Centrality and Distribution of Partitions according to the Transfer Distance

Lucile Belgacem[*], Olivier Hudry[†]

Abstract

The comparison of partitions is a central topic in clustering, as well for comparing partitioning algorithms as for classifying nominal variables. In this paper, we deal with the transfer distance between partitions, defined as the minimum number of transfers of one element from its class to another (eventually empty) necessary to turn one partition into the other one. We study the distribution of partitions according to their transfer distances to some reference partitions. Then we design criteria to define when two partitions can be considered as far or as close. Last, we look for the features of the partitions which can be considered as central, i.e. partitions such that the other partitions are relatively close to them.

Key words : Partition, distance, clustering, transfer graph, centrality

1 Introduction

Establishing classes and partitions of elements in large sets is a main methodological issue in decision making and has important applications in several domains as classification [1], molecular biology [8, 14], social networks [2, 15], electronic circuits [7], and so on. Indeed, the structure of partition is involved for instance when we want to decide how to gather entities (which can be objects, persons, projects, ...) in such a way that the entities belonging to a same cluster look globally similar according to some given criteria while the clusters of the partition look dissimilar according to the same criteria.

[*] Orange Labs FT R&D, 38-40 rue du général Leclerc, 92794 Issy-les-Moulineaux Cedex 9, France.

[†] École nationale supérieure des télécommunications 46, rue Barrault, 75634 Paris cedex 13, France. hudry@telecom-paristech.fr

The design of such a partition often requires the use of distances between partitions. The literature abounds in indices designed by multiple authors to compare two partitions P and Q defined on the same set X. Some of the most used indices are based on the pairs of elements of X. Two elements x and y can be gathered or separated in P and Q. The two partitions agree on (x, y) if these elements are simultaneously gathered or separated in P and Q. On the other hand there is a disagreement if x and y are gathered in P and separated in Q or conversely. Among these distance indices, we can mention for instance: the Rand index [18], the Rand index corrected for chance by Hubert and Arabie [9], the Jaccard index [10], the Wallace index [22], and the normalized index of Lerman [12]. Other authors proposed distances based on the partition lattice by considering the number of pairs of elements in the same class or the number of classes in the join $P \vee Q$ or the meet $P \wedge Q$ of P and Q [3].

Some other distances are defined as the minimum number of modifications of the classes (augmentation, removal, mergence or division) to transform P into Q, or conversely. These distances are called *Minimum Length Sequence Metrics* [6]. The simplest of these distances is 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*. This distance was first proposed in 1965 by S. Régnier [19] to study partitions. In [6], W. Day specifies that the transfer distance 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 $O(\max(|P|, |Q|)^3)$ (where $|P|$ and $|Q|$ denote respectively the number of classes of the partitions P and Q). Some results about the maximum values of the transfer distance between partitions can be found in [4, 5, 8].

We begin this work in Section 2 by recalling some definitions and notation about the transfer distance, and explaining how it can be computed. The methodology used for the experimental studies done in this paper is presented in Section 3. In Section 4, we study the distributions of partitions defined on the same set with respect to the transfer distance to a given partition. Then, from the light of this study, we propose in Section 5 some criteria assessing the closeness between partitions. Finally, we study in Section 6 the centrality of partitions according to the transfer distance (a partition will be qualified as *central* if the other partitions on the same set are relatively close to it).

2 Notation and 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, $1 \leqslant i \leqslant p$, 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, $1 \leqslant i \leqslant p$, and the classes of P' will be noted C'_j, $1 \leqslant j \leqslant p'$. Without loss of generality, let us assume that $p \leqslant p'$.

Let $\mathcal{M}$ be the set of mappings defined from $\{1, ..., p\}$ to $\{1, ..., p'\}$. We define the *concordance* between P and P' by:

$$c(P, P') = \max_{\sigma \in \mathcal{M}} \sum_{i=1}^{p} \mid C_i \cap C'_{\sigma(i)} \mid .$$

The value $c(P, P')$ represents the maximum number of elements that are in a class of P and in its corresponding class in P' over the mappings of $\mathcal{M}$. In other words, it is the maximum number of well-classified elements.

