Bureau
d'économie
théorique
et appliquée
(BETA)
UMR 7522

Documents
de travail

« Data games : sharing public goods with exclusion »

(second version)

Auteurs

Pierre Dehez, Daniela Tellone

Document de Travail n° 2011 - **04**

Janvier 2011

**Faculté des sciences
économiques et de
gestion**
Pôle européen de gestion et
d'économie (PEGE)
61 avenue de la Forêt Noire
F-67085 Strasbourg Cedex

Secétariat du BETA
Géraldine Manderscheidt
Tél. : (33) 03 68 85 20 69
Fax : (33) 03 68 85 20 70
g.manderscheidt@unistra.fr
http://cournot2.u–strasbg.fr/beta

cnrs
Nancy-Université
Université Nancy 2
UNIVERSITÉ DE STRASBOURG

Data games: Sharing public goods with exclusion

Pierre Dehez* and Daniela Tellone**

November 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 whose Shapley comes out as a natural solution. The main result concerns the regular structure of the core that enables a simple characterization of the nucleolus. However, compared to the Shapley value, the nucleolus defines compensations that appear to be less appropriate in the context of data sharing. Our analysis is inspired by the problem faced by the European chemical firms within the regulation program REACH that 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 Theo Driessen, Pier Mario Pacini, Eve Ramaekers, Gisèle Umbhauer, Myrna Wooders and two anonymous referees 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). According to this regulation, manufacturers and importers are required to collect safety information on the properties of the chemical substances they produce, import or use, and to register that 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! The European Commission encourages firms 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.[3] These inputs are non-rival but excludable goods i.e. public goods with exclusion such as knowledge, data or information, patents or copyrights.[4] The question is how to compensate the firms for the inputs they contribute. The problem can be framed within a cost game to which standard cost sharing rules can be applied, in particular the nucleolus, the Shapley value as well as 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 value of the missing data. It will be shown that data games, and the corresponding surplus sharing games, form an interesting class of transferable utility games on which the core, the nucleolus and the Shapley value can be characterized in a simple and straightforward way.

Data games are essential, subadditive and decreasing. They can be decomposed into a sum of elementary data games, one for each data. Data games have a non-empty core 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 living beings.

[2] The page echa.europa.eu/reach_en.asp offers guidance for the implementation of REACH. The compensation formula that is proposed there is analyzed by Béal et al. (2010) who also consider and analyze other formulas.

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

[4] 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."

coalition can object when no one is asked to pay. We shall see that the core actually limits the extent of compensation and, in some situations, 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 data pay 1/*n* each to the data holder. 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 they get back 1/*t* each. The *n*–*t* players without data pay 1/*n* each and the *t* data holders receive $1/t-1/n$ each: 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 and it differs from the nucleolus that in this case 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 while not being 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. The asymmetric (or weighted) Shapley value is defined and analyzed in Section 7 with a particular attention to the case where some players are assigned a zero weight. Weighted charge sharing rules are defined and applied to data games in Section 8. It is shown that they produce core allocations for any choice of weights. Section 9 is devoted to the particular situation where data sets form a partition of the complete data set, a situation that fits joint ventures involving patents or copyrights.[5] Data games are shown to be concave in that particular case, with the consequence that the Shapley value and the nucleolus coincide. Concluding remarks are offered in the last section.

2. Preliminaries: 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 any coalition 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*.[6] An *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:

$$V(S) = \sum_{i \in S} C(i) - C(S) \tag{1}$$

Notation: The letters *n, s, t,…* denote the size of the sets *N, S, T,…* 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…* instead of $\{i,j,k,...\}$. 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 real valued functions defined on the subsets of some finite set N. $G(N)$ is a vector space. The collection of $2^n - 1$ "unanimity" games

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

defined for all $T \subset N, T \neq \varnothing$, forms a basis of $G(N)$. These games have been introduced by Shapley in 1953 to prove existence and uniqueness of the value. Here we shall use the basis formed by the collection of $2^n - 1$ "fixed cost" games

$$\begin{aligned} e_T(S) &= 1 \quad \text{if } S \cap T \neq \varnothing \\ &= 0 \quad \text{if not} \end{aligned} \tag{2}$$

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: $e_T = u_T^*$. They are used by Dehez (2011) to characterize the weighted Shapley value in terms of the allocation of fixed costs, along the lines suggested by Shapley (1981b).

