



Maximum transfer distance between partitions

***Valeur maximum de la distance
de transferts entre partitions***

Irène Charon
Lucile Denoeud
Alain Guénoche
Olivier Hudry

2005D003

mai 2005

Département Informatique et Réseaux
Groupe Mathématiques de l'Informatique et des Réseaux

Maximum transfer distance between partitions

I. Charon¹, L. Dencœud^{1,2}, A. Guénoche³ and O. Hudry^{1,2}

¹ École nationale supérieure des télécommunications,
46, rue Barrault, 75634 Paris cedex 13

charon@enst.fr, denoeud@enst.fr, hudry@enst.fr

² CERMSEM CNRS-UMR 8095, MSE, Université Paris 1 Panthéon-Sorbonne,
106-112, boulevard de l'Hôpital, 75647 Paris cedex 13

³ Institut de Mathématiques de Luminy, 163, avenue de Luminy, 13009 Marseille
guenoche@iml.univ-mrs.fr

Abstract: In this paper, we study a distance over the partitions of a finite set. Given two partitions P and Q , this distance is defined as the minimum number of transfers of an element from one class to another, in order to transform P into Q . We recall the algorithm to evaluate this distance and we give some formulae for the maximum distance value between two partitions having exactly or at most p and q classes, for given p and q .

Key words: Partitions, Distance, Transfer

1 Introduction

Establishing classes and partitions of elements in large sets is a methodological obstacle in several important domains as:

- molecular biology, especially for clustering proteins from expression data [8] or from direct interactions considered as graphs ([13], [19], [9]),
- social networks ([18]) in which communities are searched, for instance from common authors of articles ([4], [15]), or links between web pages ([14]),
- electronic circuits, for which partitioning is established for VLSI design ([2]) or graph drawing ([6]).

In such domains, the sets to cluster have a very large size, several thousands or tens of thousands elements, and computer scientists study new clustering methods for partitioning the whole set or extracting homogenous classes. They get some satisfying results with specific data, but rarely try to validate their methodological strategy performing a rigorous simulation protocol. One of the possible reasons is the difficulty to compare partitions on the same set. The

classical Rand index ([16]), based on element pairs simultaneously joined or separated, or the Rand index corrected for chance ([10]), as many other indices, may give identical values to partition pairs very close or very distant ([7]).

A simple protocol to evaluate a clustering algorithm is to measure its ability to recover partitions initially introduced in the data; for instance several more or less separated clusters of points in an Euclidean space ([20]), or vertex subsets in a graph having a proportion of edges larger than the average on the whole graph. So a set X with a "natural" partition P can be established. Applying any partitioning algorithm, generally another partition Q is obtained. The question is to compare P and Q , both partitions having not necessarily the same number of classes.

An article of Day ([5]) provides ten distances defined as the minimum number of modifications of the classes (augmentation, removal, mergence and division) to transform P into Q , or reciprocally. These distances are denoted *Minimum Length Sequence Metrics*. Day only compares the complexity of the computation of these metrics and also defines a partial ordering on the whole set (two metrics are comparable if and only if the values of the first one are systematically lower than or equal to the values of the other).

In this paper, we study the simplest of these distances (the R-metric), defined as the minimum number of augmentations and removals of single elements to transform P into Q . These two operations correspond to a transfer of one element from its class to another, which can be empty, and this distance will be denoted in the following as the *transfer distance*. It was already used by Régnier [17] to study partitions. To compute its values, Day just specifies that it is a minimum cost flow metric "since its computation is equivalent to the solution of a minimum cost flow problem on a suitably defined graph" and concludes that this metric is computable in $O(\max(|P|, |Q|)^3)$.

In the second section, we establish equivalent definitions of the transfer distance. The remaining part of the paper is devoted to the computation of the maximum value of this distance, in order to normalize it. In Section 3, we establish upper bounds for the distance value between a p -class partition and another one with q classes, for given integers p and q , and we show that these bounds can always be reached, building partitions achieving these values. We do the same in Section 4 for partitions with upper-bounded numbers of classes (notice that it is the same as fixing the numbers of classes, as in Section 3, but after having relaxed the hypothesis of the non-emptiness of the classes from the definition of a partition). Conclusions can be found in Section 5.

2 Notations and transfer distance definitions

Let X be a finite set of n elements. Let us recall that a partition P on X is a set of p non empty disjoint classes X_i , $1 \leq i \leq p$, such that $\bigcup_{i=1}^p X_i = X$.

Let P and P' be two partitions on X of respectively p and p' classes. The classes of P will be noted C_i , $1 \leq i \leq p$, and the classes of P' will be noted C'_j , $1 \leq j \leq p'$. We will represent the set X and these two partitions by an array in

which the rows are the classes of P and the columns the classes of P' (Fig. 1). The squares represent the sets $C_i \cap C'_j$, which may be empty.

