2008/10

Data games: sharing public goods with exclusion

Pierre Dehez and Daniela Tellone

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

Data games: Sharing public goods with exclusion

Pierre Dehez* and Daniela Tellone**

January 2010

Abstract

A group of firms decides to cooperate on a project that requires a combination of inputs held by some of them. These inputs are non-rival but excludable goods i.e. public goods with exclusion such as knowledge, data or information, patents or copyrights. We address the question of how firms should be compensated for the inputs they contribute. We show that this problem can be framed within a cost sharing game. Standard allocation rules like the Shapley value, the nucleolus or accounting formulas can then be applied and their outcomes compared. Our analysis is inspired by the problem faced by the European chemical firms within the regulation program REACH, which requires submission by 2018 of a detailed analysis of the substances they produce, import or use.

JEL: C71, H41, M41

Keywords: cost sharing, Shapley value, core, nucleolus

This is a revised version of CORE Discussion Paper 2008-10. The authors are grateful to Alexandre Bailly for having drawn their attention to the data and cost sharing problem faced by the European chemical industry. Thanks are due to Pier Mario Pacini, Eve Ramaekers and Gisèle Umbhauer and Myrna Wooders for useful comments on earlier versions. Financial support from the Belgian *Interuniversity Pole of Attraction* (PAI) program is gratefully acknowledged.

---

* CORE (University of Louvain) and BETA (CNRS – Universities of Strasbourg and Nancy)
Email: pierre.dehez@uclouvain.be

** CEREC (Facultés universitaires Saint-Louis, Bruxelles)
Email: tellone@fusl.ac.be

1. Introduction

The present paper was initially motivated by the data sharing problem faced by the EU chemical industry, following the regulation imposed by the European Commission under the acronym "REACH" (**R**egistration, **E**valuation, **A**uthorization and Restriction of **Ch**emical substances). Manufacturers and importers are required to collect safety information on the properties of their chemical substances and to register the information in a central database run by the European Chemicals Agency (ECHA). This is a huge program. There are indeed about 30,000 substances and an average of 100 parameters for each substance! Firms are encouraged to cooperate by sharing the data they have collected over the past.[1] To implement this data sharing problem, a compensation mechanism is needed.[2]

This problem can be put in general terms as follows. A group of firms decides to cooperate on a project that requires the combination of various inputs held by some of them. These inputs are non-rival but excludable goods i.e. public goods with exclusion such as knowledge, data or information, patents or copyrights.[3] The question is how to compensate the firms for the inputs they contribute. The problem can be framed within a cost sharing game for which the value of the grand coalition is *zero*. These games are therefore *compensation games* to which standard cost allocation rules can be applied, in particular the nucleolus, the Shapley value and simple accounting rules.

In what follows we shall use the term "data" and "players" for expository reasons and talk about "data (sharing) games". Data games are defined on the basis of the *replacement cost* of the inputs involved e.g. the present cost of duplicating the data or developing alternative technologies. The cost associated to a coalition is then simply the cost of the missing data.

Data games are essential, subadditive and monotone decreasing. They can be decomposed into a sum of elementary data games, one for each data. They are concave in the case where data sets form a partition of the complete data set, a situation that fits joint ventures involving complementary inputs like patents or copyrights.[4] Data games have non-empty cores as the no-compensation allocation is always in the core. Indeed the cost associated to the grand coalition is zero and the costs associated to coalitions are non-negative. As a consequence, no

[1] Beyond the cost reducing motivation, the idea is to avoid unnecessary replications of analysis involving animals.

[2] The page echa.europa.eu/reach_en.asp offers guidance for the implementation of REACH. The choice of the compensation mechanism is however left open.

[3] 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."

[4] See Katz (1995) for a discussion of joint ventures involving complementary inputs.

coalition can object when no one is asked to pay. We shall see that the core limits the extent of compensation and, in some situations, it even excludes any compensation.

To illustrate the compensation problem, let us consider the case of a single data worth 1. If there are two players and the data is not available, each player should pay 1/2. If the data is held by a single player, a fair compensation would require the player without the data to pay 1/2 to the other player. Equivalently, each player pays 1/2 but the player holding the data gets 1 back. By the same argument, if there are n players, only one holding the data, the players without the data would each pay $1/n$ to the player holding the data. This allocation is actually the Shapley value as well as the nucleolus of the associated cost game. However the two solutions differ once two players or more hold the data. Assume that $t \geq 2$ players hold the data. Extending the previous rule suggests that each of them should get back $1/t$ i.e. the $n–t$ players without data should each pay $1/n$ and the t data holders should each receive $1/t - 1/n$. The worth of the data is uniformly distributed among all players and is uniformly redistributed among the data holders. This is the Shapley value of the associated cost game but it differs from the nucleolus that excludes any compensation. This is actually a property of the core when more than one player hold the data, a property that results from the competition among data holders. Surprisingly, the allocation resulting from the equal charge accounting rule happens to be precisely the nucleolus.

In some situations there may be reasons to treat players asymmetrically, independently of the initial distribution of data. For instance, firms engaged in a joint project may have different sizes as measured, for instance, by their market shares. Such situations can be accommodated by using the asymmetric Shapley value for which exogenous weights are assigned to players. The case where some players are assigned a *zero* weight is of particular interest in a context of data sharing. Some players may indeed hold data while not being otherwise part of the joint project. This is the case in REACH where independent laboratories, like university laboratories, hold relevant data on chemical substances but are not part of the submission process.

