreement measures ignore the relationships between data-points. Fig. 2 illustrate the effect of including the structure on the agreement of two clusterings, by a revisit to example of Fig. 1. An intuitive way to incorporate the structure of the data in measuring the agreement between clusterings is to simply modify the memberships of nodes in clusters (*e.g.*, weighted by degree) and/or the overlap of two clusters (*e.g.*, sum of edges instead on nodes) to also consider the relationships between the data points. This is possible only if the agreement measure is flexible and has no restricting assumptions on the marginals of the contingency table, or the form of data-points' memberships in clusters. Which is the case for CAI, but not for the other classic measures.

Applicability of CAI

In this section, we compare our CRI and CMI derivations, against other overlapping alternatives, within their most common application, *i.e.*, external evaluation of community detection algorithms. The experimental settings are as follows. First, we generate a set of benchmark datasets using a generator which synthesizes networks with built-in ground-truth communities. Then, these datasets are clustered with different community detection algorithms. Finally, the results obtained from different algorithms are compared against the ground-truth in these benchmarks, using a clustering agreement index. Therefore, we obtain a ranking of algorithms per each agreement index. Our goal here is to show that these rankings are generally similar but disagree in some cases; in particular the rankings obtained from the current overlapping extensions of NMI and CMI (our overlapping extension) disagree the most when the number of clusters is much higher/lower than the ground-truth.

In more detail, we use the overlapping LFR benchmark generators (Lancichinetti, Fortunato, and Kertesz 2008), to synthesize benchmarks with varying fraction of overlapping nodes (10 realizations for each setting to report the average). The overlapping community detection methods included in

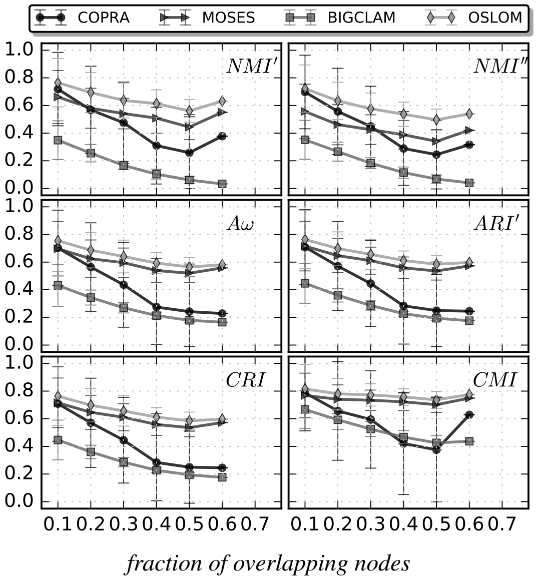


Figure 3: In each sub-plot, agreement of the results with the ground-truth is measured using the corresponding agreement index, plotted as a function of the fraction of overlapping nodes.

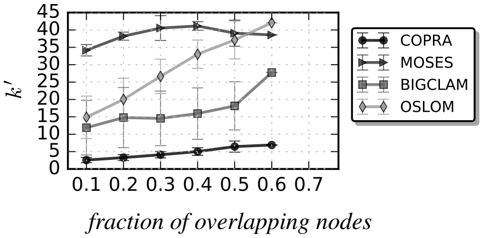


Figure 4: Size of clusters found by each community detection algorithm, corresponding to the results in Fig. 3.

this comparison are: COPRA (Gregory 2010), MOSES (McDaid and Hurley 2010), OSLOM (Lancichinetti et al. 2011), and BIGCLAM (Yang and Leskovec 2013). We apply the overlapping extensions of NMI, *i.e.*, NMI' by Lancichinetti, Fortunato, and Kertesz (2008) and NMI'' by McDaid, Greene, and Hurley (2011); the adjusted omega index ($A\omega$) by Collins and Dent (1988); the δ-based formulations for the ARI by Rabbany and Zaïane (2015), to compare them against the ARI and NMI overlapping extensions presented in this paper, *i.e.*, CRI and CMI derived from our CAI generalization. Fig. 3, shows the resulted ranking of algorithms according to each agreement measure, plotted as different sub-plots. In other words, each subplot provides a comparison of the community detection methods according to the corresponding clustering agreement index.

We can see in Fig. 3 that overall, the rankings of algorithms are consistent according to different agreement measures. This is expected, since these indexes are similar or in case of ARI_δ and CRI, identical. We can however observe the differences between set-matching based extensions of NMI (subplots in the top row), with the NMI extension presented here (CMI). For example, we can see that according to both CMI and CRI, performances of OSLOM and MOSES algorithms are close, however, the previous overlapping extensions of NMI (NMI' and NMI''), rank OSLOM significantly higher. This is because OSLOM finds relatively less communities, as shown in Fig. 4, and hence its communities are better matched with the ones in the ground-truth, whereas MOSES's communities are finer grained and not properly matched with the communities in the ground-truth, when using a set-matching based agreement index. We can observe a similar pattern in the comparison of BIGCLAM and COPRA; *i.e.*, BIGCLAM finds too few communities and its performance plot is consequently shifted much lower by the NMI' and NMI'', when compared to CMI and other agreement measures. Therefore, in general CMI seems to be a more accurate overlapping extension for the NMI when compared to the NMI' and NMI''.