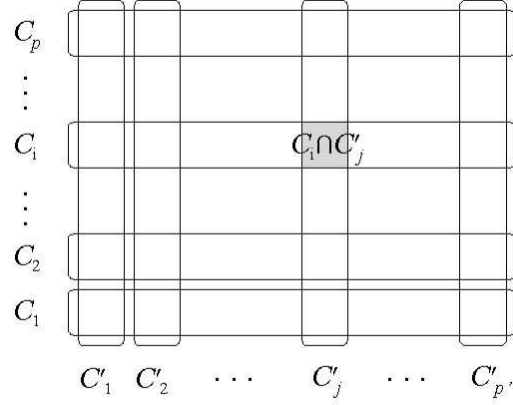


Figure 1: Partitions P et P'

We note $q = \max(p, p')$. Let \tilde{P} and \tilde{P}' be the systems of subsets of X with q subsets obtained respectively from P and P' by adding respectively $q - p$ and $q - p'$ empty subsets C_{p+1}, \dots, C_q and $C'_{p'+1}, \dots, C'_q$ respectively to the partitions P and P' . Let Σ be the set of one-to-one mappings on the set $\{1, \dots, q\}$.

Then, given a mapping σ in Σ , the number of well-classified elements, that is to say the number of elements that are in a class of \tilde{P} and in its corresponding class in \tilde{P}' according to σ , is given by:

$$c_\sigma(P, P') = \sum_{i=1}^q |C_i \cap C'_{\sigma(i)}|$$

and we may notice that $c_\sigma(P, P') = c_{\sigma^{-1}}(P', P)$.

We define the number of transfers between P and P' following σ as the number of elements that must be moved to turn P into P' :

$$t_\sigma(P, P') = n - c_\sigma(P, P')$$

and we may notice that $t_\sigma(P, P') = t_{\sigma^{-1}}(P', P)$. This definition is equivalent to the following ones:

$$t_\sigma(P, P') = \sum_{i=1}^q (|C_i| - |C_i \cap C'_{\sigma(i)}|)$$

$$t_\sigma(P, P') = \sum_{i=1}^q (|C'_i| - |C_{\sigma^{-1}(i)} \cap C'_i|)$$

$$t_\sigma(P, P') = \frac{1}{2} \sum_{i=1}^q |C_i \Delta C'_{\sigma(i)}| = \frac{1}{2} \sum_{i=1}^q (|C_i| + |C'_{\sigma(i)}| - 2 |C_i \cap C'_{\sigma(i)}|)$$

where $C_i \Delta C'_{\sigma(i)}$ stands for the symmetric difference between C_i and $C'_{\sigma(i)}$.

The *transfer distance* between P and P' is defined as:

$$\theta(P, P') = \min_{\sigma \in \Sigma} t_\sigma(P, P')$$

and the complementary notion of *concordance* as:

$$c(P, P') = n - \theta(P, P') = \max_{\sigma \in \Sigma} c_\sigma(P, P').$$

We may notice the following relations:

$$\theta(P, P') = \theta(P', P) \text{ and } c(P, P') = c(P', P).$$

Let Υ , Γ , Λ and Δ be mappings from $\{1, \dots, q\}^2$ to \mathbb{N} defined as:

$$\begin{aligned} \Upsilon(i, j) &= |C_i \cap C'_j| \\ \Gamma(i, j) &= |C_i| - |(C_i \cap C'_j)| \\ \Lambda(i, j) &= |C'_j| - |(C_i \cap C'_j)| \\ \Delta(i, j) &= |C_i \Delta C'_j| = |C_i| + |C'_j| - 2 |C_i \cap C'_j|. \end{aligned}$$

Let $K_{q,q}$ be the complete bipartite graph having the classes of \tilde{P} and \tilde{P}' as vertices (for simplicity, we will identify here the classes and their indices).

Theorem 1 *Let σ be a one-to-one mapping between the classes of \tilde{P} and the classes of \tilde{P}' . The following statements are equivalent :*

1. σ minimizes the number of transfers $t_\sigma(P, P')$;
2. σ defines a maximum perfect matching in $K_{q,q}$ weighted by Υ ; the weight of this perfect matching is $c(P, P')$;
3. σ defines a minimum perfect matching in $K_{q,q}$ weighted by Γ ; the weight of this perfect matching is $\theta(P, P')$;
4. σ defines a minimum perfect matching in $K_{q,q}$ weighted by Λ ; the weight of this perfect matching is $\theta(P, P')$;
5. σ defines a minimum perfect matching in $K_{q,q}$ weighted by Δ ; the weight of this perfect matching is $2\theta(P, P')$.

Proof : The equivalence between 1. and 2. has been proved by Day ([5]). This theorem arises simply from the following relations:

$$\begin{aligned}\theta(P, P') &= \min_{\sigma \in \Sigma} (n - \sum_{i=1}^q \Upsilon(i, \sigma(i))) = \min_{\sigma \in \Sigma} \sum_{i=1}^q \Gamma(i, \sigma(i)) \\ &= \min_{\sigma \in \Sigma} \sum_{i=1}^q \Lambda(i, \sigma(i)) = \min_{\sigma \in \Sigma} \frac{1}{2} \sum_{i=1}^q \Delta(i, \sigma(i)).\end{aligned}$$

■

Example 1 Let $P = (1, 2, 3|4, 5, 6|7, 8, 9)$ and $P' = (1, 3, 5, 6|2, 7, 9|4|8)$. The two tables corresponding to intersections and symmetric differences are given. One optimal mapping is in bold. It leads to $\theta(P, P') = 4$.

