Implementación y diseño de mecanismos

Profesores examinados

Como ejemplo del problema de la implementación y del diseño de mecanismos, consideremos la siguiente historia, basada en un ejemplo que Eric Maskin (2007) presenta en http://nobelprize.org/nobel_prizes/economics/laureates/2007/maskin-slides.pdf (discurso de aceptación del Premio Nobel de Economía de 2007).

Un profesor puede ser de 4 tipos:

x = exigente,
y = despreocupado,
z = bromista y
v = motivador.

Los estudiantes del curso que imparte el profesor tienen preferencias sobre el tipo de profesor y, sobre la base de estas preferencias, evalúan al profesor en una encuesta docente. Para simplificar, supongamos que sólo hay dos estudiantes: él y ella (también podria asumirse que, para algún $k > 1$, hay $2k$ estudiantes que las preferencias de él y $2k$ estudiantes con las preferencias de ella).

Tanto él como ella pueden ser de dos tipos: del tipo interesado t_i por la asignatura o del tipo no interesado t_n. De las cuatro combinaciones posibles de estos dos tipos, supongamos que sólo dos son posibles (o tienen probabilidad positiva): $w = (t_i, t_i)$ y $w' = (t_n, t_n)$. Podemos identificar cada combinación posible de tipos con un estado del mundo: en el estado w, ambos estudiantes están interesados en la asignatura; en el estado w', ninguno está interesado. Las preferencias de los estudiantes en cada estado son las siguientes.

Estado w	Orden de preferencia de él: $x \to y \to z \to v$	De ella: $v \to y \to z \to x$
Estado w'	Orden de preferencia de él: $v \to x \to z \to y$	De ella: $y \to x \to z \to v$

En la encuesta docente, los estudiantes han de puntuar al profesor, de 1 a 4, sabiendo el tipo del profesor. Cada estudiante da 4 puntos al profesor si su tipo es el más preferido; 3 si su tipo es el segundo más preferido; 2 si es el tercero; y 1 si es el menos preferido (por tanto, si el tipo del profesor ocupa la posición k en el orden de preferencias del estudiante, éste le da $5 - k$ puntos).

La puntuación del profesor en cada estado es la suma de los puntos que recibe de él y de ella en ese estado. El objetivo (el resultado deseado) del profesor es escoger el tipo que, en cada estado, le dé la puntuación más alta posible. El profesor sabe cuáles son las preferencias de él y de ella en cada estado, pero ignora en qué estado está (el tipo de estudiante es información privada). Dado que el profesor sabe qué preferencias tendrían los estudiantes en cada estado, sabe que en el estado w el tipo de profesor que recibe máxima puntuación es y. En w, el tipo y obtendría 6 puntos; x i v obtendrían 5 cada uno; y z obtendría 4. En w', el tipo mejor puntuado sería x.

Por tanto, el profesor desea ser de tipo y (despreocupado) cuando el estado del mundo es w (la asignatura interesa a los estudiantes) y ser de tipo x (exigente) cuando el estado del mundo es w' (la asignatura no interesa a los estudiantes). Su deseo es, pues, implementar y en w i x en w'. El problema del profesor consiste en diseñar un mecanismo mediante el cual las decisiones de los estudiantes generen el resultado pretendido por el profesor: obtener máxima puntuación.

Un mecanismo se llama directo cuando, en el juego que induce el mecanismo, las estrategias de cada jugador consisten en revelar su tipo. En un mecanismo directo, el profesor preguntaría a un estudiante si le interesa la asignatura o no. Si la revelación del estudiante escogido es sincera, cuando el estado del mundo es w, el profesor lo sabrá y escogerá ser del tipo y; y cuando el estado es w', el profesor también lo sabrá y escogerá ser del tipo x. Y problema resuelto.

Por desgracia para el profesor, este mecanismo no incentiva a decir la verdad (técnicamente, no es compatible con los incentivos) cuando los estudiantes saben cuál es el propósito del profesor (ser y en w y ser x en w'). Para comprobar que ambos estudiantes tendrían incentivo a mentir, supongamos que el profesor decide preguntarle a él.

Si el estado es w, él sabe que afirmando "Me interesa la asignatura" el profesor asumirá que el estado es w y sabe que el profesor escogerá ser y. En cambio, él sabe que diciendo "No me interesa la asignatura", el profesor, asumiendo que el estado es w', escogerá ser x. Dado que, en el estado w, él prefiere que el profesor sea del tipo x a que sea del tipo y, él tiene incentivo a mentir en el estado w diciendo que están en el w'.

Si el estado es w', por el mismo razonamiento que en el caso anterior, él sabe que revelando la verdad (diciendo "No me interesa"), el profesor será del tipo x y que mintiendo (diciendo "Me interesa"), el profesor será del tipo y. Puesto que, en el estado w', él prefiere que el profesor sea del tipo x a que sea del tipo y, él no tiene, en el estado w', incentivo a mentir. Por tanto, en w' dirá que están en w'.

Pero entonces el profesor se enfrenta con una dificultad: sea cual sea el estado del mundo, a él le conviene siempre decir que el estado del mundo es w'. Conclusión: lo que diga él no es fiable. En ese caso, el profesor puede dirigirse a ella y preguntarle qué le parece la asignatura.

En vista de lo anterior, la opinión que exprese ella tampoco es fiable para el profesor. La conclusión final es que el mecanismo directo consistente en preguntar a los estudiantes no permite implementar (cuando los estudiantes escogen mejores respuestas) el resultado deseado por el profesor. La Fig. 2 muestra un mecanismo que sí que lo conseguiría cuando el concepto de solución escogido es el de equilibrio de Nash.

