Math. Sci. hum / Mathematics and Social Sciences (50e année, n° 197, *2012*(1), p. 83-97)

NP-HARDNESS OF THE COMPUTATION OF A MEDIAN EQUIVALENCE RELATION IN CLASSIFICATION (RÉGNIER'S PROBLEM)[1]

Olivier HUDRY[2]

In memoriam Jean-Pierre Barthélemy

RÉSUMÉ – NP-difficulté de la détermination d'une relation d'équivalence médiane en classification (problème de Régnier)

Étant donnée une collection Π de relations d'équivalence (ou partitions), le problème de Régnier consiste à déterminer une relation d'équivalence qui minimise l'éloignement par rapport à Π. L'éloignement est fondé sur la distance de la différence symétrique et mesure le nombre de désaccords entre Π et la relation d'équivalence considérée. Une telle relation d'équivalence minimisant l'éloignement est appelée une relation d'équivalence médiane de Π. On montre ici la NP-difficulté du problème de Régnier, c'est-à-dire du calcul d'une relation d'équivalence médiane d'une collection Π de relations d'équivalence, du moins quand le nombre de relations d'équivalence de Π est suffisamment grand.

MOTS CLÉS – Agrégation de relations, Classification, Complexité, Distance de la différence symétrique, NP-complétude, Partition, Problème de Régnier, Problème de Zahn, Relation d'équivalence, Relation médiane

ABSTRACT – *Given a collection Π of equivalence relations (or partitions), Régnier's problem consists in computing an equivalence relation which minimizes the remoteness from Π. The remoteness is based on the symmetric difference distance and measures the number of disagreements between Π and the considered equivalence relation. Such an equivalence relation minimizing the remoteness is called a median equivalence relation of Π. We prove the NP-hardness of Régnier's problem, i.e. the computation of a median equivalence relation of a collection of equivalence relations, at least when the number of equivalence relations of Π is large enough.*

KEYWORDS – Aggregation of relations, Classification, Complexity, Equivalence relation, NP-completeness, Median relation, Partition, Régnier's problem, Symmetric difference distance, Zahn's problem

1. INTRODUCTION

Among several other scientific fields, Jean-Pierre Barthélemy was interested in the theory of algorithmic complexity and in classification (he was the vice-president of the French-speaking society of classification – SFC – in 1992-1993 and then the president of the SFC in 1994-1995 [SFC, 2012]). Two of his books give evidence of this: *Algorithmic Complexity and Communication Problems*, with Gérard Cohen and Antoine

[1] Research supported by the ANR project "Computational Social Choice" n° ANR-09-BLAN-0305-03.

[2] Télécom ParisTech, 46, rue Barrault, 75634 Paris Cedex 13, olivier.hudry@telecom-paristech.fr

Lobstein [Barthélemy *et al.*, 1992], and *Éléments de classification* with François Brucker [Brucker, Barthélemy, 2007]. The issues dealt with in this paper belong to the intersection of these two fields, since they deal with complexity results of problems arising from classification, namely Zahn's problem [Zahn, 1964] and Régnier's problem [Régnier, 1965] (for references on classification, see for instance [Arabie *et al.*, 1996], [Barthélemy *et al.*, 1995], [Barthélemy, Monjardet, 1981], [Brossier, 2003], [Brucker, Barthélemy, 2007], [Everitt *et al.*, 2011], [Mirkin, 1996], [Romesburg, 2004]).

Imagine that we want to partition a finite set X of n objects into clusters so that the objects in a same cluster look like similar while the objects of two different clusters look like dissimilar. Of course, we must specify what we mean by "similar" and "dissimilar". For this, assume that we have p criteria (R_1, R_2, ..., R_p). Each criterion R_k ($1 \leq k \leq p$) is in fact a binary relation defined as a subset of the Cartesian product $X \times X$. It is usually assumed that R_k is at least *reflexive* (x is in relation with itself with respect to R_k) and *symmetric* (x and y are in relation with respect to R_k if and only if y and x are in relation with respect to R_k). When two elements x and y of X are in relation with respect to R_k, we consider that x and y are similar with respect to the k^{th} criterion. In some context, we may also assume that R_k is *transitive* (if x and y are in relation with respect to R_k and if y and z are also in relation with respect to R_k, then x and z must be in relation with respect to R_k; in other words, if x and y are similar as well as y and z, then x and z are similar too, still with respect to R_k). If R_k is simultaneously reflexive, symmetric and transitive, then it is an *equivalence relation* or equivalently a *partition*. Indeed, partitioning X is the same as defining an equivalence relation on X since the clusters of the partition will provide the equivalence classes of the equivalence relation and conversely.

With this context:

- the problem considered by C.T. Zahn [1964] is the one for which we have only one criterion ($p = 1$) which is associated with a reflexive and symmetric relation R defined on X and we look for an equivalence relation defined on X fitting R "as well as possible",
- the problem considered by S. Régnier [1965] is the one for which we have p criteria (with $p \geq 1$) which are equivalence relations defined on X and we look for an equivalence relation defined on X summarizing these equivalence relations "as well as possible",

where "as well as possible" will be specified in Section 2.

M. Křivánek and J. Morávek [1986] proved that Zahn's problem is NP-hard (more precisely, the decision problem associated with Zahn's problem – see below – is NP-complete). From this result, we can deduce the NP-hardness of Régnier's problem, as done in [Barthélemy, Leclerc, 1995]. As this complexity result is not much detailed in [Barthélemy, Leclerc, 1995], the aim of this paper consists in making explicit the links between Zahn's problem and Régnier's problem from the complexity point of view.