[5] The problem of sharing an information protected by a patent has been studied in a cooperative framework by Muto, Potters and Tijs (1989) for the case of a single owner and from a profit sharing point of view.

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

Marginal costs play a central role in cost allocation. Given a coalition *S* and a player *i* in *S*, the *marginal cost* of player *i* with respect to coalition *S* is defined by $C(S) - C(S \setminus i)$. Let Σ_n be the set of all players' permutations. To each permutation $\sigma = (i_1,...,i_n) \in \Sigma_n$ we associate the vector of marginal costs $\mu(\sigma)$ whose elements are given by:

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

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

It is easily seen that this defines a cost allocation.

A cost game (N, C) is *symmetric* if the players are substitutable: only the size of a coalition determines its cost. It is *increasing* (resp. *decreasing*) if $S \subset T \Rightarrow C(S) \leq C(T)$ (resp. $\geq$). It is *essential* if $C(N) < \sum_{i \in N} C(i)$. It is *subadditive* if $S \bigcap T = \varnothing \;\Rightarrow\; C(S \bigcup T) \leq C(S) + C(T)$. It is *concave* if $C(S \bigcup T) \leq C(S) + C(T) - C(S \bigcap T)$ for all *S* and *T*.[7] Hence concavity implies subadditivity. 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 sharing situations and associated cost games

Given a set M_0 of data and a set $N = \{1,...,n\}$ of players, a *data sharing situation* is defined by a collection of sets $M = (M_1,...,M_n)$ where $M_i \subset M_0$ specifies the data held by player *i*, and a cost vector *d* where $d_h > 0$ is the cost of *reproducing* data *h*. We assume that each data is held by at least one player: $\bigcup_{i \in N} M_i = M_0$.[9] There are no further restrictions: players may hold no data, $M_i = \varnothing$, or hold the complete data set $M_i = M_0$. We denote by *DS*(*N*) the set of data sharing situations (M, d) on a given set *N* of players.

If $M_S = \bigcup_{i \in S} M_i$ is the set of data held by coalition *S*, the cost associated with a coalition is the value of acquiring the missing data:

$$C(S) = \sum_{h \in M_0 \setminus M_S} d_h = v_0 - \sum_{h \in M_S} d_h \quad \text{for all } S \neq \varnothing \qquad (3)$$

[7] Equivalently, a cost game (*N,C*) is concave if, for all *i*, the marginal costs *C*(*S*) – *C*(*S**i*) are non increasing with respect to set inclusion.

[8] The Shapley value is indeed the average of the marginal cost vectors while the core of a concave game is the polyhedra whose vertices are the marginal cost vectors (a necessary and sufficient condition for concavity).

[9] As indicated in the concluding section, that assumption actually entails no loss of generality.

where $v_0 = \Sigma_{h\in M_0} d_h$ is the value of the complete data set. This defines the cost game (N,C) – called *data game* – associated to the data sharing situation $(M,d) \in DS(N)$. We denote by $DG(N) \subset G(N)$ the set data cost functions on a given set N of players. Because $C(N) = 0$, data games are pure "compensation" games.

Example 1 Consider the data sharing situation involving 3 players and defined by the data sets $M_1 = \varnothing$, $M_2 = \{1,2\}$ and $M_3 = \{2,3\}$, and by cost vector $d = (6, 9, 12)$. The corresponding data game (N,C) and associated surplus game (N,V) as defined by (1) are then given by:

$$C(1) = d_1 + d_2 + d_3 = v_0 = 27 \qquad V(1) = V(2) = V(3) = 0$$

$$C(2) = C(12) = d_3 = 12 \qquad V(12) = V(13) = 27$$

$$C(3) = C(13) = d_1 = 6 \qquad V(23) = 18$$

$$C(23) = C(123) = 0 \qquad V(123) = 45$$

Lemma 1 Data games are subadditive and decreasing. Data games where at least one player does not hold the complete data set $(M_i \neq M_0$ for some $i)$ are essential.