The transfer distance, noted t, is the complementary notion of the concordance:

$$t(P, P') = \min_{\sigma \in \mathcal{M}} (n - \sum_{i=1}^{p} \mid C_i \cap C'_{\sigma(i)} \mid) = n - c(P, P').$$

The value $t(P, P')$ represents the minimum number of transfers of one element from its class to another (eventually empty) necessary to turn P into P', or conversely P' into P.

According to the transfer distance definition, its computation comes down to determining a mapping maximizing the number of well-classified elements.

Let Υ be the function from $\{1, ..., p\} \times \{1, ..., p'\}$ to $\mathbb{N}$ defined by:

$$\Upsilon(i, j) = \mid C_i \cap C'_j \mid .$$

Let $K_{p,p'}$ be the complete bipartite graph having the classes of P and P' as vertices. We recall that a matching M in a graph $G = (V, E)$ is a subset of E such that two elements of M (i.e. two edges of G) are disjoint: $\forall \{x, y\} \in M$, $\forall \{z, t\} \in M$ with $\{x, y\} \neq \{z, t\}$, we must have $x \neq z,\ x \neq t,\ y \neq z,\ y \neq t$.

W. Day proved in [6] that a matching σ maximizes the number of well-classified elements ($\sum_{i=1}^{p} \mid C_i \cap C'_{\sigma(i)} \mid$) if and only if σ defines a maximum matching in $K_{p,p'}$ weighted by Υ; the weight of this matching is $c(P, P')$.

So, computing $t(P, P')$ is the same as solving the *weighted matching problem* on $K_{p,p'}$, also known as the *assignment problem* in Operations Research. This problem can be solved by the Hungarian algorithm [11], whose complexity is in $O(p'^3)$. The interested reader will find details in [13].

3 Methodology

In order to study the transfer distance between partitions on a set X of n elements, two strategies may be considered.

The exhaustive enumeration of $\mathcal{P}_n$ can be done for $n \leqslant 12$ following the process NexEqu described in [16]. Each partition is coded by a vector containing for each element the index of its class. The first considered partition is the one class partition. Then the algorithm builds iteratively the following partitions according to the lexicographic order of this code, until reaching the partition with n classes.

For $n > 12$, it becomes intractable to enumerate all the partitions of $\mathcal{P}_n$. In order to sample the set $\mathcal{P}_n$, we need to draw partitions with a uniform distribution. Such a uniform drawing has been proposed in [17]. It is based on the following well-known relation [20]:

$$B_n = \sum_{k=0}^{n-1} \left(\begin{array}{c} n-1 \\ k \end{array} \right) .B_k$$

where B_n denotes the n^{th} Bell number, that is the number of partitions in $\mathcal{P}_n$.

Following this process, we are able to draw a sample E of partitions of $\mathcal{P}_n$. If we want to estimate a proportion Π of partitions having a certain property, according to the central limit theorem we have (see for instance [21]):

$$Prob\left(f - \mathcal{N}_{(0,1)}(\rho)\sqrt{\frac{\Pi(1-\Pi)}{|E|}} \leqslant \Pi \leqslant f + \mathcal{N}_{(0,1)}(\rho)\sqrt{\frac{\Pi(1-\Pi)}{|E|}}\right) = \rho$$

where $\mathcal{N}_{(0,1)}$ denotes the standard normal distribution, $|E|$ is the size of the sample and f is the observed frequency of the property in E.

For instance if we want a confidence interval for Π with confidence level ρ equal to 95% (which corresponds to $\mathcal{N}_{(0,1)}(\rho) = 1,96$) and such that $0.99f \leqslant \Pi \leqslant 1.01f$, we need to draw less than 10000 partitions ($\Pi(1-\Pi)$ has been upper-bounded by $\frac{1}{4}$), which is quite feasible. Thus, in the following experiments, we are going to draw 10000 partitions of $\mathcal{P}_n$ in order to study some characteristics of the partitions. Let us notice that this number does not depend on the value of n.