The paper is organized as follows. Cost games are introduced in Section 2. Section 3 is devoted to the definition and properties of data games. The core of a data game is defined in Section 4 where it is shown to be a regular simplex. The nucleolus and the Shapley value are defined and analyzed in the subsequent two sections. It is shown that they coincide in the partition case. The asymmetric (or weighted) Shapley value is defined in Section 7, including the case where some players are assigned a zero weight. Weighted charge allocation rules are defined and applied to data games in Section 8. It is shown that they produce core allocations for any choice of weights. Concluding remarks are offered in the last section.

2. Cost 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 benefit of all coalitions is also known. This defines a real-valued function C on the subsets of N. Assuming $C(\varnothing) = 0$, a pair (N,C) defines a *cost game*.[5] A *sharing rule* φ associates a cost allocation $y = \varphi(N,C)$ to any cost game (N,C) such that $\sum_{i=1}^{n} y_i = C(N)$. The *dual* (N,C^*) of a cost game (N,C) is defined by $C^*(S) = C(N) - C(N \setminus S)$. The natural *surplus game* (N,v) associated with a cost game (N,C) is defined by:

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

Cost allocation y and surplus allocations x are related by the identities

$$y_i + x_i = C(i),\ i = 1,\ldots,n.$$

Notation: The letters $n, s, t,\ldots$ denote the size of the sets $N, S, T,\ldots$ For a vector y, $y(S)$ denotes the sum over S of its coordinates. Sums over empty sets are equal to zero. Coalitions are identified as $ijk\ldots$ instead of $\{i,j,k\}\ldots$ For any set S, $S \setminus i$ denotes the coalition from which player i has been removed.

We denote by $G(N)$ the set of all functions defined on the subsets of some finite set N. $G(N)$ is a vector space. The collection of $2^n - 1$ games

$$u_T(S) = 1 \quad \text{if } T \subset S$$
$$= 0 \quad \text{if not}$$

defined for all $T \subset N, T \neq \varnothing,$ forms a basis of $G(N)$.[6] Here we shall use the basis formed by the collection of $2^n - 1$ games[7]

$$e_T(S) = 1 \quad \text{if } S \cap T \neq \varnothing$$
$$= 0 \quad \text{if not}$$

defined for all $T \subset N, T \neq \varnothing.$ These games have been introduced by Kalai and Samet (1987) as duals of the unanimity games: $e_T = u_T^*$.

Given a coalition S and a player i in S, the *marginal cost* of player i to coalition S is defined by $C(i) - C(S \setminus i)$. Marginal costs play a central role in cost allocation. Let Π_n be the set of all

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

[6] These *unanimity* games were introduced by Shapley in 1953 to prove existence and uniqueness of the value.

[7] The elementary games e_T are normalized *fixed costs games*: coalitions containing players from T entail a fixed cost equal to 1. They are used in Dehez (2009) to characterize the weighted Shapley value in terms of the allocation of fixed costs, along the lines suggested by Shapley (1981b).

players' permutations. To each permutation $\pi = (i_1,...,i_n) \in \Pi_n$ we associate the vector of marginal costs $\mu(\pi)$ whose elements are given by:

$$\mu_{i_1}(\pi) = C(i_1) - C(\varnothing) = C(i_1)$$

$$\mu_{i_k}(\pi) = C(i_1,...,i_k) - C(i_1,...,i_{k-1}) \quad (k = 2,...,n)$$

It is easily seen that it defines a cost allocation:

$$\sum_{i=1}^{n} \mu_i(\pi) = C(N)$$

A cost game (N,C) is *essential* if $C(N) < \sum_{i \in N} C(i)$. It is *subadditive* if $S \bigcap T = \varnothing$ implies $C(S \bigcup T) \leq C(S) + C(T)$. It is *concave* if $C(S \bigcup T) \leq C(S) + C(T) - C(S \bigcap T)$ for all *S* and *T*. Hence concavity implies subadditivity: concavity is a stronger form of economies of scale than subadditivity. Equivalently, a cost game (N,C) is concave if, for all *i*, the marginal costs $C(i) - C(S \setminus i)$ are non-increasing with respect to set inclusion. The surplus game associated with a subadditive (resp. concave) cost game is super-additive (resp. convex) and the total surplus to be divided is positive if the cost game is essential. Most solution concepts agree on the class of concave cost games as was proved by Shapley (1971) and Maschler, Peleg and Shapley (1972, 1979): the core is the unique stable set (in the sense of von Neumann and Morgenstern) and it coincides with the bargaining set (with respect to the grand coalition); the kernel and the nucleolus coincide; the Shapley value is centrally located in the core.[8]

3. Data games

Let M_i be the set of data held by player *i* and $M = \bigcup_{i=1}^{n} M_i$ be the set of all data. Players may hold no data $(M_i = \varnothing)$ or hold the complete data set $(M_i = M)$. We denote by M_S the set of data held by coalition *S* i.e. $M_S = \bigcup_{i \in S} M_i$. We assume that $M_i \neq M$ for some i. If c_h is the cost of *reproducing* data *h*, the cost associated with a coalition is the cost of acquiring the missing data:

$$C(S) = \sum_{h \in M \setminus M_S} c_h = d_0 - \sum_{h \in M_S} c_h \quad \text{for all } S \neq \varnothing \tag{1}$$

where $d_0 = \Sigma_{h \in M} c_h$. This defines a class of cost games that we call "data games". Because $C(N) = 0$, data games are pure "compensation" games. We assume that $c_h > 0$ for all $h \in M$.