Proof $M_0 \neq M_i$ for some i implies $\sum_{i\in N} C(i) = nv_0 - \sum_{i\in N}\sum_{h\in M_i} d_h > 0.$ Essentiality then follows from $C(N) = 0.$ To verify subadditivity, assume $S \bigcap T = \varnothing.$ We then have:

$$C(S) + C(T) = 2v_0 - \sum_{h\in M_S} d_h - \sum_{h\in M_T} d_h = C(S \bigcup T) + v_0 - \sum_{h\in M_S \cap M_T} d_h \geq C(S \bigcup T)$$

To verify that a data game is decreasing, let $S \subset T, S \neq \varnothing.$ We then have $M_S \subset M_T$ and as a consequence

$$C(T) - C(S) = \sum_{h\in M_S} d_h - \sum_{h\in M_T} d_h \leq 0. \qquad \bullet$$

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 sharing situation is a situation where there is a single data. An elementary data game (N,C_h) can then be associated to each data h:

$$\begin{aligned} C_h(S) &= 0 \quad \text{if } S \bigcap T_h \neq \varnothing \\ &= d_h \quad \text{if } S \bigcap T_h = \varnothing \end{aligned} \tag{4}$$

for all $S \subset N, S \neq \varnothing.$ Clearly any data game as defined by (3) can be decomposed into a sum of elementary data games (4):

$$\sum_{h\in M_0} C_h(S) = \sum_{h\in M_0 \setminus M_S} d_h = C(S)$$

and elementary data games can be written in terms of fixed cost games (2):

$$C_h(S) = (1 - e_{T_h}(S))\, d_h \tag{5}$$

4. Imputations and core allocations

An *imputation* y is an individually rational cost allocation: $y(N)=C(N)$ and $y(i)\leq C(i)$ for all $i\in N$. We denote by $I(N,C)$ the set of imputations of the cost game (N,C), a nonempty subset of $\mathbb{R}^n$ that has dimension $n-1$ if the game is essential.

Imputations y of the cost game (N,C) and imputations x of the associated surplus game (1) are related by the following identities:

$$x_i + y_i = C(i) \quad i=1,...,n \tag{6}$$

The *core* is the set of imputations y against which no coalition can object:

$$\mathbb{C}(N,C)=\{y\in\mathbb{R}^n \mid y(N)=0 \text{ and } y(S)\leq C(S) \text{ for all } S\subset N\} \tag{7}$$

i.e. no coalition pays more that its stand-alone cost.[10] In general, the core is a *convex polyhedron*, possibly empty, whose dimension does not exceed $n-1$.[11] Data games being subadditive, the set of imputations is non-empty. The core of a data game is non-empty: it always contains the trivial allocation 0 = (0, 0,…, 0) defined by the absence of compensation. Indeed, $C(N)=0$ and $C(S)\geq 0$ 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. Given a data sharing situation $(M,d)\in DS(N)$ we denote by $\overline{M}_0\subset M_0$ the subset of data held by single players and by $\overline{M}_i = M_i \cap \overline{M}_0$ the subset of data player i is alone to hold. On that basis, we define $\overline{v}_i = \sum_{h\in\overline{M}_i} d_h$ and $\overline{v}_0 = \sum_{h\in\overline{M}_0} d_h$. In particular $\overline{v}_i = C(N\setminus i)$.

Proposition 1 Consider the data sharing situation $(M,d)\in DS(N)$ and the associated cost function $C\in DG(N)$. If $\overline{M}_0=\varnothing$, $\mathbb{C}(N,C)=\{0\}$. If instead $\overline{M}_0\neq\varnothing$, $\mathbb{C}(N,C)$ is the regular and full dimensional simplex whose n vertices $(\pi^1,...,\pi^n)$ are:

$$\begin{aligned}
&\pi^1 = (\overline{v}_0-\overline{v}_1, -\overline{v}_2,..., -\overline{v}_n)\\
&\pi^2 = (-\overline{v}_1, \overline{v}_0-\overline{v}_2,..., -\overline{v}_n)\\
&...\\
&\pi^n = (-\overline{v}_1, -\overline{v}_2,..., \overline{v}_0-\overline{v}_n)
\end{aligned} \tag{8}$$

[10] The core was introduced by Gillies (1953). Equivalently, an allocation y belongs to the core *if and only if* $y(S)\geq C(N)-C(N\setminus S)$ 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).