For this, Section 2 specifies some definitions and notations. Section 3 provides the way of computing the remoteness. The complexity results can be found in Section 4 and the conclusion in Section 5.

2. DEFINITIONS AND NOTATIONS

Let $X = \{1, 2, ..., n\}$ be a finite set with n elements; we assume in the following that n is greater than or equal to 2. A binary relation R defined on X is a subset of the Cartesian product $X \times X$. If (x, y) belongs to R, then we write xRy; otherwise we write $x\overline{R}y$. Basic properties that R may fulfil are:

- *reflexivity*: R is reflexive if, for any $x \in X$, we have xRx;
- *irreflexivity*: R is irreflexive if, for any $x \in X$, we have $x\overline{R}x$;
- *symmetry*: R is symmetric if, for any $(x, y) \in X^2$ with $x \neq y$, we have the equivalence $xRy \Leftrightarrow yRx$;
- *transitivity*: R is transitive if, for any $(x, y, z) \in X^3$ with $x \neq y \neq z \neq x$, we have the implication (xRy and yRz) $\Rightarrow xRz$.

From these basic properties, we may define more sophisticated relations, as the structure of *equivalence relation*: an equivalence relation is a reflexive, symmetric and transitive relation.

From the point of view of the theory of NP-completeness (see [Barthélemy *et al.*, 1992] or [Garey, Johnson, 1979] for references on this theory), reflexivity does not matter: the results would remain the same if we require the reflexivity property, or the irreflexivity property, or if we require nothing about reflexivity or irreflexivity (see [Hudry, 2008]).

In the following, we shall be interested in reflexive and symmetric relations and in equivalence relations; $\mathcal{E}$ will denote the set of equivalence relations defined on X.

A *profile* $\Pi = (R_1, R_2, ..., R_p)$ defined on X is a collection (or multi-set) of p binary relations R_k ($1 \leq k \leq p$) defined on X, where p is a positive integer. In our context, p denotes the number of criteria and R_k ($1 \leq k \leq p$) describes the similarities according to the k^{th} criterion. Note that these relations are not necessarily distinct: two different criteria may induce the same relation.

To define what we meant by "as well as possible" in Section 1, we use the *symmetric difference distance* δ between two binary relations R and R' both defined on X. This distance is defined by:

$$\delta(R, R') = \left| R \,\Delta\, R' \right|,$$

where Δ denotes the usual *symmetric difference between sets*. This distance, which owns good axiomatic properties (see [Barthélemy, 1979] and [Barthélemy, Monjardet, 1981]), measures the number of disagreements between R and R':

$$\delta(R, R') = \left| \left\{ (x, y) \in X^2 : [xRy \text{ and } x\overline{R}'y] \text{ or } [x\overline{R}y \text{ and } xR'y] \right\} \right|.$$

The *remoteness* [Barthélemy, Monjardet, 1981] $\rho(\Pi, R)$ between a profile $\Pi = (R_1, R_2, \ldots, R_p)$ and a binary relation R is defined by:

$$\rho(\Pi, R) = \sum_{k=1}^{p} \delta(R_k, R).$$

So, the remoteness $\rho(\Pi, R)$ measures the total number of disagreements between Π and R. A *median equivalence relation* (also called *central partition* by S. Régnier [1965] or still *consensus partition* – see [Hudry, Monjardet, 2010] for consensus theories; for references on the median procedure and on the use of ordered sets in classification, see for instance [Barthélemy, Leclerc, 1995], [Barthélemy *et al.*, 1986], [Barthélemy, Monjardet, 1981, 1988], [Hudry *et al.*, 2006]) is a relation E^* belonging to $\mathcal{E}$ and minimizing ρ:

$$\rho(\Pi, E^*) = \min_{E \in \mathcal{E}} \rho(\Pi, E).$$

This formulation generalizes Régnier's problem and Zahn's problem: indeed, Régnier's problem is the computation of a median equivalence relation of a profile of p equivalence relations, while Zahn's problem is the computation of a median equivalence relation of a profile reduced to only one reflexive and symmetric relation ($p = 1$ in this case). A third problem related to these two ones is the case for which the p binary relations R_k ($1 \leq k \leq p$) of the profile are reflexive and symmetric (but not necessarily transitive).

In the sequel, we shall pay attention also to the values taken by p with respect to n (the cardinality of X) in Régnier's problem, in order to bring partial answers to the following question: what is the minimum number of p for which Régnier's problem is NP-hard? We shall see in particular that the parity of p plays a role in the way to summarize a profile of p symmetric relations thanks to a matrix (the *majority matrix*, see below).

More precisely, we are going to consider mainly the three decision problems specified below:

Name: Zahn's decision problem (noted ZDP below)
Data: a finite set X, a reflexive and symmetric relation S defined on X; an integer h;
Question: does there exist an equivalence relation E defined on X with $\delta(S, E) \leq h$?

Name: Aggregation of an odd number of equivalence relations into an equivalence relation (Régnier's decision problem for an odd number of equivalence relations, noted O-RDP below)
Data: a finite set X, a positive odd integer p, a profile $\Pi = (E_1, E_2, \ldots, E_p)$ of p equivalence relations defined on X; an integer h;
Question: does there exist an equivalence relation E defined on X with $\rho(\Pi, E) \leq h$?