In what follows we shall consider two examples involving three players and three data, with common cost vector. Only the distribution of data among players will change. Player 1 will however be assumed to hold no data in both examples.

[8] More precisely, the Shapley value is the average of core vertices, accounting for multiplicity.

Example 1 The game associated with the data sets $M_1 = \varnothing$, $M_2 = \{1,2\}$ and $M_3 = \{2,3\}$, and cost vector $c = (6, 9, 12)$, is given by:

$C(1) = c_1 + c_2 + c_3 = d_0 = 27$

$C(2) = C(12) = c_3 = 12$

$C(3) = C(13) = c_1 = 6$

$C(23) = C(123) = 0$

As a matter of illustration, the vector of marginal costs associated with the permutation $\pi = (3,1,2)$ is given by $\mu(\pi) = (6, 0, -6)$.

Lemma 1 Data games are essential, subadditive and monotonically *decreasing*.

Proof $M \neq M_i$ for some i implies $\sum_{i \in N} C(i) = nd_0 - \sum_{i \in N} \sum_{h \in M_i} c_h > 0$. Essentiality then follows from $C(N) = 0$. To verify subadditivity, assume $S \bigcap T = \varnothing$. We then have:

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

If $S \subset T, S \neq \varnothing,$ we have $M_S \subset M_T$ and $C(T) - C(S) = \sum_{h \in M_S} c_h - \sum_{h \in M_T} c_h \leq 0$. ●

Let $T_h = \{i \in N \mid h \in M_i\}$ and $t_h = |T_h|$ denote the subset of players holding data *h* and the size of T_h respectively. An "elementary" data game (N, C_h) can be associated to each data *h*:

$$\begin{aligned} C_h(S) &= 0 \quad \text{if } S \bigcap T_h \neq \varnothing \\ &= c_h \quad \text{if } S \bigcap T_h = \varnothing \end{aligned} \tag{2}$$

for all $S \subset T, S \neq \varnothing$. Clearly a data game as defined by (1) can be decomposed into a sum of elementary data games:

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

and elementary data games can be written in terms of fixed cost games:

$$C_h(S) = (1 - e_{T_h}(S))\, c_h \tag{3}$$

Let's consider the particular case where data sets form a *partition* of *M*:

$M_i \bigcap M_j = \varnothing$ for all $i \neq j$

The T_h's are then singletons and the value of the data held by a coalition *S* can be written as

$$\sum_{h \in M_S} c_h = \sum_{i \in S} \sum_{h \in M_i} c_h = \sum_{i \in S} d_i$$

where $d_i = \sum_{h \in M_i} c_h$ is the value of the data held by player *i*. Using (1), a "partition" data game (*N*,*C*) is then simply defined by:

$$C(S) = d_0 - \sum_{i \in S} d_i \tag{4}$$

Example 2 The data sets $M_1 = \varnothing$, $M_2 = \{1\}$ and $M_3 = \{2,3\}$ form a partition of *M*. The game associated with the cost vector $c = (6, 9, 12)$ is given by:

$C(1) = d_0 = 27$

$C(2) = C(12) = c_2 + c_3 = 21$

$C(3) = C(13) = c_1 = 6$

$C(23) = C(123) = 0$

Lemma 2 Partition data games are concave.

Proof We first show that an elementary data game (N, C_h) such that $T_h = \{i\}$ for some $i \in N$ is concave. Consider two coalitions *S* and *T*. If they have a non-empty intersection, we have:

$$C(S \cup T) + C(S \cap T) - C(S) - C(T) = 0$$

whether $i \in S \cup T$ or not. If instead *S* and *T* are disjoint coalitions, $C(S \cap T) = 0$ and

$$C(S \cup T) - C(S) - C(T) = -c_h < 0$$

whether $i \in S \cup T$ or not. Being the sum of *m* concave games, a partition data game is concave. ●

The surplus game (*N*,*v*) associated with the partition data game (N, C) as defined by (4) is given by:

$$v(S) = (s-1)d_0 \quad \text{for all } S \neq \varnothing \tag{5}$$

It is a convex and *symmetric* game.

4. The core

An *imputation y* is an individually rational cost allocation:

$$y(N) = C(N) \quad \text{and} \quad y(i) \leq C(i) \quad \text{for all } i \in N$$

We denote by $I(N, v)$ the set of imputations of the game (N, v). It is a subset of $\mathbb{R}^n$ of dimension $n-1$. The *core* is the set of imputations *y* against which no coalition can object:

$$\mathbb{C}(N, C) = \{y \in \mathbb{R}^n \mid y(N) = 0 \quad \text{and} \quad y(S) \leq C(S) \quad \text{for all } S \subset N\} \tag{6}$$

i.e. no coalition pays more that its stand-alone cost.[9] In general, the core is a *convex polyhedron*, possibly empty, whose dimension does not exceed $n-1$.[10] On the class of concave cost game, the core is the convex hull of its marginal cost vectors.[11]

The set of imputations of the data game (N,C) defined by the data sets $(M_1,...,M_n)$ and cost vector $c=(c_1,...,c_m)$ is given by:

$$I(N,C)=\{y\in\mathbb{R}^n \mid y(N)=0 \text{ and } y_i\le d_0-d_i \text{ for all } i\in N\}$$

Data games being essential, the set of imputations is non-empty and there are cost allocations that are better than the individual costs for all players. The core of a data game is non-empty: it always contains the trivial allocation $0=(0,0,\ldots,0)$ defined by the absence of compensation. Indeed, $C(N)=0$ and $C(S)\ge 0$ for all $S\subset N$. Furthermore the core of a data game has a simple and regular structure that depends only on the data held by single players. In particular, no player can expect a compensation if he/she is not alone to hold some data.

