**Noname manuscript No.**
(will be inserted by the editor)

# Generalization of Clustering Agreements and Distances for Overlapping Clusters and Network Communities

**Reihaneh Rabbany · Osmar R. Zaïane**

Received: date / Accepted: date

**Abstract** A measure of distance between two clusterings has important applications, including clustering validation and ensemble clustering. Generally, such distance measure provides navigation through the space of possible clusterings. Mostly used in cluster validation, a normalized clustering distance, a.k.a. agreement measure, compares a given clustering result against the ground-truth clustering. The two widely-used clustering agreement measures are Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). In this paper, we present a generalized clustering distance from which these two measures can be derived. We then use this generalization to construct new measures specific for comparing (dis)agreement of clusterings in networks, a.k.a. communities. Further, we discuss the difficulty of extending the current, contingency based, formulations to overlapping cases, and present an alternative algebraic formulation for these (dis)agreement measures. Unlike the original measures, the new co-membership based formulation is easily extendable for different cases, including overlapping clusters and clusters of inter-related data. These two extensions are, in particular, important in the context of finding communities in complex networks.

**Keywords** Clustering Agreement; Cluster Evaluation; Cluster Validation; Network Clusters; Community Detection; Overlapping Clusters

## 1 Introduction

A cluster distance, accordance, similarity, or divergence has different applications. *Cluster validation* is the most common usage of cluster distance measures. In particular, in external evaluation, a clustering algorithm is validated on a set of benchmark datasets by comparing the similarity of its results against the ground-truth clusterings. Another notable application is *ensemble, or consensus clustering*, where results of different clustering algorithms on the same dataset are aggregated. A notion of distance between alternative clusterings is used in modeling and formulating this aggregation, i.e. to find a clustering that has the minimum average distance

Department of Computing Science, University of Alberta
E-mail: {rabbanyk,zaiane}@ualberta.ca

to the alternative clusterings[1]. Another closely related application is multi-view clustering (Cui et al., 2007), where the objective is to find different clusterings of the same dataset, which are usually in different sub-spaces of the data, and could represent different views of that dataset. In the same context, one might be interested to find the sub-spaces that result in different/similar clusterings.

Clustering distance measures are well-studied and widely-used in cluster validation, where they measure the (dis)agreement between clustering results and the ground-truth clustering. A normalized measure is used to average agreements over different benchmark datasets; such as Rand Index ($RI$), Adjusted Rand Index ($ARI$), Normalized Mutual Information ($NMI$), or normalized Variation of Information ($VI$). These agreement measures are classified into two families of pair counting and information theoretic measures, and are often studied separately. In this paper, we demonstrate that both these families are measuring the agreements based on the same principle. In more details, these measures are all defined based on the contingency table of the two given clusterings, i.e. their pair-wise cluster overlaps; and they measure the average dispersion in this table. More specifically, we present a *generalized clustering distance* defined also based on the contingency table, which has a generative function $\varphi$. We then show that the representatives of both these families can be generated from the proposed generalized distance, using specific $\varphi$ functions. This generalized distance also provides the means for constructing new clustering agreement measures.

Moreover, unlike the original definitions, this generalized formula does not require the clusters to be disjoint and nor does it require them to cover all the data-points; the latter is particularly useful in case there are outliers, or missing cluster labels. The former, however, does not help extending to overlapping cases. In fact, there is an inherent difficulty in extending any contingency based formula for *overlapping clusters*; since the contingency table can not differentiate between the natural overlaps in the data and the cluster overlaps used for measuring the (dis)agreement. To tackle this issue, we propose to measure the agreements between two given clusterings directly based on the co-memberships of data-points in their clusters, instead of the overlaps between their clusters. More specifically, we define the *Clustering Co-Membership Difference Matrix ($\Delta$)*, based on which clustering distance could be quantified. In particular, we present two normalized forms for $\Delta$, denoted by $RI_\delta$ and $ARI_\delta$, which are overlapping counterparts for $RI$ and $ARI$, and will reduce to the original measures in case of disjoint clusters.

Last but not least, we point out the fact that all the current clustering agreement measures only consider memberships of data-points in clusters, and overlook any relations between the data-points or any attributes associated with them. This is in particular problematic when comparing clusterings in the *complex networks*, since the structure of the data, represented by the relationships between data-points, is of great importance. Whereas, these agreement measures are being widely-used in the evaluation and comparison of network clustering methods[2]. Here, we discuss the effect of neglecting these relations, i.e. links in the networks, and derive extensions of our generalized formulae which incorporate the structure of the data in measuring clustering agreements.

[1] Refer to Aggarwal and Reddy (2014), Chapter 23 on clustering validation measures (in particular the section on external clustering validation measures); and Chapter 22 on cluster ensembles (in particular the section on measuring similarity between clustering solutions).

[2] a.k.a. community mining; refer to Fortunato (2010) for a survey.

## 2 Clustering Agreement Measures: Short Survey

Consider a dataset $D$ consisting of $n$ data items, $D = \{d_1, d_2, d_3 \dots d_n\}$. A partitioning $U$ partitions $D$ into $k$ mutually disjoint subsets, $U = \{U_1, U_2 \dots U_k\}$; where $D = \cup_{i=1}^{k} U_i$ and $U_i \cap U_j = \emptyset$, $\forall i \neq j$. There are several measures defined to examine the similarity, a.k.a agreement, between two partitionings of the same dataset. More formally, let $V$ denote another partitioning of the dataset $D$, $V = \{V_1, V_2 \dots V_r\}$. Clustering agreement measures are originally introduced based on counting the pairs of data items that are in the same/different partition in $U$ and $V$. Each pair $(d_i, d_j)$ of data items is classified into one of four groups based on their co-memberships in $U$ and $V$; which results in the following pair-counts.

| | Same in $V$ | Different in $V$ |
|---|---|---|
| Same in $U$ | $M_{11} = TP$ | $M_{10} = FP$ |
| Different in $U$ | $M_{01} = FN$ | $M_{00} = TN$ |

Here, $M_{11}/M_{00}$ counts the number of pairs that are in the same/different partitions in both $U$ and $V$. $M_{10}/M_{01}$ sums up those that belong to the same/different partitions in $U$ but are in different same/partitions according to $V$. When one of these partitionings, for instance $V$, is the true partitioning i.e. the ground-truth, these could also be referred to as the true/false positive/negative scores.

These pair counts are often derived using the following contingency table a.k.a. confusion table (Hubert and Arabie, 1985). The contingency table is a $k \times r$ matrix of all the possible overlaps between each pair of clusters in $U$ and $V$, where its $ij^{th}$ element shows the intersection of cluster $U_i$ and $V_j$, i.e. $n_{ij} = |U_i \cap V_j|$.

| | $V_1$ | $V_2$ | $\dots$ | $V_r$ | marginal sums |
|---|---|---|---|---|---|
| $U_1$ | $n_{11}$ | $n_{12}$ | $\dots$ | $n_{1r}$ | $n_{1.}$ |
| $U_2$ | $n_{21}$ | $n_{22}$ | $\dots$ | $n_{2r}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $U_k$ | $n_{k1}$ | $n_{k2}$ | $\dots$ | $n_{kr}$ | $n_{k.}$ |
| marginal sums | $n_{.1}$ | $n_{.2}$ | $\dots$ | $n_{.r}$ | $n$ |

The last row and column show the marginal sums of $n_{i.} = \sum_j n_{ij}$, and $n_{.j} = \sum_i n_{ij}$. In case of disjoint clusters, we further have $n_{i.} = |U_i|$, and $n_{.j} = |V_j|$. The pair counts can then be computed using the following formulae.

$$M_{10} = \sum_{i=1}^{k} \binom{n_{i.}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}, \quad M_{01} = \sum_{j=1}^{r} \binom{n_{.j}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}$$

$$M_{11} = \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2}, \quad M_{00} = \binom{n}{2} + \sum_{i=1}^{k}\sum_{j=1}^{r} \binom{n_{ij}}{2} - \sum_{i=1}^{k} \binom{n_{i.}}{2} - \sum_{j=1}^{r} \binom{n_{.j}}{2}$$

These *pair counts* have been used to define a variety of different clustering agreement measures (Manning et al., 2008). Here, we briefly explain the most common measures; the reader can refer to Albatineh et al. (2006) for a complete survey.

Considering co-membership of data-points in the same or different clusters as a binary variable, *Jaccard* agreement between clustering $U$ and $V$ is defined as:

$$J = \frac{TP}{(FP + FN + TP)} = \frac{M_{11}}{(M_{01} + M_{10} + M_{11})}$$

*Rand Index* is defined similarly to Jaccard, but it also values pairs that belong to different clusters in both partitionings, i.e. $RI = \frac{(M_{11}+M_{00})}{(M_{11}+M_{01}+M_{10}+M_{00})}$, which gives:

$$RI = 1 + \frac{1}{n^2-n}(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - (\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2)) \tag{1}$$

*F-measure* is a weighted mean of the precision, $P = {}^{M_{11}}/_{(M_{11}+M_{10})}$, and recall, $R = {}^{M_{11}}/_{(M_{11}+M_{01})}$, i.e. $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$; where $\beta$ indicates how much recall is more important than precision, with the two common values of 2 and ${}^1/_2$.

There is also a family of *information theoretic* based measures. These measures consider the overlaps between clusters in $U$ and $V$, as a joint distribution of two random variables, i.e. the cluster memberships in $U$ and $V$. The entropy of cluster $U$, $H(U)$, the joint entropy of $U$ and $V$, $H(U,V)$, their mutual information, $I(U,V)$, and their *Variation of Information* (Meilă, 2007), $VI(U,V)$ are then defined as:

$$H(U) = -\sum_{i=1}^{k}\frac{n_{i.}}{n}\log(\frac{n_{i.}}{n}), \quad H(V) = -\sum_{j=1}^{r}\frac{n_{.j}}{n}\log(\frac{n_{.j}}{n})$$
$$H(U,V) = -\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}}{n}\log(\frac{n_{ij}}{n}), \quad I(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}}{n}\log(\frac{n_{ij}/n}{n_{i.}n_{.j}/n^2})$$
$$VI(U,V) = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}}{n}\log(\frac{n_{i.}n_{.j}/n^2}{n_{ij}^2/n^2}) \tag{2}$$

All the pair counting measures defined here have a fixed range of $[0,1]$. The above information theoretic measures, however, do not have a fixed range. For example, the mutual information ranges between $(0, logk]$, and the range for $VI$ is $[0, 2\log\max(k,r)]$ (Wu et al., 2009). Having a fixed range, i.e. being *normalized*, is a desired property for clustering agreement indexes, since we often require to compare/average agreements over different datasets. Consequently, different normalized forms for the mutual information are defined; refer to (Vinh et al., 2010) for a survey. The most commonly used forms are:

$$NMI_\Sigma = \frac{2I(U,V)}{H(U)+H(V)} \quad \text{and} \quad NMI_\surd = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{3}$$

