APPENDIX A

Proof of Theorem 1

First, we introduce a new way to derive the expected value of mutual information in case of random partitions and under hyper-geometric distribution assumption and then we use the expected value to prove (13). Consider a pair of clusters P_i and G_j. The probability that an object in P_i exists in G_j is m_j / N. Accordingly, the number of objects in both P_i and G_j is simplified as: $n_{ij}=n_i \times (m_j / N)$. Then, the expected value can be calculated according to (7) as:

$$E(MI) = E\left\{\sum_i \sum_j \frac{n_{ij}}{N} \log\left(\frac{N \times (n_i \times m_j / N)}{n_i \times m_j}\right)\right\} \\ = E\left\{\sum_i \sum_j \frac{n_{ij}}{N} \log(1)\right\} = 0 \tag{24}$$

According to (2), AMI=NMI which confirms the result from [9]. Applying $max(\mathrm{MI})=(H(P) + H(G))/2$ as an option for normalization [22], [17], we can write:

$$AMI = NMI = \frac{2 \times MI(P,G)}{H(P)+H(G)} \tag{25}$$

Since $E(H(P)) = H(P)$ and $E(H(G)) = H(G)$ under hyper-geometric distribution assumption, the expected value of VI (8) is derived as:

$$E(VI) = H(P) + H(G) \tag{26}$$

VI is a dissimilarity measure and $min(\mathrm{VI}) = 0$ when the two partitions are equal. Therefore, the adjusted variation of information according to (2) is:

$$AVI = \frac{VI}{H(P)+H(G)} \tag{27}$$

An upper bound for VI is $H(P) + H(G)$ and therefore (27) also represents the normalized variation of information. We simplify AVI_s and NVI_s using (8) as follows:

$$AVI_s = NVI_s = \frac{2 \times MI(P,G)}{H(P)+H(G)} \tag{28}$$

From (25) and (28), we see that the adjusted mutual information and adjusted variation of information are equal to their normalized forms, and thus, theorem 1 is proven.

APPENDIX B

Proof of Theorem 2

Suppose that in a matching, m_1 is paired to $n_i < n_1$ and n_1 is paired to $m_j < m_1$ (case *a*). We show that if we change the matching so that m_1 is paired to n_1 and m_j is paired to n_i (case *b*), higher similarity is achieved. The total similarities for these two cases (*a* and *b*) are:

$$S_a = \frac{m_1 \times (n_i / N))}{\max(m_1, n_i)} + \frac{m_j \times (n_1 / N))}{\max(m_j, n_1)} \\ S_b = \frac{m_1 \times (n_1 / N))}{\max(m_1, n_1)} + \frac{m_j \times (n_i / N))}{\max(m_j, n_i)} \tag{29}$$

where S_a is the original pairing and S_b is the new pairing after changing the pairs for m_1 and m_j. Six different situations may happen:

1. $m_1 > m_j > n_1 > n_i$

$$\left[S_a = \frac{1}{N}(n_1 + n_i)\right] = \left[S_b = \frac{1}{N}(n_1 + n_i)\right]$$

2. $m_1 > n_1 > m_j > n_i$

$$\left[S_a = \frac{1}{N}(n_i + m_j)\right] < \left[S_b = \frac{1}{N}(n_1 + n_i)\right]$$

3. $m_1 > n_1 > n_i > m_j$

$$\left[S_a = \frac{1}{N}(n_i + m_j)\right] < \left[S_b = \frac{1}{N}(n_1 + m_j)\right]$$

4. $n_1 > m_1 > n_i > m_j$

$$\left[S_a = \frac{1}{N}(n_i + m_j)\right] < \left[S_b = \frac{1}{N}(m_1 + m_j)\right]$$

5. $n_1 > n_i > m_1 > m_j$

$$\left[S_a = \frac{1}{N}(m_1 + m_j)\right] = \left[S_b = \frac{1}{N}(m_1 + m_j)\right]$$

6. $n_1 > m_1 > m_j > n_i$

$$\left[S_a = \frac{1}{N}(m_j + n_i)\right] < \left[S_b = \frac{1}{N}(m_1 + n_i)\right] \tag{30}$$

Considering all the above situations, pairings (m_1, n_i) and (n_1, m_j) must be changed to (n_1, m_1) and (m_j, n_i) to achieve higher similarity. We can apply this proof recursively to all the smaller clusters as well. Hence, the two largest clusters must be always paired and then the next two largest and so on in order to achieve maximum total similarity with a random partition. This proves the theorem 2.

