DATA GAMES

SHARING PUBLIC GOODS WITH EXCLUSION

Pierre Dehez* and Daniela Tellone**

ABSTRACT

A group of agents considers collaborating on a project which requires putting together elements owned by some of them. These elements are pure public goods with exclusion i.e. nonrival but excludable goods like for instance knowledge, data or information, patents or copyrights. The present paper addresses the question of how should agents be compensated for the goods they own. It is shown that this problem can be framed as a cost sharing game – called "data game" – to which standard cost sharing rules like the Shapley value or the nucleolus can then be applied and compared.

JEL: C71, D46, M41
Keywords: cost sharing, compensation, Shapley value

The authors are grateful to Alexandre Bailly from TRASYS for having drawn their attention to the cost sharing problem faced by the European chemical industry within EU regulation "REACH". Thanks are due to Eve Ramaekers, Pier Mario Pacini and the participants of the CORE Workshop on Welfare Economics for useful comments and suggestions.

---

* CORE, Université catholique de Louvain, Voie du roman pays 34, 1348 Louvain-la-Neuve, Belgium. Tel: 0032-10-472934. Fax: 0032-10-474301. Email: pierre.dehez@uclouvain.be

** CEREC, Facultés universitaires Saint-Louis, Boulevard du jardin botanique 38, 1000 Bruxelles, Belgium. Tel: 0032-2-2117939. Fax: 0032-2-2117873. Email: tellone@fusl.ac.be

1. Introduction

Imagine the following situation. A group of agents considers collaborating on a project which requires putting together elements owned by some of them. These elements are *pure* public goods with exclusion[1] i.e. nonrival but excludable goods like for instance knowledge, data or information, patents or copyrights.[2] The question is not to share the cost of these goods because they are already available. Their costs are sunk. The question is instead to possibly compensate some of the agents who own these goods, knowing that any *additional* cost has to be shared independently. This problem can be framed as a cost sharing game in which the value of the grand coalition is zero: these games are *compensation games* to which standard cost allocation rules like the Shapley value or the nucleolus can be applied.

These games are defined on the basis of the *replacement cost* of the goods involved e.g. the cost of acquiring today the data or the cost of developing alternative technologies. In what follows we shall keep the term "data" for expository reason and talk about "data games".

We first define data games and analyze their general properties. Data games are essential, monotone *decreasing* and subadditive. Their core is nonempty as it always contains the no compensation allocation: no coalition of players can object when no one is asked to pay. We then consider two special classes of data games.

In the first class, individual datasets are nested and, as a consequence, at least one player owns the complete dataset. Although this does not fit many actual situations, any data game can be written as a sum of nested data games by working data by data. Furthermore, nested data games are "reverse" airport games, a property which is used to compute the compensation rule derived from the Shapley value.

In the second class, individual datasets form a partition of the complete dataset. This case applies to many actual situations and fits perfectly the case of patents or copyrights.

The core of both nested and partition data games happens to have a very simple structure: it is a regular simplex. As a consequence, the nucleolus coincides with the core centroid. Partition data games are concave implying that the Shapley compensation belongs to the core and, given the regular structure of the core, it coincides with the nucleolus. Instead, nested data games are generally not concave and core allocations involve compensations only if a single player owns the complete dataset in which case only that player is compensated. As a consequence, the allocation derived from the Shapley value does generally not belong to the core because it often compensates other players as well.

The paper is organized as follows. Cost games in general and data games in particular, together with their associated surplus sharing games, are defined in Section 2. Section 3 is devoted to the core which is further characterized for the two special classes of data games, nested data games and partition data games. The compensation rule derived from the Shapley value is defined in Section 4 and compared to the sharing rules derived from the nucleolus and from accounting methods. Concluding remarks are offered in the last section.

---

[1] To quote Drèze (1980, p.6): "Public goods with exclusion are public goods … the consumption of which by individuals can be controlled, measured and subjected to payment or other contractual limitations."

[2] The origin of the present paper is the cost sharing problem faced by the European chemical industry which must submit detailed analysis for about 30.000 substances it produces, a requirement imposed by EU under the acronym "REACH".

2. Data games

2.1 Cost sharing games

A set $N = \{1,\ldots,n\}$ of players, $n \geq 2$, have a common project and face the problem of dividing its cost. The cost of realizing the project to the exclusive benefit of the members of any coalition is also known.

This defines a real-valued function C – a cost function – on the subsets of N with $C(\varnothing) = 0$.[3] A pair (N,C) defines a cost game and the cost to be divided is $C(N)$.

A sharing rule φ associates a cost allocation $y = \varphi(N,C)$ to any cost game (N,C) such that

$$\sum_{i \in N} y_i = C(N)$$

Notation: The letters $n, s, t,\ldots$ will denote the size of the sets $N, S, T,\ldots$ For a vector y, $y(S)$ will denote the sum over S of its coordinates. Coalitions will be identified as $ijk\ldots$ instead of $\{i,j,k\}\ldots$ For any set S, $S\backslash i$ will denote the coalition out of which player i has been removed.