Proposition 1 Let (N,C) be the game defined by the data sets $(M_1,...,M_n)$ and cost vector $c=(c_1,...,c_m)$. If $\bar{M}=\{h\in M \mid t_h=1\}\neq\varnothing$, then $\mathbb{C}(N,C)$ is the regular and full dimensional simplex whose *n* vertices $(v^1,...,v^n)$ are given by:

$$\begin{aligned} v^1&=(\bar{d}_0-\bar{d}_1,-\bar{d}_2,...,-\bar{d}_n)\\ v^2&=(-\bar{d}_1,\bar{d}_0-\bar{d}_2,...,-\bar{d}_n)\\ &\ldots\\ v^n&=(-\bar{d}_1,-\bar{d}_2,...,\bar{d}_0-\bar{d}_n) \end{aligned} \tag{7}$$

where $\bar{d}_0=\sum_{h\in\bar{M}}c_h$, $\bar{d}_i=\sum_{h\in\bar{M}_i}c_h$ and $\bar{M}_i=M_i\bigcap\bar{M}$. If $\bar{M}=\varnothing$, $\mathbb{C}(N,C)=\{0\}$.

Proof Using (6), the core can be written simply as

$$\mathbb{C}(N,C)=\{y\in\mathbb{R}^n \mid y(N)=0 \text{ and } y_i\ge-\bar{d}_i \text{ for all } i=1,...n\} \tag{8}$$

Indeed, if $y\in\mathbb{C}(N,C)$ we have $y(N\setminus i)\le C(N\setminus i)=\bar{d}_i$ and therefore $y_i\ge-\bar{d}_i$. If *y* satisfies (8) and $S\subset N$, we have:

$$y(N\setminus S)\ge-\sum_{i\in N\setminus S}\bar{d}_i$$

[9] The core was introduced by Gillies (1953). Equivalently, an allocation *y* belongs to the core *if and only if* $y(S)\ge C(N)-C(N\setminus S)$ for all $S\subset N$. There is *no cross-subsidization* in the sense that every coalition pays at least its marginal cost. See Faulhaber (1975).

[10] A *polyhedron* (or polyhedral set) in $\mathbb{R}^n$ is the intersection of a finite number of closed half spaces of $\mathbb{R}^n$. See Grünbaum (2003).

[11] See Shapley (1971). The core of a concave cost game coincides with the *Weber set* defined as the set of all *random order values*. See Weber (1988).

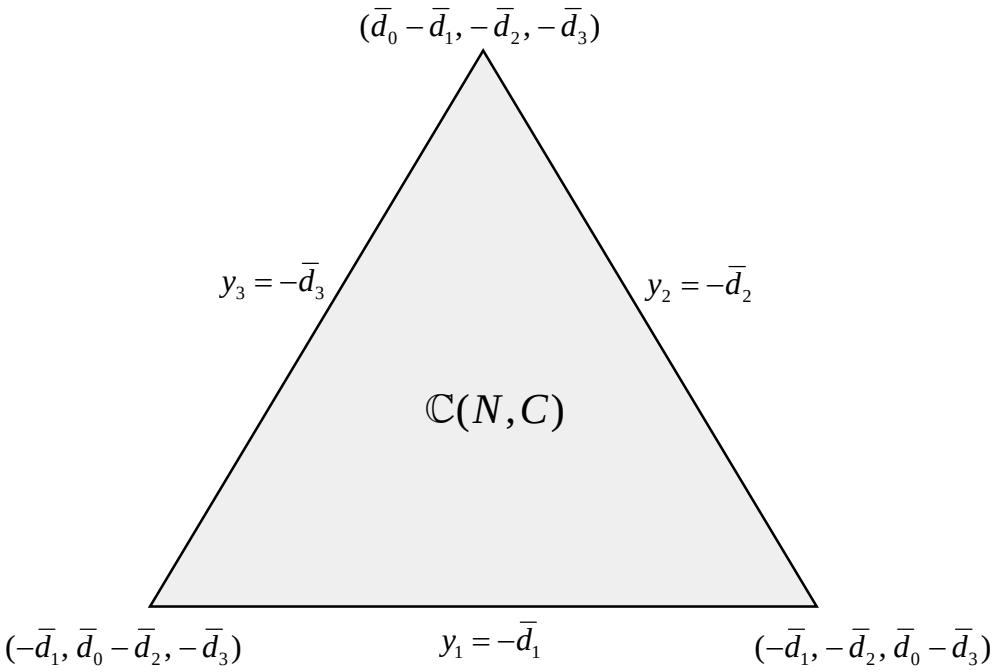
and therefore $y(S) \leq \sum_{i \in N \setminus S} \bar{d}_i \leq \sum_{h \in M \setminus M_S} c_h = C(S)$ for all $S \subset N$.