[11] 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).

Proof Using (7), the core can be written simply as

$$\mathbb{C}(N,C) = \{y \in \mathbb{R}^n \mid y(N) = 0 \text{ and } y_i \geq -\bar{v}_i \text{ for all } i = 1,...n\} \qquad (9)$$

Indeed, if $y \in \mathbb{C}(N,C)$ we have $y(N \setminus i) \leq C(N \setminus i) = \bar{v}_i$ and therefore $y_i \geq -\bar{v}_i$. If now y satisfies (9) and $S \subset N$, $S \neq \varnothing$, we have:

$$y(N \setminus S) \geq -\sum_{i \in N \setminus S} \bar{v}_i = -C(S)$$

$y(N) = 0$ then implies $y(S) \leq \sum_{i \in N \setminus S} \bar{v}_i$ where $\sum_{i \in N \setminus S} \bar{v}_i \leq \sum_{h \in M_0 \setminus M_S} d_h = C(S)$ for all $S \subset N$.

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

Once each data is held by more than two players, the core reduces to the no compensation allocation. Furthermore, no player can expect a compensation if he/she is not alone to hold some data.

Hence, if $\bar{M}_0 \neq \varnothing$, 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 center of gravity is simply the average of its vertices, a property that will be used later to define the nucleolus. Its n facets have dimension $n-2$ and are given by:[13]

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

as illustrated by the figure below for $n = 3$.

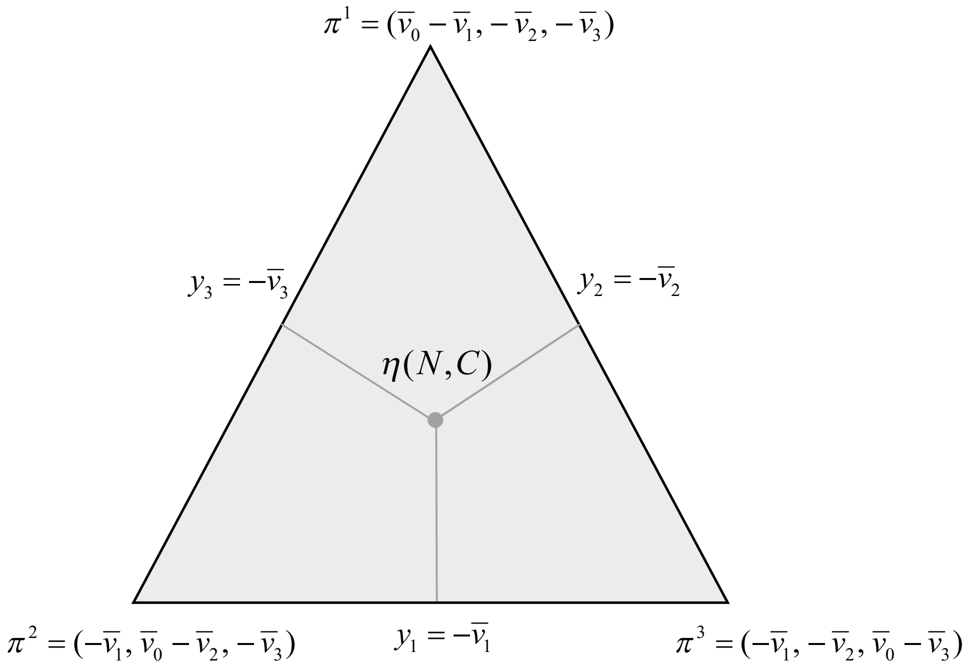
For the game defined in Example 1, $\bar{v} = (0,6,12)$ and the core is given by:

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

Its vertices are $\pi^1 = (18,-6,-12)$, $\pi^2 = (0,12,-12)$ and $\pi^3 = (0,-6,6)$, and its center of gravity is the allocation $(6,0,-6)$.

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

[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). Cost games whose core are regular simplices are 1-*concave* games. See Driessen (1985).