Fig. 1

él \ *ella*	c	d
a	y	z
b	v	x

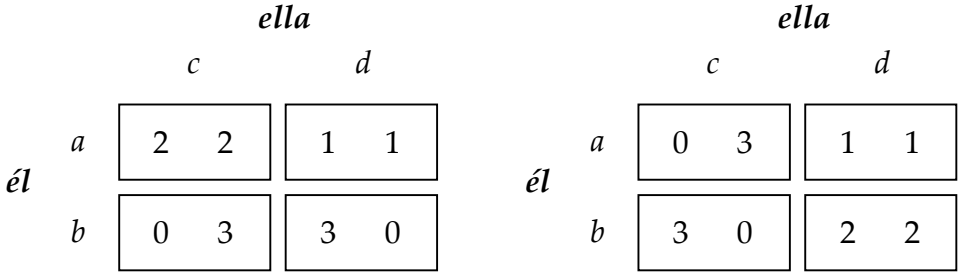
En este mecanismo, el profesor le dice a él que escoja entre a o b i, a ella, que escoja entre c i d. Lo que representen a, b, c i d es irrelevante: a y c pueden consistir en levantar la mano derecha y b y d en levantar la izquierda; o en escribir las letras a, b, c y d en un papel… lo que sea. Por tanto, $\{a, b\}$ es el conjunto de mensajes de él y $\{c, d\}$ el conjunto de mensajes de ella.

Aparte de indicar el conjunto de mensajes (o estrategias) de cada jugador, el mecanismo debe establecer qué resultado produce cada combinación de mensajes. El conjunto de combinaciones de mensajes viene dado por el producto cartesiano $\{a, b\} \times \{c, d\}$. Los elementos de este conjunto son (a, c), (a, d), (b, c) y (b, d). Cada uno de estos elementos puede asociarse con una casilla de la matriz de la Fig. 1. La función de resultados r completa la descripción del mecanismo asociando un resultado con cada elemento de $\{a, b\} \times \{c, d\}$. En este ejemplo, el conjunto de resultados es el conjunto $\{x, y, z, v\}$ de tipos del profesor. El mecanismo de la Fig. 1 es tal que $r(a, c) = y$, $r(a, d) = z$, $r(b, c) = v$ i $r(b, d) = x$. Por ejemplo, $r(a, c) = y$ significa que si él escoge a y ella escoge c entonces el profesor decide ser del tipo y.

El mecanismo de la Fig. 1 puede interpretarse del modo siguiente. Por un lado, el profesor le da a él el poder de determinar si el profesor escoge ser un tipo del conjunto $\{y, z\}$ o si escoge ser un tipo del conjunto $\{v, x\}$: escogiendo a, él fuerza al profesor a ser y o z; escogiendo b, él hace que el profesor se limite a ser v o x. Por otro lado, el profesor le da a ella el el poder de determinar si el profesor escoge ser un tipo del conjunto $\{y, v\}$ (lo que ella consigue seleccionando c) o si escoge ser un tipo del conjunto $\{z, x\}$ (lo que ella consigue seleccionando d).

El mecanismo de la Fig. 1 se transforma en un juego cuando añadimos las preferencias que tienen los estudiantes sobre los resultados. Para visualizar con más claridad el juego resultante, tomemos las siguientes representaciones numéricas de las preferencias de los estudiantes: la utilidad del resultado más preferido es 3, la del segundo más preferido 2, la del tercero 1 y la del menos preferido, 0 (la función de utilidad sería $u(\alpha) = 4 - k$ si el resultado α ocupa la posición k en el orden de preferencia representado). Por ejemplo, la funció de utilidad de él en el estado w sería $u(x) = 3$, $u(y) = 2$, $u(z) = 1$ y $u(v) = 0$.

La Fig. 2 muestra el juego entre los estudiantes en cada estado, en donde el primer número en los vectores de pagos representa la utilidad (o pago) de él y el segundo representa el de ella.



estado w Fig. 2 *estado w'*

El juego de la Fig. 2 es un juego bayesiano trivial porque cada estudiante sabe cuál es el tipo del otro estudiante. Ello permite resolver todo el juego resolviendo cada matriz por separado. Si el estado es w, ambos estudiantes lo saben y saben que el mecanismo del profesor induce el juego de la matriz izquierda en la Fig. 2. El único equilibrio de Nash de este juego (con estrategias pura o mixtas) és $[a, c]$, puesto que c es una estrategia fuertemente dominante para ella y, dada c, la mejor respuesta de él es a.

Si el estado es w', ambos estudiantes lo saben y saben que el mecanismo del profesor induce el juego de la matriz derecha en la Fig. 2. El único equilibrio de Nash de este juego (con estrategias pura o mixtas) és $[b, d]$, puesto que b es una estrategia fuertemente dominante para él y, dada b, la mejor respuesta de ella es d.

Como consecuencia, si el estado es w, los estudiantes juegan $[a, c]$, que produce el resultado y deseado por el profesor cuando el estado es w. Y si el estado es w', los estudiantes juegan $[b, d]$, que produce el resultado x deseado por el profesor cuando el estado es w'. Conclusión: el mecanismo de la Fig. 1 permite implementar (mediante equilibrios de Nash) el resultado deseado por el profesor: ser y en w y ser x en w'.

Funciones de elección social

Sea $N = \{1, \dots, n\}$ un conjunto de n individuos, sea A un conjunto de alternativas (o resultados) y, para cada individuo $i \in N$, sea L_i el conjunto de preferencias que se asume que i puede tener sobre los elementos del conjunto A. Las preferencias se asume que son ordenaciones lineales: todos los elementos de A pueden listarse en un ránking en el que ninguna alternativa es indiferente a otra. Sea $L = L_1 \times L_2 \times \dots \times L_n$ el conjunto de perfiles de preferencias, esto es, el conjunto de todas las maneras de asignar una preferencia a cada individuo.