Name: Aggregation of an even number of equivalence relations into an equivalence relation (Régnier's decision problem for an even number of equivalence relations, noted E-RDP below)
Data: a finite set X, a positive even integer p, a profile $\Pi = (E_1, E_2, \ldots, E_p)$ of p equivalence relations defined on X; an integer h;
Question: does there exist an equivalence relation E defined on X with $\rho(\Pi, E) \leq h$?

To study the links between ZDP and O-RDP or E-RDP from the complexity point of view, we need some extra notations.

Let R be a reflexive and symmetric relation defined on X. We denote by $C = \left(c_{xy}\right)_{(x,y)\in X^2}$ the *characteristic matrix associated with* R, i.e. the matrix defined, for any pair x and y of elements of X, by $c_{xy} = 1$ if we have xRy and $c_{xy} = 0$ otherwise. Similarly, if $\Pi = (R_1, R_2, \ldots, R_p)$ is a profile of p reflexive and symmetric relations defined on X, for $1 \leq k \leq p$, let $\left(c^k_{xy}\right)_{(x,y)\in X^2}$ be the characteristic matrix of R_k: c^k_{xy} is equal to 1 if we have xR_ky and to 0 otherwise. Note the equalities $c_{xx} = 1$ and $c^k_{xx} = 1$ for any x belonging to X and any k between 1 and p since the relations R and R_k are assumed to be reflexive, and the equalities $c_{xy} = c_{yx}$ and $c^k_{xy} = c^k_{yx}$ for any elements x and y of X and any k between 1 and p since the relations R and R_k are assumed to be symmetric.

Now, for any pair x and y of elements of X, let $m^{\Pi}_{xy} = 2\sum_{k=1}^{p} c^k_{xy} - p$ denote twice the number of relations R_k of Π for which we have xR_ky (i.e., in our context, the number of criteria for which x and y are considered as similar) minus the total number p of relations. Note that the quantities m^{Π}_{xy} have the same parity as p and that they range between $-p$ and p: a positive (respectively negative) value of m^{Π}_{xy} means that x and y are similar (respectively dissimilar) for at least half the criteria; a value of m^{Π}_{xy} equal to p (respectively $-p$) means that x and y are similar for all (respectively none of) the criteria; more generally, the larger the value of m^{Π}_{xy}, the more similar x and y. In particular we have, for any x: $m^{\Pi}_{xx} = p$, since all the considered relations are reflexive. Similarly, as all the relations R_k for $1 \leq k \leq p$ are assumed to be symmetric, we have $m^{\Pi}_{yx} = m^{\Pi}_{xy}$ for any x and any y. In the following, the matrix $M^{\Pi} = \left(m^{\Pi}_{xy}\right)_{(x,y)\in X^2}$ will be called the *majority matrix* of the profile Π. Note that, for Zahn's problem, Π contains only one reflexive and symmetric relation S: $p = 1$ and $\Pi = (S)$; then the majority matrix M^{Π} of Π contains only 1's and –1's, and is equal to $2C_S - 1_{nn}$, where C_S denotes the characteristic matrix of S and 1_{nn} is the (n, n)-matrix of which all the entries are equal to 1.

Last, we shall use the following four kinds of equivalence relations:

- the equivalence relation X^2 which contains only one class (which gathers all the elements of X); note that the majority matrix of a profile containing only one copy of X^2 is the matrix 1_{nn};
- the equivalence relation U which contains n classes (each class contains only one element of X); note that the majority matrix of a profile containing only one copy of U is the matrix $2I_{nn} - 1_{nn}$, where I_{nn} denotes the (n, n)-identity matrix (in fact, I_{nn} is the characteristic matrix of U);
- for i with $1 \leq i \leq n$, the equivalence relation U_i contains two classes: the first one contains only the element i of X, the other one contains all the other elements of X;

- for i and j with $1 \leq i \leq n$, $1 \leq j \leq n$ and $i \neq j$, U_{ij} contains $n - 1$ classes: the first one contains the two elements i and j of X, each other class contains only one element of X (different from i and j).

3. STATEMENT OF THE REMOTENESS

Given a profile Π, Lemma 1 provides a link between the remoteness ρ and the entries of the majority matrix of Π.

LEMMA 1. *Let $\Pi = (R_1, R_2, \ldots, R_p)$ be a profile of p reflexive and symmetric relations defined on X. Let R be a reflexive and symmetric relation also defined on X and with $(c_{xy})_{(x,y)\in X^2}$ as its characteristic matrix. Then we have:*

$$\rho(\Pi, R) = \lambda_\Pi - \sum_{x \neq y} m_{xy}^{\Pi} \cdot c_{xy}$$

where λ_Π does not depend on R and where the m_{xy}^{Π}'s are the entries of the majority matrix M^{Π} of Π.

Proof. By the definition of the remoteness, we have:

$$\rho(\Pi, R) = \sum_{k=1}^{p} \delta(R_k, R).$$

Remember that $\delta(R_k, R)$ measures the number of disagreements between R and R_k:

$$\delta(R_k, R) = \left| \left\{ (x, y) \in X^2 : [xR_k y \text{ and } x\bar{R}y] \text{ or } [x\bar{R}_k y \text{ and } xRy] \right\} \right|.$$

This can be stated thanks to the quantities c_{xy}^k and c_{xy}, where $(c_{xy}^k)_{(x,y)\in X^2}$ is the characteristic matrix of R_k ($1 \leq k \leq p$):

$$\delta(R_k, R) = \sum_{(x,y)\in X^2} \left| c_{xy}^k - c_{xy} \right|.$$

Because the quantities c_{xy}^k and c_{xy} are equal to 1 or 0, we have also:

$$\delta(R_k, R) = \sum_{(x,y)\in X^2} \left| c_{xy}^k - c_{xy} \right| = \sum_{(x,y)\in X^2} \left(c_{xy}^k - c_{xy} \right)^2$$

$$= \sum_{(x,y)\in X^2} c_{xy}^k + \sum_{(x,y)\in X^2} (1 - 2c_{xy}^k) \cdot c_{xy} .$$