Translating the core by adding the vector $\bar{b} = (\bar{d}_1, ..., \bar{d}_n)$, we obtain the standard simplex $\{y \in \mathbb{R}^n_+ \mid y(N) = \bar{d}_0\}$.[12] If $\bar{M} \neq \varnothing$, positivity of $\bar{d}_0$ ensures full dimensionality. Core vertices are obtained by subtracting the vector $\bar{b}$. If $\bar{M} = \varnothing$, $\bar{d}_i = 0$ for all $i = 0,1,...,n$ and $\mathbb{C}(N,C) = \{0\}$. ●

Hence, if $\bar{M} \neq \varnothing$, the core of a data game is a regular simplex of dimension $n-1$ i.e. an equilateral triangle for n = 3, a regular tetrahedron for n = 4,… Its facets have dimension $n-2$ and are given by:[13]

$$F_i = \{y \in \mathbb{R}^n \mid y(N) = 0,\ y_i = -\bar{d}_i\} = \{y \in \mathbb{R}^n \mid y(N) = 0,\ y(N \setminus i) = \bar{d}_i\} \quad i = 1,...n$$

as illustrated by the figure below.



For the game defined in Example 1, the core is the set of allocations (y_1, y_2, y_3) such that:

$$\mathbb{C}(N,C) = \{y \in \mathbb{R}^n \mid y_1 + y_2 + y_3 = 0,\ y_1 \geq 0,\ y_2 \geq -6,\ y_3 \geq -12\}$$

[12] The unit simplex $\Delta_n = \{y \in \mathbb{R}^n_+ \mid y(N) = 1\}$ is obtained by dividing by $\bar{d}_0$.

[13] A *simplex* in $\mathbb{R}^n$ is the convex hull of *n* affinely independent vectors. A simplex is a polyhedral set. A *facet* is a maximal proper face of a polyhedral set. See Grünbaum (2003).

Consider an elementary data game (N, C_h). By Proposition 1, if $t_h = 1$, say $T_h = \{1\}$, the core is the regular simplex with vertices

$v^1 = (0, ..., 0)$

$v^2 = (-c_h, c_h, 0, ..., 0)$

...

$v^n = (-c_h, 0, ..., 0, c_h)$

If instead $t_h \geq 2$, the core reduces to $\{0\}$, the absence of compensation resulting from the competition between data holders. It is easily verified that the vertices of a data game (N, C) are the sum of the vertices of the elementary games associated to the data in $\bar{M}$. Hence we have the following result.

Corollary 1 The core of a data game is the sum of the cores of the elementary games associated to the data held by single players:

$$\mathbb{C}(N, C) = \sum_{h \in \bar{M}} \mathbb{C}(N, C_h) \quad \text{if } \bar{M} \neq \varnothing$$
$$= \{0\} \qquad\qquad \text{if } \bar{M} = \varnothing$$

In the partition case, the core is the convex hull of the marginal cost vectors. This is a consequence of concavity. Given that $\bar{M} = M$, there are *n* distinct marginal cost vectors, each with multiplicity $(n-1)!$. They are given by the vertices of the core defined in (7), with $d_i = \bar{d}_i$ for all $i = 0, 1, ..., n$. v^i is the marginal cost vector corresponding to the permutations where player *i* is *first*.

5. The nucleolus

Given an imputation $y \in I(N, C)$ and a coalition $S \subset N$ $(S \neq \varnothing, N)$, we define the "excess"

$$e(y, S) = y(S) - C(S)$$

as the difference between what coalition *S* contributes under *y* and its cost. An imputation *y* belongs to the core if $e(y, S) \leq 0$ for all $S \subset N$ $(S \neq \varnothing, N)$. The *least core* and the *nucleolus* are solution concepts concerned with the minimization of these excesses. The least core is the set of imputations that minimize the largest excess:

$$\text{Min}_{y \in I(N,v)} \text{Max}_{\substack{S \subset N \\ S \neq \varnothing, N}} e(y, S)$$

It has dimension at most $n - 2$. If the core is non-empty, the least core is obviously a subset of the core. The nucleolus introduced by Schmeidler (1969) goes further by comparing excesses lexicographically so as to eventually retain a unique allocation.

For a data game (N,C) such that $\bar{M} \neq \varnothing$, the core is given by (8) and the definition of the least core then simplifies to:

$$\text{Min}_{y \in I(N,v)} \text{Max}_{i \in N}\ e(y, N \setminus i)$$

where $e(y, N \setminus i) = y(N \setminus i) - \bar{d}_i = -(y_i + \bar{d}_i)$. The least core is therefore uniquely defined by the equations:

$$y(N) = 0 \text{ and } y_i + \bar{d}_i = a \quad (i = 1,...n) \tag{9}$$

for some real *a*. Solving (9), we get:

$$y_i = a - d_i \text{ with } a = \frac{\bar{d}_0}{n}$$

The least core being uniquely defined, it coincides with the nucleolus. It is also centre of gravity defined as the average of core vertices. It is located at equal distance from core facets.[14]

Proposition 2 Let (N,C) be the game defined by the data sets $(M_1,...,M_n)$ and cost vector $c = (c_1,...,c_m)$. If $\bar{M} = \{h \in M \mid t_h = 1\} \neq \varnothing$, the nucleolus is given by the average of the vertices of its core:

$$\eta_i(N,C) = \frac{\bar{d}_0}{n} - \bar{d}_i \quad (i = 1,...n) \tag{10}$$

where $\bar{d}_0 = \sum_{h \in \bar{M}} c_h$, $\bar{d}_i = \sum_{h \in \bar{M}_i} c_h$ and $\bar{M}_i = M_i \cap \bar{M}$. If $\bar{M} = \varnothing$, $\eta(N,C) = 0$.