2.2 Data games

We denote by $M = \{1,\ldots, m\}$ the set of existing data and by d_h the cost of *reproducing* data $h \in M$, with $d_h > 0$ for all $h \in M$. We denote by $M_i \subset M$ the subset of data owned by player i and by M_S the dataset owned by S:

$$M_S = \bigcup_{i \in S} M_i$$

We assume that $M_N = M$ and $M_i \neq M$ for some $i \in N$. This includes the possibility that some players do not own data ($M_i = \varnothing$) or do own the complete dataset ($M_i = M$).

The cost associated to a coalition is the cost of reproducing the data it does *not* own:

$$C(S) = \sum_{h \in M \backslash M_S} d_h = d_0 - \sum_{h \in M_S} d_h \qquad (1)$$

where $d_0 = \sum_{h \in M} d_h$ is the replacement cost of the complete dataset.

This defines a cost game (N,C) that we call "data game". Because $C(N) = 0$, data games are "compensation games".

In what follows we shall consider examples involving four players and four data, with a common cost vector $d = (6,4,10,12)$. The cost of the complete dataset is then $d_0 = 32$. Only the distribution of data among players will change. Player 1 will however be assumed to own no data in all these examples.

[3] See for instance Young (1985) or Moulin (1988, 2003).

Example 1 Consider the following datasets $M_1 = \varnothing$, $M_2 = \{1\}$, $M_3 = \{1,2\}$ and $M_4 = \{3,4\}$. The corresponding data game is given by:

$C(1) = d_0 = 32$
$C(2) = C(12) = d_2 + d_3 + d_4 = 26$
$C(3) = C(13) = C(23) = C(123) = d_3 + d_4 = 22$
$C(4) = C(14) = d_1 + d_2 = 10$
$C(24) = C(124) = d_2 = 4$
$C(34) = C(134) = C(234) = C(1234) = 0$

Because $M \neq M_i$ for some i, data games are *essential*:

$$\sum_{i=1}^{n} C(i) = nd_0 - \sum_{i \in N} \sum_{h \in M_i} d_h > 0$$

Data games are monotonically *decreasing*: for any coalitions S and T such that $S \subset T$,

$$M_S \subset M_T \text{ and } C(T) - C(S) = \sum_{h \in M_S} d_h - \sum_{h \in M_T} d_h \leq 0$$

Data games are *subadditive*: for any *disjoint and nonempty* coalitions S and T,

$$C(S) + C(T) = 2d_0 - \sum_{h \in M_S} d_h - \sum_{h \in M_T} d_h = C(S \cup T) + d_0 - \sum_{h \in M_S \cap M_T} d_h \geq C(S \cup T)$$

Concavity is a stronger form of scale economies.[4] However data games are generally not concave except in special cases as we shall see.

We denote by k_i the cost of reproducing the data owned by player i:

$$k_i = \sum_{h \in M_i} d_h$$

and by $c_i = C(i)$ the cost of reproducing the data player i does *not* own. The c_i's and the k_i's are related by the equations

$$c_i + k_i = d_0 \tag{2}$$

and the c_i's satisfy the inequalities

$$0 \leq c_i \leq d_0 \text{ and } 0 < \sum_{i=1}^{n} c_i \leq (n-1)d_0$$

[4] A set function f is *concave* if $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. Hence concavity implies subadditivity. Equivalently, a set function f is concave if, for all i, the marginal costs $f(S) - f(S \backslash i)$ are non increasing with respect to set inclusion.

2.3 Surplus sharing games

It will be useful to consider the division of the surplus generated by the grand coalition. For any coalition $S \subset N$, we denote by $v(S)$ the gain generated by coalition S if it forms:

$$v(S) = \sum_{i \in S} C(i) - C(S) \tag{3}$$

The characteristic v function defines the surplus (sharing) game (N,v) associated to the cost game (N,C). In particular $v(i) = 0$ and $v(N) > 0$ (for essential cost games). If the cost game (N,C) is subadditive (resp. concave) then the surplus game (N,v) is superadditive (resp. convex), and vice versa.

Shares z in the surplus $v(N)$ and shares y in the cost $C(N)$ are linked by the following identities:

$$z_i + y_i = C(i) \quad i = 1,\ldots,n \tag{4}$$

The surplus game associated to the data game (1) is defined by

$$v(S) = \sum_{i \in S} c_i - \sum_{h \in M \setminus M_S} d_h = \sum_{i \in S} \sum_{h \in M \setminus M_i} d_h - \sum_{h \in M \setminus M_S} d_h$$

and the complete surplus to be divided is

$$v(N) = \sum_{i=1}^{n} c_i$$

Clearly $v(i) = 0$ for all i. The surplus game associated to Example 1 is given by:

$v(12) = v(13) = v(14) = d_0 = 32$
$v(23) = d_2 + 2d_3 + 2d_4 - d_3 - d_4 = 26$
$v(24) = d_1 + 2d_2 + d_3 + d_4 - d_2 = 32$
$v(34) = d_1 + d_2 + d_3 + d_4 - 0 = 32$
$v(123) = d_0 + d_2 + 2d_3 + 2d_4 - d_3 - d_4 = 58$
$v(124) = d_0 + d_1 + 2d_2 + d_3 + d_4 - d_2 = 64$
$v(134) = d_0 + d_1 + d_2 + d_3 + d_4 = 64$
$v(234) = d_1 + 2d_2 + 2d_3 + 2d_4 = 58$

with $v(1234) = d_0 + d_1 + 2d_2 + 2d_3 + 2d_4 = 90$.