4 Distribution of the partitions with respect to the transfer distance

In order to evaluate the distance between two partitions we can compute the transfer distance, but how to interpret this value? To what extent does it correspond to close partitions or far partitions? We try to deal with this difficult issue by studying experimentally the transfer distance.

Let us focus on two cases: $n = 12$ and $n = 100$. These values have been chosen as the maximum values allowing respectively an enumeration of $\mathcal{P}_n$ and a random drawing defined on $\mathcal{P}_n$ (due to computing limitations). For each case, we consider some reference partitions and we study the distributions of the partitions of $\mathcal{P}_n$ according to their transfer distances to the reference partitions. In other words, considering a partition of reference P defined on the set X, we try to evaluate, for any $t \in [0; t_{max}(P)]$, the values of $|\tau_t(P)|$, $\tau_t(P)$ being defined as the set of partitions of $\mathcal{P}_n$ at exactly t transfers from P.

Table 1: Characteristics of the ten reference partitions

$n = 12$	P1	P2	P3	P4	P5	average
numb. of classes	7	6	6	5	4	5.55
Cardinalities	3|2|2|2|1|1|1	4|3|2|1|1|1	2|2|2|2|2|2	4|3|2|2|1	3|3|3|3	-
$n = 100$	R1	R2	R3	R4	R5	average
numb. of classes	35	31	28	27	25	28.6
av. cardinalities	2.9	3.2	3.6	3.7	4	3.5

For $n = 12$, there are $|\mathcal{P}_{12}| = 4213597$ partitions that can be enumerated in a reasonable time. For $n = 100$ we draw randomly 10 000 partitions of $\mathcal{P}_{100}$ (let us recall that $\mathcal{P}_{100} = B_{100} \approx 10^{115}$) using the drawing presented in Section 3. For each value of n, we chose to study the case of five partitions of reference (labelled P1, ..., P5 for $n = 12$ and R1, ..., R5 for $n = 100$) that have been determined beforehand (P1, P2 and P4 have been drawn randomly, P3 and P5 have been chosen for their regularities; R1, ..., R5 have all been drawn randomly). The reference partitions are described in Table 1.

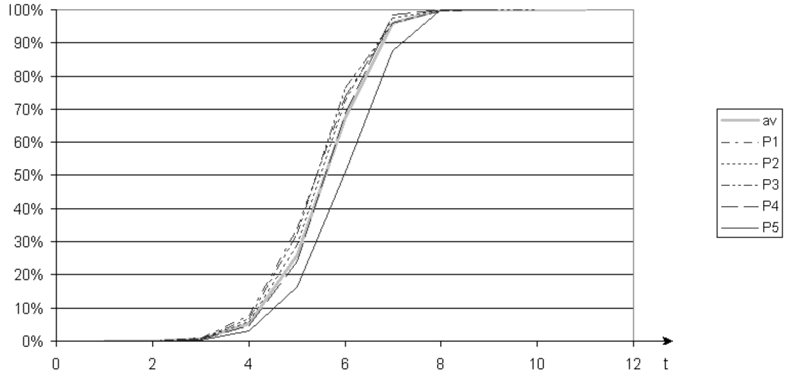

Figure 1: Distributions of the transfer distance to five reference partitions and average distribution on $\mathcal{P}_{12}$

Figures 1 and 2 represent the distributions of the partitions according to their transfer distances to the reference partitions for $n = 12$ and $n = 100$. The x-axis represents the number of transfers t, and the y-axis the cumulated percentages of the distributions. We have also displayed the average distribution when considering 10000 pairs of random partitions.