From this and from the fact that the considered relations are reflexive, we obtain:

$$\rho(\Pi, R) = \sum_{k=1}^{p} \delta(R_k, R)$$

$$= \sum_{k=1}^{p} \sum_{(x,y)\in X^2} c_{xy}^k + \sum_{k=1}^{p} \sum_{(x,y)\in X^2} (1 - 2c_{xy}^k) \cdot c_{xy}$$

$$= \sum_{k=1}^{p} \sum_{(x,y)\in X^2} c_{xy}^k + \sum_{x\in X} \sum_{k=1}^{p} (1 - 2c_{xx}^k) \cdot c_{xx} + \sum_{x\neq y} \sum_{k=1}^{p} (1 - 2c_{xy}^k) \cdot c_{xy}$$

$$= \lambda_\Pi - \sum_{x\neq y} m_{xy}^\Pi \cdot c_{xy}$$

where $\lambda_\Pi = \sum_{k=1}^{p} \sum_{(x,y)\in X^2} c_{xy}^k - np$ is a constant for any given profile Π. ♦

Lemma 2 specifies the expression of the remoteness ρ when applied to a profile obtained as the concatenation of two profiles.

LEMMA 2. *Let* $\Pi_1 = (R_1, R_2, \ldots, R_{p_1})$ *(respectively* $\Pi_2 = (S_1, S_2, \ldots, S_{p_2})$*) be a profile of* p_1 *(respectively* p_2*) reflexive and symmetric relations defined on X and let* Π *be the profile obtained as the concatenation of* Π_1 *and* Π_2*:* $\Pi = (R_1, R_2, \ldots, R_{p_1}, S_1, S_2, \ldots, S_{p_2})$*. Then we have, for any relation R defined on X:*

$$\rho(\Pi, R) = \rho(\Pi_1, R) + \rho(\Pi_2, R).$$

Proof. By the definition of the remoteness, we have:

$$\rho(\Pi, R) = \sum_{k=1}^{p_1} \delta(R_k, R) + \sum_{i=1}^{p_2} \delta(S_k, R) = \rho(\Pi_1, R) + \rho(\Pi_2, R). \quad ♦$$

Similarly, Lemma 3 specifies the expression of the majority matrix of a profile obtained as the concatenation of two profiles.

LEMMA 3. *Let* Π_1 *and* Π_2 *be two profiles of reflexive and symmetric relations defined on X and let* Π *be the profile obtained as the concatenation of* Π_1 *and* Π_2*. Then the majority matrix of* Π *is the sum of the majority matrices of* Π_1 *and* Π_2*.*

Proof. Remember (see Section 2) that the entries m_{xy}^Π of the majority matrix of the profile Π are equal to twice the number of relations of Π for which x and y are together minus the number of relations belonging to Π. Hence the result. ♦

4. COMPLEXITY OF RÉGNIER'S PROBLEM

As said above, M. Křivánek and J. Morávek [1986] studied the complexity of Zahn's problem. More precisely, they proved the following theorem (see also [Brucker, Barthélemy, 2007]):

THEOREM 4. *The decision problem ZDP associated with the aggregation of one reflexive and symmetric relation into an equivalence relation is NP-complete.*

From the NP-completeness of ZDP, we are going to prove the NP-completeness of O-RDP and of E-RDP. This result, stated first by J.-P. Barthélemy and B. Leclerc in [1995], is based on a construction designed by B. Debord [1987], allowing to build a profile Π of equivalence relations from a profile Π' of reflexive and symmetric relations with $\rho(\Pi, R) = \rho(\Pi', R) + \lambda$, for any reflexive and symmetric relation R and where λ does not depend on R. Unfortunately, this construction is not utterly correct and contains some mistakes (see a corrected construction in [Hudry, 2012]), though this does not invalidate the complexity result of [Barthélemy, Leclerc, 1995]. Moreover, when the profile Π' contains only one reflexive and symmetric relation, as it will be the case for us, it is possible to design a construction involving a smaller number of equivalence relations than in Debord's construction (Debord's construction may involve $O(n^3)$ equivalence relations, while the transformations of Theorems 8 and 9 involve only $O(n^2)$ equivalence relations). We are going to detail this more efficient construction below. We first state some lemmas, useful to reach this aim.

LEMMA 5. *Let i and j be integers with $1 \leq i < j \leq n$. Let $M_{ij}^+ = \left(m_{ij}^+(x,y)\right)_{(x,y)\in X^2}$ be the symmetric matrix defined by:*

1. $m_{ij}^+(i,j) = m_{ij}^+(j,i) = 2$;
2. *for any integer x with $1 \leq x \leq n$, $m_{ij}^+(x,x) = 2$;*
3. *for x and y with $(x,y) \neq (i,j)$, $(x,y) \neq (j,i)$ and $x \neq y$, $m_{ij}^+(x,y) = 0$.*

Then M_{ij}^+ is the majority matrix of the profile $\Pi_{ij}^+ = (U_{ij}, X^2)$.