$\pi^1 = (\bar{v}_0 - \bar{v}_1, -\bar{v}_2, -\bar{v}_3)$
$y_3 = -\bar{v}_3$
$y_2 = -\bar{v}_2$
$\eta(N,C)$
$\pi^2 = (-\bar{v}_1, \bar{v}_0 - \bar{v}_2, -\bar{v}_3)$
$y_1 = -\bar{v}_1$
$\pi^3 = (-\bar{v}_1, -\bar{v}_2, \bar{v}_0 - \bar{v}_3)$

The core of a 3-player data game and its nucleolus

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

$$\pi^1 = (0, ..., 0)$$

$$\pi^2 = (-d_h, d_h, 0, ..., 0)$$

$$...$$

$$\pi^n = (-d_h, 0, ..., 0, d_h)$$

If instead $t_h \geq 2$, the core reduces to $\{0\}$: the absence of compensation reflects 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 $\bar{M}_0$. 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:

$$\begin{aligned} \mathbb{C}(N,C) &= \sum_{h \in \bar{M}_0} \mathbb{C}(N, C_h) \text{ if } \bar{M}_0 \neq \varnothing \\ &= \{0\} \qquad\qquad\quad \text{if } \bar{M}_0 = \varnothing \end{aligned}$$

5. The nucleolus

Given an imputation $y \in I(N,C)$ and a coalition $S \subset N\ (S \neq \emptyset, N),$ 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 $S \subset N\ (S \neq \emptyset, N).$ 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:

$$\mathrm{Min}_{y \in I(N,C)} \mathrm{Max}_{\substack{S \subset N \\ S \neq \emptyset, N}} e(y,S)$$

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.

Proposition 2 Consider the data sharing situation $(M,d) \in DS(N)$ and the associated cost function $C \in DG(N).$ If $\bar{M}_0 = \emptyset,$ $\eta(N,C) = 0.$ If instead $\bar{M}_0 \neq \emptyset,$ the nucleolus is given by the average of the vertices of its core:

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

Proof If $\bar{M}_0 \neq \emptyset,$ the core is given by (9) and the definition of the least core then simplifies to:

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

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

$$y(N) = 0 \text{ and } y_i + \bar{v}_i = a \quad (i = 1,...n) \tag{11}$$

for some real a. Solving (11), we get:

$$y_i = a - \bar{v}_i \text{ with } a = \frac{\bar{v}_0}{n}$$

The least core being uniquely defined, it coincides with the nucleolus. ●

The nucleolus is also center of gravity defined as the average of core vertices.[14] It is located at equal distance from the core's facets.[15] Hence a player is compensated *if and only if* the value

[14] Notice that the Shapley value and the nucleolus have a similar formula. The nucleolus can be viewed as the restriction of the Shapley value to the data held by single players.

[15] The center of gravity of the core has been proposed as a possible core selection by Gonzáles-Díaz and Sánchez-Rodríguez (2007).

of the data he or she is *alone* to hold exceeds the per capita value of the data held by *single* players. For the game defined in Example 1, the nucleolus is the allocation $(6,0,-6)$.

6. The Shapley value

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

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

It is the unique *additive* sharing 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).[16]

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 Consider the data sharing situation $(M,d)\in DS(N)$ and the associated cost function $C\in DG(N)$. Its Shapley value is given by:

$$\varphi_i(N,C)=\frac{v_0}{n}-\sum_{h\in M_i}\frac{d_h}{t_h}\quad (i=1,...,n) \tag{12}$$

where $t_h=|T_h|, T_h=\{i\in N\mid h\in M_i\}$.

Proof For any subset $T\subset N$, the Shapley value of the fixed cost game (N,e_T) as defined by (2) 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 (5), the value of an elementary data game (N,C_h) is then given by:

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

[16] There are alternative axiomatizations. They are reviewed by Moulin (2003). The nucleolus satisfies symmetry and dummy but not additivity.

Knowing that data games can be written as a sum of elementary data games, the value of the data game (N,C) is given by (12) as a consequence of additivity. ●

Hence the value of the complete data set is uniformly allocated among all players and the value 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 value 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.

7. The asymmetric Shapley value

The *weighted* Shapley value allows asymmetries between players to be taken into account.[17] 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:

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

where $w(T)$ is the weight of coalition *T*. The symmetric case corresponds to $w_i = 1.$ Using (5), the value of the elementary data game (N,C_h) associated with weights $(w_1,...,w_n)$ is given by:

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

[17] 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).