2.4 Marginal cost vectors

Marginal cost vectors play an important role in the core of concave games and in the computation of the Shapley value. Let $\prod$ be the set of all players' permutations. To each permutation $\pi = (i_1,\ldots,i_n) \in \prod$, we associate the vector $t(\pi)$ whose element i_k is given by:

$$t_{i_1}(\pi) = C(i_1)$$

$$t_{i_k}(\pi) = C(i_1,\ldots,i_k) - C(i_1,\ldots,i_{k-1}) \text{ for } k = 2,\ldots,n$$

The player who is first pays his/her own cost. In example 1, the marginal cost vector associated to the permutation $\pi = (1,2,3,4)$ is given by:

$$t(\pi) = (32, -6, -4, -22)$$

We shall now analyze two special classes of data games.

2.5 Nested data games

Let us assume that players can be ordered in such a way that datasets are *nested* i.e. $i < j$ implies $M_i \subset M_j$. Then $M_n = M$ and $d_0 = k_n$. Furthermore, the individual cost parameters c_i's satisfy the following inequalities:

$$c_1 \geq c_2 \geq ... \geq c_{n-1} \geq c_n = 0 \tag{5}$$

and the corresponding data game (N,C) is simply defined by

$$C(S) = \text{Min}_{i \in S} c_i \quad \text{for all } S \subset N, S \neq \varnothing \tag{6}$$

with $C(\varnothing) = 0$.

This game looks like a "reversed" airport game.[5] The cost function C can indeed be written as $C(S) = -C_0(S)$ where $C_0(S) = \text{Max}_{i \in S}(-c_i)$. However C_0 does not define a proper airport game because $C_0(S) \leq 0$ for all S. In particular, the cost function C_0 is superadditive.

Using (2) nested data games can alternatively be written in terms of the k_i's:

$$C(S) = k_n - \text{Max}_{i \in S} k_i \tag{7}$$

where $k_n = d_0$ is a fixed cost which applies to all players and coalitions, and the cost function $K(S) = \text{Max}_{i \in S} k_i$ defines a proper airport game. Indeed the k_i's satisfy the inequalities

$$0 \leq k_1 \leq k_2 \leq ... \leq k_{n-1} \leq k_n \quad \text{and} \quad k_n > 0$$

Example 2 Consider the datasets $M_1 = \varnothing$, $M_2 = \{1\}$, $M_3 = \{1,2,3\}$ and $M_4 = \{1,2,3,4\}$. The corresponding data game is given by:

$C(1) = d_0 = 32$
$C(2) = C(12) = d_2 + d_3 + d_4 = 26$
$C(3) = C(13) = C(23) = C(123) = d_4 = 12$

and $C(S) = 0$ for all the other coalitions (i.e. all coalitions including player 4).

Actually, by working data by data, any data game (N,C) can be written as a sum of nested data games. For each $h \in M$, define the cost game (N,C_h) by

$$C_h(S) = 0 \quad \text{if } h \in M_S$$

$$C_h(S) = d_h \quad \text{if } h \notin M_S$$

for all $S \subset N$, $S \neq \varnothing$, and $C_h(\varnothing) = 0$.

[5] See Littlechild and Owen (1973) and Littlechild and Thomson (1977) for a definition and analysis of airport games.

Clearly (N,C_h) is an elementary nested data game and

$$\sum_{h\in M} C_h(S) = \sum_{h\in M\setminus M_S} d_h = C(S) \tag{8}$$

2.6 The partition case

Let us assume that the datasets form a partition of M i.e. $M_i \cap M_j = \varnothing$ for all $i \neq j$. As a consequence,

$$\sum_{h\in M_S} d_h = \sum_{i\in S}\sum_{h\in M_i} d_h = \sum_{i\in S} k_i$$

and, using (1) the corresponding data game (N,C) is then simply defined by:

$$C(S) = d_0 - \sum_{i\in S} k_i \tag{9}$$

to be compared to (7) where d_0 coincides with k_n. Here instead $d_0 = \sum_{i\in N} k_i$ and $C(S) = \sum_{i\in N/S} k_i$.

Partition data games are *concave*. Indeed, $C(S) - C(S\backslash i) = -k_i$ for all $i \in S$ and all $S \neq \{i\}$ while $C(i) - C(\varnothing) = c_i = d_0 - k_i$. Hence, for any given player, marginal costs associated to proper coalitions are constant (and negative). Altogether they are non-increasing.

Example 3 Consider the following datasets $M_1 = \varnothing$, $M_2 = \{1\}$, $M_3 = \{2\}$ and $M_4 = \{3,4\}$. The corresponding data game is given by:

$C(1) = d_0 = 32$
$C(2) = C(12) = d_2 + d_3 + d_4 = 26$
$C(3) = C(13) = d_1 + d_3 + d_4 = 28$
$C(4) = C(14) = d_1 + d_2 = 10$
$C(23) = C(123) = d_3 + d_4 = 22$
$C(24) = C(124) = d_2 = 4$
$C(34) = C(134) = d_1 = 6$
$C(234) = C(1234) = 0$

Interestingly, the surplus game turns out to be *symmetric* in the partition case. Indeed, the value of a coalition depends only on its size:

$$v(S) = \sum_{i\in S} c_i - d_0 + \sum_{i\in S} k_i = (s-1)d_0$$

As a consequence, any sharing rule satisfying symmetry ("equal treatment of equals") allocates the total surplus equally.