Beside having a fixed range, a clustering agreement measure should also return the same value, usually zero, for agreement no better than random[3] (Vinh et al., 2010; Hubert and Arabie, 1985). *Correction for chance* is adjusting a measure to have a constant expected value for agreements due to chance. This adjustment is done based on an upper bound on the measure, $Max[M]$, and its expected value, $E[M]$, using the following formula:

$$AM = \frac{M - E[M]}{Max[M] - E[M]} \tag{4}$$

The *Adjusted Rand Index (ARI)* is proposed by Hubert and Arabie (1985), assuming that the contingency table is constructed randomly when the marginals are fixed, i.e. the size of the clusters in $U$ and $V$ are fixed. With this assumption, $RI$ is

[3] In other words, it should have a constant baseline, i.e. the expected value of agreements between two random clusterings of a same dataset. If not constant, an example of 0.7 agreement value can be both a strong (when baseline is 0.2) or a weak (when baseline is 0.6) agreement.

a linear transformation of $\sum_{i,j}\binom{n_{ij}}{2}$, and $E\left(\sum_{i,j}\binom{n_{ij}}{2}\right) = \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}/\binom{n}{2}$. Hence, adjusting $RI$ with upper bound 1 results in the following formula:

$$ARI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r}\binom{n_{ij}}{2} - \sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{j=1}^{r}\binom{n_{.j}}{2}/\binom{n}{2}}{\frac{1}{2}[\sum_{i=1}^{k}\binom{n_{i.}}{2} + \sum_{j=1}^{r}\binom{n_{.j}}{2}] - \sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{j=1}^{r}\binom{n_{.j}}{2}/\binom{n}{2}} \tag{5}$$

There is also an approximate formulation (Hubert and Arabie, 1985; Albatineh et al., 2006) for this expectation defined as $E(\sum_{i,j} n_{ij}^2) = \sum_i n_{i.}^2 \sum_j n_{.j}^2/n^2$, which results in a slightly different formula for the $ARI$, i.e.

$$ARI' = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r} {n_{ij}}^2 - \sum_{i=1}^{k} {n_{i.}}^2 \sum_{j=1}^{r} {n_{.j}}^2/n^2}{\frac{1}{2}[\sum_{i=1}^{k} {n_{i.}}^2 + \sum_{j=1}^{r} {n_{.j}}^2] - \sum_{i=1}^{k} {n_{i.}}^2 \sum_{j=1}^{r} {n_{.j}}^2/n^2} \tag{6}$$

There are several variations of pair counting agreement measures, such as Gamma, Hubert, Pearson, etc. These measures, however, become similar or even equivalent after correction for chance (Albatineh et al., 2006). Furthermore, it has been shown that after correction for chance they also become equivalent to one of the statistical inter-rater agreement indices[4] (Warrens, 2008a). Specifically, the equivalence of *Cohen's kappa*, one the most widely used inter-rater agreement index, and the $ARI$ is proved by Warrens (2008b). Cohen's kappa is a chance corrected index of association defined for assessing the agreement between two raters, who categorize data into $k$ categories.

Correction for chance of the information theoretic measures is discussed by Vinh et al. (2009). More specifically, they propose *Adjusted Mutual Information* as $AMI = {}^{(I - E[I])}/_{(Max[I] - E[I])}$ using Equation 4; where different forms of $AMI$ are derived using different upper bounds on $I$ as $Max[I]$, which are:

$$I(U,V) \leq \min(H(U),H(V)) \leq \sqrt{H(U)H(V)} \leq \tfrac{H(U)+H(V)}{2} \leq \max(H(U),H(V)) \leq H(U,V)$$

In particular, $AMI_{sum}$, i.e. $AMI$ with upper bound of ${}^{[H(U)\,+\,H(V)]}/_2$, is equivalent to the Adjusted form for Variation of Information. On the other hand, the expected value, $E[I]$, is derived assuming the sizes of the clusters are fixed, i.e. similar to the $ARI$'s assumption on the hypergeometric model of randomness, as:

$$E[I(U,V)] = \sum_{i,j}\sum_{\mathtt{m}=\max(n_{i.}+n_{.j}-n,1)}^{min(n_{i.},n_{.j})} \frac{\mathtt{m}}{n}\log\left(\frac{n\mathtt{m}}{n_{i.}n_{.j}}\right)\frac{n_{i.}!n_{.j}!(n-n_{i.})!(n-n_{.j})!}{n!\mathtt{m}!(n_{i.}-\mathtt{m})!(n_{.j}-\mathtt{m})!(n-n_{i.}-n_{.j}+\mathtt{m})!}$$

This formulation includes big factorials, therefore is computationally complex; which makes $AMI$ less practical when compared to the $ARI$.

[4] The well-studied *inter-rater agreement indices* in statistics are defined to measure the agreement between different coders, or judges on categorizing the same data. Examples are the goodness of fit, chi-square test, the likelihood chi-square, kappa measure of agreement, Fisher's exact test, Krippendroff's alpha, etc. (see test 16 in (Cortina-Borja, 2012)). These statistical tests are also defined based on the contingency table which displays the multivariate frequency distribution of the (categorical) variables.

## 3 Generalization of Clustering Agreement Measures

Both families of pair counting and information theoretic measures quantify the agreement between two clusterings based on their contingency table. Here, we generalize them, and show that they are both measuring the (normalized) sum of the divergences in rows (and columns) of this table; whereas the perfect agreement occurs if the sum is zero[5]. Our generalization is symmetric and is defined based on the relation between the Rand Index ($RI$) and the Variation of Information ($VI$), which are respectively a representative for the pair counting and information theoretic families.

**Proposition 1** *VI (RI) of two partitionings is proportional to the conditional entropies (variances) of memberships in them (see Appendix A.1 for proof), i.e.*

$$VI(U,V) = H(U|V) + H(V|U) \quad \textit{and} \quad RI(U,V) \propto Var(U|V) + Var(V|U)$$

Based on this proposition, we define the generalized distance for clusterings as:

**Definition 1** Generalized Clustering Distance ($\mathcal{D}$)

$$\mathcal{D}^{\eta}_{\varphi}(U,V) = \mathcal{D}^{\eta}_{\varphi}(U||V) + \mathcal{D}^{\eta}_{\varphi}(V||U), \quad \mathcal{D}^{\eta}_{\varphi}(U||V) = \sum_{v\in V}\left[\varphi(\sum_{u\in U}\eta_{uv}) - \sum_{u\in U}\varphi(\eta_{uv})\right]$$

where $\eta_{uv}$ quantifies the similarity between the two clusters of $u \in U$ and $v \in V$, i.e. $\eta : 2^V \times 2^U \to \mathbb{R}$; and $\varphi : \mathbb{R} \to \mathbb{R}$ is a non-linear function, such that $\varphi(\sum x) \neq \sum \varphi(x)$, which is used to quantify the divergence or dispersion in a set of numbers.

This generalized formula can be extended to define novel clustering distances, using the flexibility that the $\varphi$ and $\eta$ functions provide[6]. More specifically, $\eta$ is any function that transforms the two given clusterings into a contingency table. The distance between the clusterings is then measured by adding up the dispersion in each row (and column) of this table. Function $\varphi$ is used to quantify how disperse are the values in each row (and column), i.e. to measure the divergence from a spike distribution, observed if the corresponding clusters are perfectly matched.

**Corollary 1** *$\mathcal{D}$ is bounded if $\varphi$ is a positive superadditive function (proof in Appendix A.2), i.e.*

$$\varphi(x) \geq 0 \ \wedge \ \varphi(x+y) \geq \varphi(x) + \varphi(y) \implies 0 \leq \mathcal{D}^{\eta}_{\varphi}(U||V) \leq \varphi(\sum_{v\in V}\sum_{u\in U}\eta_{uv})$$

Using this bound as a normalizing factor, we define:

**Definition 2** Normalized Generalized Clustering Distance ($\mathcal{ND}$)

$$\mathcal{ND}^{\eta}_{\varphi}(U,V) = \frac{\mathcal{D}^{\eta}_{\varphi}(U,V)}{NF(U,V)}, \quad NF(U,V) = \varphi(\sum_{v\in V}\sum_{u\in U}\eta_{uv})$$

[5] Which happens when the clusterings are identical, and only the order of the clusters is permuted, i.e. the distribution of overlaps in each row/column of the contingency table has a single spike on the matched cluster and is zero elsewhere.

[6] For example, we introduce an extension of this generalization for clusterings of nodes in graphs, a.k.a. communities, in the following section.

Here, we first show that both the Rand Index ($RI$) and the (normalized) Variation of Information ($VI$) generate from the proposed normalized distance ($\mathcal{ND}$). Then, we introduce the adjusted form for $\mathcal{ND}$, and show that the Adjusted Rand Index ($ARI$) and Normalized Mutual Information ($NMI$) both derive from this adjusted form. More specifically,

**Identity 1** *The Variation of Information (Equation 2) derives from $\mathcal{ND}$ if we set $\varphi(x) = x \log x$, and $\eta$ as the overlap size: $\eta_{uv} = |u \cap v|$ (proof in Appendix A.3), i.e.*

$$\mathcal{ND}^{|\cap|}_{x \log x}(U, V) \equiv \frac{VI(U, V)}{\log n}$$

**Identity 2** *The Rand Index (Equation 1) derives from $\mathcal{ND}$ if we set $\varphi(x) = \binom{x}{2}$, and $\eta$ as the overlap size (proof in Appendix A.4), i.e.*

$$\mathcal{ND}^{|\cap|}_{\binom{x}{2}}(U, V) \equiv 1 - RI(U, V), \text{ also } \quad \mathcal{ND}^{|\cap|}_{x^2}(U, V) \equiv 1 - RI'(U, V)$$

Similar to the Identity 2, in the rest of this paper, we consider clustering agreement ($\mathcal{I}$) and normalized distance ($\mathcal{ND}$) interchangeably, i.e. using $\mathcal{I} = 1 - \mathcal{ND}$.

We further adjust the generalized distance to take its maximum, i.e. one, if $U$ and $V$ are independent. Assume $P_{U,V}(u, v) = \eta_{uv} / \sum_{uv} \eta_{uv}$ as the joint probability distribution with the marginals of $P_U(u) = \sum_v P_{U,V}(u, v) = \eta_{.v} / \sum_{uv} \eta_{uv}$ and $P_V(v) = \eta_{u.} / \sum_{uv} \eta_{uv}$. Then the independence condition for $U$ and $V$, i.e. $P_{U,V}(u, v) = P_U(u) P_V(v)$, translates into $\eta_{uv} = \eta_{u.} \eta_{.v} / \sum_{uv} \eta_{uv}$. On the other hand, from Definition 1, we have:

$$\mathcal{D}^{\eta}_{\varphi}(U, V) = \sum_{v \in V} \varphi(\eta_{.v}) + \sum_{u \in U} \varphi(\eta_{u.}) - 2 \sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})$$

Therefore, we define the adjusted distance as:

**Definition 3** Adjusted Generalized Clustering Distance ($\mathcal{AD}$)

$$\mathcal{AD}^{\eta}_{\varphi} = \frac{\mathcal{D}^{\eta}_{\varphi}(U, V)}{NF(U, V)}, \quad NF = \sum_{v \in V} \varphi(\eta_{.v}) + \sum_{u \in U} \varphi(\eta_{u.}) - 2 \sum_{u \in U} \sum_{v \in V} \varphi \left( \frac{\eta_{.v} \eta_{u.}}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}} \right)$$

**Identity 3** *The Normalized Mutual Information (Equation 3) derives from $\mathcal{AD}$, if we set $\varphi(x) = xlogx$, and $\eta$ as the overlap size: $\eta_{uv} = |u \cap v|$ (proof in Appendix A.5), i.e.*

$$\mathcal{AD}^{|\cap|}_{xlogx}(U, V) \equiv 1 - NMI_{sum}(U, V)$$

**Identity 4** *The Adjusted Rand Index of Equation 5 and Equation 6 derive from $\mathcal{AD}$, if we set $\varphi(x) = x(x-1)$ and $\varphi(x) = x^2$ respectively, where $\eta$ is the overlap size, (proof in Appendix A.5), i.e.*

$$\mathcal{AD}^{|\cap|}_{x^2}(U, V) \equiv 1 - ARI'(U, V), \quad \mathcal{AD}^{|\cap|}_{\binom{x}{2}}(U, V) \cong 1 - ARI(U, V)$$

This line of generalization is similar to the works in Bergman Divergence and $f$-divergences. For example, the mutual information and variance are proved to be special cases of Bergman information (Banerjee et al., 2005). The (reverse) KL divergence and Pearson $\chi^2$ are shown to be $f$-divergences when the generator is $x \log x$ and $(x-1)^2$ respectively (Nielsen and Nock, 2014). Beside this analogy, our generalized measure is different from these divergences. One could consider our proposed measure as an (adjusted normalized) conditional Bergman entropy for clusterings. This relation is however non-trivial and is out of scope of this paper.