Υ	1,3,5,6	2,7,9	4	8	Δ	1,3,5,6	2,7,9	4	8
1,2,3	2	1	0	0	1,2,3	3	4	4	4
4,5,6	2	0	1	0	4,5,6	3	6	2	4
7,8,9	0	2	0	1	7,8,9	7	2	4	2
\emptyset	0	0	0	0	\emptyset	4	3	1	1

To the maximum mapping corresponds the transfer series :

$$\begin{aligned}(1, 2, 3|4, 5, 6|7, 8, 9) &\rightarrow (1, 3|4, 5, 6|2, 7, 8, 9) \rightarrow (1, 3, 5|4, 6|2, 7, 8, 9) \\ &\rightarrow (1, 3, 5, 6|4|2, 7, 8, 9) \rightarrow (1, 3, 5, 6|4|2, 7, 9|8).\end{aligned}$$

So, evaluating $\theta(P, P')$ is the same as solving the weighted matching problem on $K_{q,q}$, also known as the assignment problem in Operations Research. We are not going to detail the algorithm known as the Hungarian algorithm ([11], [12]) and designed to solve this problem. Its complexity is in $O(q^3)$. The interested reader will find details in [1].

3 Minimum concordance between two partitions with given numbers of classes

In this section, we determine, for three given integers p, q and n , with $p \leq q \leq n$, the maximum distance value between two partitions P and Q of respectively p and q classes and defined on a same set X of order n . To establish the upper bounds, we use the complementary notion of concordance introduced in the previous section. Let Σ be the set of all the possible mappings between the p classes of P and q classes of Q such that all the classes of P are matched, that is to say the set of injective functions σ from $\{1, \dots, p\}$ to $\{1, \dots, q\}$ (the set of permutations of $\{1, \dots, p\}$ when $p = q$). With respect to σ , the class C_i of P corresponds to the class $C'_{\sigma(i)}$ of Q .

According to the previous definition of the transfer distance, the number of well-classified elements considering σ is now given by :

$$c_\sigma(P, Q) = \sum_{i=1}^p |C_i \cap C'_{\sigma(i)}|$$

and the concordance between P and Q is :

$$c(P, Q) = n - \theta(P, Q) = \max_{\sigma \in \Sigma} c_\sigma(P, Q).$$

To maximize θ over the set of all the partitions having p and q classes is the same as to minimize c . Let $c_{\min}(p, q)$ be its smallest value:

$$c_{\min}(p, q) = \min\{c(P, Q) : P \text{ with } p \text{ classes and } Q \text{ with } q \text{ classes}\}.$$

Our main result is given in the following theorem :

Theorem 2 *Let X be a finite set of n elements. Let p and q be two integers with $p \leq q \leq n$. The minimum concordance $c_{\min}(p, q)$ between two partitions defined on X and with p and q classes is given by:*

- If $n \leq p + q - 2$,

$$c_{\min}(p, q) = p + q - n.$$

- If $p + q - 1 \leq n \leq (p - 1)q$,

$$c_{\min}(p, q) = \left\lceil \frac{n + q - p}{q} \right\rceil.$$

- If $(p - 1)q < n$,

$$c_{\min}(p, q) = \left\lceil \frac{n}{q} \right\rceil.$$

We are going to prove this theorem in two steps: first we prove that the values given above are lower bounds for the concordance, then we build two partitions which reach these bounds. We will split up each step into a series of lemmas with respect to the values of n , q and p .

3.1 Lower bounds of the concordance

In this section, we will first propose a general lower bound of the concordance for all p , q and n , and then more specific ones for $n \leq (p - 1)q$. In each case we state a lemma and prove it.

3.1.1 A general lower bound

Lemma 1 Let P and Q be two partitions of respectively p and q classes with $p \leq q \leq n$. Then:

$$c(P, Q) \geq \left\lceil \frac{n}{q} \right\rceil.$$

Proof :

Consider the q mappings between P and Q denoted σ_j , $j \in \{1, \dots, q\}$, defined, $\forall i \in \{1, \dots, p\}$, by :

- if $i + j - 1 \leq q$, $\sigma_j(i) = i + j - 1$
- If $i + j - 1 > q$, $\sigma_j(i) = i + j - 1 - q$.

The mapping σ_j is given by the shadowed squares on Fig. 2.