Using (4), the corresponding compensation is given by

$$y_i = c_i - \mu_c \quad i = 1,\ldots,n \tag{10}$$

where

$$\mu_c = \frac{v(N)}{n} = \frac{n-1}{n} d_0 = \frac{1}{n}\sum_{i=1}^{n} c_i$$

A player pays *if and only if* his/her cost exceeds the mean upgrading cost μ_c. This is the *equal surplus sharing rule*. Alternatively, the compensation defined by (10) can be written as:

$$y_i = \frac{d_0}{n} - k_i \quad i = 1,...,n$$

i.e. a player pays *if and only if* the per capita cost of the complete dataset exceeds the cost of the data he/she owns.

In Example 3, c = (32,26,28,10) and μ_c = 24. Hence, the resulting compensation is given by y = (8,2,4,–14): player 4 is compensated by the other three players.

3. The core

3.1 Nonemptiness of the core of data games

Individual rationality is the minimal requirement to impose to a cost allocation y

$$y_i \leq C(i) \text{ for all } i \in N$$

i.e. no player should pay more than his or her "stand alone" cost. This defines an *imputation*. Extending the argument to coalitions is a stronger requirement: the *core* is the set of allocations y against which no coalition can object

$$y(S) \leq C(S) \text{ for all } S \subset N \tag{11}$$

i.e. no coalition pays more that its stand alone cost. Equivalently, an allocation y is in the core if and only if

$$y(S) \geq C(N) - C(N \backslash S) \text{ for all } S \subset N$$

i.e. there is *no cross-subsidization*: each coalition pays at least its marginal cost.[6]

In general, the core is a *convex polyhedron*, possibly empty, whose dimension does not exceed n–1. The core of a data game as defined by (1) is *always* nonempty. Indeed, $C(S) \geq 0$ for all $S \subset N$ and, as a consequence, the trivial allocation defined by the absence of compensation y_0 = (0,0,…,0) belongs to the core of any data game. The following proposition concerns the case where some players own the complete dataset.

Proposition 1 If one but only one player owns the complete dataset, only that player is possibly compensated in core allocations and there are core allocations different from y_0. If two players (or more) own the complete dataset, the core reduces to the singleton $\{y_0\}$.

[6] See Faulhaber (1975).

Proof Assume that $M_n = M$ and let y be a core allocation. Applying (11), $y(N/i) \le C(N/i) = 0$ for all $i \neq n$. Combining this with $y(N) = 0$, we get $y_i \ge 0$ for all $i \neq n$ and $y_n \le 0$.

If $M_i \neq M$ for all $i \neq n$, the allocation $y_b = (b,b,\ldots,(1\text{-}n)b)$ where $b = C(N/n)/n > 0$ belongs to the core. Indeed consider a coalition $S \subset N$:

$$\text{if } n \notin S,\ y_b(S) = sb < nb = C(N \setminus n) \le C(S)$$

$$\text{if } n \in S,\ y_b(S) = (s-n)b \le 0 = C(S)$$

Assume that $M_{n-1} = M$ as well. Then $y_n \ge 0$ because $y(N/i) \le C(N/i) = 0$ for all $i \neq n\text{-}1$. Hence $y_n = 0$ and $y(N/n) = 0$. This is possible only if $y = y_0$. à

Remark 1 Actually the core reduces to y_0 whenever *each data* is owned by at least two players. Indeed, in this case $C(N\backslash i) = 0$ for all $i \in N$.

We shall now investigate the structure of the core for the two special classes of data game, nested and partition data games. It turns out to be a regular simplex in both cases i.e. an equilateral triangle for $n = 3$, a regular tetrahedron for $n = 4$,…

3.2 The core of a nested data game

In the nested case, only the last player – who owns the complete dataset – is possibly compensated in the core and it reduces to $\{y_0\}$ *if and only if* more than one player own the complete dataset, independently of the other cost parameters. These are consequences of Proposition 1. Furthermore, the core has a very simple structure which depends on a single parameter, c_{n-1}, the largest compensation player n can expect to receive.

Proposition 2 The core of the nested data game (6) is a regular simplex whose n vertices are:

$$(a,0,\ldots,0,-a),\ (0,a,0,\ldots,0,-a),\ \ldots,(0,0,\ldots,a,-a) \text{ and } (0,\ldots,0)$$

where $a = c_{n-1}$.

Proof We first show that y belongs to the core *if and only if* there exist $\lambda_1,\ldots,\lambda_{n-1}$ such that:

$$\sum_{i=1}^{n-1} \lambda_i \le 1,\ 0 \le \lambda_i \le 1 \text{ and } y_i = \lambda_i a \quad i = 1,\ldots,n-1 \tag{12}$$

If y is an allocation defined by (12), the following inequalities hold for all $S \subset N$

$$\sum_{i \in S} y_i = a \sum_{i \in S} \lambda_i \le a \le C(S) \quad \text{if } n \notin S$$

$$\sum_{i \in S} y_i = a \sum_{\substack{i \in S \\ i \neq n}} \lambda_i - a \sum_{i=1}^{n-1} \lambda_i = -a \sum_{\substack{i=1 \\ i \notin S}}^{n-1} \lambda_i \le 0 = C(S) \quad \text{if } n \in S$$

Hence y belongs to the core.