Una función de elección social (FES) es una función $f : L \to A$ que asigna, a cada perfil de preferencias, una alternativa. Una FES representa un método de toma de decisiones colectivas: si los individuos tienen las preferencias del perfil $(P_1, P_2, \dots, P_n)$ entonces $f(P_1, P_2, \dots, P_n) \in A$ es la alternativa escogida.

La regla de Borda (propuesta por Jean-Charles de Borda en el último cuarto del siglo XVIII) permite construir una FES. Con A teniendo m elementos, la puntuación de la alternativa $a \in A$ en la preferencia P_i se define como m si a ocupa la primera posición en el ránking P_i, $m - 1$ si ocupa la segunda posición, $m - 2$ si ocupa la tercera… y 1 si ocupa la última posición. Por tanto, la puntuación de a en P_i es $m + 1$ menos la posición que a ocupa en P_i. La puntuación de a en el perfil de preferencias $(P_1, \dots, P_n)$ es la suma de la puntuación que a recibe en cada preferencia. Sea $(a_1, \dots, a_m)$ una ordenación lineal arbitraria de los m miembros del conjunto A de alternativas. La regla f que asigna a cada perfil de preferencias la alternativa con máxima puntuación que aparece antes en el orden $(a_1, \dots, a_m)$ es una FES. Con $N = \{1, 2, 3, 4\}$ y $A = \{a, b, c\}$ sea el siguiente perfil de preferencias (en donde se tiene, por ejemplo, $a\ P_1\ b\ P_1\ c$: 1 prefiere a a b i b a c).

		P_1	P_2	P_3	P_4
3 puntos	$\longrightarrow$	a	b	c	c
2 puntos	$\longrightarrow$	b	c	b	b
1 punto	$\longrightarrow$	c	a	a	a

La puntuación de a es $3 + 1 + 1 + 1 = 6$; la de b es $2 + 3 + 2 + 2 = 9$; y la de c es $1 + 2 + 3 + 3 = 9$. Tomando el ránking (a, c, b), la regla f sería tal que $f(P_1, P_2, P_3, P_4) = c$.

El problema de la implementación

El problema de implementar una función de elección social consiste en diseñar un mecanismo cuyos resultados, para cada perfil de preferencias, coincidan con el resultado escogido por la FES para ese perfil de preferencias. La interpretación es que la FES representa una toma de decisiones centralizada: todos los individuos revelan su información privada (sus preferencias sobre A) ante un coordinador y el coordinador aplica la FES para escoger un elemento de A. Sin embargo, la actuación del coordinador sólo puede llevarse a cabo mediante la colaboración de los individuos, ya que el coordinador ignora las preferencias de los individuos. Implementar la FES consiste en dar a los individuos la capacidad de informar al coordinador de manera que la información que los individuos revelan permita al coordinador tomar la decisión que resultaría de aplicar la FES en el caso en que se conocieran las preferencias de los individuos.

Mecanismo

Un mecanismo (o forma de juego) consiste en cuatro elementos. Primero, un conjunto N de individuos. Segundo, un conjunto A de alternativas (o resultados). Tercero, para cada individuo $i \in N$, un conjunto M_i de estrategias (o mensajes). Y cuarto, con $M = M_1 \times M_2 \times \ldots \times M_n$, una función de resultados $r : M \to A$ que especifica cuál es el resultado asociado con cada combinación de mensajes que escogen los individuos. Para abreviar, un mecanismo se identifica en ocasiones con el par (M, r), entendiendo que el conjunto N de individuos está implícito en la descripción de M y que el conjunto A de alternativas está implícito en la descripción de r.

Juego asociado con un mecanismo

Supongamos que los individuos del mecanismo tienen preferencias sobre el conjunto de resultados A del mecanismo y que, para cada individuo i, u_i es una función (de utilidad) sobre A que representa numéricamente sus preferencias: para todo a y b en A, $u_i(a) > u_i(b)$ si, y sólo si, i prefiere a a b. En ese caso, el mecanismo induce un juego simultáneo en el que: (i) el conjunto de jugadores es el mismo que el conjunto N de individuos del mecanismo; (ii) el conjunto de estrategias de cada jugador es su conjunto de mensajes en el mecanismo; y (iii) para cada jugador i, el pago $u_i(r(m))$ asociado con un combinación $m \in M = M_1 \times M_2 \times \ldots \times M_n$ de mensajes es la utilidad que la función de utilidad u_i atribuye a la alternativa $r(m)$, que es el resultado que se obtiene, según el mecanismo, cuando cada jugador $j \in N$ escoge el mensaje m_j.

Solución de equilibrio de un juego simultáneo

Un perfil (o vector) de estrategias de un juego simultáneo es una asignación de una estrategia a cada jugador. Si M_i es el conjunto de estrategias del jugador i en el juego y hay n jugadores, el conjunto de perfiles de estrategias es el producto cartesiano $M = M_1 \times M_2 \times \ldots \times M_n$. Una solución de equilibrio de un juego simultáneo es un conjunto de perfiles de estrategias del juego que: (i) constituyen un equilibrio de Nash; y (ii) posiblemente satisfacen alguna otra condición.

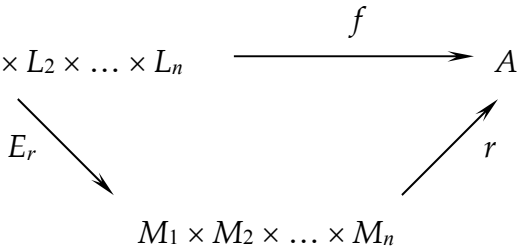
Por ejemplo, los equilibrios dominantes de Nash son aquellos equilibrios de Nash formados por estrategias débilmente dominantes (y, por tanto, son equilibrios en los que ninguna estrategia es débilmente dominada).