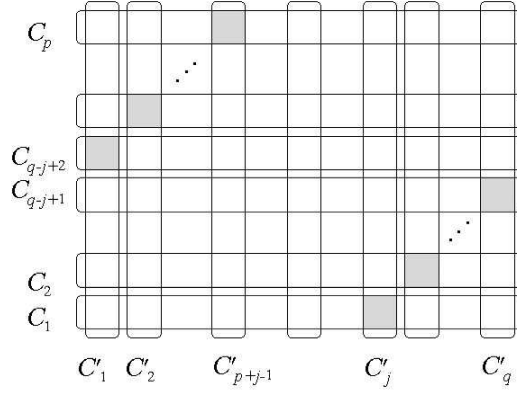


Figure 2: The mappings σ_j

The associated concordances c_j are given by the following expression:

$$\forall j \in \{1, \dots, q\}, c_j = \sum_{i=1}^p |C_i \cap C'_{\sigma_j(i)}|.$$

If we sum over j , we obtain :

$$\begin{aligned} \sum_{j=1}^q c_j &= \sum_{j=1}^q \sum_{i=1}^p |C_i \cap C'_{\sigma_j(i)}| = \sum_{i=1}^p \sum_{j=1}^q |C_i \cap C'_{\sigma_j(i)}| \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{q-i+1} |C_i \cap C'_{i+j-1}| + \sum_{j=q-i+2}^q |C_i \cap C'_{i+j-1-q}| \right) \\ &= \sum_{i=1}^p \left(\sum_{k=i}^q |C_i \cap C'_k| + \sum_{k=1}^{i-1} |C_i \cap C'_k| \right) = \sum_{i=1}^p \sum_{k=1}^q |C_i \cap C'_k|. \end{aligned}$$

It gives :

$$\sum_{j=1}^q c_j = \sum_{i=1}^p |C_i| = n.$$

This equality implies that at least one of the c_j 's is greater than or equal to the average $\frac{n}{q}$:

$$\exists j_0 \in \{1, \dots, q\} \text{ such that } c_{j_0} \geq \frac{n}{q}.$$

Thus we have found a lower bound of the concordance :

$$c(P, Q) = \max_{\sigma} c_{\sigma}(P, Q) \geq c_{j_0} \geq \frac{n}{q}$$

and since $c_{j_0} \in \mathbb{N}$, we get :

$$c(P, Q) \geq \left\lceil \frac{n}{q} \right\rceil.$$

■

3.1.2 Case $n \leq p + q - 2$

We set $n = q + \alpha$, with $0 \leq \alpha \leq p - 2$. In this case, we find a better lower bound of the concordance. We get the following lemma :

Lemma 2 *Let P and Q be two partitions of respectively p and q classes with $n \leq p + q - 2$. Then:*

$$c(P, Q) \geq p + q - n.$$

Proof :

Consider a partition Q . Since the classes of a partition cannot be empty, there is at least one element e_j in each class C'_j , and as $n = q + \alpha$, it remains α elements spread over the q classes. These α elements belong to at most α classes of the partition P . Therefore the q elements e_j belong to at least $p - \alpha$ classes of P since, otherwise, at least one class of P would be empty.

Thus we can find a mapping σ in which each of these $p - \alpha$ classes of P corresponds to an appropriate class of Q such that there is at least one well-classified element. The value of the corresponding concordance is then greater than or equal to $p - \alpha$.

Therefore we have proved the lemma:

$$\text{if } n \leq p + q - 2, \quad c(P, Q) \geq c_{\sigma}(P, Q) \geq p - \alpha = p + q - n.$$

■

3.1.3 Case : $p + q - 1 \leq n \leq pq - q$

Again, we propose a better lower bound of the concordance, specific to this case.

Lemma 3 Let P and Q be two partitions of respectively p and q classes with $p + q - 1 \leq n \leq pq - q$. Then:

$$c(P, Q) \geq \left\lceil \frac{n + q - p}{q} \right\rceil.$$

Proof : We suppose that we know the best mapping between P and Q , that is to say a mapping achieving the maximum number of well-classified elements. We may represent it by the shadowed squares on Fig. 3 (if necessary we may change the indices of the classes of Q in order to get this representation).

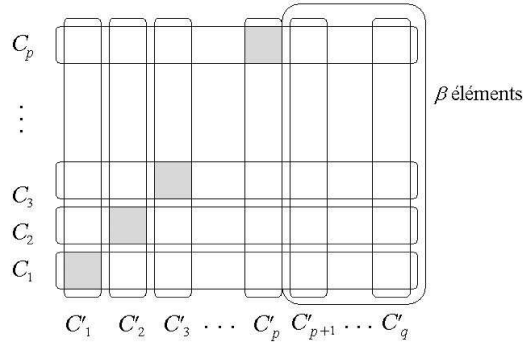


Figure 3: The partitions P and Q for the case $p + q - 1 \leq n \leq pq - q$.

Let c denote the concordance between P and Q : $c = c(P, Q)$; c is given by:

$$c = \sum_{i=1}^p |C_i \cap C'_i|.$$

We must have:

$$\forall j \in \{1, \dots, q - p\}, \quad \forall i \in \{1, \dots, p\}, \quad |C_i \cap C'_{p+j}| \leq |C_i \cap C'_i|$$

otherwise a better mapping would associate C_i to some C'_{p+j} instead of C'_i . Thus since $|C'_{p+j}| = \sum_{i=1}^p |C_i \cap C'_{p+j}|$ we obtain:

$$\forall j \in \{1, \dots, q - p\}, \quad |C'_{p+j}| \leq \sum_{i=1}^p |C_i \cap C'_i| = c.$$

We are now going to consider two cases :