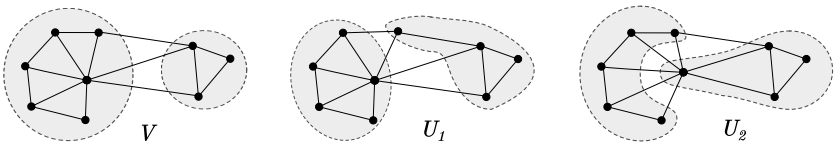



Fig. 1: An example graph clustered in three different ways: by clustering $V$ (i.e. true clustering), and by $U_1$ and $U_2$ (i.e. two candidate clusterings). Considering only the number of nodes in the overlaps and ignoring the edges, $U_1$ and $U_2$ have the same contingency table with $V$, i.e. $|\cap|(U_1, V) = |\cap|(U_2, V) = \{\{5, 0\}, \{1, 3\}\}$. Therefore, they have the same agreement with $V$, regardless of the choice of the agreement measure: $ARI$, $NMI$, etc. However if considering the edges, $U_1$ is more similar to the true clustering $V$. This could be enforced using an alternative overlap function that incorporates edges, such as the degree weighted overlap function, by which we get: $\Sigma d(U_1, V) = \{\{18, 0\}, \{3, 9\}\}$ and $\Sigma d(U_2, V) = \{\{14, 0\}, \{7, 9\}\}$; or the edge based variation, which gives: $\xi(U_1, V) = \{\{7, 0\}, \{0, 3\}\}$ and $\xi(U_2, V) = \{\{4, 0\}, \{0, 3\}\}$.

### 3.1 Extension for Inter-related Data

The common clustering agreement measures introduced in the previous section, only consider memberships of data-points in clusters, and *overlook the attributes of individual data-points or any relations between them.* This overlook is problematic, as also mentioned by a few previous works. For example, Zhou et al. (2005) illustrate the issue of ignoring the distances between data-points[7] when comparing clusterings; and propose a measure which incorporates the distances between the representatives of clusters. This overlook is in particular important when comparing *clusterings of nodes within information networks.* An information network encodes relationships between data-points, and a clustering on such network forms sub-graphs. Using the original clustering agreement measures, we only consider the nodes in measuring the clustering distance. One should however also consider edges when comparing two sub-graphs; see Figure 1 for a clarifying example.

To incorporate the structure of the data in our generalized distance (Definition 1), we simply modify the overlap function $\eta$. The generator overlap function for the original measures ($RI$, $VI$, $ARI$, and $NMI$) is $|\cap| : \eta_{uv} = \sum_{i \in u \cap v} 1$; which counts the number of common nodes. Therefore, the first intuitive modification to incorporate the structure is to consider a degree weighted function as:

$$\Sigma d : \eta_{uv} = \sum_{i \in u \cap v} d_i$$

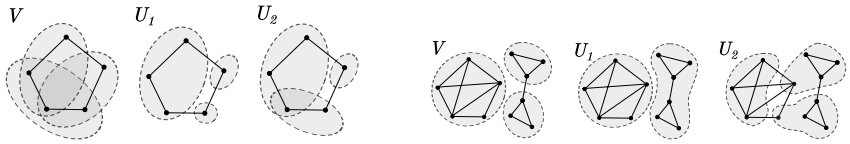
Using this $\eta$, well-connected nodes with higher degree weigh more in the distance. Another possibility is to alter $\eta$ to directly assess the structural similarity of these sub-graphs by counting their common edges, as:

$$\xi : \eta_{uv} = \sum_{i,j \in u \cap v} A_{ij}$$

One can consider many other alternatives for measuring the overlaps based on the application at hand[8]. We revisit and delve deeper in this topic in Section 4, after providing an alternative formulation for the clustering distance or agreement measures.

[7] Which are measured based on the properties and attributes of data-points.

[8] Refer to the supplementary materials for more information, available at: `https://github.com/rabbanyk/CommunityEvaluation`




(a) Omega example: the extended pair-counts matrices of $U_1$ and $U_2$ with V are respectively $\{\{3,0,0\},\{1,1,1\},\{2,0,1\}\}$, and $\{\{3,0,0\},\{3,2,0\},\{0,2,0\}\}$. In the latter, the second row corresponds to the pairs of nodes which are together in one cluster in $V$; where 3, 2, and 0 of them are respectively clusters together in 0, 1, and 2 clusters of $U_2$.

(b) Matching example: using the original formulation we have $NMI(U_1, V) = 0.78$ and $NMI(U_2, V) = 0.71$; whereas the overlapping version results in $NMI'(U_1, V) = 0.61$ and $NMI'(U_2, V) = 0.62$ with Lancichinetti et al. (2008a) extension and $NMI''(U_1, V) = 0.53$ and $NMI''(U_2, V) = 0.61$ with the extension proposed by McDaid et al. (2011).

Fig. 2: Example for **the limitation of Omega index** on the left: the pair-counts table for $U_1$ and $U_2$ with V have the same trace, and therefore $U_1$ and $U_2$ have the same agreement with $V$ according to the Omega index; whereas $U_2$ should have been ranked more similar. Example for **the problem of matching** on the right: using the set matching based measures, such as the overlapping version of the $NMI$, clustering $U_2$ is in higher agreement with $V$, while the non-overlapping version of $NMI$ suggests the opposite[9].

### 3.2 Extension for Overlapping Clusters

There are several non-trivial extensions of the clustering agreement measures for the crisp overlapping clusters. Notably, Collins and Dent (1988) proposed the *Omega index* as a generalization of the (adjusted) rand index; which expands the $2\times2$ pair-counts table $\{\{M_{00}, M_{10}\}, \{M_{01}, M_{11}\}\}$; i.e. $M_{ij}$ counts the pairs of data-points which appeared together in $i$ clusters of $U$ and $j$ clusters of $V$. Similar to the $RI$, trace of this matrix, i.e. $\sum_i M_{ii}$, gives the agreement index, which is further adjusted for chance using the marginals of $M$. The Omega index reduces to the $(A)RI$ if the clusterings are disjoint. However, it only considers the pairs that appeared in the *exact* same number of clusters together, which is not ideal. Figure 2a provides an example for the issue of this limitation. Another commonly used measure is *the overlapping extension of* $NMI$ proposed by Lancichinetti et al. (2008a). This extension does not reduce to the original $NMI$ if the clusterings are disjoint. Moreover, it assumes a matching between clusters in $U$ and $V$, and only compares the best matched clusters. Therefore, it suffers from the "problem of matching" (Meilă, 2007), which is an inherent problem for any index defined based on the best matching of clusters[10]; Figure 2b gives a visualized example.

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 (e.g. $V$ in Figure 3); which are common in clusters in networks, a.k.a communities. However, the bonding concept presented by Brouwer (2008) is similar to the main idea behind our extension for overlapping cases, which we introduce in the next section.

[9] Here we used a disjoint example to be able to compare the results quantitatively with the original $NMI$; the same problem however exists for any matching based measure, regardless of the overlapping or disjoint.

[10] Other examples of matching based agreements are the Balanced Error Rate with alignment, average $F1$ score, and Recall measures used in (Yang and Leskovec, 2013; Mcauley and Leskovec, 2014; McDaid et al., 2011).

The extension of the proposed $\mathcal{D}$ formula (Definitions 1, 2, and 3) for overlapping clusters is not straightforward. The $(\mathcal{A}/\mathcal{N})\mathcal{D}$ formula is indeed bounded for overlapping clusters, and reduces to the original formulation if we have disjoint covering clusters. However, the current formulation is not appropriate for comparing overlapping clusters, since it treats overlaps as variations and penalizes them. Consider an extreme example when we are comparing two identical clusterings, and therefore we should have $(\mathcal{A}/\mathcal{N})\mathcal{D} = 0$ (i.e. the perfect agreement); this is true if there is no overlapping nodes, however as the number of overlapping nodes increases, $(\mathcal{A}/\mathcal{N})\mathcal{D}$ also increases (i.e. the agreement decreases). This is an inherent problem in any agreement measure formulated based on the contingency/confusion table, since overlaps in the data are confused with the overlaps between the matched clusters. We overcome this problem by proposing an alternative formulation for the clustering agreement measures, presented in the next section. More generally, the difficulty of computing the agreement of different clusterings, and in particular their extension for general cases such as overlapping clusters, comes from the fact that there is no matching between the clusters from the two clusterings. Therefore, one should consider all the permutations (e.g. using the contingency table), or only consider the best matching, which is cursed with the "problem of matching" as discussed earlier. Alternatively, we propose an algebraic formulation which takes the permutation out of the equation.

## 4 Algebraic Formulation for Clustering Distance

We first show that the proposed generalized formulae in Section 2 can be reformulated in terms of matrices. These formulae are defined based on the contingency table of $U$ and $V$, which we obtain from the $\eta$ overlap function. We can denote this contingency table with a $k \times r$ matrix, i.e. $N_{k\times r}$. Then, we can rewrite $\mathcal{D}$ (Definition 1) and $\mathcal{ND}$ (Definition 2) as follows:

$$\mathcal{D}_\varphi = \left[\mathbf{1}\varphi(N\mathbf{1}^T) - \mathbf{1}\varphi(N)\mathbf{1}^T\right] + \left[\varphi(\mathbf{1}N)\mathbf{1}^T - \mathbf{1}\varphi(N)\mathbf{1}^T\right], \quad \mathcal{ND}_\varphi = \frac{\mathcal{D}_\varphi}{\varphi(\mathbf{1}N\mathbf{1}^T)}$$

where $\mathbf{1}$ is a vector of ones with appropriate shape so that the matrix-vector product is valid, i.e. $\mathbf{1}N = [n_{.1}, n_{.2}, \dots n_{.r}]$, and $N\mathbf{1}^T = [n_{1.}, n_{2.}, \dots n_{k.}]^T$; and $\varphi$ is applied element-wise to the given matrix. In the same manner, we can reformulate $\mathcal{AD}$ (Definition 3) as:

$$\mathcal{AD}_\varphi = \frac{\mathcal{D}_\varphi}{\frac{1}{2}[\mathbf{1}\varphi(N\mathbf{1}^T) + \varphi(\mathbf{1}N)\mathbf{1}^T] - E}, \qquad E = \mathbf{1}\varphi(\frac{(N\mathbf{1}^T) \times (\mathbf{1}N)}{\mathbf{1}N\mathbf{1}^T})\mathbf{1}^T$$

Naturally, the identities proved in Section 2 still hold; for example, when $\eta$ gives the overlap sizes, the normalized Variation of Information derives from $\mathcal{D}_{\varphi(x)=x\log x}$ (Identity 1); and $1 - \mathcal{ND}_{\varphi(x)=\binom{x}{2}}$ is equivalent to the rand index (Identity 2).