If y is an element of the core, we have successively:

$$0 \le y_{n-1} \le a$$

$$0 \le y_{n-2} \le y_{n-2} + y_{n-1} \le a$$

...

$$0 \le y_1 \le \sum_{i=1}^{n-1} y_i \le a$$

i.e. $0 \le y_i \le a$ for all $i = 1,\ldots,n\text{-}1$. The λ_i's defined by $\lambda_i = \frac{y_i}{a}$ then satisfy (12).

Consequently, the core is the convex hull of the vectors $(a,0,\ldots,0,-a)$, $(0,a,0,\ldots,0,-a)$, $\ldots$,$(0, 0,\ldots, a, -a)$ and $(0,\ldots,0)$. It is a simplex whose regularity follows from the fact that all vertices are connected to each other by line segments of identical length $2^{1/2}a$. à

Remark 2 The proof of Proposition 2 reveals that an allocation belongs to the core *if and only if* $0 \le y_i \le c_{n\text{-}1}$ for all $i \ne n$ and $-c_{n\text{-}1} \le y_n \le 0$.

Normalizing and translating – by dividing by $c_{n\text{-}1}$ and adding the vector $(0,0,\ldots,1)$ – the core is transformed into the standard unit simplex $\Delta_n = \{x \in \tilde{\mathrm{N}}^n \mid x \ge 0, \Sigma_i x_i = 1\}$. Applying this transformation to the nested data games defined by the cost vector $(c_1,\ldots,c_{n\text{-}1})$, we obtain the equivalent cost game defined by:

$$C(S) = 1 \text{ if } n \in S$$

$$C(S) = \mathrm{Min}_{i \in S} \frac{c_i}{c_{n-1}} \text{ if } n \notin S$$

whose core is indeed Δ_n.

The core being a regular simplex, its centroid or centre of gravity[7] is simply defined by the average of its vertices:

$$\hat{y} = \left(\frac{c_{n-1}}{n}, \frac{c_{n-1}}{n}, \ldots, \frac{c_{n-1}}{n}, -(n-1)\frac{c_{n-1}}{n} \right)$$

which coincides with the nucleolus (and the least core).[8] Only the last player is compensated and the n-1 other players all contribute the same amount. Given the coordinates of its centre, the diameter of the core is equal to $2(1\text{-}1/n)^{1/2}\ c_{n\text{-}1} < 2^{1/2}\ c_{n\text{-}1}$ and it has full dimension *if and only if* $c_{n\text{-}1} > 0$.

In example 2, $c_{n\text{-}1} = c_3 = 14$ and the resulting compensation is (3½,3½,3½,–10½).

[7] See Gonzales-Diaz and Sanchez-Rodriguez (2007) for a general definition of the core centroid.

[8] The nucleolus is a single-value solution introduced by Schmeidler (1969). Intuitively the idea is to minimize the loss incurred by coalitions suffering the highest loss – the loss of a coalition being measured by the difference between the amount it pays and its cost. The nucleolus is always defined and belongs to the core if nonempty. For a definition of the least core see Maschler M. et al. (1979) where it is shown that nucleolus is the lexicographic centre of the core if nonempty.

3.3 The core in the partition case

Partition data games are concave and, as a result, the core is the convex hull of the marginal cost vectors associated to the $n!$ players' permutations. Actually the core is again a regular simplex.

Proposition 3 The core of the partition data game (10) is a regular simplex of full dimension whose n vertices are:

$$(c_1,c_2-d_0,\ldots,c_n-d_0),\ (c_1-d_0,c_2,\ldots,c_n-d_0),\ \ldots\ ,(c_1-d_0,c_2-d_0,\ldots,c_n).$$

Proof If player i is first in a given permutation he/she pays its cost c_i. Otherwise he/she saves the cost of the data he/she owns $d_0 - c_i$. Hence there are n *distinct* marginal cost vectors each with a multiplicity equal to $(n-1)!$ and the vector t associated to the permutations where player i is first is defined by $t_i = c_i$ and $t_j = c_j - d_0$ for all $j \neq i$.

It is a simplex whose regularity follows from the fact that all vertices are connected to each other by line segments of identical length $2^{1/2}d_0$. Positivity of d_0 ensures the full dimensionality of the core. à

Translating and normalizing – by substracting the vector $(c_1-d_0,c_2-d_0,\ldots,c_n-d_0)$ and dividing by d_0 – the core is transformed into the standard unit simplex Δ_n. Applying this transformation to the partition data games defined by the cost vector $(c_1,\ldots,c_{n-1})$, we obtain the equivalent "constant" cost game defined by:

$$C(S)=1 \text{ for all } S\subset N$$

whose core is indeed Δ_n.

Remark 3 In the extreme case where $M_n = M$ and $M_i = \varnothing$ for all $i \neq n$, the resulting game is nested and its core is given by Proposition 2 with $a = d_0$.