Case 1 : the inequalities are strict :

$$\forall j \in \{1, \dots, q - p\} \quad |C'_{p+j}| < c.$$

Then

$$\forall j \in \{1, \dots, q-p\} \quad |C'_{p+j}| \leq c-1 \quad \text{since } |C'_{p+j}| \in \mathbb{N}.$$

Let $\beta = \sum_{j=1}^{q-p} |C'_{p+j}|$ (see Fig 3). So, if we sum over j , we get :

$$\beta \leq (c-1)(q-p).$$

We are now going to consider only the first p classes of Q . Let Y be the set containing the $n-\beta$ elements of these p classes. We can apply the general bound found previously to the partitions P and Q restricted to Y . We obtain a lower bound for c :

$$c \geq \frac{n-\beta}{p}.$$

Then we get, using the previous upper bound of β :

$$c \geq \frac{n-(c-1)(q-p)}{p} \Leftrightarrow c \geq \frac{n+q-p}{q}.$$

Case 2 : at least one of the inequalities is tight :

$$\exists j_0 \in \{1, \dots, q-p\} \text{ such that } |C'_{p+j_0}| = c$$

Since for any $j \in \{1, \dots, q-p\}$ and for any $i \in \{1, \dots, p\}$, $|C_i \cap C'_{p+j}| \leq |C_i \cap C'_i|$, we must then have :

$$\forall i \in \{1, \dots, p\} \quad |C_i \cap C'_{p+j_0}| = |C_i \cap C'_i|.$$

It involves that

$$\forall (i, k) \in \{1, \dots, p\}^2, i \neq k, \quad |C_k \cap C'_i| \leq |C_i \cap C'_i|.$$

Indeed, otherwise, we could associate C_k with C'_i and C_i with C'_{p+j_0} and the mapping could be better.

Then we get

$$\forall i \in \{1, \dots, p\} \quad |C_i \cap C'_i| \neq 0.$$

Indeed if $|C_{i_0} \cap C'_{i_0}| = 0$, we would have, for any $k \in \{1, \dots, p\}$, $|C_k \cap C'_{i_0}| = 0$, which implies that $C'_{i_0} = \emptyset$, a contradiction with the non-emptiness of the classes of a partition.

Finally, since for any $i \in \{1, \dots, p\}$, $|C_i \cap C'_i| \geq 1$, we get the following lower bound of c :

$$c = \sum_{i=1}^p |C_i \cap C'_i| \geq p.$$

Conclusion :

The two cases give

$$c \geq \min\left(p, \frac{n+q-p}{q}\right).$$

We can easily see that $p \geq \frac{n+q-p}{q}$ since $n \leq pq - q$. Therefore we finally obtain a specific lower bound for the concordance : $c(P, Q) \geq \frac{n+q-p}{q}$. That is to say, since $c(P, Q) \in \mathbb{N}$,

$$c(P, Q) \geq \left\lceil \frac{n+q-p}{q} \right\rceil.$$

■

3.2 Construction of partitions achieving the lower bound of the concordance

We are now going to build, for each case, two partitions which reach the previous lower bounds.

3.2.1 Case $n \leq p + q - 2$

Lemma 4 *There exist two partitions P and Q on X of respectively p and q classes with $n \leq p + q - 2$ such that*

$$c(P, Q) = p + q - n.$$

Proof :

We set $n = q + \alpha$, with $0 \leq \alpha \leq p - 2$.

We consider the partitions P and Q on X given by Figure 4. Each element of X is represented by a cross, we can see that there are $q + \alpha = n$ elements and that there is no empty class.

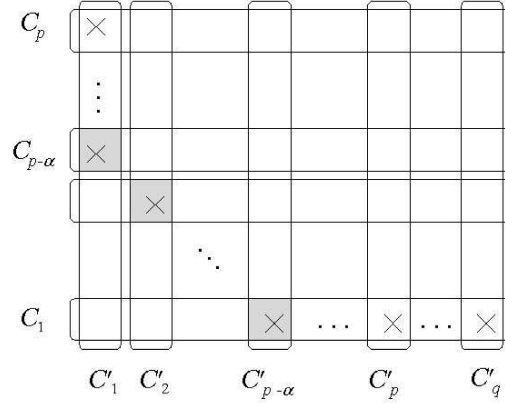


Figure 4: Two partitions P and Q with $c(P, Q) = p - n + q$.

Let σ be any mapping between P and Q . For $p - \alpha \leq i \leq p$, $|C_i \cap C'_1| = 1$ and, for $2 \leq j \leq q$, $|C_i \cap C'_j| = 0$. Thus $\sum_{i=p-\alpha}^p |C_i \cap C'_{\sigma(i)}| \leq 1$. Moreover, for $2 \leq i \leq p - \alpha - 1$, $|C_i \cap C'_{\sigma(i)}| \leq |C_i| = 1$ and similarly $|C_1 \cap C'_{\sigma(1)}| \leq 1$.