These formulations based on the contingency matrix, as also discussed in Section 3.2, are only appropriate for disjoint clusters. Therefore in the following, we propose another reformulation, which is not defined based on the contingency matrix, and is valid for both disjoint and overlapping cases. In more details, let $U_{n\times k}$ denote a general representation for a clustering of a dataset with $n$ data-points;

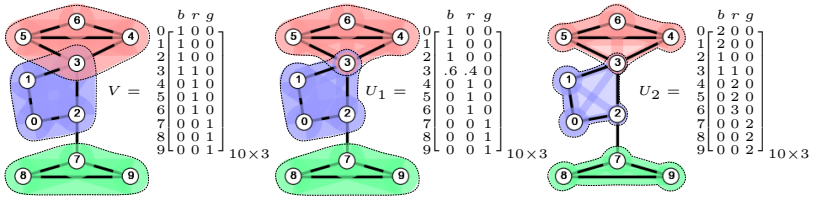



Fig. 3: Example of general matrix representation for a clustering: $V$ and $U_1$ are the classic overlapping clusters with crisp, and soft memberships respectively. Node 3 fully belongs to both blue and red clusters in $V$, wherein $U_1$, it belongs 60% to the blue cluster and 40% to the red cluster. This representation is general in a sense that it could encode membership of nodes to clusters in any form, with no assumptions on the matrix.

i.e. $u_{ik}$ represents the memberships of node $i$ in the $k^{th}$ cluster of $U$. Different constraints on this representation derive different cases of clustering[11]; see Figure 3 for examples. Then, the contingency matrix of $U$ and $V$ obtained by $\eta = |\cap|$ (e.g. in Identity 1), which indicates the size of overlaps between all pairs of clusters in $U_{d\times k}$ and $V_{d\times r}$, can also be derived from:

$$N = (U^T V)_{k\times r} = (V^T U)^T_{k\times r}$$

On the other hand, there is an analogy between co-membership and overlap, i.e. $(UU^T)_{ij}$ denotes in how many clusters node $i$ and $j$ appeared together, and $(U^T U)_{ij}$ denotes how many nodes clusters $i$ and $j$ have in common. Inspired by this analogy, we propose to measure the distance between clusterings directly by comparing their co-membership matrices, i.e. $(UU^T)_{n\times n}$ *v.s.* $(VV^T)_{n\times n}$, instead of their contingency/overlap matrix, i.e. $(U^T V)_{k\times r}$. More specifically, we consider the clustering co-membership difference as follows, and then show that both the rand index ($RI$) and the adjusted rand index ($ARI$) derive from different normalization of this difference.

**Definition 4** Clustering Co-Membership Difference Matrix ($\Delta$)

$$\Delta(U, V) = (UU^T - VV^T)_{n\times n}$$

To calculate the distance between $U$ and $V$, we need to quantify $\Delta$ using a matrix function: $\mathbb{R}^{n\times n} \to \mathbb{R}$, e.g. a matrix norm. In particular, the $RI$ and $ARI$ are different normalized form of $\Delta$ when we use the Frobenius matrix norm, i.e.

**Identity 5** *The Rand Index of Equation 1 derives from $\Delta$, i.e.*

$$\frac{\|\Delta(U,V)\|_F^2}{\mathfrak{m}\times n(n-1)} \equiv 1 - RI(U,V)\,, \quad \textit{also} \quad \frac{\|\Delta(U,V)\|_F^2}{\mathfrak{m}\times n^2} \equiv 1 - RI'(U,V)$$

*where $\|.\|_F^2$ sums the squared values of the given matrix, a.k.a. squared Frobenius norm; and $\mathfrak{m} = [max(max(UU^T), max(VV^T))]^2$, which is equal to one for disjoint clusters.*

[11] For crisp clusters (a.k.a strict membership), $u_{ik}$ is restricted to $0, 1$ (1 if node $i$ belongs to cluster $k$ and 0 otherwise); whereas for probabilistic clusters (or soft membership), $u_{ik}$ could be any real number in $[0, 1]$. Fuzzy clusters usually assume an additional constraint that the total membership of a data-point is equal to one, i.e. $u_{i.} = \sum_k u_{ik} = 1$. Which should also be true for disjoint clusters, since each data-point can only belong to one cluster.

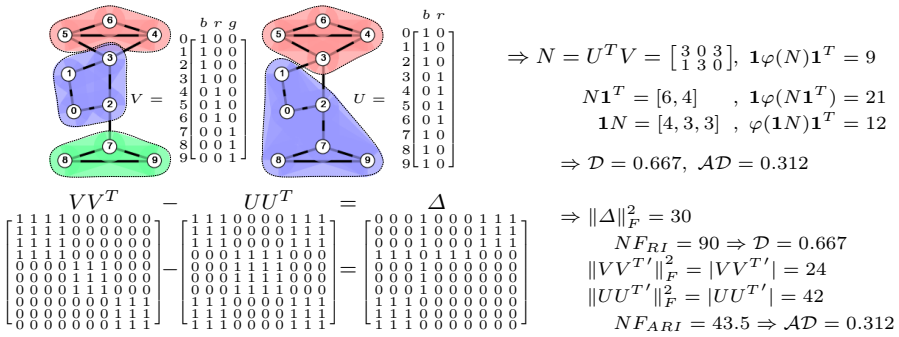



Fig. 4: Example for contingency v.s. co-membership based formulation. The $(A)RI$ is first derived from the contingency table $N$, using $\mathcal{D}$ formula where $\varphi(x) = x(x-1)/2$. Then same results are derived from the comparison of co-membership matrices $UU^T$ and $VV^T$, using the alternative formulation of $\mathcal{D}$, where $A'_{n\times n} = A - \mathbf{I}_n$.

The normalization factor for the Rand Index in the Identity 5, assumes an unlikely worse case scenario when all pairs are in disagreements. The $ARI$ normalization in Identity 6, on the other hand, adopts the expected difference when $UU^T$ and $VV^T$ are independent; refer to the Appendix A.6 for details on the derivation of these normalizing factors and proofs of these two Identities.

**Identity 6** *The Adjusted Rand Index of Equation 6 derives from $\Delta$, i.e.*

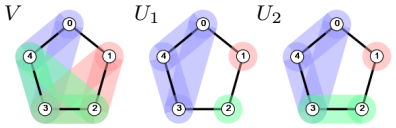
$$\frac{\|\Delta(U,V)\|_F^2}{\|UU^T\|_F^2 + \|VV^T\|_F^2 - 2(|UU^T||VV^T|)/n^2} \equiv 1 - ARI'(U,V)$$

*where $|.|$ is the sum of all elements in the matrix.*

The $ARI$ of Equation 5 derives from the same formula, if we set the diagonal elements of the co-membership matrices to zero, i.e. $(UU^T)' = UU^T - \mathbf{I}_n$. Since the original $ARI$ formula counts only the co-memberships of different nodes, i.e. $(i, j)$ where $i \neq j$; whereas, $ARI'$ also considers the co-memberships for each single node with itself in different clusters, which is more suitable for overlapping cases.

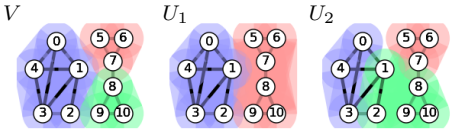
The $\Delta$-based formulations for $RI$ (Identity 5) and $ARI$ (Identity 6), denoted respectively by $RI_\delta$ and $ARI_\delta$ hereafter, not only are identical to the original formulations if the clusterings are disjoint (see Figure 4 for an example), but are also valid for overlapping cases (see Figure 5 for examples). Unlike the prior contingency based formulations (i.e. Definition 1, 2, and 3), $RI_\delta$ and $ARI_\delta$ do not need to consider all permutations of the matched clusters using their pair-wise overlaps, and hence will not confuse the natural overlaps in the data with the overlaps of matched clusters used to compute the agreements. For the extreme example discussed earlier in Section 3.2, $RI_\delta$ and $ARI_\delta$ always return 1 if the clusterings are identical, regardless of the amount of the overlapping nodes. Figure 5 shows the other two test case examples from Section 3.2, and compares the results from $RI_\delta$ and $ARI_\delta$ with the other alternative overlapping measures (i.e. *Omega* index and two overlapping versions of $NMI$); where unlike these alternatives, the new $\delta$-based measures rank the agreements correctly.

It is worth mentioning that, the Omega Index($\omega$) (Collins and Dent, 1988) can also be derived from comparing the co-membership matrices. In more details, if




| | $\omega$ | $A\omega$ | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ |
|---|---|---|---|---|---|---|
| $(V, U_1)$ | 0.5 | 0.22 | 0.80 | 0.25 | 0.90 | 0.32 |
| $(V, U_2)$ | 0.5 | 0.19 | 0.88 | 0.49 | 0.94 | 0.58 |

(a) Revisit to Figure 2a. Reported in the table are values for Omega index ($\omega$), and its adjusted version ($A\omega$), followed by our exact and approximate (marked by $'$) $\delta$-based $(A)RI$, derived from the proposed clustering co-memberships distance $\Delta$.




| | $NMI$ | $NMI'$ | $NMI''$ | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ |
|---|---|---|---|---|---|---|---|
| $(V, U_1)$ | 0.78 | 0.61 | 0.53 | 0.84 | 0.66 | 0.85 | 0.70 |
| $(V, U_2)$ | 0.71 | 0.62 | 0.61 | 0.78 | 0.47 | 0.80 | 0.57 |

(b) Revisit to Figure 2b. Here our exact and approximate co-membership based formulations are in agreement with the original non-overlapping $NMI$, and give a higher similarity score to $U_1$. Whereas the two overlapping $NMI$ extensions state the opposite.