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 :

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

The following proposition is an immediate consequence of additivity.

Proposition 4 Consider the data sharing situation $(M,d) \in DS(N)$ and the associated cost function $C \in DG(N)$. Given *positive* weights $(w_1,...,w_n)$, the weighted Shapley value of the data game (N,C) associated to the data sharing situation (M,d) is given by:

$$\varphi_i(N,C,w) = \frac{w_i}{w(N)} v_0 - \sum_{h \in M_i} \frac{w_i}{w(T_h)} d_h \quad (i = 1,...,n) \tag{13}$$

In Example 1, the value associated with the weights (1, 1, 2) is given by the allocation (6.75,–2.25,–4.5) to be compared to the allocation (9,–1.5,–7.5) under equal weights.

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.

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.[18]

As a consequence, we have the following proposition.

Proposition 5 Zero weight players are compensated for a data they hold *if and only if* no player with positive weight 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 $i \in Z_w$.

[18] For more details, see Dehez (2011) where the weighted Shapley value is axiomatized along the lines proposed by Shapley (1981b).

The resulting value of an elementary game (N,C_h) is then unchanged for positive weight players while, for zero weight players, we get:

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

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

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

8. Accounting rules

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

$$\theta_i(N,C,\alpha) = MC_i + \alpha_i(C(N) - \sum_{j=1}^{n} MC_j) \quad (i = 1,...,n) \tag{14}$$

where the weights α belong to the unit simplex Δ_n and $MC_i = C(N) - C(N\backslash i)$ is the "separable cost" of player *i*. Weights may be exogenously given or may depend on the cost function.[19]

We shall restrict our attention to the case where weights are exogenous. θ is then an *additive* (actually linear) rule that does not satisfy the dummy axiom. If weights are equal, it satisfies the symmetry axiom and (14) defines the "equal charge" sharing rule.

Proposition 6 Consider the data sharing situation $(M,d) \in DS(N)$ and the associated cost function $C \in DG(N)$. If $\bar{M}_0 = \varnothing$, $\theta(N,C,\alpha) = 0$. If instead $\bar{M}_0 \neq \varnothing$, the weighted charge accounting rule (14) leads to the following allocation:

$$\theta_i(N,C,\alpha) = \alpha_i \, \bar{v}_0 - \bar{v}_i \quad (i = 1,...,n) \tag{15}$$

Proof This is an immediate consequence of the equations $MC_i = -C(N \setminus i) = -\bar{v}_i$ for all $i = 1,...,n$. ●

Corollary 2 Applied to data games, the weighted charge accounting rule (15) defines a core allocation *for any choice of weights*. Furthermore the *equal* charge rule defines an allocation that coincides with the nucleolus (10).

[19] Other accounting rules use endogenous weights like for instance the "*separable costs remaining benefits*" rule (SCRB). They are reviewed by Béal et al. (2010) and applied to REACH data sharing.

Actually, the core being the convex hull of its vertices (8), 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.

9. A particular case: partition data games

As explained in the introduction, the case where data sets form a partition of the complete data set M_0 is of particular interest: $M_i \cap M_j = \varnothing$ for all $i \neq j$ and each data can then be associated to a single player i.e. the T_h's are singletons. Furthermore, $\bar{M}_i = M_i$ and $\bar{v}_i = v_i$ for all $i = 0,1,...,n$. As a consequence, a "partition" data game (N,C) can simply be written as:

$$C(S)=v_0-\sum_{i\in S} v_i \ \text{ where } v_i=\sum_{h\in M_i} d_h.$$

The surplus game (N,V) associated to the partition data game (N,C) as defined by (1) is given by:

$$V(S)=(s-1)v_0 \ \text{ for all } S\neq\varnothing \qquad (16)$$

It is a symmetric game.