Implementación de una función de elección social

Sea (N, M, A, r) un mecanismo y $P = (P_1, \dots, P_n) \in L$ un perfil de preferencias sobre A. Sea $E_r(P)$ el conjunto de equilibrios de Nash seleccionados por la solución de equilibrio E en el juego asociado con el mecanismo (N, M, A, r) y el perfil de preferencias P.

Sea $f : L \to A$ una FES. El mecanismo (N, M, A, r) implementa la función de elección social f mediante la solución de equilibrio E si, para todo perfil de preferencias $P \in L$ y para todo $m \in E_r(P)$, $r(m) = f(P)$. Cuando existe un mecanismo (N, M, A, r) que implementa f mediante E se dice que f es implementable mediante la solución E [y el mecanismo (N, M, A, r)].

Fig. 3



La Fig. 3 describe en qué consiste la implementación. El punto de partida es la parte superior: la FES f que escoge una alternativa del conjunto A tomando como input las preferencias de los individuos. Implementar la FES f consiste en construir el camino inferior: asignar un conjunto de mensajes $M_1, M_2, \dots, M_n$ a cada individuo y definir una función de resultados r (esto es, construir un mecanismo) de manera que, para cada perfil de preferencias $P = (P_1, P_2, \dots, P_n)$, embutir cada perfil de mensajes $m \in M_1 \times M_2 \times \dots \times M_n$ seleccionado por la solución de equilibrio E_r (en el juego asociado con el mecanismo y el perfil de preferencias P) en la función de resultados produce la misma elección $r(m)$ que la elección $f(P)$ que genera la función de elección social f cuando las preferencias son P. Por tanto, para todo perfil de preferencias $P \in L_1 \times L_2 \times \dots \times L_n$, $f(P) = r[E_r(P)]$: la vía inferior (la vía descentralizada) replica el resultado de la vía superior (la vía centralizada).

¿Qué FES son implementables? El Teorema de Gibbard-Satterthwaite (uno de los teoremas fundamentales en teoría económica) establece que, en esencia, sólo las FES dictatoriales lo son.

Función de elección social dictatorial

Una función de elección social $f : L \to A$ es dictatorial si existe un individuo $i \in N$ tal que, para todo perfil de preferencias $P \in L$, $f(P)$ es el resultado más preferido por i en la preferencia P_i.

Una FES dictatorial es consistente con la interpretación de que un individuo (siempre el mismo) determina el resultado: la FES siempre escoge la alternativa más preferida por ese individuo (denominado "dictador").

Función de elección social Paretoeficiente

Una función de elección social $f : L \to A$ es Paretoeficiente si, para todo perfil de preferencias $P \in L$ y todo par de alternativas $a \in A$ y $b \in A \setminus \{a\}$, si se tiene que, para todo $i \in N$, $a \, P_i \, b$ entonces $f(P) \neq b$.

Una FES es Paretoeficiente si no escoge una alternativa b que todos los individuos consideran menos preferida que otra alternativa a. De hecho, si todos prefieren a a b y se escogiera b, todos mejorarían pasando a escoger a en lugar de b.

Función de elección social no manipulable

Una función de elección social $f : L \to A$ es no manipulable (*strategy-proof*) si no existen perfil de preferencias P, individuo i y preferencia Q_i del individuo tal que $f(Q_i, P_{-i}) \, P_i \, f(P_i, P_{-i})$.

La no manipulabilidad de una FES expresa la siguiente idea. Supongamos que las preferencias auténticas de los individuos vienen dadas por el perfil de preferencias P. Entonces, para que una FES f sea no manipulable, no puede existir ningún individuo i y ninguna preferencia falsa Q_i tal que la alternativa $f(Q_i, P_{-i})$ que la FES selecciona cuando i miente es preferida por i (según su auténtica preferencia P_i) a la alternativa $f(P_i, P_{-i})$ que la FES escoge cuando i revela la verdad. La no manipulabilidad significa que ningún individuo tiene nunca incentivo a mentir. Por tanto, la no manipulabilidad hace que la revelación honesta sea una estrategia dominante. De hecho, que una FES sea no manipulable es equivalente a que sea implementable honestamente mediante equilibrios dominantes (basta con considerar el mecanismo directo (N, A, L, f) en el que la función de resultados es la propia FES).

Teorema de Gibbard (1973) - Satterthwaite (1975) (TGS)

Sea $f : L \to A$ una FES en donde A tiene al menos tres elementos y en donde L contiene todos los perfiles de preferencias posibles. Entonces f es Paretoeficiente y no manipulable si, y sólo si, f es dictatorial.

El TGS nos dice que no hay mucho que sea implementable mediante equilibrios dominantes cuando se exige Paretoeficiencia, que todas las preferencias sean posibles y tener al menos tres alternativas entre las que elegir: sólo las reglas de elección dictatoriales lo son.

El Teorema de Gibbard-Satterthwaite para el caso de 2 individuos y 3 alternativas

Éste es el caso más sencillo de validez del TGS. Ilustremos la prueba con el siguiente ejemplo. Un profesor da a los estudiantes de su curso la posibilidad de escoger el sistema de evaluación de entre un conjunto de tres alternativas, a, b y c. Los estudiantes se organizan escogiendo un representante (R1) entre los repetidores del curso y escogiendo otro representante (R2) entre los no repetidores, de modo que el sistema propuesto al profesor dependa de las preferencias de estos dos representantes. Sea $\alpha\beta\gamma$ la manera de expresar la preferencia del individuo i tal que $\alpha \, P_i \, \beta \, P_i \, \gamma$ (α es la alternativa más preferida, β la segunda más preferida y γ la menos).