Fig. 5: Revisit to the examples of Figure 2. On the left we see that Omega index ($\omega$) is unable to differentiate between $U_1$ and $U_2$, whereas its adjusted version even gives higher score to $U_1$, which is the opposite of what we expect. The fact that $U_2$ is more similar to $V$ is captured by our $\delta$-based $(A)RI$. On the right we see an example of disagreement between the original $NMI$ and its two set-matching based extensions for overlapping cases. Here since the problem is disjoint, $(A)RI_\delta$ gives same results as the original $(A)RI$.

we define $\Omega = [UU^T == VV^T]$, i.e. $\Omega_{ij} = 1$ if $(UU^T)_{ij} == (VV^T)_{ij}$ and zero otherwise; and assuming $f_A(i)$ denotes the frequency of value $i$ in matrix $A$, then $\omega$ and its adjusted version $(A\omega)$ can be calculated as:

$$\omega = |\Omega| - tr(\Omega), \quad A\omega = \frac{\omega - E[\omega]}{1 - E[\omega]}, \quad where \quad E[\omega] = \sum_{i=0}^{min(r,k)} f_{UU^T}(i) f_{VV^T}(i)$$

Similarly, we can compare the co-membership matrices of $UU^T$ and $VV^T$ in other way, e.g. using matrix divergences (Dhillon and Tropp, 2007; Kulis et al., 2009); or considering other normalized[12] forms of $\Delta$ . In our experiments, we examine these two variations:

$$\mathcal{D}_{norm} = \frac{\|UU^T - VV^T\|_F^2}{\|UU^T\|_F^2 + \|VV^T\|_F^2}, \quad I_{\sqrt{tr}} = \frac{tr(UU^T VV^T)}{\sqrt{tr((UU^T)^2) tr((VV^T)^2)}} = \frac{|UU^T \circ VV^T|}{\|UU^T\|_F^2 \|VV^T\|_F^2}$$

We conclude this section by presenting the extension of these algebric reformulations for *network clustering*. Let $N$ denote the structure of the graph $G$ as an incidence matrix, i.e. $N_{ik} = \sqrt{A_{ij}}$ if node $i$ is incident with edge $k = (i, j)$, and zero otherwise. Assuming a clustering as a transformation which assigns each data-point to one of its $k$ clusters, i.e. $U : n \mapsto k$ , we can incorporate the structure by measuring the distance between the transformed data by $U$ and $V$ as:

$$\mathcal{D}_\perp(U, V|G) = \mathcal{D}(N^T U, N^T V)$$

Figure 6 provides an intuitive example for this transformation. This transformation generates overlaps, hence only the new reformulations, which are valid for overlapping clusters, are applicable as $\mathcal{D}$; e.g. the $ARI_\delta$. One should also note that, this is very similar to counting the edges using overlap function $\xi$ introduced earlier in Section 3.1.

[12] It is also worth pointing out that in some applications, such as ensemble or multi-view clustering, we may not need the normalization and a measure of distance may suffice.

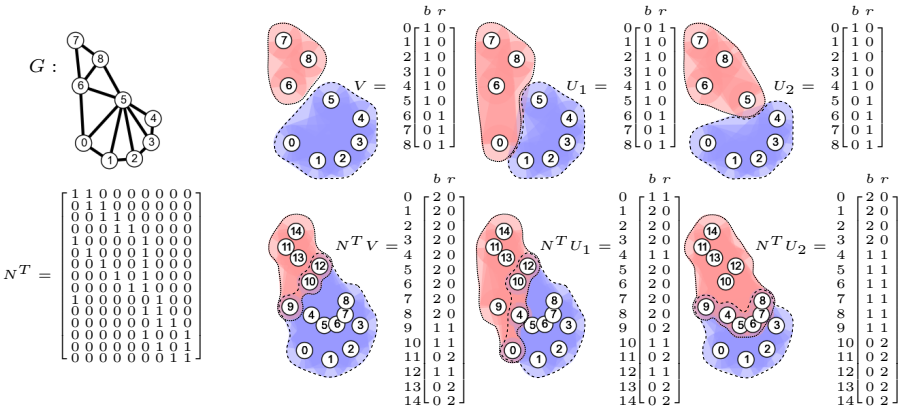


Fig. 6: A revisits to the example of Figure 1. Top) In the original data and considering only nodes, $U_1$ and $U_2$ have the same agreement with $V$. Since both $U_1$ and $U_2$ have one node clustered differently than $V$. Bottom) Transformed data using corresponding clusterings correctly identifies that $U_1$ is closer to $V$ compared to $U_2$. Note that the transformed data is similar to the line graph (edges as nodes) of the original data.

| $I:$ | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{tr}}$ |
|---|---|---|---|---|---|---|
| $(V, U_1)$ | 0.778 | 0.556 | 0.802 | 0.604 | 0.695 | 0.815 |
| $(V, U_2)$ | 0.778 | 0.556 | 0.802 | 0.604 | 0.695 | 0.815 |
| $C_\perp(V, U_1\|G)$ | 0.926 | 0.744 | 0.928 | 0.752 | 0.799 | 0.923 |
| $C_\perp(V, U_2\|G)$ | 0.857 | 0.417 | 0.859 | 0.435 | 0.708 | 0.844 |
| $C_+(V, U_1\|G)$ | 0.889 | 0.773 | 0.901 | 0.797 | 0.843 | 0.904 |
| $C_+(V, U_2\|G)$ | 0.833 | 0.660 | 0.900 | 0.776 | 0.832 | 0.885 |
| $(N, V)$ | 0.750 | 0.500 | 0.979 | 0.327 | 0.512 | 0.662 |
| $(N, U_1)$ | 0.750 | 0.491 | 0.979 | 0.337 | 0.503 | 0.668 |
| $(N, U_2)$ | 0.639 | 0.264 | 0.977 | 0.275 | 0.481 | 0.616 |

Table 1: Results of different agreement measures for the test case of Figure 1 (and Figure 6). For example looking at $ARI'_\delta$, from the structure independent version we have $ARI'_\delta(V, U_1) = ARI'_\delta(V, U_2) = 0.604$; whereas when considering the structure, both $C_\perp ARI'_\delta$ and $C_+ ARI'_\delta$ rank $U_1$ in higher agreement with $V$ compared to $U_2$, i.e. $C_\perp ARI'_\delta(V, U_1|G) = 0.752 > C_\perp ARI'_\delta(V, U_2|G) = 0.435$ and $C_+ ARI'_\delta(V, U_1|G) = 0.797 > C_+ ARI'_\delta(V, U_2|G) = 0.776$.

Alternatively, we can assume each edge as a cluster of two nodes, and measure the distance of a clustering from the underlying structure of the graph. Consequently, the structure dependent distance of $U$ and $V$ can be defined as a combination of $\mathcal{D}(U, N)$, $\mathcal{D}(V, N)$ and $\mathcal{D}(U, V)$, for example:

$$\mathcal{D}_+(U, V|G) = \alpha\mathcal{D}(U, V) + (1 - \alpha)|\mathcal{D}(U, N) - \mathcal{D}(V, N)|, \quad \alpha = 0.5$$

Table 1, Table 2 and Table 3 compare the structure dependent and independent measures for our earlier test case examples in Figure 1, Figure 2a, and Figure 2b, respectively. The first two rows of these tables show the structure independent measures, the next four rows are the structure based versions, and the last three rows show the agreement of each clustering directly with the structure.

In particular in Table 1, we see that unlike the original structure independent measures which result in the same agreement for $U_1$ and $U_2$, all the structure based extensions correctly give higher agreement score to $U_1$ compared to $U_2$. We can also see that $U_1$, when compared to $U_2$, has in fact more agreement with the structure of the underlying graph.

| $I:$ | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{tr}}$ |
|---|---|---|---|---|---|---|
| $(V, U_1)$ | 0.800 | 0.245 | 0.902 | 0.318 | 0.532 | 0.764 |
| $(V, U_2)$ | 0.875 | 0.490 | 0.942 | 0.577 | 0.663 | 0.894 |
| $C_\perp(V, U_1\|G)$ | 0.856 | 0.186 | 0.868 | 0.211 | 0.536 | 0.860 |
| $C_\perp(V, U_2\|G)$ | 0.913 | 0.427 | 0.924 | 0.483 | 0.672 | 0.961 |
| $C_+(V, U_1\|G)$ | 0.775 | 0.556 | 0.919 | 0.617 | 0.720 | 0.859 |
| $C_+(V, U_2\|G)$ | 0.863 | 0.712 | 0.954 | 0.765 | 0.824 | 0.945 |
| $(N, V)$ | 0.850 | 0.333 | 0.933 | 0.528 | 0.682 | 0.816 |
| $(N, U_1)$ | 0.600 | 0.200 | 0.870 | 0.444 | 0.590 | 0.771 |
| $(N, U_2)$ | 0.700 | 0.400 | 0.900 | 0.576 | 0.666 | 0.822 |

Table 2: Results of different agreements for the omega example of Figure 2a.

| $I:$ | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{tr}}$ |
|---|---|---|---|---|---|---|
| $(V, U_1)$ | 0.836 | 0.660 | 0.851 | 0.703 | 0.705 | 0.840 |
| $(V, U_2)$ | 0.782 | 0.471 | 0.802 | 0.567 | 0.626 | 0.721 |
| $C_\perp(V, U_1\|G)$ | 0.900 | 0.790 | 0.906 | 0.806 | 0.768 | 0.902 |
| $C_\perp(V, U_2\|G)$ | 0.857 | 0.564 | 0.862 | 0.607 | 0.667 | 0.798 |
| $C_+(V, U_1\|G)$ | 0.855 | 0.708 | 0.922 | 0.793 | 0.839 | 0.866 |
| $C_+(V, U_2\|G)$ | 0.818 | 0.556 | 0.897 | 0.716 | 0.782 | 0.804 |
| $(N, V)$ | 0.945 | 0.865 | 0.977 | 0.620 | 0.615 | 0.814 |
| $(N, U_1)$ | 0.818 | 0.621 | 0.970 | 0.502 | 0.589 | 0.707 |
| $(N, U_2)$ | 0.800 | 0.506 | 0.968 | 0.485 | 0.552 | 0.702 |

Table 3: Results of different agreements for the matching example of Figure 2b.

Table 2 and Table 3 extend the results presented in Figure 5 for the two overlapping test case examples. Here we see that the results of the structure dependent measures are consistent with the structure independent measures; however, the structure dependent agreements become stronger than the independent versions in Table 3, while getting weaker in Table 2. This is due to the fact that the clusterings in Figure 5b better correspond with the structure of the underling graph, when compared to the clusterings of Figure 5a.

## 5 Experimental Results on Community Mining Evaluation

Here, we examine the clustering agreement measures, introduced in this paper, in the context of community mining evaluation; which is their most common application. More specifically, we select a set of common community mining methods, which discover clusters in a given network based on different methodologies. We then rank their performance according to different clustering agreement measures; which compare the results of these methods with the ground-truth clustering. However, the purpose here is not to compare the general performance of community mining methods, but rather to show different comparisons/rankings we obtain using different agreement measures.

In more details, datasets are generated using LFR (Lancichinetti et al., 2008b) benchmarks[13], which are commonly used in the evaluation and comparison of community mining algorithms. The selected methods are: Louvain (Blondel et al., 2008), WalkTrap (Pons and Latapy, 2005), PottsModel (Ronhovde and Nussinov,

[13] Parameters are chosen similar to the experiments by Lancichinetti and Fortunato (2009), i.e. networks with 1000 nodes, average degree of 20, max degree of 50, and power law degree exponent of -2; where the size of communities follows a power law distribution with exponent of -1, and ranges between 20 to 100 nodes. Results for other parameter settings, including smaller sized communities, 10 to 50, could be found in the supplementary materials.