The core of partition data games has the same structure than the core of nested data games. Being a regular simplex, its centroid is again the average of its vertices

$$\hat{y}_i = c_i - \frac{n-1}{n}d_0 = \frac{d_0}{n} - k_i \quad i=1,...,n$$

where used has been made of equation (2). Indeed c_i appears n times and $-d_0$ appears $(n-1)$ times. This is the equal surplus allocation – as defined by (10) – which also coincides with the least core and the nucleolus.

Given the coordinates of its centre, the diameter of the core is equal to $2(1-1/n)^{1/2}\ d_0 < 2^{1/2}\ d_0$. Hence the size of the core depends only on d_0 – the cost of the complete dataset.

4. The Shapley value

4.1 The cost allocation derived from the Shapley value

For a general cost game (*N*,*C*), the cost allocation derived from the Shapley value is simply the average marginal cost vector

$$\varphi_i(N,C) = \frac{1}{n!}\sum_{\pi\in\Pi} t_i(\pi)$$

This formula is obtained by first computing the Shapley value of the associated surplus game (3) and then using identity (4).

The Shapley value is the unique *additive* sharing rule which satisfies *symmetry* and *dummy*.[9] There exist alternative axiomatizations of the Shapley value.[10] In the context of cost sharing it is shown that the Shapley sharing rule is the unique rule which allocates fixed costs fairly.[11] The Shapley value is individually rational for subadditive cost games and belongs to the core for concave cost games.[12]

We have seen that any data game can be written as a sum of nested data games. This fact will be used to obtain a simple formula based on the additivity of the Shapley value.

4.2 Shapley compensation in the nested case

One way to derive the Shapley compensation is to use the definition of a nested data game given by (7) and the cost allocation formula for the airport game (*N*,*K*). Additivity and symmetry of the Shapley value then imply:

$$y_i = \frac{k_n}{n} - \varphi_i(N,K)$$

i.e.

$$\begin{aligned} y_1 &= \frac{k_n}{n} - \frac{k_1}{n} = \frac{c_1}{n} \\ y_2 &= \frac{k_n}{n} - \left(\frac{k_1}{n} + \frac{k_2 - k_1}{n-1}\right) = \frac{c_1}{n} + \frac{c_2 - c_1}{n-1} \\ &\ldots \\ y_n &= \frac{k_n}{n} - \left(\frac{k_1}{n} + \frac{k_2 - k_1}{n-1} + \ldots + \frac{k_{n-1} - k_{n-2}}{2} + k_n - k_{n-1}\right) = \frac{c_1}{n} + \ldots + \frac{c_{n-1} - c_{n-2}}{2} + c_n - c_{n-1} \end{aligned} \tag{13}$$

with $c_n = 0$. The resulting allocation looks like the airport cost allocation derived from the Shapley value. This is consistent with the fact that $C(S) = -\mathrm{Max}_{i\in S}(-c_i)$.

[9] These are the original axioms used by Shapley (1953, 1981): players with identical marginal costs pay the same amount (symmetry or "equal treatment of equals") and players with zero marginal costs pay nothing (dummy). Instead the nucleolus satisfies symmetry and dummy but not additivity.

[10] See Moulin (2003).

[11] See Dehez (2006).

[12] The core is then typically large and the Shapley value is located somewhere in its centre. See Shapley (1971).

In example 2, $c = (32,26,12,0)$ and the resulting Shapley compensation is given by:

$$y_1 = \frac{32}{4} = 8,\ \ y_2 = 8 - \frac{6}{3} = 6,\ \ y_3 = 6 - \frac{14}{2} = -1,\ \ y_4 = -1 - \frac{12}{1} = -13$$

This allocation does not belong to the core because two players are compensated. The following proposition clarifies the relationship between the core and the Shapley value in nested data games.

Proposition 4 In a nested data game, the allocation y derived from the Shapley value belongs to the core *if and only if* $y_{n-1} \geq 0$.

Proof We already know that $y_{n-1} \geq 0$ if y belongs to the core. Let y be the allocation derived from the Shapley value. It is such that $y_1 \geq y_2 \geq \ldots \geq y_n$. Hence $y_{n-1} \geq 0$ implies $y_i \geq 0$ for all $i \neq n$ and $y_n \leq 0$. Furthermore $y_n = y_{n-1} - c_{n-1} \geq - c_{n-1}$ implies $y(N\backslash n) \leq c_{n-1}$ and $y_i \leq c_{n-1}$ for all $i \neq n$. From Remark 2, we can then conclude that y belongs to the core. à

This proposition gives a single condition on the cost parameters c_i's such that the Shapley value belongs to the core in the nested case. For $n = 3$, that condition reduces to $c_1 \leq 3c_2$. For $n = 4$, it becomes $c_1 + 2c_2 \leq 6c_3$.

Formula (13) has a simple recursive structure and the Shapley compensation y can be written simply as:

$$y = A.c$$

where A is a $n \times n$ triangular matrix whose elements are defined by:

$$a_{ij} = \frac{-1}{(n-j+1)(n-j)} \text{ if } j < i \text{ and } a_{ii} = \frac{1}{n-i+1}$$

with $a_{ij} = 0$ otherwise. For $n = 4$, the matrix A is given by:

$$A = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ -1/12 & 1/3 & 0 & 0 \\ -1/12 & -1/6 & 1/2 & 0 \\ -1/12 & -1/6 & -1/2 & 1 \end{bmatrix}$$

These matrices are overlapping, starting from the lower right element 1.[13] For instance, if $n = 5$, the first column starts with 1/5, followed by -1/20.[14] The condition under which the allocation derived from the Shapley value belongs to the core can then be written as:

$$-\sum_{j=1}^{n-2} a_{n-1\,j}\, c_j \leq \frac{c_{n-1}}{2}$$

[13] Actually the element a_{nn} is arbitrary because $c_n = 0$. It is equal to 1 in the matrix defining the Shapley value of an airport game.