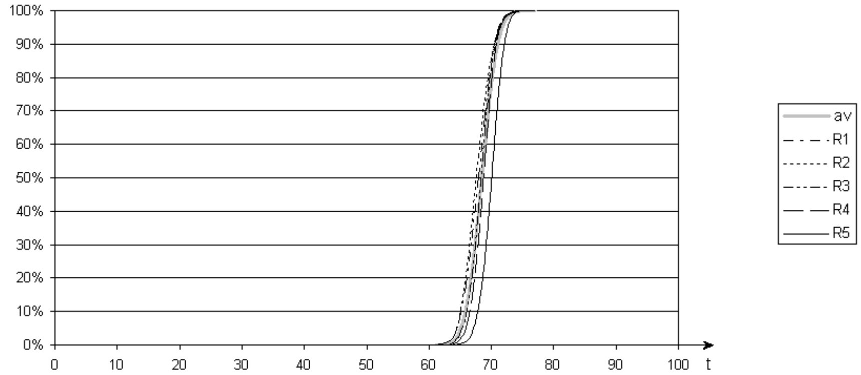

Figure 2: Distributions of the transfer distance to five reference partitions and average distribution on $\mathcal{P}_{100}$

We notice that the six graphs represented on both Figures 1 and 2 are very similar. For both values of n, the five distributions according to the reference partitions are almost identical, and very close to the average distribution. The percentages remain near zero until a relatively high number of transfers, then increase quickly and regularly. It is quite surprising that almost all partitions are located at a large number of transfers from the reference partitions, close to the maximum values of the transfer distance (between 3 and 8 for $n = 12$ and between 60 and 75 for $n = 100$ while the maximum values are respectively around 9 and around 92). For $n = 100$, partitions from less than 60 transfers from the reference partitions are non-existent in proportion in $\mathcal{P}_{100}$, although those partitions are very numerous.

This study allows to observe that the partitions are concentrated on a narrow transfer interval (the values of the transfer distance are very close to the average transfer distance). We remark moreover that this interval corresponds to large values of transfer, close to the maximum values. This specific behaviour of the distributions of the partitions according to the transfer distance leads us to define some criteria to evaluate the closeness between two partitions.

5 Close partitions in terms of transfers

The previous study permits to define some threshold values specifying the closeness or the remoteness between two partitions.

We define, given a partition P and a threshold $\alpha \in [0; 1]$, two critical values denoted

t_α^- and t_α^+ by:

$$\frac{|\{Q \in \mathcal{P}_n, t(P,Q) \leqslant t_\alpha^-\}|}{|\mathcal{P}_n|} \leqslant \alpha \text{ and } \frac{|\{Q \in \mathcal{P}_n, t(P,Q) \leqslant t_\alpha^- + 1\}|}{|\mathcal{P}_n|} > \alpha$$

$$\frac{|\{Q \in \mathcal{P}_n, t(P,Q) \geqslant t_\alpha^+\}|}{|\mathcal{P}_n|} \leqslant \alpha \text{ and } \frac{|\{Q \in \mathcal{P}_n, t(P,Q) \geqslant t_\alpha^+ - 1\}|}{|\mathcal{P}_n|} > \alpha.$$

In other words, t_α^- (resp. t_α^+) is the largest (resp. the smallest) number of transfers such that the proportion of partitions in $\mathcal{P}_n$ being at less than t_α^- (resp. at more than t_α^+) transfers from P is lower than α.

Definition 1 *Let P and Q be two partitions of the set X of n elements. Let t_α^- and t_α^+ be the critical values at threshold α computed following the previous definitions. We will say that Q is close to P at threshold α if $t(P,Q) \leqslant t_\alpha^-$, and conversely that Q is far from P at threshold α if $t(P,Q) \geqslant t_\alpha^+$.*

Example 1 Let us consider again the case $n = 100$ and let us assume that one wants to compare the partition R5 (see Table 1 for the description of this partition) with another partition Q. We choose the value $\alpha = 5\%$ as threshold.

We compute by sampling the distributions of the partitions of $\mathcal{P}_{100}$ according to the transfer distance to R5. We obtain respectively the following cumulated percentages for each value of $t \in [63; 76]$: 0.01%, 0.07%, 0.26%, 0.99%, 3.72%, 11.07%, 25.31%, 47.15%, 71.34%, 88.53%, 97.24%, 99.54%, 99.98%, 100.00%. In this case, we have then $t_{5\%}^- = 67$ since the proportion of partitions being at less than 67 transfers from R5 is equal to $3.72\% \leqslant 5\%$ whereas the proportion of partitions being at less than 68 transfers from R5 is equal to $11.07\% > 5\%$. Similarly, we deduce that $t_{5\%}^+ = 73$. According to Definition 1, we will consider that R5 and Q are close at threshold 5% if $t(\text{R5}, Q) \leqslant 67$, and that $R5$ and Q are far at threshold 5% if $t(\text{R5}, Q) \geqslant 73$.