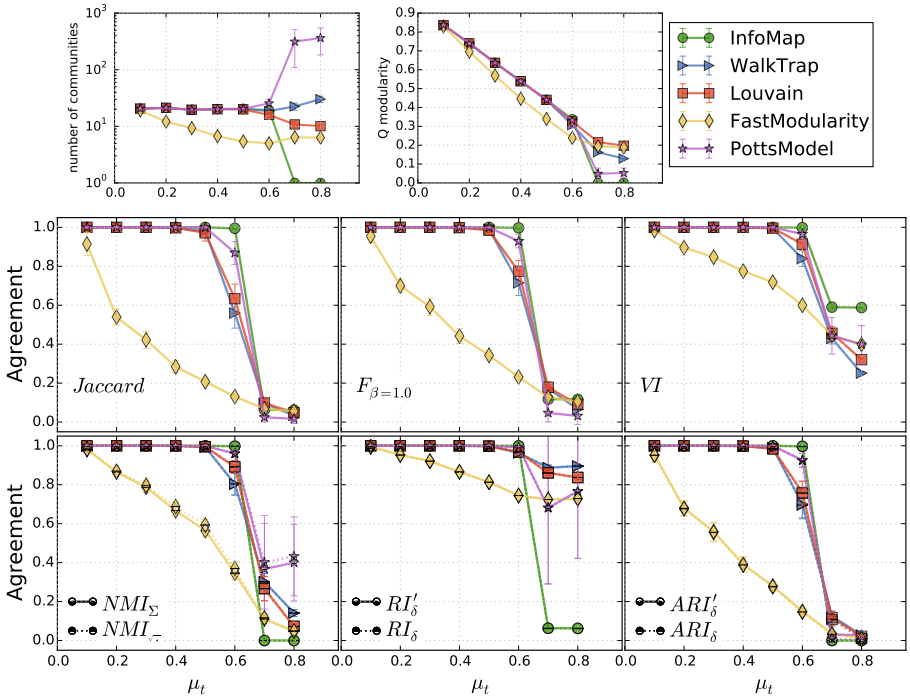



Fig. 7: The agreement of results from different community detection algorithms with the ground-truth in *unweighted LFR benchmarks*, plotted as a function of the mixing parameter ($\mu_t$). In the last three plots, similar measures are overlaid to show they are highly similar.

2009), FastModularity (Newman, 2004), and InfoMap (Rosvall and Bergstrom, 2008) for disjoint clusterings of Section 5.1 and 5.2; and COPRA (Gregory, 2010), MOSES (McDaid and Hurley, 2010), OSLOM (Lancichinetti et al., 2011), and BIGCLAM (Yang and Leskovec, 2013) for overlapping clusterings in Section 5.3. The authors' original implementations are used for the methods, with no parameter tuning (defaults are used); and the reported results are averaged over ten runs.

### 5.1 Classic Measures

Figure 7 shows the rankings of the selected algorithms with respect to six common agreement measures[14]. *First*, we can see that the rankings are overall consistent, which is expected since these indices are measuring the agreement with similar principle, as shown with our generalizations. *Second*, from the plot for $NMI$ we can observe its bias in favor of the large number of clusters (Vinh et al., 2009); i.e. for large mixing parameters, when PottsModel algorithms detects significantly more communities, $NMI_{\sum}$ and $NMI_{\sqrt{\cdot}}$ rank the PottsModel significantly higher; which is not true according to all the other measures. The opposite bias is also observed in the plots of $VI$, where the algorithm that finds significantly less communities is ranked significantly higher, i.e. Infomap. Based on these observation, we advise against using these two measures, particularly if the number of discovered clusters/communities might be very different from the ground-truth. *Third*,

[14] Similar trends are observed for other variations of the agreement measures which can be found in the supplementary materials.

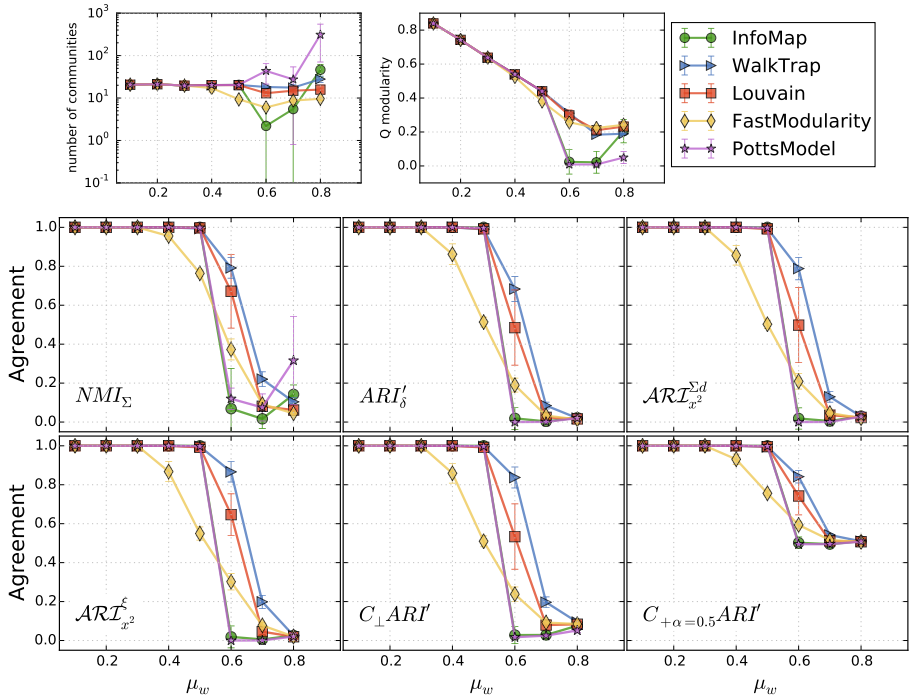



Fig. 8: Comparison of the agreement indexes on *weighted LFR benchmark*, when the mixing parameter for weights varies and the mixing parameter for topology is fixed to 0.5.

from the plot for $ARI$ we can see that there is no clear difference between the rankings obtained by $ARI$ of Equation 5 and Equation 6, which are plotted as $ARI_\delta$ and $ARI'_\delta$ respectively[15]. The latter is less commonly used, whilst its extended form presented in Section 4 is more appropriate for overlapping cases; hence we recommend using $ARI'$ in general.

### 5.2 Structure Dependent Measures

Figure 8 compares the selected methods over *weighted* LFR benchmarks. These methods all result in clusterings which correspond well with the underlying structure, therefore the effect of ignoring the structure in comparing the clusterings is less apparent. The rankings are overall consistent, however, we can still observe the difference between the structure dependent and independent agreement measures, which became clear only with the presence of weights. In more details, we can see that the Walktrap method is performing better according to all the measures, however its superiority is more significant according to the structure dependent measures: i.e. $\mathcal{ARI}^{\xi}_{x^2}$ , and $\mathcal{ARI}^{\Sigma d}_{x^2}$ introduced in Section 3.1, and $C_\perp ARI'$ introduced in Section 4. On the other hand, $C_+ ARI'$ does not decrease to zero similar to the other measures; since both the results and the ground-truth are with the same distance from the structure. Being less consistent with the other measures, we recommend using the three former structure dependent forms.

[15] The $\delta$ subscript indicates that the $ARI$ is computed based on our $\delta$-based formulation, which is equivalent to the original $ARI$ in this experiment, since communities are covering all nodes and non-overlapping (Identity 6).

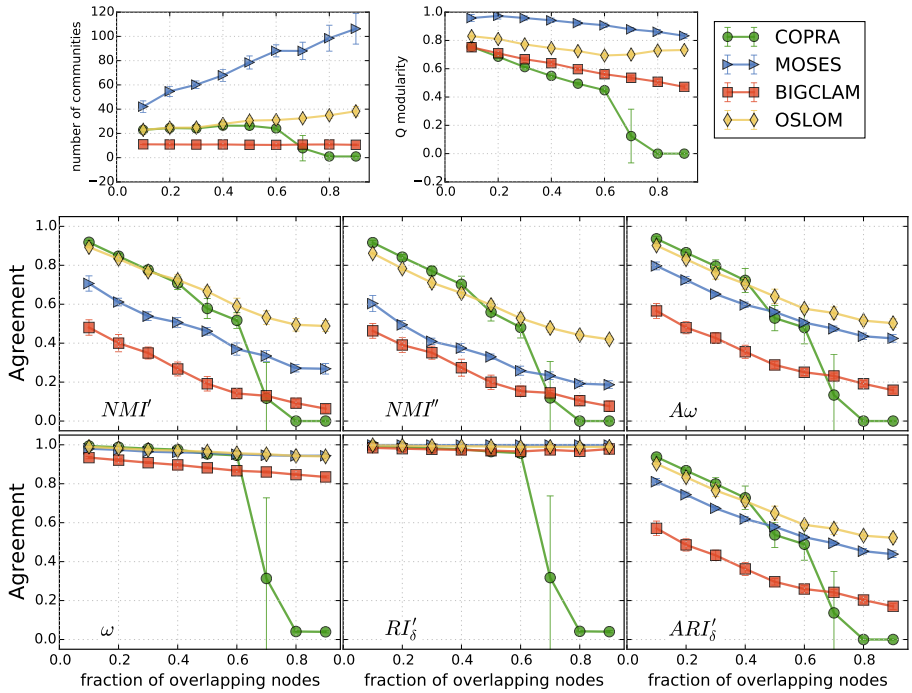



Fig. 9: Comparison of agreement indexes on unweighted *overlapping LFR benchmark*, where the fraction of overlapping nodes varies, the mixing parameter for topology is fixed to 0.1, and the maximum number of communities a node can belong to is limited to 2; similar to experiments in (Lancichinetti et al., 2011).

### 5.3 Overlapping Measures

Figure 9 shows the comparison of the selected methods based on the different overlapping agreement indexes, which are: the overlapping extensions of NMI, i.e. $NMI'$ by Lancichinetti et al. (2008a) and $NMI''$ by McDaid et al. (2011); the omega index ($\omega$), and its adjusted version ($A\omega$); and our $\delta$-based formulations for the $RI$ and $ARI$, i.e. $RI'_\delta$, and $ARI'_\delta$, which naturally extend to overlapping cases. Obtained rankings are generally consistent, similar to the previous experiments. *First,* we can see that the unadjusted measures, i.e. $\omega$ and $RI'_\delta$, can not properly differentiate between the different algorithms, which makes them unfavorable. *Secondly,* we observe the "problem of matching" with the overlapping extensions of $NMI$, described earlier in Section 3.2. In more details, MOSES algorithm results in finer grained and thus more communities, which are more likely to not get matched/compared with the communities in the ground-truth, when applying a set-matching based agreement measure. Therefore it gets unfairly penalized, and is ranked significantly lower than the OSLOM. Their difference, however, is less significant according to both adjusted omega, $A\omega$, and our overlapping extension of $ARI$, $ARI'_\delta$; whereas the quality of the MOSES results are higher than OSLOM according to the Q-modularity of Newman (2004). *Lastly,* in this setting the difference between $A\omega$ and $ARI'_\delta$ is not as clear as it could be, since each node can only belong to at most two communities; whereas the difference becomes clear if a node can belong to many communities, refer to Section 4 for more details.