[14] The elements of any of the first n-1 columns sum up to 0 and the elements of any row sum up to 1/n.

In the particular case where a subset T of t players own the complete dataset while the other players own no data i.e. $M_i = \varnothing$ for all $i \in N\backslash T$ and $M_i = M$ for all $i \in T$, we have:

$$\varphi_i(N,C) = \frac{d_0}{n} \text{ for all } i \notin T$$
$$\varphi_i(N,C) = \frac{d_0}{n} - \frac{d_0}{t} \text{ for all } i \in T$$

This can be used to obtain a formula for the Shapley compensation for general data games. Indeed, using (8) and the additivity of the Shapley value, we have:

$$C(S) = \sum_{h\in M} C_h(S) \Rightarrow \varphi_i(N,C) = \sum_{h\in M} \varphi_i(N,C_h) \quad i = 1,...,n$$

where

$$\varphi_i(N,C_h) = \frac{d_h}{n} \quad \text{for all } i \notin T_h$$
$$\varphi_i(N,C_h) = \frac{d_h}{n} - \frac{d_h}{t_h} \quad \text{for all } i \in T_h$$

Here t_h is the number of players owing data h. Hence,

$$\varphi_i(N,C) = \frac{1}{n}\sum_{h\in M_i} d_h\left(1-\frac{n}{t_h}\right) + \frac{1}{n}\sum_{h\in M\backslash M_i} d_h = \frac{1}{n}\sum_{h\in M} d_h - \sum_{h\in M_i}\frac{1}{t_h}d_h$$

The compensation derived from the Shapley value for a general data game (N,C) is then given by the following simple formula:

$$\varphi_i(N,C) = \frac{d_0}{n} - \sum_{h\in M_i}\frac{d_h}{t_h} \quad i = 1,...,n \tag{14}$$

Remark 4 This formulation shows that what a player pays (resp. receives) decreases (resp. increases) with the cost of the data he/she owns. It also increases (resp. decreases) with the number of players owning the same data.

In Example 1, $d = (6,4,10,12)$ and $t = (2,1,1,1)$. The resulting Shapley compensation is given by:

$$y_1 = 8,\ y_2 = 8 - \frac{6}{2} = 5,\ y_3 = 8 - \frac{6}{2} - \frac{4}{1} = 1,\ y_4 = 8 - \frac{10}{1} - \frac{12}{1} = -14$$

4.2 Shapley compensation in the partition case

In the partition case, we already know that the compensation derived from the Shapley value is given by (10):

$$\hat{y}_i = c_i - \frac{n-1}{n}d_0 = \frac{d_0}{n} - k_i \quad i = 1,...,n$$

It is a consequence of the symmetry of the associated surplus game and it is consistent with the particular structure of marginal costs in the partition case, as described in Proposition 3 where, for each player i, c_i appears $n!$ times and $-d_0$ appears $(n\text{-}1)(n\text{-}1)!$ It is also consistent with (14) because $t_h = 1$ for all h in the case of a partition.

Remark 5 In the extreme case where only one player owns some data, $M_n = M$ and $M_i = \varnothing$ for all $i \neq n$, the Shapley compensation is defined by $y_i = d_0/n$ for $i = 1,\ldots,n\text{-}1$.

We have seen that partition data games are concave. As a consequence, the compensation derived from the Shapley value belongs to the core. Actually the Shapley compensation coincides with the compensations derived from any symmetric sharing rules, in particular the nucleolus, the core centroid and the equal surplus.

4.3. Accounting methods

There exist various accounting methods for dividing joint costs based on players' marginal costs computed with respect to the grand coalition. They are of the form:

$$\varphi_i(N,C) = MC_i + \alpha_i(C(N) - \sum_{j=1}^{n} MC_j) \qquad (15)$$

where $MC_i = C(N) - C(N\backslash i)$ is sometime called the "separable costs" of player i and the weights α_i are such that $0 \leq \alpha_i \leq 1$ for all i and $\Sigma_i\, \alpha_i = 1$.

The "*equal charge*" method uses equal weights $\alpha_i = 1/n$ for all i. Another is the "*separable costs remaining benefits*" method (SCRB)[15] in which the weights are given by:

$$\alpha_i = \frac{b_i}{b(N)}$$

where $b_i = C(i) - MC_i$ is the "remaining benefits" of player i and $b(N) = \Sigma_i\, b_i$.

For a nested data game as defined by (6), we have:

$MC_i = 0$ for all $i \neq n$

$MC_n = -\, c_{n\text{-}1}$

Applying (15), the corresponding cost allocation is given by:

$y_i = \alpha_i\, c_{n\text{-}1}$ for all $i \neq n$

$y_n = (\alpha_n - 1)\, c_{n\text{-}1}$

From Remark 2, we can conclude that it defines a core allocation for any choice of weights while the nucleolus corresponds to equal weights.