Hence a player is compensated *if and only if* the cost of the data he or she is *alone* to hold exceeds the per capita cost of the data held by *single* players. For the game defined in Example 1, the nucleolus is the allocation $(6,0,-6)$. In the partition case, $\bar{M} = M$ and the nucleolus becomes:

$$\eta_i(N,C) = \frac{d_0}{n} - d_i \quad (i = 1,...n) \tag{11}$$

For the game defined in Example 2, the nucleolus is the allocation $(9,3,-12)$.

[14] The centre of gravity of the core has been introduced as a solution concept by Gonzales-Diaz and Sanchez-Rodriguez (2007).

6. The Shapley value

The (symmetric) Shapley value of a cost game (N,C) is the average marginal cost vector:

$$\varphi(N,C) = \frac{1}{n!}\sum_{\pi\in\Pi_n}\mu(\pi)$$

It is the unique *additive* allocation rule on $G(N)$ that satisfies *symmetry* (players with identical marginal costs are *substitutes* and pay the same amount) and *dummy* (players with zero marginal costs are *dummies* and pay nothing). Additivity, symmetry and dummy are the original axioms introduced by Shapley (1953, 1981a).[15]

The value defines an imputation for subadditive cost games and belongs to the core of concave cost games. Because data games can be written as sums of elementary games, computation of the value is straightforward as a consequence of additivity.

Proposition 3 The Shapley value of a data game (N,C) defined by the data sets $(M_1,...,M_n)$ and cost vector $c=(c_1,...,c_m)$ is given by:

$$\varphi_i(N,C) = \frac{d_0}{n} - \sum_{h\in M_i}\frac{c_h}{t_h} \quad (i=1,...,n) \tag{12}$$

where $d_0 = \sum_{h\in M} c_h$ and $t_h = |T_h|, T_h = \{i\in N \mid h\in M_i\}$.

Proof The Shapley value of a fixed cost game (N,e_T) is given by:

$$\begin{aligned}\varphi_i(N,e_T) &= \frac{1}{t} \quad \text{for all } i\in T\\ &= 0 \quad \text{for all } i\notin T\end{aligned}$$

Indeed players outside T are dummies and players in T are substitutes. The Shapley value is a linear operator. Using (3), the value of an elementary data game (N,C_h) is then given by:

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

Data games can be written as a sum of elementary data games. By additivity, the value of the data game (N,C) defined by $(M_1,...,M_n)$ and $(c_1,...,c_m)$ is given by (12). •

[15] There are alternative axiomatizations in particular the one proposed by Young (1985a). They are reviewed by Moulin (2003). The nucleolus satisfies symmetry and dummy but not additivity.

Hence the cost of the complete data set is uniformly allocated among all players and the cost of each data is uniformly redistributed to the players holding it. In Example 1, the Shapley value is the allocation (9,–1.5,–7.5) to be compared to the allocation (6,0,–6) derived from the nucleolus.

Remark 1 According to the Shapley value, what a player receives decreases with the number of players holding the same data. Furthermore, that amount increases with the cost of the data he or she holds. The same is true for the nucleolus (10) but only with respect to the data that player is alone to hold.

The Shapley value of a partition data game is easily derived, either by setting $t_h = 1$ in (12) or by using its definition as the average marginal cost vector:

$$\varphi_i(N,C) = \frac{1}{n}\sum_{j=1}^{n} v_i^j = \frac{1}{n} d_0 - d_i \tag{13}$$

knowing that the n marginal cost vectors have the same multiplicity. The only players to be compensated are those endowed with a data set whose value exceeds the per capita value of the complete data set.

The following proposition is an immediate consequence of the equality between marginal cost vectors and core vertices in the partition case.

Proposition 4 The Shapley value and the nucleolus of a partition data game coincide.

Any allocation rule satisfying symmetry produces the same allocation. Indeed, the associated surplus game (5) is symmetric and the total surplus $(n-1)\,d_0$ is then divided equally:

$$x_i = \frac{n-1}{n} d_0$$

The resulting cost allocation is the nucleolus (11) and the Shapley value (13):

$$y_i = C(i) - \frac{n-1}{n} d_0 = (d_0 - d_i) - \frac{n-1}{n} d_0 = \frac{1}{n} d_0 - d_i$$

In Example 2, the marginal costs vectors are given by:

$$v^1 = (27, -6, -21)$$

$$v^2 = (0, 21, -21)$$

$$v^3 = (0, -6, 6)$$

Taking the average, we get the allocation $(9, 3, -12)$. It is indeed the nucleolus previously computed.