APPENDIX C

Triangular Inequality Proof for the Simplified form of PSI

Let P_1, P_2 and P_3 be three partitions with K_1, K_2 and K_3 clusters, and $K_{12}=max(K_1,K_2)$, $K_{23}=max(K_2,K_3)$, $K_{13}=max(K_1,K_3)$. Let n_i, n_j and n_k be the number of objects in clusters i, j and k in P_1, P_2 and P_3 respectively. We denote the number of shared objects between clusters by n_{ij}, n_{jk} and n_{ik}. The simplified distance form of PSI, for P_1 and P_2, according to (20) is:

$$D_{12} = \frac{K_{12} - S_{12}}{K_{12} - 1}$$

Lemma. $D_{12}+D_{23} \geq D_{13}$ (31)

Proof. We define $D'_{12} = K_{12} - S_{12}$, $D'_{23} = K_{23} - S_{23}$ and $D'_{13} = K_{13} - S_{13}$ and prove first that: $D'_{12}+D'_{23} \geq D'_{13}$ which is equivalent to

$$K_{12} - S_{12} + K_{23} - S_{23} \geq K_{13} - S_{13} \tag{32}$$

We consider three possible situations and simplify (32):

(1) $K_1 \geq K_{23}$: $S_{12} + S_{23} \leq K_{23} + S_{13}$
(2) $K_3 \geq K_{12}$: $S_{12} + S_{23} \leq K_{12} + S_{13}$
(3) $K_2 \geq K_{13}$: $S_{12} + S_{23} \leq K_2 + (K_2 - K_{13}) + S_{13}$

In the case (3), since $K_2 \geq K_{13}$, it is sufficient to prove $S_{12} + S_{23} \leq K_2 + S_{13}$. Since $K_{23} \geq K_2$ and $K_{12} \geq K_2$, for all cases it is sufficient to prove:

$$S_{12} + S_{23} \leq K_2 + S_{13} \tag{33}$$

According to the definitions (14) and (15), we divide the inequality (33) into K_2 sub-inequalities by considering each cluster j in P_2 on the left. Each sub-inequality is of the form:

$$\frac{n_{ij}}{\max(n_i, n_j)} + \frac{n_{jk}}{\max(n_j, n_k)} \leq 1 + \frac{n_{ik}}{\max(n_i, n_k)} \tag{34}$$

Clusters i and k from P_1 and P_3 which are the pairs for cluster j are not necessarily a pair in comparing P_1 and P_3. Since S_{13} is derived according to perfect matching, we can consider another matching of P_1 and P_3 in which i and k are paired. If (33) holds in this case, it will also be true for S_{13} which is the maximum possible similarity.

If the cluster j has a pair cluster only in P_1 or P_3, it is trivial to prove (34). If it has pair clusters both in P_1 and P_3, and $n_{ij} + n_{jk} \leq n_j$, proving (34) is trivial as well since the left side of the inequality is smaller than one. Note that if the clusters i and k do not have any shared objects, $n_{ij} + n_{jk} \leq n_j$. So we prove (34) when $n_{ij} + n_{jk} > n_j$. Considering a minimum value for n_{ik} as $n_{ij} + n_{jk} - n_j$, we rewrite (34) as follows:

$$\frac{n_{ij}}{\max(n_i, n_j)} + \frac{n_{jk}}{\max(n_j, n_k)} \leq 1 + \frac{n_{ij} + n_{jk} - n_j}{\max(n_i, n_k)} \tag{35}$$

Three possible cases are:

(1) $n_j \geq max(n_i, n_k)$: By replacing $max(n_i, n_j)$ and $max(n_j, n_k)$ by n_j and after simplifications, we have:

$$(n_{ij}+n_{jk}-n_j)(n_j-\max(n_i,n_k)) \geq 0$$

which is always true in this case.

(2) $n_i \geq max(n_j, n_k)$: We replace $max(n_i, n_j)$ and $max(n_i, n_k)$ by n_i. Since $max(n_j, n_k) \geq n_j$, it is sufficient to prove (35) by replacing $max(n_j, n_k)$ by n_j. The equivalent inequality derived after simplification:

$$(n_i - n_j)(n_j - n_{jk}) \geq 0$$

is always true.

(3) $n_k \geq max(n_i, n_j)$: The same proof in the case (2) can be applied.

The lemma (31) can now be represented as:

$$\frac{K_{12} - S_{12}}{K_{12} - 1} + \frac{K_{23} - S_{23}}{K_{23} - 1} \geq \frac{K_{13} - S_{13}}{K_{13} - 1} \tag{36}$$

We consider three possible cases:

(1) $K_1 \geq K_{23}$: It is sufficient to prove (36) if K_{23} in denumerator is replaced by K_1. So we simplify (36) as follows:

$$\frac{K_{12} - S_{12}}{K_1 - 1} + \frac{K_{23} - S_{23}}{K_1 - 1} \geq \frac{K_{13} - S_{13}}{K_1 - 1}$$

Since $K_1 \geq 2$, The denumerators can be canceled and the inequality is true according to (32).

(2) $K_3 \geq K_{12}$: The same inference as the case (1) can be performed by replacing K_{12} with K_3.

(3) $K_2 \geq K_{13}$: By simplifying (36), the following equivalent inequality is resulted:

$$S_{12} + S_{23} \leq 2K_2 - \frac{(K_{13} - S_{13})(K_2 - 1)}{K_{13} - 1} \tag{37}$$

Using (32), it is sufficient to prove:

$$K_2 + S_{13} \leq 2K_2 - \frac{(K_{13} - S_{13})(K_2 - 1)}{K_{13} - 1}$$

After simplification we have:

$$S_{13}(K_2 - K_{13}) \geq (K_2 - K_{13})$$

According to (14), $S_{13} \geq 0$, and therefore the above inequality is true.

According to the cases (1), (2) and (3), the inequalities (36) and consequently (31) hold, thus, the lemma is proven.