Proof. It is straightforward to check that the majority matrix of (U_{ij}, X^2) is indeed M_{ij}^+: the elements i and j are together twice in Π_{ij}^+, while the other pairs of distinct elements are together once in Π_{ij}^+; the diagonal entries are equal to the number of equivalence relations of the profile, i.e. 2. ♦

LEMMA 6. *Let S be a reflexive and symmetric relation defined on X, and let $M^{(S)} = \left(m_{ij}^{(S)}\right)_{(i,j)\in X^2}$ be the majority matrix of the profile (S) reduced to S. Then, there exists a profile $\Pi_0(S)$ of an even number of equivalence relations such that the non-diagonal entries of the majority matrix of $\Pi_0(S)$ are the non-diagonal entries of $M^{(S)} + 1_{nn}$. Moreover, $\Pi_0(S)$ contains at most $n(n-1)$ equivalence relations and half of the equivalence relations of $\Pi_0(S)$ are equal to X^2.*

Proof. As noticed above (see Section 2), all the entries of $M^{(S)}$ belong to {–1, 1} and all the diagonal entries of $M^{(S)}$ are equal to 1. Then all the entries of $M^{(S)} + 1_{nn}$ belong to {0, 2} and all the diagonal entries of $M^{(S)} + 1_{nn}$ are equal to 2. Let α_S denote the number of entries $m_{ij}^{(S)}$ of $M^{(S)}$ with $i < j$ and $m_{ij}^{(S)} = 1$; note that α_S is between 0 (if S is the equivalence relation U) and $n(n-1)/2$ (if S is the equivalence relation X^2).

If α_S is not equal to 0, then $M^{(S)} + 1_{nn}$ can be written as follows:

$$M^{(S)} + 1_{nn} = \sum_{i<j \text{ and } m_{ij}^{(S)}=1} M_{ij}^{+} + (2 - 2\alpha_S) I_{nn}.$$

According to Lemma 5, M_{ij}^{+} is the majority matrix of (U_{ij}, X^2). So, thanks to Lemma 3, $\sum_{i<j \text{ and } m_{ij}^{(S)}=1} M_{ij}^{+}$ is the majority matrix of the profile $\Pi_0(S)$ obtained by concatenating the α_S profiles (U_{ij}, X^2) for i and j with $i < j$ and $m_{ij}^{(S)} = 1$. Note that the non-diagonal entries of the majority matrix of $\Pi_0(S)$ and the ones of $M^{(S)} + 1_{nn}$ are then the same. Moreover, $\Pi_0(S)$ contains $2\alpha_S \leq n(n-1)$ equivalence relations of which half of them are equal to X^2.

If α_S is equal to 0, then $M^{(S)} + 1_{nn}$ is equal to $2I_{nn}$, which is the majority matrix of the profile $\Pi_0(S) = (U, X^2)$.

In both cases, we obtain the result stated in Lemma 6. ♦

LEMMA 7. *Let S be a reflexive and symmetric relation defined on X, and let $M^{(S)}$ be the majority matrix of the profile (S) reduced to S. Then, there exists a profile $\Pi_1(S)$ of an odd number of equivalence relations such that the non-diagonal entries of the majority matrix of $\Pi_1(S)$ are the non-diagonal entries of $M^{(S)}$. Moreover, $\Pi_1(S)$ contains at most $n(n-1) - 1$ equivalence relations.*

Proof. With the same notations as for Lemma 6 and its proof, we know by Lemma 6 that there exists a profile $\Pi_0(S)$ of $p_S = \max(2\alpha_S, 2)$ equivalence relations such that:

- the non-diagonal entries of the majority matrix of $\Pi_0(S)$ are the non-diagonal entries of $M^{(S)} + 1_{nn}$;
- $p_S/2$ of the equivalence relations of $\Pi_0(S)$ are equal to X^2.

Consider the profile $\Pi_1(S)$ of $p_S - 1$ equivalence relations obtained from $\Pi_0(S)$ by removing one copy of X^2. As the majority matrix of the profile reduced to X^2 is 1_{nn}, the non-diagonal entries of the majority matrix of $\Pi_1(S)$ are the non-diagonal entries of $M^{(S)}$, by Lemma 3. Hence the result, since $\Pi_1(S)$ contains $p_S - 1$ equivalence relations with $p_S \leq n(n-1)$. ♦

We may now study the complexity of Régnier's problem. We begin with the case for which the number of relations is odd.

Observe that an equivalence relation E defined on X and with q classes may be described as a vector v of n integers belonging to $\{1, 2, ..., q\}$: the i^{th} component of v specifies the number of the equivalence class of E which the i^{th} element of X is assigned to. As an equivalence class cannot be empty, all the values between 1 and q must appear in v.

THEOREM 8. *O-RDP is NP-complete*.

Proof. We proceed in two steps: we first show that O-RDP belongs to NP; then we prove the NP-completeness of O-RDP by reducing ZDP to O-RDP in polynomial time (ZDP $\prec$ O-RDP).

To show that O-RDP belongs to NP, consider any instance $I = (X, p, \Pi = (E_1, E_2, \ldots, E_p), h)$ as defined in Section 2 with p odd and assume that we are given a vector v^* of n integers supposed to define an equivalence relation E^* on X satisfying the inequality $\rho(\Pi, E^*) \leq h$. We want to check the two properties:

- E^* is indeed an equivalence relation;
- the remoteness of E^* from Π is at most h.