Supongamos que cada representante puede adoptar cualquier orden lineal sobre $A = \{a, b, c\}$ como preferencia. Esto hace que cada representante tenga una de las 6 preferencias *abc*, *acb*, *bac*, *bca*, *cab* y *cba*. La combinación de estas 6 preferencias produce el conjunto L de 36 perfiles de preferencias, en donde la preferencia representada por la primera de las dos columnas en cada casilla es la del represetante R1. Este conjunto se representa en la Fig. 4, en donde cada casilla se corresponde con un perfil de preferencias.

Fig. 4

R1 R2	f	R1 R2	f	R1 R2	f	R1 R2	f	R1 R2	f	R1 R2	f
a a *b b* *c c*	→ *a*	*a a* *c b* *b c*	→ *a*	*b a* *a b* *c c*	→	*b a* *c b* *a c*	→	*c a* *a b* *b c*	→	*c a* *b b* *a c*	→
a a *b c* *c b*	→ *a*	*a a* *c c* *b b*	→ *a*	*b a* *a c* *c b*	→	*b a* *c c* *a b*	→	*c a* *a c* *b b*	→	*c a* *b c* *a b*	→
a b *b a* *c c*	→	*a b* *c a* *b c*	→	*b b* *a a* *c c*	→ *b*	*b b* *c a* *a c*	→ *b*	*c b* *a a* *b c*	→	*c b* *b a* *a c*	→
a b *b c* *c a*	→	*a b* *c c* *b a*	→	*b b* *a c* *c a*	→ *b*	*b b* *c c* *a a*	→ *b*	*c b* *a c* *b a*	→	*c b* *b c* *a a*	→
a c *b a* *c b*	→	*a c* *c a* *b b*	→	*b c* *a a* *c b*	→	*b c* *c a* *a b*	→	*c c* *a a* *b b*	→ *c*	*c c* *b a* *a b*	→ *c*
a c *b b* *c a*	→	*a c* *c b* *b a*	→	*b c* *a b* *c a*	→	*b c* *c b* *a a*	→	*c c* *a b* *b a*	→ *c*	*c c* *b b* *a a*	→ *c*

La flecha "→" apunta a la alternativa que la FES selecciona cuando el perfil de preferencias es el indicado en la casilla. La Fig. 4 indica implicaciones inmediatas del hecho de suponer que la FES es Paretoeficiente: en todos aquellos perfiles en que ambos representantes están de acuerdo en que una alternativa dada es la más preferida, ésa debe ser la alternativa escogida por la FES. Toda FES Paretoeficiente debe asignar los valores indicados en la Fig. 4. Con ello, la Paretoeficiencia reduce el problema de asignar valores a 36 casillas a uno de asignarlos a 24.

Supongamos que los representantes no sólo desean recurrir a una FES f Paretoeficiente que realice una elección para cada uno de los 36 perfiles de preferencia posibles, sino que también desean que la FES sea no manipulable, esto es, que ningún representante obtenga un sistema de evaluación más preferido mintiendo sobre su preferencia que revelando la preferencia real. Por el TGS sólo hay dos FES que cumplen esos requisitos: la FES f_1 que siempre escoge la alternativa más preferida por R1 o la FES f_2 que siempre escoge la alternativa más preferida por R2. Comprobémoslo. Sea f una FES Paretoeficiente y no manipulable.

Consideremos primero la casilla remarcada en la Fig. 4. Esta casilla representa el perfil de preferencias (*abc*, *bac*). Por Paretoeficiencia, no puede ser que $f(abc, bac) = c$, puesto que ambos

representantes prefieren *a* (o *b*) a *c*. Por tanto, *f* sólo puede asignar *a* o *b* a esta casilla. Supongamos que es *a* (comprueba qué pasaría si fuera *b*). Esta elección se indica en la Fig. 5.

Fig. 5

a b *b a* → *a* *c c*	*a b* *c a* → *b c*	*b b* *a a* → *b* *c c*	*b b* *c a* → *b* *a c*	*c b* *a a* → *b c*	*c b* *b a* → *a c*

Pasemos ahora a la casilla remarcada en la Fig. 5. Cuando nos encontramos en esta casilla, la presunción es que las preferencias auténticas son las de la casilla: la preferencia de R1 es *acb* y la de R2 es *bac*. Como en todas las casillas, hay sólo tres posibilidades: *f* selecciona *a*, *b* o *c*. Supongamos que selecciona *c*. Esto es, *f*(*acb*, *bac*) = *c*. Entonces R1 podría manipular *f*, diciendo que su preferencia no es *acb* sino *abc*, ya que *f*(*abc*, *bac*) = *a* (como acaba de asumirse) y R1 (según la preferencia auténtica *acb* asumida en la casilla remarcada de la Fig. 5) tiene *acb* como preferencia auténtica. Así pues, diciendo que su preferencia es *abc* en lugar de *acb*, R1 consigue que la regla pase de elegir *c* = *f*(*acb*, *bac*) a elegir *a* = *f*(*abc*, *bac*). Dado que, según su preferencia auténtica *acb*, R1 prefiere *a* a *c*, *f* sería manipulable, lo que contradice la hipótesis de que no lo es. El mismo razonamiento demuestra que *f*(*acb*, *bac*) no puede ser *b*. Conclusión: *f*(*acb*, *bac*) = *a*. Este nuevo valor descubierto de la FES *f* se indica en la Fig. 6.