Example 2 Consider the data sharing situation involving 3 players and defined by the data sets $M_1 = \varnothing$, $M_2 = \{1\}$ and $M_3 = \{2,3\}$. Given the cost vector $d = (6, 9, 12)$, the data game (N,C) and associated surplus game (N,V) are defined by:

$C(1) = v_0 = 27$	$V(1) = V(2) = V(3) = 0$
$C(2) = C(12) = d_2 + d_3 = 21$	$V(12) = V(13) = V(23) = 27$
$C(3) = C(13) = d_1 = 6$	$V(123) = 54$
$C(23) = C(123) = 0$	

Data games are in general not concave, as shown by Example 1.

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

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

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

The surplus game defined by (16) being symmetric, *equal division* is the natural allocation under "equal treatment of equals". Each player then receives an equal share of the total surplus:

$$x_i = \frac{1}{n} v(N) = \frac{1}{n}(n-1)v_0$$

Using (6), the corresponding cost allocation is given by:

$$y_i = C(i) - x_i = C(i) - \frac{1}{n}(n-1)v_0 = \frac{v_0}{n} - v_i$$

Applied to Example 2, we get $y = (9,3,-12)$: player 3 is compensated by the other two players.

Proposition 7 The Shapley value and the nucleolus of a partition data game coincide with the equal division allocation.

Proof This is an immediate consequence of the fact that both the Shapley value and the nucleolus are sharing rules satisfying the symmetry axiom (equal treatment of equals). ●

Alternatively, we observe that in the partition case there are *n distinct* marginal cost vectors (μ_1,...,μ_n), each with multiplicity $(n-1)!$ where μ_i is the marginal cost vector corresponding to the permutations where player *i* is *first*. By concavity, the core is the ployhedron whose vertices are the marginal cost vectors i.e. $\mu^i = \pi^i$ where π^i is obtained by replacing v_i by $\overline{v}_i$ in (8). Proposition 7 then follows from the definition of the Shapley value as the average marginal cost vector.

Proposition 8 In the partition case, the weighted charge accounting rule (15) coincide with the weighted Shapley value (17), when weights are positive and satisfy $w_i = \alpha_i$ for all $i \in N$.

Proof Using (13), the asymmetric Shapley value of a partition data game is given by:

$$\varphi_i(N,C,w) = \frac{w_i}{w(N)} v_0 - v_i \quad (i = 1,...,n) \tag{17}$$

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

10. Concluding remarks

The question that comes up immediately concerns the choice of the sharing 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, the core restricts 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 some coalitions may 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 shows. Consider a situation where only two players hold data, say players *n* and *n*-1, and their data sets 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 d_1 – the value of the missing data – while all the other players including player $n-1$ may be asked to pay up to d_1. The nucleolus goes further by imposing that the $n-1$ first players pay the same amount, namely d_1 / n. This is to be compared with the allocation derived from the Shapley value. Using (12) we get:

$$y_i = \frac{v_0}{n} \quad (i = 1,...,n-2)$$

$$y_{n-1} = \frac{v_0}{n} - \sum_{h=2}^{m} \frac{d_h}{2} = -\frac{n-2}{2n} v_0 + \frac{d_1}{2}$$

$$y_n = \frac{v_0}{n} - \sum_{h=2}^{m} \frac{d_h}{2} - d_1 = -\frac{n-2}{2n} v_0 - \frac{d_1}{2}$$

It is definitely more acceptable: players without data pay the per capita value 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 value 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.[20] 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 value of the data he or she holds. One should however keep in mind that these parameters measure the present cost of *reproducing* the data and not the actual cost that has been sunk in the past.

We have assumed that the data needed were all held by some players. There is actually no loss of generality in doing so. Indeed, if this was not the case, the value of the data not previously held is a fixed cost: any coalition has to acquire these data and support the cost.

[20] 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.

That fixed cost would then be distributed uniformly or, possibly, according to some given weights.

Our analysis covers data sharing situations where data sets are *nested*. It has not considered explicitly because it hardly applies to actual data sharing situations. One possible illustration could be a situation where firms are running R&D programs that are at different stages of development. In the nested case, $M_i \subset M_{i-1}$ for $i = 2,...,n$ and $M_1 = M_0$. Defining $c_i = v_0 - v_i$ the associated data game is given by:

$$C(S) = Min_{i \in S}\ c_i$$

where the c_i's satisfy $c_i \geq c_{i-1}$ and $C(N) = c_1$. This is a kind of "reverse" airport game that has been applied to the provision of indivisible public goods by Dehez (2010), a context in which the nucleolus appears to be a more appropriate solution than the Shapley value.