Hence $c_\sigma(P, Q) = \sum_{i=1}^p |C_i \cap C'_{\sigma(i)}| \leq 1 + (p - \alpha - 2) + 1 = p - \alpha = p - n + q$. Moreover, any mapping $\hat{\sigma}$ extending the partial mapping given by the shadowed squares on Figure 4 clearly gives $c_{\hat{\sigma}}(P, Q) = p + q - n$. Thus, by lemma 2, $\hat{\sigma}$ is a maximum mapping between P and Q , and $c(P, Q) = p + q - n$. ■

3.2.2 Case: $p + q - 1 \leq n \leq pq - q$

Notice that in this case q is greater than 1.

Lemma 5 *There exist two partitions P and Q on X of respectively p and q classes with $p + q - 1 \leq n \leq pq - q$ such that*

$$c(P, Q) = \left\lceil \frac{n + q - p}{q} \right\rceil.$$

Proof:

We build P and Q as follows: we put one element of X in $C_i \cap C'_1$, for $i \in \{2, \dots, p\}$. Then we spread the remaining elements uniformly over the sets $C_1 \cap C'_j$ for $j \in \{1, \dots, q\}$. There is no empty class since $p + q - 1 \leq n$. These two partitions are represented in Figure 5.

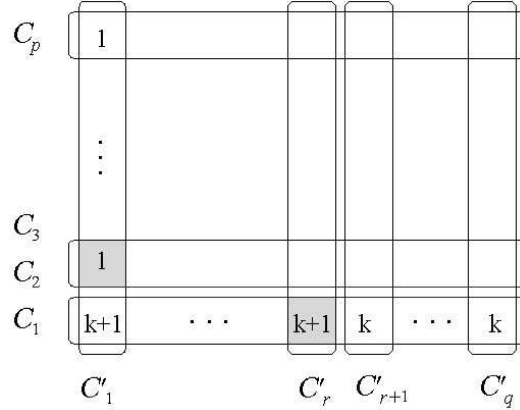


Figure 5: Two partitions P and Q with $c(P, Q) = \lceil (n + q - p)/q \rceil$.

We are now going to compute the concordance between these two partitions. We must first compute the number of elements in each $C_1 \cap C'_j$. We have spread the $n - p + 1$ elements of C_1 over the C'_j 's. We set $n - p + 1 = kq + r$, with $k \geq 1$ and $r \in \{0, \dots, q - 1\}$. There are $k + 1$ elements in the $C_i \cap C'_j$'s with $j \leq r$, and k elements in the others. Therefore we can distinguish two cases :

If $r = 0$, the classes of Q contain k elements except C'_1 ; any maximum mapping associates C_1 with one of the C'_j 's, $j > 1$ and C'_1 with one of the C_i 's, $i \geq 2$. It gives

$$c = k + 1.$$

If $r = 1$, any maximum mapping associates C_1 with one of the C'_j 's, $1 \leq j \leq r$ and C'_q with one of the C_i 's, $i \geq 2$ or C_1 with C'_1 and the other classes of P with other classes of Q . It gives

$$c = k + 1.$$

If $r > 1$, any maximum mapping associates C_1 with one of the C'_j 's, $2 \leq j \leq r$ and C'_1 with one of the C_i 's, $i \geq 2$. It gives

$$c = k + 2.$$

It remains to show that these values correspond to the lower bound stated previously.

We set $b = \left\lceil \frac{n+q-p}{q} \right\rceil$:

$$b = \left\lceil \frac{kq + r - 1 + q}{q} \right\rceil = \left\lceil \frac{(k+1)q + r - 1}{q} \right\rceil.$$

If $r \in \{0, 1\}$, as $q > 1$,

$$b = k + 1.$$

If $r > 1$,

$$b = k + 2.$$

Thus in every case

$$c(P, Q) = b = \left\lceil \frac{n+q-p}{q} \right\rceil.$$

■

3.2.3 Case $(p-1)q < n$

Lemma 6 *There exist two partitions P and Q on X of respectively p and q classes with $(p-1)q < n$ such that*

$$c(P, Q) = \left\lceil \frac{n}{q} \right\rceil.$$

Proof :

We are going to distinguish two cases.

Case 1: $(p-1)q < n < pq$

We set $n = (p-1)q + r$, with $r \in \{1, \dots, q-1\}$. We build P and Q as follows (see Figure 6): we put one element in each $C_i \cap C'_j$, for $i \in \{1, \dots, p-1\}$, $j \in \{1, \dots, q\}$. Then, for any $j \in \{1, \dots, r\}$, we put one of the r remaining elements in $C_p \cap C'_j$.

Any maximum mapping associates C_p with one of the C'_j 's for $1 \leq j \leq r$, and the other C_i 's ($1 \leq i \leq p-1$) with $p-1$ classes C'_j 's still available. It gives

$$c(P, Q) = p = \left\lceil \frac{n}{q} \right\rceil.$$

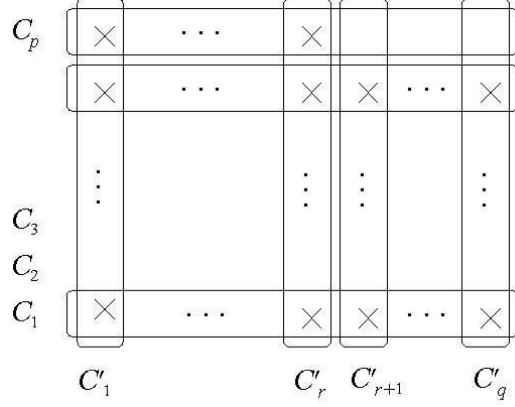


Figure 6: Two partitions P and Q with $c(P, Q) = \lceil n/q \rceil$ for $(p-1)q < n < pq$.