Fig. 6

a b *b a* → *a* *c c*	*a b* *c a* → *a* *b c*	*b b* *a a* → *b* *c c*	*b b* *c a* → *b* *a c*	*c b* *a a* → *b c*	*c b* *b a* → *a c*
a b *b c* → *c a*	*a b* *c c* → *b a*	*b b* *a c* → *b* *c a*	*b b* *c c* → *b* *a a*	*c b* *a c* → *b a*	*c b* *b c* → *a a*

Consideremos ahora la casilla remarcada en la Fig. 6. Por Paretoeficiencia, no puede escogerse *c*. Así que hay dos posibilidades: *f*(*abc*, *bca*) = *a* o *f*(*abc*, *bca*) = *b*. Asumamos la segunda: *f*(*abc*, *bca*) = *b*. Situémonos en la casilla con preferencias (*abc*, *bac*), que es la casilla justo encima de la remarcada en la Fig. 6. Por la hipótesis inicial, *f*(*abc*, *bac*) = *a*, tal y como indica la Fig. 6. Si ahora R2 anunciara la preferencia *bca* en lugar de la que se presume auténtica en esa casilla (la preferencia *bac*), la FES escogería *b*, ya que se ha asumido que *f*(*abc*, *bca*) = *b*. Por tanto, R2 podría manipular la FES si las preferencias auténticas fueran (*abc*, *bac*): revelando *bac*, resulta *a*; revelando en su lugar *bca*, resulta *b*, que es preferida por R2 a *a*. Puesto que *f* es no manipulable, no puede ser que *f*(*abc*, *bca*) = *b*. Como resultado, *f*(*abc*, *bca*) = *a*. Esto se indica en la Fig. 7.

Fig. 7

a b *b c* → *a* *c a*	*a b* *c c* → *b a*	*b b* *a c* → *b* *c a*	*b b* *c c* → *b* *a a*	*c b* *a c* → *b a*	*c b* *b c* → *a a*

Como en el caso de la Fig. 5, el valor de la función *f*(*acb*, *bca*) en la casilla remarcada ha de ser *a*. Si no fuera así, R1 podría declarar la preferencia *abc* en lugar de la presumida auténtica *acb* y pasar de obtener *f*(*acb*, *bca*) ≠ *a* a obtener *f*(*abc*, *bca*) = *a*, lo que permitiría a R1 conseguir su opción

más preferida mintiendo. Dado que esto violaría la no manipulabilidad, ha de tenerse $f(acb, bca) = a$.

Fig. 8

$a\ c$ $b\ a \rightarrow a$ $c\ b$	$a\ c$ $c\ a \rightarrow a$ $b\ b$	$b\ c$ $a\ a \rightarrow$ $c\ b$	$b\ c$ $c\ a \rightarrow$ $a\ b$	$c\ c$ $a\ a \rightarrow c$ $b\ b$	$c\ c$ $b\ a \rightarrow c$ $a\ b$
$a\ c$ $b\ b \rightarrow a$ $c\ a$	$a\ c$ $c\ b \rightarrow a$ $b\ a$	$b\ c$ $a\ b \rightarrow$ $c\ a$	$b\ c$ $c\ b \rightarrow$ $a\ a$	$c\ c$ $a\ b \rightarrow c$ $b\ a$	$c\ c$ $b\ b \rightarrow c$ $a\ a$

Con un razonamiento análogo se demuestra que, para los 12 perfiles de preferencias de las dos columnas de la izquierda en la Fig. 4, la FES escoge precisamente la alternativa más preferida por R1: *a*. Se trata de comprobar que lo mismo pasa en las dos columnas centrales (en las que la FES escogerá *b*) y en las dos de la derecha (donde escogerá *c*).

Comenzando con las dos columnas centrales, tomemos la casilla remarcada en la Fig. 8. Por Paretoeficiencia, no puede seleccionarse *a*. Así que $f(bac, cba) \in \{b, c\}$. Si $f(bac, cba) = c$, entonces R1 podría declarar, en lugar de la preferencia auténtica *bac*, la preferencia *acb*. En tal caso, tal y como indica la Fig. 8, se obtendría $f(acb, cba) = a$, que es una alternativa preferida por R1 a *c* cuando la preferencia auténtica de R1 es la de la casilla remarcada en la Fig. 8 (preferencia *bac*). Por ello, *f* sería manipulable, contradiciendo la hipótesis de que no lo es. Así que $f(bac, cba) = b$.

El TGS no es necesariamente cierto si la FES se define en un dominio restringido (cuando no todas las preferencias son posibles). Un ejemplo de implementación, mediante equilibrios dominantes, de FES que no son dictatoriales es el mecanismo de Groves-Clarke, en el que las preferencias admisibles son las representables mediante funciones de utilidad cuasi-lineales.

Referencias

Gibbard, Allan (1973): "Manipulation of voting schemes: A general result", *Econometrica* 41, 587–601.

Myerson, Roger B. (2007): "Perspectives on mechanism design in economic theory", Nobel Lecture, http://nobelprize.org/nobel_prizes/economics/laureates/2007/myerson_lecture.pdf.

Satterthwaite, Mark (1975): "Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions", *Journal of Economic Theory* 10, 187–217.

El mecanisme de Groves-Clarke

Celebracions

Els estudiants de Política Industrial han aprovat tots l'assignatura i es plantegen fer una celebració. L'opció a és una microfesta on només hi participin els estudiants. L'opció b és muntar una macrofesta on hi pugui assistir tothom que ho vulgui. Per a cada estudiant i, la utilitat (neta) de l'opció $c \in \{a, b\}$ és $u_i(c) = v_i(c) - c_i(c)$, on $v_i(c)$ representa el benefici que c proporciona a i i $c_i(c)$ representa el cost de finançar l'opció c que ha d'assumir l'estudiant i.