[15] See Young (1985).

For a partition data game as defined by (9), we have:

$$MC_i = -k_i \quad \text{for all } i \in N$$

Applying (15), the corresponding cost allocation is given by:

$$y_i = \alpha_i d_0 - k_i \quad \text{for all } i \in N$$

This allocation belongs to the core for any choice of weights. Indeed we have:

$$C(S) - y(S) = (d_0 - \sum_{i \in S} k_i) - (d_0 \sum_{i \in S} \alpha_i - \sum_{i \in S} k_i) = d_0(1 - \sum_{i \in S} \alpha_i) \geq 0 \quad \text{for all } S \subset N$$

For $\alpha_i = 1/n$, y is the equal surplus allocation (10) which coincides with the Shapley value, the core centroid and the nucleolus. This is also the allocation resulting from the SCRB method. Indeed we have:

$$b_i = c_i + k_i = d_0 \quad \text{for all } i$$

and consequently $\alpha_i = 1/n$.

6. Concluding remarks

The Shapley value is the natural sharing rule to be used in cost sharing as well as in the compensation framework considered here. The fact that the resulting allocation may not belong to the core because it involves cross subsidization should not be a reason to dismiss the Shapley value as a fair compensation mechanism because what the core suggests is often not fair. This appears forcefully in the situation where only two players own data, say players n and n-1, and the datasets they own differ only by a single data, say data 1:

$$M_n = \{1,\ldots,m\} \text{ and } M_{n-1} = \{2,\ldots,m\}$$

In this case, the core imposes that only player n may be compensated with an amount not exceeding d_1, while all the other players may be asked to pay up to d_1, including player n-1. The nucleolus goes even further by imposing that the n-1 first players pay the same amount, d_1/n. This is to be compared with the allocation derived from the Shapley value. Using for instance (14) we get:

$$y_i = \frac{d_0}{n} \quad \text{for } i = 1,\ldots n-2$$
$$y_{n-1} = -\frac{n-2}{2n} d_0 + \frac{d_1}{2} = \frac{d_0}{n} + \frac{d_1}{2} - \frac{d_0}{2}$$
$$y_n = -\frac{n-2}{2n} d_0 - \frac{d_1}{2} = \frac{d_0}{n} - \frac{d_1}{2} - \frac{d_0}{2}$$

This is definitely more acceptable from a fairness point of view: the players without data pay the per capita cost of the complete dataset while players n and n-1 are both compensated, the difference between what they receive being precisely equal to the cost of the missing data.

In actual cost sharing problems, like the one faced by the European chemical industry, there must be an agreement on the compensation formula *and* on the value of the costs parameters. Reaching a consensus on the cost parameters is clearly the most difficult part in particular because, under the Shapley value, we know from Remark 4 that what a player pays decreases with the cost of the data he/she owns. One should however keep in mind that these cost parameters measure the present cost of *reproducing* the data and not the actual cost that has been sunk in the past.

References

Dehez, P. (2006), Fair division of fixed costs defines the Shapley value, Revised CORE Discussion Paper 2006/115.

Drèze, J.H. (1980), Public goods with exclusion, *Journal of Public Economics* 13,5-24.

Faulhaber, G. (1975), Cross-subsidization: Pricing in public enterprises, *American Economic Review* 65, 966-977.

Gonzáles-Díaz J. and Sánchez-Rodríguez E. (2007), A natural selection from the core of a TU game: The core-centre, *International Journal of Game Theory* 36, 27-46.

Littlechild S. C. and Owen G. (1973), A simple expression for the Shapley value in a special case, *Management Science* 20, 370-372.

Littlechild S. C. and Thomson G. F. (1977), Aircraft landing fees: A game theory approach, *Bell Journal of Economics* 8, 186-204.

Maschler M., Peleg B. and Shapley L.S. (1979), Geometric properties of the kernel, nucleolus and related solution concepts, *Mathematics of Operations Research* 4, 303-338.

Moulin, H. (1988), *Axioms of cooperative decision making*, Cambridge University Press, Cambridge.

Moulin, H. 2003. *Fair division and collective welfare*. The MIT Press, Cambridge.

Shapley, L.S. (1953), A value for n-person games, In Kuhn H. and Tucker A.W. (eds.), *Contributions to the Theory of Games II*, Princeton University Press, Princeton, 307-317. Reprinted in: Roth A.E. (ed., 1988), *The Shapley value. Essays in honor of Lloyd Shapley*, Cambridge University Press, Cambridge.

Shapley, L. S. (1971), Cores of convex games. *International Journal of Game Theory* 1, 11-26.

Shapley, L.S. (1981), Valuation of games In Lucas W.F. (ed.). *Game Theory and its Applications*, Proceedings of Symposia in Applied Mathematics 24, American Mathematical Society, Providence, Rhode Island.

Schmeidler, D. (1969), The nucleolus of a characteristic function game, *SIAM Journal of Applied Mathematics* 17, 1163-1170.

Young, P.Y. (1985), Cost allocation. In Young P.Y. (ed.), *Fair allocation*, Proceedings of Symposia in Applied Mathematics 33, American Mathematical Society, Providence, Rhode Island.