Since the interval $[t_\alpha^-; t_\alpha^+]$ is narrow with respect to $[0, t_{max}]$ (for instance, for $n = 100$ and $\alpha = 5\%$, it is equal to [67 ; 73]), there are few transfer values for which we are not able to give an interpretation. According to these observations, the transfer distance proved to be a well discriminant distance to compare partitions. Indeed, two pairs of partitions being respectively at 67 and 73 transfers will be treated very differently following the proposed criteria since the first value corresponds to close partitions whereas the second one corresponds to far partitions. This distinction would not have been possible without the distributions study since the transfer values 67 and 73 are both high and close to each other, and therefore do not let suspect a significant difference.

As a drawback, to the contrary, the proposed criteria cannot distinguish between two partitions being at few transfers one from each other and two partitions being at a great number of transfers, as long as these two values remain lower than t_α^- (for instance, for $n = 100$, the number of transfers $t = 1$ and $t = 60$). Indeed, those two values of the transfer distance represent an insignificant proportion of the partitions, and cannot be discriminated from each other although they are very different. Thus in order to complete the information given by the criteria, it may be necessary to consider also the numerical values of the transfer distance especially if one wants to compare close partitions.

6 Centrality of a partition

In the previous section, we have presented the distributions of the partitions according to their transfer distances to some reference partitions. Although all the computed distributions seemed to be very similar, we can nevertheless notice some small fluctuations. If we focus for instance on the case $n = 100$, the average transfer distance $\bar{t}$ observed for the five reference partitions are respectively 68.48, 68.25, 68.90, 69.27 and 70.55.

These values reflect the *centrality* of the considered partition P in $\mathcal{P}_{100}$, which we define herein according to the distribution of partitions with respect to their transfer distances to P. We have observed in Section 4 that a large majority of the partitions of $\mathcal{P}_n$ are concentrated in a narrow interval $[t_\alpha^-; t_\alpha^+]$ of transfers from P, centered around the average value $\bar{t}$. The closer to P the partitions of $\mathcal{P}_n$ are located (the values $[t_\alpha^-; t_\alpha^+]$ are small), the more central P can be considered, and conversely the farther from P the partitions of $\mathcal{P}_n$ are located (the values $[t_\alpha^-; t_\alpha^+]$ are large), the more eccentric P can be considered.

From a practical point of view, the centrality is expressed by the position of the distribution graph according to the y-axis (the closer to the axis, the more central the considered partition). We can say for instance that R2 is more central than R1, which is more central than R3, which is more central than R4, and that R5 is the most eccentric partition among those five (see Figure 2).

In the following, we are going to study this notion of centrality more deeply, by answering the question: what types of partitions are central or eccentric? We will see that this property depends on two main parameters : the number of classes in the partition and the balance of the classes.

6.1 Impact of the number of classes

In this subsection, we study the impact of the number of classes on the centrality of a partition. In this aim, we consider seven partitions of $\mathcal{P}_{100}$ with several numbers of classes (1, 5, 10, 20, 30, 50, and 100), but all with well-balanced classes (we try as far as possible to spread uniformly the 100 elements over the classes). We estimate the distributions of partitions according to the transfer distance to these seven reference partitions. The distributions are represented in Figure 3 in cumulated percentages, as well as the average distribution when considering 10000 pairs of partitions of $\mathcal{P}_{100}$.