Els estudiants adopten la següent regla de decisió (on el sumatori comprèn tots els estudiants): si $\sum_i u_i(a) > \sum_i u_i(b)$ aleshores es tria l'opció a; en cas contrari, es tria l'opció b. Suposem que l'objectiu sigui implementar aquesta regla: que quan $\sum_i u_i(a) > \sum_i u_i(b)$ es triï a i que quan $\sum_i u_i(a) \leq \sum_i u_i(b)$ es triï b. L'inconvenient és que cada u_i és informació privada: només i sap quina és la seva funció v_i (la funció c_i se suposa ja determinada pel col·lectiu d'estudiants).

L'inconvenient es resol dissenyant un mecanisme (directe) que indueixi els estudiants a revelar la utilitat real que li proporciona cada opció. Per a eliminar tota consideració estratègica a l'hora de revelar utilitats, es proposa que la implementació de la regla sigui mitjançant equilibris dominants. Això és, que revelar l'autèntica utilitat (dir la veritat) sigui sempre (revelin el que revelin els altres) una millor resposta per a cada estudiant. El mecanisme de Groves-Clarke (atribuït a Theodore Groves i Edward H. Clarke) ofereix una solució a aquest problema, ja que és un mecanisme que incentiva a tot estudiant a revelar la utilitat que assigna a cada opció.

El mecanisme de Groves (1973) - Clarke (1971) (MGC)

L'MGC s'entén aplicat per un agent coordinador (que podria ser un dels estudiants) que segueix mecànicament i fidel les 3 etapes en què s'organitza el mecanisme.

• Etapa 1: revelació. Cada estudiant i declara al coordinador els valors d'utilitat $\hat{u}_i(a)$ i $\hat{u}_i(b)$ que i atribueix a cada opció (atès que els valors $c_i(a)$ i $c_i(b)$ s'entenen coneguts per tothom, revelar els valors relatius a $u_i(a)$ i $u_i(b)$ equival a revelar els valors relatius a $v_i(a)$ i $v_i(b)$). Els valors $\hat{u}_i(a)$ i $\hat{u}_i(b)$ no tenen perquè coincidir amb els valors autèntics $u_i(a)$ i $u_i(b)$: cada estudiant decideix lliurement quins valors declarar.

• Etapa 2: decisió. El coordinador determina les sumes dels valors revelats per a cada opció. Si $\sum_i \hat{u}_i(a) > \sum_i \hat{u}_i(b)$, el coordinador declara que l'opció a seguir és a; si $\sum_i \hat{u}_i(a) \leq \sum_i \hat{u}_i(b)$, declara que és b.

• Etapa 3: transferències. A banda dels pagaments $c_i(a)$ i $c_i(b)$ que cada estudiant i hauria de fer per a costejar cada opció, el coordinador dicta que cada estudiant i ha pagar addicionalment l'import T_i calculat de la següent manera. Sigui i un estudiant, sigui c l'opció que se selecciona a l'etapa 2 i sigui d l'opció que es triaria a l'etapa 2 si i no participés en el mecanisme (si i no hi participés, el valors $\sum_{j \neq i} \hat{u}_j(a)$ i $\sum_{j \neq i} \hat{u}_j(b)$ determinarien l'opció a seguir).

(i) Si $c = d$ aleshores $T_i = 0$.

(ii) Si $c \neq d$ aleshores $T_i = \sum_{j \neq i} \hat{u}_j(d) - \sum_{j \neq i} \hat{u}_j(c)$.

L'etapa 3 és la clau de l'MGC perquè elimina els incentius a revelar valoracions falses de les opcions. La condició (i) diu que l'estudiant i no ha de fer cap contribució addicional si la seva participació no altera el resultat que s'hauria produït sense la seva participació. Per exemple, suposem que a es tria a l'etapa 2. Per a què a també es triï a l'etapa 2 sense la participació d'i cal que $\sum_{j \neq i} \hat{u}_j(a) > \sum_{j \neq i} \hat{u}_j(b)$. Així doncs, si $\sum_{j \neq i} \hat{u}_j(a) > \sum_{j \neq i} \hat{u}_j(b)$ i a és l'opció triada aleshores i no ha de pagar més del valor $c_i(a)$ ja establert. La raó és que la valoració que faci i d'a o de b no afecta la decisió presa a l'etapa 2: amb ell, es tria a; sense ell, es triaria també a. El principi que justifica (i) és que si i no altera la decisió amb les seves valoracions, ja està bé amb el inicialment s'havia determinat que havia de pagar.

La condició (ii) estableix que i ha de pagar més de l'inicialment acordat $c_i(c)$ només en cas que les valoracions comunicades per i al coordinador modifiquin l'opció que s'escolliria a l'etapa 2 si i no participés. Quan això passa, i ha de pagar la pèrdua d'utilitat (el cost) que la seva participació causa als altres.