Checking that E^* is an equivalence relation can be done in $O(n)$ since it is sufficient to check that the components of v^* define a set of consecutive integers of which the minimum value is 1. Checking the inequality $\rho(\Pi, E^*) \leq h$ can be done in $O(n^2p)$ since the computation of $\delta(E_k, E^*)$ for $1 \leq k \leq p$ can be done in $O(n^2)$. So checking both properties can be done in $O(n + n^2p)$. Describing an equivalence relation E defined on X requires at least n bits (at least one bit for each element x of X in order to specify the number of the equivalence class of E which x belongs to; in fact, it requires more, but it does not matter here). So the (binary) size of I is at least np. As $n + n^2p$ can be upper-bounded by a polynomial with respect to np, then we can check the two properties in polynomial time with respect to the (binary) size of the instance I. Hence the belonging of O-RDP to NP.

We turn now to the second step: ZDP $\prec$ O-RDP. For this, consider any instance $I_Z = (X, S, h_Z)$ of ZDP as defined in Section 2. We want to transform it, in polynomial time, into an instance I_R of O-RDP admitting the same answer as I_Z. In order to define I_R, we keep the same set X on which the equivalence relations are going to be defined. Then we consider the profile $\Pi_1(S)$ as defined in Lemma 7 and the number p of equivalence relations contained in $\Pi_1(S)$ (i.e., $p_S = \max(2\alpha_S, 2) - 1$, with the same notations as in Lemma 7). Let $M^{\Pi_1(S)} = \left(m_{xy}^{\Pi_1(S)}\right)_{(x,y)\in X^2}$ and $M^{(S)} = \left(m_{xy}^{(S)}\right)_{(x,y)\in X^2}$ be respectively the majority matrix of $\Pi_1(S)$ and the majority matrix of the profile (S) containing only one relation, namely S. Thanks to Lemma 7, we know that the non-diagonal entries of $M^{\Pi_1(S)}$ and the ones of $M^{(S)}$ are the same: for $x \neq y$, $m_{xy}^{\Pi_1(S)} = m_{xy}^{(S)}$. Let $\lambda_{\Pi_1(S)}$ and $\lambda_{(S)}$ be the constants computed in Lemma 1

associated to the profiles $\Pi_1(S)$ and (S). Then we set: $h_R = h_Z + \lambda_{\Pi_1(S)} - \lambda_{(S)}$. Thus I_R is equal to $(X, p_S, \Pi_1(S), h_Z + \lambda_{\Pi_1(S)} - \lambda_{(S)})$.

Encoding S requires $n(n-1)/2$ bits in order to know, for any elements x and y of X with $1 \leq x < y \leq n$ whether we have xSy or $x\overline{S}y$. So the size of I_Z is at least $n(n-1)/2$. With respect to I_Z, the definition of I_R requires only to construct the p_S relations of $\Pi_1(S)$ and the computation of $\lambda_{\Pi_1(S)}$ and $\lambda_{(S)}$. Each equivalence relation of $\Pi_1(S)$ can be described by at most about $n.\log_2 n$ bits by specifying, for any element x of X, which equivalence class contains x (such an equivalence class can be specified by a number less than or equal to n; the encoding of this number thus requires at most about $\log_2 n$ bits). Moreover, the computation of $\lambda_{\Pi_1(S)}$ and $\lambda_{(S)}$ can be done in $O(p_S.n^2)$ and $O(n^2)$ respectively. So, the construction of I_R can be done in $O(p_S.n.\log_2 n + p_S.n^2)$, i.e. in $O(n^4)$ since p_S is upper-bounded by n^2. Hence the polynomiality of the transformation, since n^4 can obviously be upper-bounded by a polynomial in $n(n-1)/2$.

We must now check that the transformation keeps the answer: I_Z admits the answer "yes" if and only if I_R admits the answer "yes". Let E be an equivalence relation defined on X and let $(e_{xy})_{(x,y)\in X^2}$ be the characteristic matrix of E. Then we have, by Lemma 1:

$$\rho(\Pi_1(S), E) = \lambda_{\Pi_1(S)} - \sum_{x \neq y} m_{xy}^{\Pi_1(S)} \cdot e_{xy} .$$

Similarly, we have, by considering the profile (S) reduced to S:

$$\rho((S), E) = \lambda_{(S)} - \sum_{x \neq y} m_{xy}^{(S)} \cdot e_{xy} .$$

By Lemma 7, we have:

$$\sum_{x \neq y} m_{xy}^{\Pi_1(S)} \cdot e_{xy} = \sum_{x \neq y} m_{xy}^{(S)} \cdot e_{xy}$$

what involves:

$$\rho((S), E) = \rho(\Pi_1(S), E) - \lambda_{\Pi_1(S)} + \lambda_{(S)}.$$

So, as h_R is equal to $h_Z + \lambda_{\Pi_1(S)} - \lambda_{(S)}$, we have $\rho(\Pi_1(S), E) \leq h_R$ if and only if $\rho((S), E) \leq h_Z$, and thus the transformation keeps the answer since $\rho((S), E)$ and $\delta(S, E)$ are equal.

In conclusion, we have the following results: O-RDP $\in$ NP, ZDP $\prec$ O-RDP, and ZDP is NP-complete (Theorem 4). Hence the NP-completeness of O-RDP. ♦

The study of the case for which the number of relations is even is easier.

THEOREM 9. *E-RDP is NP-complete*.

Proof. The belonging of E-RDP to NP can be done exactly as the one of O-RDP (see the proof of Theorem 8: the parity of the number of relations does not play any role).

To prove that E-RDP is NP-complete, we are going to transform O-RDP into E-RDP in polynomial time: O-RDP $\prec$ E-RDP.