## 6 Conclusions and Recommendations

In this paper, we presented a generalized clustering distance, from which we can derive the two commonly used clustering agreement measures, i.e. $NMI$ and $ARI$. Not only this generalization sheds light on the relation between these two measures, but we also recommend using the derived formulae from this distance over the original formulations; since, *first*, they are identical when the original measures are defined; *second*, they require less assumptions on the clusterings and hence apply to more general cases, e.g. when there are un-clustered data-points; *third*, they can be easily altered, for example to generate specific measures for clusters in networks. The latter example is in particular important, since all of the current agreement measures overlook the structure of the data, and hence are not appropriate for comparing clusters over networks, a.k.a. communities. Using our generalization, we introduced two extensions of the $ARI$, which incorporate the structure when comparing communities, i.e. $\mathcal{ARI}_{x^2}^{\Sigma d}$ (degree weighted overlap function) and $\mathcal{ARI}_{x^2}^{\xi}$ (edge counting overlap function). We recommend using these two extensions when comparing disjoint communities.

The generalized clustering distance, similar to other contingency based measure, does not readily extend to overlapping cases. Therefore, we presented an algebraic reformulation for the $ARI$, based on the difference of co-membership matrices of the two clusterings, denoted by $ARI'_\delta$. We recommend using $ARI'_\delta$, in particular when clusters are overlapping; since, *first*, it is identical to the original measure if clusters are disjoint; *second*, it naturally extends to overlapping cases; *third*, it is more valid compared to the current alternative overlapping measures, i.e. it does not have the shortcomings of the overlapping extensions of $NMI$ or the Omega index. However, one should note that this formulation requires matrix representations, hence is harder to implement and computationally more expensive. We have provided efficient implementations, using sparse matrices, in the supplementary materials[16].

To also incorporate the structure within this algebraic reformulation, we proposed $C_\perp ARI'$, and $C_+ ARI'$. The former measures the distance between the transformed structure by each clustering; whereas the latter linearly combines the distances of each clustering to the structure, assuming the structure itself as another clustering, i.e. when each edge is considered as a single cluster. We recommend using $C_\perp ARI'$ when comparing overlapping communities, since it is more consistent with the other measures. However, this transformation requires matrix multiplications and hence is computationally expensive; whereas in our implementations, it is not as scalable as the other measures.

## References

Aggarwal CC, Reddy CK (2014) Data Clustering: Algorithms and Applications. CRC Press

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

[16] available at: `https://github.com/rabbanyk/CommunityEvaluation`

Anderson DT, Bezdek JC, Popescu M, Keller JM (2010) Comparing Fuzzy, Probabilistic, and Possibilistic Partitions. IEEE Transactions on Fuzzy Systems 18(5):906–918

Banerjee A, Merugu S, Dhillon IS, Ghosh J (2005) Clustering with Bregman Divergences. The Journal of Machine Learning Research 6:1705–1749

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008:P10,008+

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

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

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

Cortina-Borja M (2012) Handbook of parametric and nonparametric statistical procedures, 5th edn. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175(3):829–829

Cui Y, Fern X, Dy J (2007) Non-redundant multi-view clustering via orthogonalization. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp 133–142

Dhillon IS, Tropp JA (2007) Matrix nearness problems with bregman divergences. SIAM Journal on Matrix Analysis and Applications 29(4):1120–1146

Fortunato S (2010) Community detection in graphs. Physics Reports 486(35):75–174

Gregory S (2010) Finding overlapping communities in networks by label propagation. New Journal of Physics 12(10):103,018

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

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

Kulis B, Sustik MA, Dhillon IS (2009) Low-rank kernel learning with bregman matrix divergences. The Journal of Machine Learning Research 10:341–376

Lancichinetti A, Fortunato S (2009) Community detection algorithms: A comparative analysis. Physical Review E 80(5):056,117

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

Lancichinetti A, Fortunato S, Radicchi F (2008b) Benchmark graphs for testing community detection algorithms. Physical Review E 78(4):046,110

Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PloS one 6(4):e18,961

Light RJ, Margolin BH (1971) An analysis of variance for categorical data. Journal of the American Statistical Association 66(335):pp. 534–544

Manning CD, Raghavan P, Schtze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA

Mcauley J, Leskovec J (2014) Discovering social circles in ego networks. ACM Trans Knowl Discov Data 8(1):4:1–4:28

McDaid A, Hurley N (2010) Detecting highly overlapping communities with model-based overlapping seed expansion. In: Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, IEEE, pp 112–119

McDaid AF, Greene D, Hurley N (2011) Normalized mutual information to evaluate overlapping community finding algorithms. arXiv preprint arXiv:11102515

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

Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066,133

Nielsen F, Nock R (2014) On the chi square and higher-order chi distances for approximating f-divergences. Signal Processing Letters, IEEE 21(1):10–13

Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005, Springer, pp 284–293

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

Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. Physical Review E 80(1):016,109

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4):1118–1123

Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, ICML '09, pp 1073–1080

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

Warrens MJ (2008a) On similarity coefficients for 2x2 tables and correction for chance. Psychometrika 73:487–502

Warrens MJ (2008b) On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. Journal of Classification 25:177–183

Wu J, Xiong H, Chen J (2009) Adapting the right measures for k-means clustering. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '09, pp 877–886

Yang J, 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, ACM, pp 587–596

Zhou D, Li J, Zha H (2005) A new mallows distance based metric for comparing clusterings. In: Proceedings of the 22nd international conference on Machine learning, ACM, pp 1028–1035

## A Proofs

**A.1 Proof of Proposition 1:** From the definition of Variation of information we have:

$$VI(U,V) = H(U) + H(V) - 2I(U,V) = 2H(U,V) - H(U) - H(V) = \mathbf{H(V|U)} + \mathbf{H(U|V)}$$

On the other hand, we have:

$$\begin{aligned} RI(U,V) &\propto \frac{1}{n^2-n}(\sum_{i=1}^{k}[\sum_{j=1}^{r} n_{ij}^2 - (\sum_{j=1}^{r} n_{ij})^2] + \sum_{j=1}^{r}[\sum_{i=1}^{k} n_{ij}^2 - (\sum_{i=1}^{k} n_{ij})^2]) \\ &\overset{*}{\propto} \sum_{i=1}^{k}[E_j(n_{ij}^2) - E_j(n_{ij})^2] + \sum_{j=1}^{r}[E_i(n_{ij}^2) - E_i(n_{ij})^2] \\ &\overset{*}{\propto} \sum_{i=1}^{k} Var_j(n_{ij}) + \sum_{j=1}^{r} Var_i(n_{ij}) \quad \overset{**}{\propto} \quad \mathbf{Var(V|U)} + \mathbf{Var(U|V)} \qquad \square \end{aligned}$$

$(*)$ $E_j/Var_j$ shows the average/variance of values in the $j^{th}$ column of the contingency table.
$(**)$ The RI is in fact proportional to the average variance of rows/columns values in the contingency table, which we denote by conditional variance. For other forms of conditional variance for categorical data see Light and Margolin (1971).

**A.2 Proof of Corollary 1:** We first show that $0 \leq \mathcal{D}_\varphi^\eta(U||V)$ which also results in the lower bound 0 for $\mathcal{D}_\varphi^\eta(U,V)$ since, $\mathcal{D}_\varphi^\eta(U,V) = \mathcal{D}_\varphi^\eta(U||V) + \mathcal{D}_\varphi^\eta(V||U)$. From the superadditivity of $\varphi$ we have:

$$\sum_{u\in U} \varphi(\eta_{uv}) \leq \varphi(\sum_{u\in U} \eta_{uv}) \implies \sum_{v\in V} \left[ \varphi(\sum_{u\in U} \eta_{uv}) - \sum_{u\in U} \varphi(\eta_{uv}) \right] \geq 0 \implies \mathcal{D}_\varphi^\eta(\mathbf{U}||\mathbf{V}) \geq \mathbf{0}$$

Similarly for the upper bound, from positivity and super-additivity we get respectively:

$$\mathcal{D}_\varphi^\eta(U||V) = \sum_{v\in V} \varphi(\sum_{u\in U} \eta_{uv}) - \sum_{v\in V}\sum_{u\in U} \varphi(\eta_{uv}) \leq \sum_{v\in V} \varphi(\sum_{u\in U} \eta_{uv}) \leq \varphi(\sum_{v\in V}\sum_{u\in U} \eta_{uv})$$

**A.3 Proof of Identity 1:** The proof is elementary, if we write the definition for $\varphi = x\log x$, we get:

$$\begin{aligned} \mathcal{ND}_{x\log x}^{|\cap|}(U,V) &= \frac{\sum_{v\in V}\sum_{u\in U} |u\cap v| \left[\log(\sum_{u\in U} |u\cap v|) - \log(|u\cap v|)\right]}{\left(\sum_{v\in V}\sum_{u\in U} |u\cap v|\right)\log\left(\sum_{v\in V}\sum_{u\in U} |u\cap v|\right)} \\ &+ \frac{\sum_{u\in U}\sum_{v\in V} |u\cap v| \left[\log(\sum_{v\in V} |u\cap v|) - \log(|u\cap v|)\right]}{\left(\sum_{u\in U}\sum_{v\in V} |u\cap v|\right)\log\left(\sum_{u\in U}\sum_{v\in V} |u\cap v|\right)} \\ &\overset{*}{=} \frac{\sum_j^r \sum_i^k n_{ij}\left[\log(\sum_i^k n_{ij}) + \log(\sum_j^r n_{ij}) - 2\log(n_{ij})\right]}{(\sum_i^k \sum_j^r n_{ij})\log(\sum_i^k \sum_j^r n_{ij})} \\ &\overset{**}{=} \frac{1}{\log n}\sum_j^r\sum_i^k \frac{n_{ij}}{n}\log(\frac{n_{i.}n_{.j}}{n_{ij}^2}) = \frac{VI(U,V)}{\log n} \qquad \square \end{aligned}$$

$(*)$ slight change of notation, i.e. from $\sum_{u\in U}$ to $\sum_i^k$, $\sum_{v\in V}$ to $\sum_j^r$ and $|u\cap v|$ to $n_{ij}$.
$(**)$ assuming disjoint covering partitionings: $\sum_i^k\sum_j^r n_{ij} = n$, $\sum_i^k n_{ij} = n_{.j}$ and $\sum_j^r n_{ij} = n_{i.}$.

**A.4 Proof of Identity 2:** Similar to the previous proof from the definition we derive:

$$\begin{aligned} \mathcal{ND}_{\binom{x}{2}}^{|\cap|}(U,V) &\overset{*}{=} \frac{\sum_j^r\left[(\sum_i^k n_{ij})^2 - \sum_i^k n_{ij}^2\right] + \sum_i^k\left[(\sum_j^r n_{ij})^2 - \sum_j^r n_{ij}^2\right]}{(\sum_i^k\sum_j^r n_{ij})^2 - \sum_i^k\sum_j^r n_{ij}} \\ &\overset{**}{=} \frac{1}{n^2-n}[\sum_j^r (n_{.j})^2 + \sum_i^k (n_{i.})^2 - 2\sum_j^r\sum_i^k n_{ij}^2] = 1 - RI(U,V) \qquad \square \end{aligned}$$

$(*), (**)$ same as previous proof.