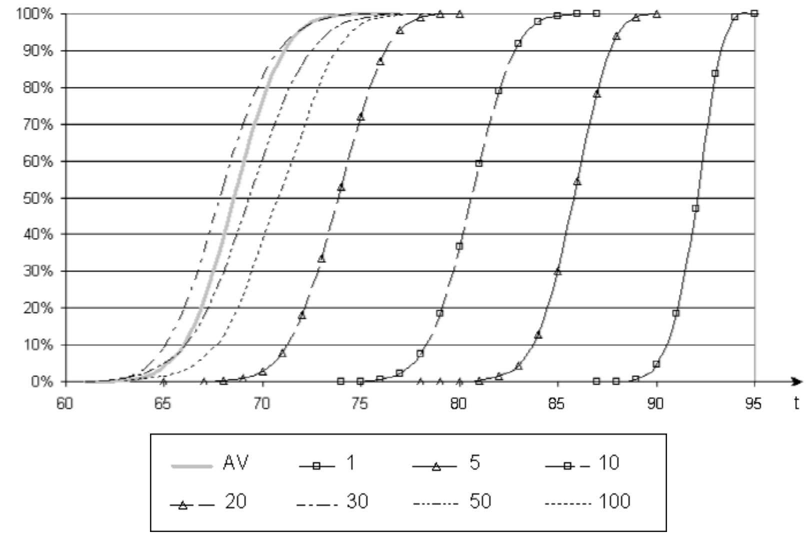

Figure 3: Distributions of the transfer distance to seven well-balanced partitions

We notice on this figure that the graphs vary a lot following the number of classes in the reference partitions. The average transfer values are respectively equal to 92.5, 86.3, 81.1, 74.3, 68.4, 69.8 and 71.2 for the partition with 1, 5, 10, 20, 30, 50, and 100 classes. We observe that the distribution for the partition with one class is very far from the average distribution as its non-zero values belong to $[88; 95]$. When the number of classes increases from 1 to 30, the partitions become more central. The partition with 30 classes is the only one which corresponds to transfer values a little smaller than those of the average distribution. Actually, we did not succeed in finding any partition more central than the well-balanced partition with 30 classes. Then beyond 30 classes the graphs become again less close to the average graph, the partition being less central.

This study shows that the number of classes has a great impact on the centrality of a partition. The centrality seems to be unimodal according to this criterion: partitions with small number of classes are very eccentric, that is very far in average from other partitions of $\mathcal{P}_{100}$; partitions having a number of classes near the average are the most central (the average number of classes for a partition of $\mathcal{P}_{100}$ is 28.6, see Table 1), and partitions with larger number of classes are a little less central.

6.2 Impact of the balance of classes

Let us now study the effect of the balance of the classes on the centrality of a partition. We consider again the case $n = 100$ and five reference partitions having 30 classes, but with the balance of the classes varying: from the most balanced partition (each class contains the same number of elements) to the most unbalanced partition (one class contains almost all the elements, the other classes contain only one element). The estimated distributions are represented in Figure 4. The reference partitions are labelled by the value of the standard deviation of the classes cardinalities, which reflects the balance of the classes.

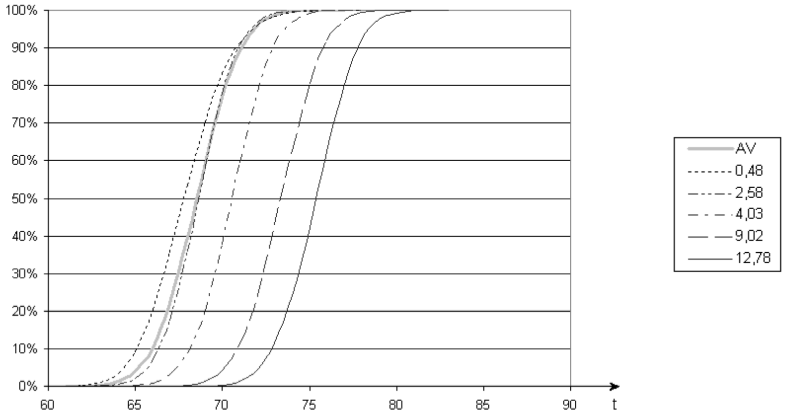

Figure 4: Distributions of the transfer distances to partitions with more or less well-balanced classes

We can observe that the partitions are more central if the standard deviations are smaller. The better-balanced the classes of the reference partitions, and the more central the partition. The variation of the distributions with respect to the standard deviation is a little less considerable than with respect to the number of classes but remains quite sizeable: the average transfer distance for the five reference partitions with 30 classes are respectively equal to 68.38, 69.13, 71.03, 73.83 and 75.84 when the standard deviation increases.