For this, consider any instance $I_O = (X_O, p_O, \Pi_O, h_O)$ of O-RDP as defined in Section 2. We want to transform it, in polynomial time, into an instance $I_e = (X_e, p_e, \Pi_e, h_e)$ of E-RDP admitting the same answer as I_O. In order to define I_e, we keep the same set X_O on which the equivalence relations are going to be defined. Then Π_e is obtained by the duplication of Π_O (in other words, we concatenate Π_O with itself), what involves $p_e = 2p_O$. Last we set $h_e = 2h_O$.

This transformation is obviously polynomial (and even linear) and keeps the answer, thanks to Lemma 2. Indeed we have, for any equivalence relation E: $\rho(\Pi_e, E) = 2\rho(\Pi_O, E)$. So $\rho(\Pi_e, E)$ is less than or equal to h_e if and only if $\rho(\Pi_O, E)$ is less than or equal to h_O.

The NP-completeness of O-RDP (Theorem 8) yields the one of E-RDP. ♦

We may also pay attention to the aggregation of p reflexive and symmetric relations into an equivalence relation, as in the next theorem, still easier than the previous one. (Note that Y. Wakabayashi [1986, 1998] proved the NP-completeness of the following problem when the number p of relations is large enough with respect to n.)

THEOREM 10. *Let p be any positive integer. The following problem is NP-complete.*

Name: Aggregation of p reflexive and symmetric relations into an equivalence relation (ApSRER)
Data: a finite set X, a profile $\Pi = (S_1, S_2, \ldots, S_p)$ of p reflexive and symmetric relations defined on X; an integer h;
Question: does there exist an equivalence relation E defined on X with $\rho(\Pi, E) \leq h$?

Proof. The belonging of ApSRER to NP is easy and is left to the reader.

To prove that ApSRER is NP-complete, we transform ZDP into ApSRER in polynomial time: ZDP $\prec$ ApSRER.

For this, consider any instance $I_Z = (X, S, h_Z)$ of ZDP as defined in Section 2. In order to transform it, in polynomial time, into an instance of ApSRER admitting the same answer as I_Z, we keep the same set X on which the equivalence relations are going to be defined, we consider the profile Π containing S exactly p times and we set $h = ph_Z$.

This transformation is obviously polynomial (and even linear), since p is fixed, and keeps the answer, thanks to Lemma 2. Indeed we have, for any equivalence relation E: $\rho(\Pi, E) = p\delta(S, E)$. So $\rho(\Pi, E)$ is less than or equal to h if and only if $\delta(S, E)$ is less than or equal to h_Z.

The NP-completeness of ZDP (Theorem 4) yields the one of ApSRER. ♦

As said above, as the reflexivity or irreflexivity properties do not change anything to the complexity results, the previous complexity results can be extended to the cases where the considered relations must be irreflexive or where nothing is specified about reflexivity or irreflexivity.

5. CONCLUSION

The results of the previous section show that Régnier's problem, i.e. the aggregation of p equivalence relations into a median equivalence relation, is an NP-hard problem, and remains so even if we fix the parity of p. From a practical point of view, this involves that the computation of an exact solution may require a prohibitive CPU time; then the application of heuristics (as done in [de Amorim *et al.*, 1992]; see also [Guénoche, 2011] for a more recent reference or [Guénoche, 2012] in this special issue) may be more realistic. We may observe anyway that the range of the number p of equivalence relations involved in the proof is n^2 (instead of n^3 with the corrected version – see [Hudry, 2012] – of the construction designed by B. Debord [1987]). Is it possible to design a polynomial transformation involving less equivalence relations?

For instance, instead of considering $M^{(S)} + 1_{nn}$ in the proof of Lemma 6, we may decompose $M^{(S)} - 1_{nn}$ thanks to matrices M^{-}_{ij} defined as M^{+}_{ij} but with –2 instead of 2 for the two non-diagonal entries not equal to 0, and associate a profile of equivalence relations to M^{-}_{ij}. Unfortunately, as the entries equal to 1 or to –1 do not play the same role in a majority matrix, the profiles of equivalence relations associated with M^{-}_{ij} in Debord's corrected construction [Hudry, 2012] require a greater number of equivalence relations than the profiles associated with M^{+}_{ij} (the ratio is about n). Then, this strategy usually does not improve qualitatively the sizes of the associated profiles (to obtain smaller profiles thanks to this strategy, it is necessary that the number of entries of $M^{(S)}$ equal to –1 is less than or equal to n; but, on the opposite, if $M^{(S)}$ contains many entries equal to –1, then this strategy leads to profiles with about n^3 equivalence relations; this is why we do not develop such a construction here).

Indeed, an interesting question would be to determine the complexity of Régnier's problem when p is smaller than n^2, in particular when p is a constant: is there a constant p for which the aggregation of p equivalence relations into a median equivalence relation remains NP-hard? Theorem 10 shows that if we consider a profile of p reflexive and symmetric relations instead of a profile of p equivalence relations, then the problem is NP-hard for any value of p greater than or equal to 1. Obviously, for $p = 1$, Régnier's problem (the aggregation of one equivalence relation into a median equivalence relation) is polynomial: the unique equivalence relation of the profile is a median equivalence relation. Similarly, for $p = 2$, Régnier's problem is polynomial: indeed, consider a profile $\Pi = (E_1, E_2)$ of two equivalence relations; it is easy to check that $E_1 \cap E_2$ (the relation keeping only the unanimous pairs) is a median equivalence relation of Π since $E_1 \cap E_2$ gathers exactly all the elements x and y of X for which the entries m^{Π}_{xy} are positive (for the contribution of Jean-Pierre Barthélemy to the study of the unanimity rule – also called Pareto rule –, see [Barthélemy, 1976] and [Monjardet, 2012]). But what about greater values of p?