### A.5 Proof of Identity 3 and 4:

$$\mathcal{AD}^{\eta}_{\varphi} = \frac{\sum_{v\in V}\varphi(\eta_{.v}) + \sum_{u\in U}\varphi(\eta_{u.}) - 2\sum_{v\in V}\sum_{u\in U}\varphi(\eta_{uv})}{\sum_{v\in V}\varphi(\eta_{.v}) + \sum_{u\in U}\varphi(\eta_{u.}) - 2\sum_{u\in U}\sum_{v\in V}\varphi\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)}$$

$$\Rightarrow 1-\mathcal{AD}^{\eta}_{\varphi}(U,V) = \frac{\sum_{v\in V}\sum_{u\in U}\varphi(\eta_{uv}) - \sum_{u\in U}\sum_{v\in V}\varphi\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)}{\frac{1}{2}\left[\sum_{v\in V}\varphi(\eta_{.v}) + \sum_{u\in U}\varphi(\eta_{u.})\right] - \sum_{u\in U}\sum_{v\in V}\varphi\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)}$$

This formula resembles the adjuctment for chance in Equation 4, where the measure being adjusted is $\sum_{v\in V}\sum_{u\in U}\varphi(\eta_{uv})$, the upper bound used for it is $\frac{1}{2}[\sum_{v\in V}\varphi(\eta_{.v})+\sum_{u\in U}\varphi(\eta_{u.})]$, and the expectation is defined as:

$$E[\sum_{v\in V}\sum_{u\in U}\varphi(\eta_{uv})] = \sum_{u\in U}\sum_{v\in V}\varphi\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)$$

Now if we have $\varphi(xy)=\varphi(x)\varphi(y)$, which is true for $\varphi(x)=x^2$, we get:

$$E[\sum_{v\in V}\sum_{u\in U}\varphi(\eta_{uv})] = \sum_{u\in U}\sum_{v\in V}\frac{\varphi(\eta_{.v})\varphi(\eta_{u.})}{\varphi(\sum_{u\in U}\sum_{v\in V}\eta_{uv})} = \frac{\sum_{v\in V}\varphi(\eta_{.v})\sum_{u\in U}\varphi(\eta_{u.})}{\varphi(\sum_{u\in U}\sum_{v\in V}\eta_{uv})}$$

Using this expecation, if we substitute $\varphi=x^2$ we get the $ARI'$ of Equation 6, and using the $\varphi=\binom{x}{2}$ and the later reformulation of $E$, we get the original $ARI$ of Equation 5, as:

$$1-\mathcal{AD}^{|\cap|}_{\binom{x}{2}}(U,V) = \frac{\sum_{v\in V}\sum_{u\in U}\binom{|u\cap v|}{2} - E(\sum_{v\in V}\sum_{u\in U}\binom{|u\cap v|}{2})}{\frac{1}{2}\left[\sum_{v\in V}\binom{\sum_{u\in U}|u\cap v|}{2} + \sum_{u\in U}\binom{\sum_{v\in V}|u\cap v|}{2}\right] - E(\sum_{v\in V}\sum_{u\in U}\binom{|u\cap v|}{2})}$$

$$where \quad E(\sum_{v\in V}\sum_{u\in U}\binom{|u\cap v|}{2}) = \frac{\sum_{v\in V}\binom{\sum_{u\in U}|u\cap v|}{2}\sum_{u\in U}\binom{\sum_{v\in V}|u\cap v|}{2}}{\binom{n}{2}}$$

$$\Rightarrow 1-\mathcal{AD}^{|\cap|}_{\binom{x}{2}}(U,V) \overset{*,**}{=} \frac{\sum_j\sum_i\binom{n_{ij}}{2} - \sum_j\binom{n_{.j}}{2}\sum_i\binom{n_{i.}}{2}/\binom{n}{2}}{\frac{1}{2}\left[\sum_j\binom{n_{.j}}{2}+\sum_i\binom{n_{i.}}{2}\right] - \sum_j\binom{n_{.j}}{2}\sum_i\binom{n_{i.}}{2}/\binom{n}{2}} = ARI(U,V)\square$$

$(*),(**)$ same as proof of identity 1.

On the other hand for the $NMI$, we have:

$$1-\mathcal{AD}^{|\cap|}_{x\log x}(U,V) = \frac{\sum_{v\in V}\sum_{u\in U}n_{uv}\log n_{uv} - E(\sum_{v\in V}\sum_{u\in U}n_{uv}\log n_{uv})}{\frac{1}{2}\left[\sum_{v\in V}n_{.v}\log n_{.v} + \sum_{u\in U}n_{u.}\log n_{u.}\right] - E(\sum_{v\in V}\sum_{u\in U}n_{uv}\log n_{uv})}$$

$$where\ E(\sum_{v\in V}\sum_{u\in U}n_{uv}\log n_{uv}) = \sum_{u\in U}\sum_{v\in V}\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)\log\left(\frac{\eta_{.v}\eta_{u.}}{\sum_{u\in U}\sum_{v\in V}\eta_{uv}}\right)$$

$$\Rightarrow 1-\mathcal{AD}^{|\cap|}_{x\log x}(U,V) \overset{*,**}{=} \frac{\sum_j\sum_i n_{ij}\log n_{ij} - \sum_i\sum_j\frac{n_{.j}n_{i.}}{n}\log\frac{n_{.j}n_{i.}}{n}}{\frac{1}{2}\left[\sum_j n_{.j}\log n_{.j} + \sum_i n_{i.}\log n_{i.}\right] - \sum_i\sum_j\frac{n_{.j}n_{i.}}{n}\log\frac{n_{.j}n_{i.}}{n}}$$

$$= \frac{n\sum_j\sum_i\frac{n_{ij}}{n}\log\frac{n_{ij}}{n} + n\log n - \sum_i\sum_j\frac{n_{.j}n_{i.}}{n}[\log\frac{n_{.j}}{n}+\log\frac{n_{i.}}{n}+\log n]}{\frac{n}{2}\left[\sum_j\frac{n_{.j}}{n}\log\frac{n_{.j}}{n} + \sum_i\frac{n_{i.}}{n}\log\frac{n_{i.}}{n} + 2\log n\right] - \sum_i\sum_j\frac{n_{.j}n_{i.}}{n}[\log\frac{n_{.j}}{n}+\log\frac{n_{i.}}{n}+\log n]}$$

$$= \frac{-H(U,V)+\log n - \sum_i\frac{n_{i.}}{n}\sum_j\frac{n_{.j}}{n}\log\frac{n_{.j}}{n} + \sum_i\frac{n_{.j}}{n} - \sum_j\frac{n_{i.}}{n}\log\frac{n_{i.}}{n} - \sum_i\sum_j\frac{n_{.j}n_{i.}}{n^2}\log n}{\frac{1}{2}[-H(U)-H(V)] + \log n + \sum_i\frac{n_{i.}}{n}H(V) + \sum_i\frac{n_{.j}}{n}H(U) - \log n}$$

$$= \frac{-H(U,V)+H(V)+H(U)}{-\frac{1}{2}[H(U)+H(V)]+H(V)+H(U)} = \frac{I(U,V)}{\frac{1}{2}[H(U)+H(V)]} = NMI_{sum}(U,V) \quad \square$$

$(*),(**)$ same as proof of identity 1.

**A.6 Proof of Identity 5 and 6:** First we prove that in general cases we have:

$$\|UU^T - VV^T\|_F^2 = \|U^TU\|_F^2 + \|V^TV\|_F^2 - 2\|U^TV\|_F^2$$

where $\|.\|_F^2$ is squared Frob norm. This holds since we have:

$$\begin{aligned}\|UU^T - VV^T\|_F^2 &= \sum_{ij}(UU^T - VV^T)_{ij}^2 \\ &= \sum_{ij}(UU^T)_{ij}^2 + \sum_{ij}(VV^T)_{ij}^2 - 2\sum_{ij}(UU^T)_{ij}(VV^T)_{ij} \\ &= \|UU^T\|_F^2 + \|VV^T\|_F^2 - 2|UU^T \circ VV^T|\end{aligned}$$

Where the $\circ$ is element-wise matrix product, a.k.a. hadamard product, and $|.|$ is sum of all elements in the matrix[17]. The proof is complete with showing:

$$\begin{aligned}|UU^T \circ VV^T| &= tr((UU^T)^TVV^T) = tr(V^TUU^TV) = tr((U^TV)^TU^TV) = ||U^TV||_F^2 \\ ||UU^T||_F^2 &= tr((UU^T)^TUU^T) = tr(U^TUU^TU) = tr((U^TU)^TU^TU) = ||U^TU||_F^2\end{aligned}$$

Now, we can prove the identities for the cases of disjoint hard clusters, using the notation $n_{ij} = (U^TV)_{ij}$, we have $\|U^TV\|_F^2 = \sum_{ij} n_{ij}^2$ and:

$$\|U^TU\|_F^2 = \sum_{ij} < U_{.i}, U_{.j} >^2 = \sum_{ij}(\sum_k u_{ki}u_{kj})^2 \overset{*}{=} \sum_i(\sum_k u_{ki}^2)^2 \overset{**}{=} \sum_i(\sum_k u_{ki})^2 \overset{***}{=} \sum_i n_{i.}^2$$

$(*)$ with assumption that clusters are disjoint, $u_{ki}u_{kj}$ is only non-zero iff $i = j$
$(**)$ with the assumption that memberships are hard, $u_{ki}$ is either 0 or 1, therefore $u_{ki} = u_{ki}^2$
$(***)$ marginals of $N$ give cluster sizes in $U$ and $V$, i.e. $n_{i.} = \sum_j n_{ij} = \sum_k u_{ki} = |V_i|$

Therefore for disjoint hard clusters we get:

$$\|UU^T - VV^T\|_F^2 = \sum_i n_{i.}^2 + \sum_j n_{.j}^2 - 2\sum_{ij} n_{ij}^2$$

The *RI* normalization assumes that all pairs are in disagreement, i.e. $|\mathbf{1}_{n\times n}| = n^2$, since $max(UU^T) = 1$ and, $max(VV^T) = 1$. The *ARI* normalization compares $\Delta$ to the difference where the two random variable of $UU_{ij}^T$ and $VV_{ij}^T$ are independent, in which case we would have:

$$E(UU_{ij}^TVV_{ij}^T) = E((UU^T)_{ij})E((VV^T)_{ij})$$

which is calculated by:

$$\frac{\sum_{ij}((UU^T)_{ij}(VV^T)_{ij})}{n^2} = \frac{\sum_{ij}(UU^T)_{ij}}{n^2}\frac{\sum_{ij}(VV^T)_{ij}}{n^2}$$

Since $\Delta = ||UU^T - VV^T||_F^2 = ||UU^T||_F^2 + ||VV^T||_F^2 - 2Sum(UU^T \circ VV^T)$, we have $ARI = 0$ or normalized distance 1, i.e. agreement no better than chance, when this independence condition holds, i.e.:

$$Sum(UU^T \circ VV^T) = \frac{|UU^T||VV^T|}{n^2}$$

[17] This equality is also useful in the implementation to improve the scalability.