Case 2: $pq \leq n$

We set $n = kq + r$ with $k \geq p$ and $0 \leq r \leq q-1$. We put one element in each $C_i \cap C'_j$. The remaining elements are spread uniformly over the sets $C_1 \cap C'_j$, $j \in \{1, \dots, q\}$ (see Figure 7).

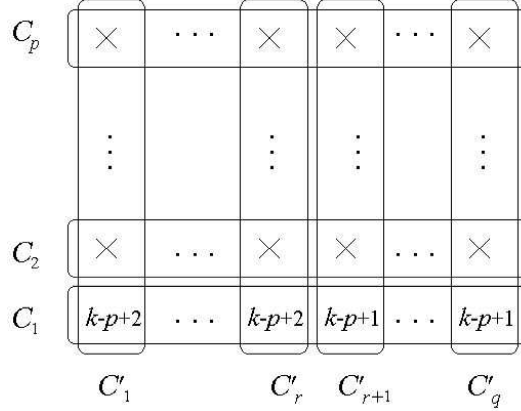


Figure 7: Two partitions P and Q with $c(P, Q) = \lceil n/q \rceil$ for $pq \leq n$.

We distinguish again two cases.

If $r = 0$, all the subsets $C_1 \cap C'_j$ ($1 \leq j \leq q$) contain $k-p+1$ elements. Any mapping gives

$$c = (k-p+1) + (p-1) = k = \left\lceil \frac{n}{q} \right\rceil.$$

If $r > 0$, the subsets $C_1 \cap C'_j$ contain $k - p + 2$ elements for $1 \leq j \leq r$, and $k - p + 1$ for $r + 1 \leq j \leq q$. The greatest concordance is obtained if we associate C_1 and a C'_j with $1 \leq j \leq r$ and the other classes of P with $p - 1$ of the remaining classes of Q . Its value is

$$c = (k - p + 2) + (p - 1) = k + 1 = \left\lceil \frac{n}{q} \right\rceil.$$

■

3.3 Proof of Theorem 2

The proof of Theorem 2 follows from the previous lemmas. More precisely, the first statement of Theorem 2 follows from Lemma 2 (lower bound) and Lemma 4 (how to reach the lower bound). Similarly, the second (respectively third) statement of Theorem 2 follows from Lemma 3 and Lemma 5 (respectively Lemmas 1 and 6).

4 Minimum concordance between two partitions with upper-bounded numbers of classes

In some cases, it is not obvious to fix the numbers of classes of the two partitions P and Q , and it is easier to upper-bound them. We consider this problem now. More precisely, given two integers p and q , we define the minimum concordance $c'_{min}(p, q)$ between two partitions with at most p and q classes by:

$$c'_{min}(p, q) = \min\{c(P, Q) : P \text{ partition of } X \text{ with at most } p \text{ classes and } Q \text{ partition of } X \text{ with at most } q \text{ classes}\},$$

where $c(P, Q)$ still denotes the concordance between P and Q as in Section 3. So, the only difference with respect to $c_{min}(P, Q)$ is that the minimum of the concordance $c(P, Q)$ is taken over the set of partitions with upper-bounded numbers of classes instead of given numbers of classes. Notice that another way of defining $c'_{min}(p, q)$ would have been to extend the definition of partitions by allowing empty classes. Then, $c'_{min}(p, q)$ could be seen as the minimum concordance over the set of subsets systems such that these subsets are mutually disjoint and cover X (but can be empty). Theorem 3 gives the solution to this new problem.

Theorem 3 *Let p and q be two integers with $p \leq q$. Let X be a finite set with n elements. The minimum concordance $c'_{min}(p, q)$ between two partitions with at most p and q classes is given by:*

$$c'_{min}(p, q) = \left\lceil \frac{n}{q} \right\rceil.$$

Proof :

Consider the two partitions P and Q defined as follows: P has only $h = 1$ class (equal to X), Q has exactly $k = \min(n, q)$ (non-empty) classes in which the elements of X are uniformly spread over: each class of Q contains $\lfloor \frac{n}{k} \rfloor$ or $\lceil \frac{n}{k} \rceil$ elements of X . Then, it is easy to compute $c(P, Q)$: $c(P, Q) = \lceil \frac{n}{k} \rceil$, and thus $c(P, Q) = \lceil \frac{n}{q} \rceil$. Hence by definition of a minimum: $c'_{min}(p, q) \leq \lceil \frac{n}{q} \rceil$.

Moreover, consider now any partitions P and Q with respectively h and k classes, with $h \leq p$ and $k \leq q$. By Lemma 1, we get: $c(P, Q) \geq \lceil \frac{n}{\max(h, k)} \rceil \geq \lceil \frac{n}{q} \rceil$. Since this inequality is true for any P and Q , it comes $c'_{min}(p, q) \geq \lceil \frac{n}{q} \rceil$. Hence the result of Theorem 3.