6.3 Conclusion

In this section, we proposed to study the centrality of the partitions of $\mathcal{P}_n$ according to the transfer distance distributions. We showed on some examples that the centrality seems to be unimodal with the number of classes in the considered partition (increasing until some threshold value then decreasing) and increasing with the balance of the classes.

The experiments have been done with other values of n and other reference patitions, and the same results have been obtained. We found no partition being much more central than the average distribution, whereas some partitions turned out to be very eccentric. We observed that the characteristics of central partitions are very close to the average ones, and correspond to very numerous partitions. To the contrary, very few partitions are located very far from the other partitions, which makes those eccentric partitions very special. Although we could not characterize precisely which partition is the most central, we may conjecture that the partition with one class is the most eccentric partition of $\mathcal{P}_n$.

References

[1] Barthélemy J-P, Monjardet B (1981) The median procedure in cluster analysis and social choice theory. Mathematical Social Sciences, 1:235-267

[2] Batagelj V, Mrvrar M, Zaversnik M (1999) Partitioning approach to visualisation of large graphs. In: Graph Drawing, 7th International Symposium, GD'99, Proceedings, Lecture Notes in Computer Science, Springer, 1731:90-97

[3] Boorman SA, Olivier DC (1973) Metrics on Spaces of Finite Trees. Journal of mathematical psychology, 10:26-59

[4] Charon I, Denœud L, Guénoche A, Hudry O (2006) Maximum transfer distance between partitions. Journal of Classification, 23(1):103-121

[5] Charon I, Denœud L, Hudry O (2007) Maximum de la distance de transfert à une partition donnée. Mathématiques et Sciences humaines, 179:45-83

[6] Day W (1981) The complexity of computing metric distances between partitions. Mathematical Social Sciences, 1:269-287

[7] de Frayssex H, Kuntz P (1992) Pagination of large scale networks; embedding a graph in $\mathbb{R}^n$ for effective partitioning. Algorithmic review, 2(3):105-112

[8] Denœud L (2006) Étude de la distance de transfert entre partitions et recherche de zones denses dans un graphe. PhD thesis, Université Paris 1, France

[9] Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification, 2:193-218

[10] Jaccard P (1908) Nouvelle recherche sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles, 44:223-270

[11] Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Res. Logist. Quart., 2:83-97

[12] Lerman IC (1988) Comparing partitions. Mathematical and statistical aspects. Classification and Related Methods of Data Analysis, Bock HH (Ed.) Elsevier Science Publishers, pp. 121-132

[13] Lovász L, Plummer MD (1986) Matching Theory. Annals of Discrete Mathematics 29, North-Holland

[14] Matsuda H, Ishihara T, Hashimoto A (1999) Classifying molecular sequences using a linkage graph with their pairwise similarities. Theoretical Computer Science, 210:305-325

[15] Newman MEJ (2001) The structure of scientific collaboration networks. Proc. Natl. Acad. Sci., USA, 98:404-409

[16] Nijenhuis A, Wilf H (1978 a) Next Partition of an n-Set. In: Combinatorial algorithms, Academic Press, New-York, pp. 88-92

[17] Nijenhuis A, Wilf H (1978 b) Random Partition of an n-Set. In: Combinatorial algorithms, Academic Press, New-York, pp. 93-98

[18] Rand WM (1971) Objective criteria for the evaluation of clustering methods. J. Amer. Stat. Assoc., 66:846-850

[19] Régnier S (1965) Quelques aspects mathématiques des problèmes de classification automatique. I.C.C. Bulletin 4

[20] Rota G-C (1964) The Number of Partitions of a Set. American Mathematical Monthly 71(5): 498-504

[21] Tijms H (2004) Understanding Probability: Chance Rules in Everyday Life. Cambridge University Press, Cambridge

[22] Wallace DL (1983) Comment. J. of the Am. Stat. Assoc., 78:569-579