7. The asymmetric Shapley value

The *weighted* Shapley value allows asymmetries between players to be taken into account.[16] We denote by $(w_1,...,w_n)$ the weights assigned to players. At this stage we assume that $w_i > 0$ for all $i \in N$. The case where some players are assigned a *zero weight* will be considered later. In a cost allocation context, w_i determines the share of player *i* in a fixed cost i.e.

$$\varphi_i(N,C,w) = \frac{w_i}{w(N)} F \quad (i = 1,...,n)$$

for the game (N,C) defined by $C(S) = F$ for all $S \subset N, S \neq \varnothing$. More generally, the value of a fixed cost game (N,e_T) is given by:

$$\begin{aligned} \varphi_i(N,e_T,w) &= \frac{w_i}{w(T)} \quad \text{for all } i \in T \\ &= 0 \quad \text{for all } i \notin T \end{aligned}$$

where $w(T)$ is the weight of coalition *T*. The symmetric case corresponds to $w_i = 1$. Using (3), the value of the elementary data game (N,C_h) associated with weights $(w_1,...,w_n)$ is given by:

$$\begin{aligned} \varphi_i(N,C_h,w) &= \frac{w_i}{w(N)} c_h - \frac{w_i}{w(T_h)} c_h \quad \text{for all } i \in T_h \\ &= \frac{w_i}{w(N)} c_h \quad \text{for all } i \notin T_h \end{aligned}$$

Remark 2 We observe that, for a given data *h*, the ratio between what two players in T_h pay or receive is equal to their weight ratio. The same applies to players outside T_h :

$$\frac{\varphi_i(N,C_h,w)}{\varphi_j(N,C_h,w)} = \frac{w_i}{w_j} \quad \text{for all } i, j \in T_h \quad \text{or} \quad \text{for all } i, j \notin T_h$$

The following proposition is an immediate consequence of additivity.

Proposition 5 Given *positive* weights $(w_1,...,w_n)$, the weighted Shapley value of the data game (N,C) defined by the data sets $(M_1,...,M_n)$ and cost vector $(c_1,...,c_m)$ is given by:

$$\varphi_i(N,C,w) = \frac{w_i}{w(N)} d_0 - \sum_{h \in M_i} \frac{w_i}{w(T_h)} c_h \quad (i = 1,...,n) \qquad (14)$$

[16] Weighted values were introduced in Shapley's Ph.D. dissertation and have been later axiomatized by himself (1981b) in a cost allocation context and by Kalai and Samet (1987). The set of all weighted values contains the core and a cost game is concave *if and only if* the set of weighted values and the core coincide. See Monderer, Samet and Shapley (1992).

The weighted value is not necessarily monotonic with respect to weights. What a player pays may well decrease while his or her weight increases. Monderer, Samet and Shapley (1992) have shown that concavity is actually a *necessary and sufficient* condition for monotonicity.

Using (14), the asymmetric Shapley value of a partition data game is given by:

$$\varphi_i(N,C,w) = \frac{w_i}{w(N)} d_0 - d_i \quad (i = 1,...,n) \tag{15}$$

Indeed $w(T_h) = w_i$ for all $h \in M_i$ in the partition case.

In the examples 1 and 2, the Shapley values associated with the weights (1, 1, 2) are given respectively by (6.75,–2.25,–4.5) and (6.75, 0.75,–7.5), to be compared to the allocations (9,–1.5,–7.5) and (9, 3,–12) under equal weights.

So far we have considered the case where weights are positive. A zero weight can be assigned to players who hold data but are not interested in completing their data set. Let $Z_w = \{i \in N \mid w_i = 0\}, Z_w \neq N,$ denote the set of zero weight players. Consider the sequences (w^ν) defined by $w_i^\nu = w_i$ for all $i \in N \setminus Z_w$ and $w_i^\nu \to 0$ for all $i \in Z_w$. Then players' permutation in which a zero weight player precedes a nonzero weight player has a zero probability limit.[17] As a consequence, we have the following proposition.

Proposition 6 Zero weight players are compensated for a data they hold *if and only if* no positive weight player holds the same data.

In particular, if a data is held exclusively by a single zero weight player, he or she receives the total value of his or her data. If a data is held exclusively by several zero weight players, the way they share the value of the data is indeterminate.

If there is no reason to discriminate among zero weight players, we may restrict ourselves to sequences (w^ν) where $w_i^\nu = t^\nu \to 0$ for all $i \in Z_w$. The resulting value of an elementary game (N, C_h) is then unchanged for positive weight players while, for zero weight players, we get:

$$\begin{aligned} \varphi_i(N,C_h,w) &= -\frac{c_h}{u_h} \quad \text{if } T_h \subset Z \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

where u_h is the number of zero weight players holding data *h*. Hence we have:

$$\varphi_i(N,C,w) = -\sum_{\substack{h \in M_i \\ T_h \subset Z}} \frac{1}{u_h} c_h \quad \text{for all } i \in Z$$

[17] See Dehez (2009) for more details.

8. Weighted charge accounting rule

There exist various accounting rules for dividing joint costs. The simplest ones are based on players' marginal costs with respect to the grand coalition:

$$\theta_i(N,C,\alpha) = MC_i + \alpha_i(C(N) - \sum_{j=1}^{n} MC_j) \quad (i = 1,...,n) \tag{16}$$

where $\alpha \in \Delta_n$ and $MC_i = C(N) - C(N\backslash i)$ is the "separable cost" of player *i*. Weights are exogenously given or depend on the cost function.[18] We shall restrict our attention to the case where weights are exogenous. θ is then an *additive* (actually linear) rule that satisfies the symmetry axiom but not the dummy axiom. If weights are equal, (16) defines the "equal charge" allocation rule.

Consider a data game (N,C). Then $MC_i = -C(N \setminus i) = -\bar{d}_i$ for all *i* and we have the following propositions.