Beyond the Shapley value and the nucleolus, it would be interesting to study other cooperative solution concepts like for instance the *tau-value* introduced by Tijs (1987). About the axiomatization of the Shapley value on the set of data sharing situations, we observe that no player can be a dummy in a data game except in the trivial situation where all players hold the complete data set. Béal et al. (2010) have proposed an alternative axiom that characterizes the Shapley value together with efficiency, symmetry and additivity (properly defined). An alternative and simple axiom could be the following: for all $(M,d) \in DS(N)$ such that $M_i = \varnothing$ or $M_i = M_0$ for all i,

$$M_i = \varnothing \text{ implies } \varphi_i(N,M,d) = \frac{v_0}{n}$$

It says that in situations where players either hold no data or hold the complete data set, players without data are asked to contribute the per capita cost of the complete data set. Interestingly, replacing v_0/n by $\overline{v}_0/n$ defines the nucleolus.

References

Béal, S., M. Deschamps, J.T. Ravix and O. Sautel (2010) "Les informations exigées par la législation REACH: analyse du partage des coûts" Mimeo, Université de Saint-Etienne.

Dehez, P. (2011) "Allocation of fixed costs: Characterization of the (dual) weighted Shapley value", revised CORE Discussion Paper 2009-35, to appear in *International Game Theory Review.*

Dehez, P. (2010) "Cooperative provision of public goods" CORE Discussion Paper 2010/26.

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

Driessen, T. (1985) "Properties of 1-convex n-person games" OR Spectrum 7, 19-26.

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, University of Princeton.

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

Grünbaum, B. (2003) Convex Polytopes, Springer Verlag: Berlin.

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.

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. (1995) Cooperative Microeconomics, Princeton University Press: Princeton.

Moulin, H. (2003) Fair Division and Collective Welfare, The MIT Press: Cambridge.

Muto, S., J. Potters and S. Tijs (1989) "Information market games" International Journal of Game Theory 18, 209-226.

Shapley, L.S. (1953) "A value for n-person games" in Contributions to the Theory of Games II by H. Kuhn and A.W. Tucker, Eds., Princeton University Press: Princeton, 307-317.

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

Shapley, L.S. (1981a) "Valuation of games" in Game Theory and its Applications by W.F. Lucas, Ed., Proceedings of Symposia in Applied Mathematics 24, American Mathematical Society: Providence - Rhode Island, 55-68.

Shapley, L.S. (1981b), Discussion comments on "Equity considerations in traditional full cost allocation practices: An axiomatic approach", in Joint Cost Allocation, by S. Moriarity, Ed., Proceeding of the University of Oklahoma Conference on Costs Allocations, Center for Economic and Management Research, University of Oklahoma, 131-136.

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

Tijs, S.H. (1987) "An axiomatization of the tau-value" Mathematical Social Sciences 13, 177-181.

Weber, R.J. (1988) "Probabilistic values for games", in The Shapley Value. Essays in Honor of Lloyd Shapley by A. Roth, Cambridge University Press: Cambridge, 101-119.

Young, H.P. (1985) Cost Allocation: Methods, Principles, Applications, North-Holland: Amsterdam.

Documents de travail du BETA

2011–**01** *La création de rentes : une approche par les compétences et capacités dynamiques*
Thierry BURGER-HELMCHEN, Laurence FRANK, janvier 2011.

2011–**02** *Le Crowdsourcing : Typologie et enjeux d'une externalisation vers la foule.*
Claude GUITTARD, Eric SCHENK, janvier 2011.

2011–**03** *Allocation of fixed costs : characterization of the (dual) weighted Shapley value*
Pierre DEHEZ, janvier 2011.

2011–**04** *Data games : sharing public goods with exclusion (2nd version)*
Pierre DEHEZ, Daniela TELLONE, janvier 2011.

La présente liste ne comprend que les Documents de Travail publiés à partir du 1er janvier 2011. La liste complète peut être donnée sur demande.
This list contains the Working Paper writen after January 2011, 1rst. The complet list is available upon request.