■

A special case is the one for which $n = p = q$. Then the minimum concordance is equal to 1. We find back the fact (which can be shown directly) that the maximum distance between any two partitions (that is, the diameter of the set of all the partitions) is equal to $n - 1$.

5 Conclusion

The previous results may be summarized as follows, when stated in terms of the distance θ . Let X be a finite set of n elements, and p and q be two integers with $p \leq q$. For $n \geq q \geq p$, the maximum distance

$$\theta_{max}(p, q) = \max\{\theta(P, Q) : P \text{ with } p \text{ classes and } Q \text{ with } q \text{ classes}\} = n - c_{min}(p, q)$$

between partitions with p and q classes is equal to:

- If $n \leq p + q - 2$,

$$\theta_{max}(p, q) = 2n - p - q.$$

- If $p + q - 1 \leq n \leq (p - 1)q$,

$$\theta_{max}(p, q) = n - \left\lceil \frac{n + q - p}{q} \right\rceil.$$

- If $(p - 1)q < n$,

$$\theta_{max}(p, q) = n - \left\lceil \frac{n}{q} \right\rceil.$$

If we consider now partitions with upper-bounded numbers of classes, the maximum distance

$$\theta'_{max}(p, q) = \max\{\theta(P, Q) : P \text{ with at most } p \text{ classes and } Q \text{ with at most } q \text{ classes}\}$$

is given by:

$$\theta'_{max}(p, q) = n - \left\lceil \frac{n}{q} \right\rceil.$$

An interesting problem would consist in studying the maximum distance between two partitions with prescribed numbers of classes and with given cardinalities of these classes. It will be the aim of our future research.

Acknowledgements

This work is supported by the CNRS ACI IMP-Bio. We would like also to thank B. Leclerc for his help about the paper by Day [5].

References

- [1] R.K. Ahuja, T.L. Magnanti and J.B. Orlin (1993) Network flows, Prentice Hall, Englewood Cliffs, New Jersey.
- [2] C.J. Alpert and A. Kang (1995) Recent direction in netlist partitioning: a survey, *Integration: the VLSI Journal*, 19, 1-2, 1-81.
- [3] V. Batagelj, M. Mrvar and M. Zaversnik (1999) Partitioning approach to visualisation of large graphs, *Lecture Notes in Computer Science* 1731, Springer, 90-97.
- [4] V. Batagelj, M. Zaversnik (2000) An $O(m)$ Algorithm for Cores Decomposition of Networks, submitted.
- [5] W. Day (1981) The complexity of computing metric distances between partitions, *Mathematical Social Sciences*, 1, 269-287.
- [6] H. de Fraisse and P. Kuntz (1992) Pagination of large scale networks; embedding a graphe in \mathbb{R}^n for effective partitioning, *Algorithmic review*, 2(3), 105-112.
- [7] L. Denceud, H. Garreta and A. Guenoche (2005) Comparison of distance indices between partitions, Proceedings of Applied Stochastic Models and Data Analysis, Ph. Lenca et al. (eds.) on CD-Rom, Brest, France.
- [8] G. Getz, E. Levine, E. Domany (2000) Coupled two-way clustering analysis of gene microarray data, *Proc. Natl. Acad. Sci. USA*, 97, 22, 12079-12084.
- [9] A. Guenoche (2004) Clustering by graph density, *Proceedings of the International Federation of the Classification Societies*, D. Banks and al. (eds), Springer, 15-23.
- [10] L. Hubert and P. Arabie (1985) Comparing partitions, *J. of Classification*, 2, 193-218.
- [11] H.W. Kuhn (1955) The Hungarian method for the assignment problem, *Naval Res. Logist. Quart.*, 2, 83-97.
- [12] H.W. Kuhn (1956) Variants on the Hungarian method for the assignment problems, *Naval Res. Logist. Quart.*, 3, 253-258.
- [13] H. Matsuda, T. Ishihara, A. Hashimoto (1999) Classifying molecular sequences using a linkage graph with their pairwise similarities, *Theoretical Computer Science*, 210, 305-325.

- [14] J. Moody (2001) Identifying dense clusters in large networks, *Social Networks*, 23, 261-283.
- [15] M. E. J. Newman (2001) The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA* 98, 404-409.
- [16] W.M. Rand (1971) Objective criteria for the evaluation of clustering methods, *J. Amer. Statist. Assoc.*, 66, 846-850.
- [17] S. Régnier (1965) Quelques aspects mathématiques des problèmes de classification automatique, I.C.C. Bulletin 4.
- [18] S.B. Seidman (1983) Network structure and minimum degree, *Social Networks*, 5, 269-287.
- [19] S. van Dongen (2000) *Graph clustering by flow simulation*, PhD thesis, University of Utrecht.
- [20] G. Youness and G. Saporta (2004) Une méthodologie pour la comparaison des partitions, *Revue de Statistique Appliquée*, 52(1), 97-120.