Proposition 7 Let (N,C) be the data game defined by the data sets $(M_1,...,M_n)$ and cost vector $(c_1,...,c_m)$. If $\bar{M} = \{h \in M \mid t_h = 1\} \neq \varnothing$, the weighted charge accounting rule (16) leads to the following allocation:

$$\theta_i(N,C,\alpha) = \alpha_i \, \bar{d}_0 - \bar{d}_i \quad (i = 1,...,n) \tag{17}$$

If $\bar{M} = \varnothing$, $\theta(N,C,\alpha) = 0$.

Corollary 2 Applied to data games, the weighted charge accounting rule (16) defines a core allocation for any choice of weights. Furthermore the equal charge rule defines an allocation that coincides with the nucleolus:

$$\theta_i(N,C,\alpha) = \frac{1}{n} \bar{d}_0 - \bar{d}_i \quad (i = 1,...,n)$$

. In the partition case, it also coincides with the Shapley value.

Actually, the core being the convex hull of its vertices (7), it can alternatively be defined as:

$$\mathbb{C}(N,C) = \{y \in \mathbb{R}^n \mid y = \theta(N,C,\alpha),\ \alpha \in \Delta_n\}$$

i.e. weights can be associated to core allocations and vice-versa. Furthermore, (17) is exactly the weighted Shapley value in the partition case given by (15), when weights are positive and satisfy $w_i = \alpha_i$ for all $i \in N$.

[18] Other accounting rules use endogenous weights like for instance the "*separable costs remaining benefits*" rule (SCRB). See Béal et al. (2009) for an application to REACH data sharing.

9. Concluding remarks

The question that comes up immediately concerns the choice of the allocation method – outside the partition case – between the nucleolus or the Shapley value. The nucleolus is a core allocation: no coalition can improve upon the proposed compensations. At the same time, it limits the extend of these compensations: only the data held by single players, if any, enter into account. The Shapley value instead takes into account the entire data distribution and compensates players for data they are not alone to hold. This seems to be fairer. The problem is that there will then be coalitions to challenge the resulting allocation. Should it be a reason to dismiss the Shapley value as a compensation mechanism? Not necessarily because what the core suggests may be unacceptable as the following example suggest. Consider a situation where only two players hold data, say players n and n-1, and the data sets they hold differ only by a single data, say data 1:

$$M_i = \varnothing \ \ (i = 1,...,n-2),\ \ M_{n-1} = \{2,...,m\} \ \text{ and } \ M_n = \{1,...,m\},$$

In this case, the core imposes that only player n may be compensated with an amount not exceeding c_1 – the cost of the missing data – while all the other players including player $n-1$ may be asked to pay up to c_1. The nucleolus goes further by imposing that the $n-1$ first players pay the same amount, namely c_1 / n. It is to be compared with the allocation derived from the Shapley value. Using (12) we get:

$$y_i = \frac{d_0}{n} \quad (i = 1,...n-2)$$

$$y_{n-1} = \frac{d_0}{n} - \sum_{h=2}^{m} \frac{c_h}{2} = -\frac{n-2}{2n} d_0 + \frac{c_1}{2}$$

$$y_n = \frac{d_0}{n} - \sum_{h=2}^{m} \frac{c_h}{2} - c_1 = -\frac{n-2}{2n} d_0 - \frac{c_1}{2}$$

It is definitely more acceptable: players without data pay the per capita cost of the complete data set 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 costs parameters.[19] Reaching a consensus on the cost parameters is clearly the most difficult part, in particular because under the Shapley value or the nucleolus, we know from Remark 1 that what a player pays decreases with the cost of the data he or she holds. 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.

[19] In that framework the firms are typically of different sizes and an agreement on weights must then also be reached. These are the weights that would be used to share the cost of *additional* data.

References

Béal S., Deschamps M., J.T. Ravix and O. Sautel (2009), Les informations exigées par la législation REACH: analyse du partage des coûts, *mimeo*, Université de Saint-Etienne.

Dehez P. (2009), The allocation of fixed costs and the weighted Shapley value, CORE *Discussion Paper* 2009/35.

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.

Gillies D.B. (1953), Some theorems on *n*-person games. *PhD Thesis*, Princeton.

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

Grünbaum B. (2003), *Convex polytopes*, 2nd edition, Springer Verlag.

Kalai E. and D. Samet (1987), On weighted Shapley values, *International Journal of Game Theory* 16, 205-222.

Katz M.L. (1995), Joint ventures as a mean of assembling complementary inputs, *Group Decision and Negotiation* 4, 383-400.

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

Maschler M., B. Peleg and L.S. Shapley (1972), The kernel and bargaining set for convex games, *International Journal of Game Theory* 1, 73-93.

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

Monderer D., D. Samet and L.S. Shapley (1992), Weighted values and the core, *International Journal of Game Theory* 21, 27-39.

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

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

Roth A.E. (ed., 1988), *The Shapley value. Essays in honor of Lloyd Shapley*, Cambridge University 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 (1988), 31-40.

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

Shapley L.S. (1981a), 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.

Shapley L.S. (1981b), Discussion comments on "Equity considerations in traditional full cost allocation practices: An axiomatic approach", in Moriarity S. (ed.), *Joint cost allocation*, Proceeding of the University of Oklahoma Conference on Costs Allocations, April 1981, Center for Economic and Management Research, University of Oklahoma.

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

Weber R.J. (1988), Probabilistic values for games, in Roth (ed. 1988), 101-119.

Young H.P. (1985a), Monotonic solutions of cooperative games, *International Journal of Game Theory* 14, 65-72.

Young H.P. (1985b), *Cost allocation: methods, principles, applications*, North-Holland, Amsterdam.