Acknowledgement. I would like to thank Alain Guénoche and Bruno Leclerc for their help. Their comments were very useful to improve the text.

BIBLIOGRAPHY

AMORIM S.G. (de), BARTHÉLEMY J.-P., RIBEIRO C.-C. (1992), "Clustering and clique partitioning: simulated annealing and tabu search approaches", *J. of Classification* 9, p. 17-42.

ARABIE P., HUBERT L.-J., DE SOETE G. (eds) (1996), *Clustering and Classification*, Singapore and River Edge (NJ), World Scientific Publishers.

BARTHÉLEMY J.-P. (1976), « Sur les éloignements symétriques et le principe de Pareto », *Mathématiques et Sciences humaines* 56, p. 97-125.

BARTHÉLEMY J.-P. (1979), « Caractérisations axiomatiques de la distance de la différence symétrique entre des relations binaires », *Mathématiques et Sciences humaines* 67, p. 85-113.

BARTHÉLEMY J.-P., COHEN G., LOBSTEIN A. (1992), *Complexité algorithmique et problèmes de communications*, Paris, Masson. English translation: *Algorithmic Complexity and Communication Problems*, London, University College London Press, 1996.

BARTHÉLEMY J.-P., LECLERC B. (1995), "The median procedure for partitions", I.J. Cox, P. Hansen, B. Julesz (eds), in *Partitioning data sets*, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 19, Providence (RI), Amer. Math. Soc., p. 3-34.

BARTHÉLEMY J.-P., LECLERC B., MONJARDET B. (1986), "On the use of ordered sets in problems of comparison and consensus of classifications", *J. of Classification* 3, p. 187-224.

BARTHÉLEMY J.-P., MONJARDET B. (1981), "The median procedure in cluster analysis and social choice theory", *Mathematical Social Sciences* 1, p. 235-267.

BARTHÉLEMY J.-P., MONJARDET B. (1988), "The median procedure in data analysis: new results and open problems", H.H. Bock (ed.), *Classification and related methods of data analysis*, Amsterdam, North Holland, p. 309-316.

BROSSIER G. (2003), « Les éléments fondamentaux de la classification », G. Govaert (ed.), *Analyse des données*, Paris, Hermès Lavoisier.

BRUCKER F., BARTHÉLEMY J.-P. (2007), *Éléments de classification*, Paris, Hermès.

DEBORD B. (1987), *Axiomatisation de procédures d'agrégation de préférences*, PhD thesis, Grenoble, University of Grenoble.

EVERITT B., LANDAU S., LEESE M., STAHL D. (2011), *Cluster Analysis*, John Wiley & Sons.

GAREY M.R., JOHNSON D.S. (1979), *Computers and intractability, a guide to the theory of NP-completeness*, New York, Freeman.

GUÉNOCHE A. (2011), "Consensus of partitions: a constructive approach", *Advanced Data Analysis and Classification* 5(3), p. 215-229.

GUÉNOCHE A. (2012), « Sur le consensus en catégorisation libre », *Mathématiques et Sciences humaines* 197, p. 65-82.

HUDRY O. (2008), "NP-hardness results on the aggregation of linear orders into median orders", *Annals of Operations Research* 163(1), p. 63-88.

HUDRY O. (2012), "Majority matrices of profiles of equivalence relations", *Electronic Notes in Discrete Mathematics*, [to appear].

HUDRY O., LECLERC B., MONJARDET B., BARTHÉLEMY J.-P. (2006), « Médianes métriques et laticielles », D. Bouyssou, D. Dubois, M. Pirlot, H. Prade (eds), *Concepts et méthodes pour l'aide à la décision*, vol. 3 : *Analyse multicritère*, Paris, Hermès, p. 271-316. English translation: "Metric and latticial medians", D. Bouyssou, D. Dubois, M. Pirlot, H. Prade (eds), *in Concepts and methods of decision-making process*, Wiley, 2009, p. 771-812.

HUDRY O., MONJARDET B. (2010), "Consensus theories. An oriented survey", *Mathematics and Social Sciences* 190, p. 139-167.

KŘIVÁNEK M., MORÁVEK J. (1986), "NP-hard problems in hierarchical-tree clustering", *Acta Informatica* 23, p. 311-323.

MIRKIN B. (1996), *Mathematical Classification and Ordination*, Dordrecht, Kluwer.

MONJARDET B. (2012), « Jean-Pierre Barthélemy et le principe de Pareto », *Mathématiques et Sciences humaines* 197, p. 9-17.

RÉGNIER S., (1965), « Sur quelques aspects mathématiques des problèmes de classification automatique », *I.C.C. Bulletin* 4, p. 175-191. Reprint : *Mathématiques et Sciences humaines* 82, 1983, p. 13-29.

ROMESBURG C. (2004), *Cluster Analysis for Researchers*, Lulu Press (North Carolina).

SOCIÉTÉ FRANCOPHONE DE CLASSIFICATION (SFC), http://www.sfc-classification.net/

WAKABAYASHI Y. (1986), *Aggregation of binary relations: algorithmic and polyhedral investigations*, PhD thesis, Augsbourg.

WAKABAYASHI Y. (1998), "The Complexity of Computing Medians of Relations", *Resenhas* 3(3), p. 323-349.

ZAHN C.T. (1964), "Approximating symmetric relations by equivalence relations", *SIAM Journal on Applied Mathematics* 12, p. 840-847.