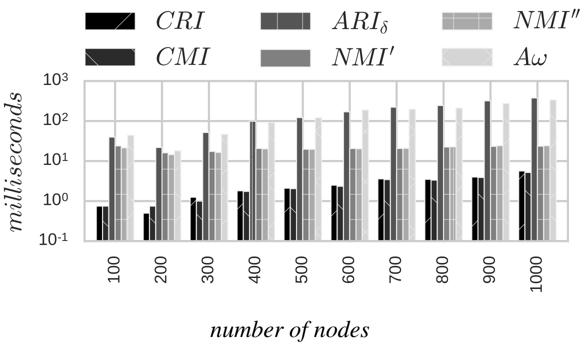


Figure 5: Experimental time comparison of CRI and CMI derived from our generalization, with ARI_δ, NMI', NMI'', and A_ω.

We further compare the time complexity of these indexes, reported in Fig. 5, where the average run time for computing different indexes is plotted as a function of number of nodes in the network. We can see that the proposed CMI and CRI overlapping derivations of CAI are much more efficient compared to the other measures, particularly when compared to the ARI_δ. This confirms our theoretical comparison presented earlier.

Conclusions

In this paper, we presented an elegant generalization of clustering agreement indexes, which i) extends to overlapping cases, ii) generalizes existing measures, iii) introduces novel measures. In particular, our CMI derivation is both more accurate and more efficient compared to the previous overlapping NMI extensions. Our generalization, called CAI, and its derivations, are scalable and flexible, and therefore, unlike the previous measures, applicable to the large datasets, such as a typical network. The implications of our proposal are however more broad, since our newly proposed generalization applies to all cases of disjoint, fuzzy, or crisp overlapping clusters. As a future work, it is interesting to investigate other sensible choices for the generative function used in our generalization, to derive new CAIs, and to study its advantages and limitations.

References

Aggarwal, C. C., and Reddy, C. K. 2014. *Data Clustering: Algorithms and Applications*.

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

Anderson, D. T.; Bezdek, J. C.; Popescu, M.; and Keller, J. M. 2010. Comparing Fuzzy, Probabilistic, and Possibilistic Partitions. *IEEE Transactions on Fuzzy Systems* 18(5):906–918.

Brouwer, R. K. 2008. Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems* 32(3):213–235.

Campello, R. R. 2010. Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters* 31(9):966–975.

Collins, L. M., and Dent, C. W. 1988. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research* 23(2):231–242.

Cui, Y.; Fern, X.; and Dy, J. 2007. Non-redundant multi-view clustering via orthogonalization. In *IEEE International Conference on Data Mining*, 133–142.

Danon, L.; Guilera, A. D.; Duch, J.; and Arenas, A. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* (09):09008.

Fowlkes, E. B., and Mallows, C. L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78(383):pp. 553–569.

Fred, A. L., and Jain, A. K. 2003. Robust data clustering. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, II–128–II–133 vol.2.

Fred, A. L., and Jain, A. K. 2005. Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27(6):835–850.

Gregory, S. 2010. Finding overlapping communities in networks by label propagation. *New Journal of Physics* 12(10):103018.

Gustafsson, M.; Hörnquist, M.; and Lombardi, A. 2006. Comparison and validation of community structures in complex networks. *Physica A Statistical Mechanics and its Applications* 367:559–576.

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

Hullermeier, E.; Rifqi, M.; Henzgen, S.; and Senge, R. 2012. Comparing Fuzzy Partitions: A Generalization of the Rand Index and Related Measures. *IEEE Transactions on Fuzzy Systems* 20(3):546–556.

Lancichinetti, A., and Fortunato, S. 2009. Community detection algorithms: A comparative analysis. *Physical Review E* 80(5):056117.

Lancichinetti, A.; Radicchi, F.; Ramasco, J. J.; and Fortunato, S. 2011. Finding statistically significant communities in networks. *PloS one* 6(4):e18961.

Lancichinetti, A.; Fortunato, S.; and Kertesz, J. 2008. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics* 11(3):20.

Leskovec, J.; Lang, K. J.; and Mahoney, M. 2010. Empirical comparison of algorithms for network community detection. In *International conference on World wide web*, 631–640.

Mcauley, J., and Leskovec, J. 2014. Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data* 8(1):4:1–4:28.

McDaid, A., and Hurley, N. 2010. Detecting highly overlapping communities with model-based overlapping seed expansion. In *ASONAM*, 112–119.

McDaid, A. F.; Greene, D.; and Hurley, N. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

Meilă, M. 2007. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis* 98(5):873–895.

Quere, R.; Le Capitaine, H.; Fraisseix, N.; and Frelicot, C. 2010. On Normalizing Fuzzy Coincidence Matrices to Compare Fuzzy and/or Possibilistic Partitions with the Rand Index. In *IEEE International Conference on Data Mining*, 977–982.

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

Strehl, A., and Ghosh, J. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3:583–617.

Vinh, N. X.; Epps, J.; and Bailey, J. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *International Conference on Machine Learning*, 1073–1080.

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

Wagner, S., and Wagner, D. 2007. *Comparing clusterings: an overview*.

Wallace, D. L. 1983. A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association* 78(383):pp. 569–576.

Wu, J.; Xiong, H.; and Chen, J. 2009. Adapting the right measures for k-means clustering. In *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, 877–886.

Yang, J., and Leskovec, J. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 587–596.