Per exemple, suposem que a se selecciona a l'etapa 2. Si la participació d'i n'altera el resultat, llavors s'ha de tenir que, sense i, es triaria b. Per tant, $\sum_{j \neq i} \hat{u}_j(a) \leq \sum_{j \neq i} \hat{u}_j(b)$: si i no participés, la regla de decisió de l'etapa 2 dictaria que, amb $\sum_{j \neq i} \hat{u}_j(a) \leq \sum_{j \neq i} \hat{u}_j(b)$, b fos l'opció escollida. Per a què les valoracions d'i modifiquin aquest resultat, cal que $\sum_{j \neq i} \hat{u}_j(a) + \hat{u}_i(a) > \sum_{j \neq i} \hat{u}_j(b) + \hat{u}_i(b)$. En conseqüència, cal que $\hat{u}_i(a) > \hat{u}_i(b)$. Així que, quan a se selecciona a l'etapa 2, l'únic cas en què (ii) s'aplica té lloc quan $\sum_{j \neq i} \hat{u}_j(a) \leq \sum_{j \neq i} \hat{u}_j(b)$ i $\hat{u}_i(a) > \hat{u}_i(b)$. Quan aquest és el cas, l'estudiant i ha de pagar $T_i = \sum_{j \neq i} \hat{u}_j(b) - \sum_{j \neq i} \hat{u}_j(a)$. Aquesta diferència és el cost que representa als altres estudiants passar de prendre l'opció b a prendre l'opció a. Sense i, s'hauria pres l'opció b, la qual cosa suposa que $\sum_{j \neq i} \hat{u}_j(b) \geq \sum_{j \neq i} \hat{u}_j(a)$. Amb i, s'hauria pres l'opció a i, atès que $\sum_{j \neq i} \hat{u}_j(b) \geq \sum_{j \neq i} \hat{u}_j(a)$, el canvi de decisió causa un perjudici a la resta d'estudiants igual a $\sum_{j \neq i} \hat{u}_j(b) - \sum_{j \neq i} \hat{u}_j(a) \geq 0$. L'etapa 3 diu que si l'individu i és decisiu (la seva intervenció altera el resultat) llavors i ha de pagar pel perjudici que la seva participació crea en els altres. Atès que el perjudici seria la diferència, $\sum_{j \neq i} \hat{u}_j(b) - \sum_{j \neq i} \hat{u}_j(a)$ és aquest mateix import el que i ha de pagar en forma de transferència (o impost) $T_i = \sum_{j \neq i} \hat{u}_j(b) - \sum_{j \neq i} \hat{u}_j(a)$.

Un exemple

Tres individus (1, 2 i 3) han de decidir entre a i b. La Fig. 1 mostra els valors $u_i(a)$ i $u_i(b)$ i el pagament T_i que, segons el mecanisme, cada individu i ha d'assumir.

Fig. 1

i	1	2	3
$u_i(a)$	2	4	6
$u_i(b)$	9	1	3
T_i	6	0	0

Fig. 2

i	1	2	3
$u_i(a)$	2	6	6
$u_i(b)$	9	1	3
T_i	–	–	–

La utilitat total d'a és $u_1(a) + u_2(a) + u_3(a) = 2 + 4 + 6 = 12$. La utilitat total de b és $u_1(b) + u_2(b) + u_3(b) = 9 + 1 + 3 = 13$. Aplicant la regla de triar l'opció amb més utilitat total, l'opció escollida seria b.

Aquesta regla és manipulable. Per exemple, si 2 canviés $u_2(a) = 4$ per $\hat{u}_2(a) = 6$ (tal com es reflecteix a la Fig. 2), l'opció seleccionada seria a. El canvi d'elecció beneficiaria a 2: abans, amb la selecció de b, la seva utilitat era $u_2(b) = 1$; ara, amb la selecció d'a, la seva utilitat seria $u_2(a) = 4$. Així, 2 té incentiu a mentir si la regla de decisió es basa en comparar utilitats totals revelades. L'MGC, aplicat a les utilitats de la Fig. 1, faria que l'opció escollida fos b amb l'afegit que 1 hauria de pagar $T_1 = 6$. Aquest seria el resultat si tothom, a l'etapa 1, declarés les seves valoracions reals. Comprovem que ningú no té incentiu a revelar una valoració diferent de la real quan els altres també revelen les valoracions reals.

• Individu 1. La utilitat neta d'1 quan revela honestament és $u_1(b) - T_1 = 9 - 6 = 3$. No hi ha manera d'1 de reduir el pagament de $T_1 = 6$ quan b és l'opció triada, perquè T_1 depèn de les utilitats revelades pels altres individus (T_1 és la utilitat total que perden els altres individus quan 1 participa al mecanisme: $u_2(a) + u_3(a) - u_2(b) - u_2(b) = 4 + 6 - 1 - 3 = 6$). L'única alternativa que 1 pot plantejar-se és declarar valors $\hat{u}_1(a)$ i $\hat{u}_1(b)$ que alterin l'opció escollida pel mecanisme. Si 1 força el canvi d'opció (de b a a), T_1 seria 0 i la utilitat neta d'1 seria $u_1(a) = 2$. Per tant, 1 obté més utilitat neta quan (declarant els altres la veritat sobre les seves valoracions) ell mateix declara honestament que quan falseja la seva declaració i força un canvi en l'opció escollida.

• Individu 2. La utilitat neta de 2 quan revela honestament és $u_2(b) = 1$. Atès que, sense 2, b encara seria l'opció escollida, $T_2 = 0$. L'incentiu per a què 2 reveli informació falsa sobre les seves valoracions només pot provenir de la possibilitat que, forçant un canvi en l'opció que tria el mecanisme, la utilitat neta de 2 augmentés. Si 2 força el canvi de b a a, $T_2 = (9 + 3) - (2 + 6) = 4$, que és la pèrdua d'utilitat total que provocaria el canvi de b a a forçat per 2. Així, la utilitat neta de 2 quan menteix (i provoca que a s'esculli en comptes de b) seria $u_2(a) - T_2 = 4 - 4 = 0$. Conclusió: 2 no millora la seva utilitat neta i, en conseqüència, no té incentiu a mentir quan els altres no ho fan. Per a l'individu 3 l'anàlisi dels incentius és anàloga a l'anàlisi del 2.

Què es fa amb els pagaments addicionals T_i del mecanisme?

En general, l'MGC genera uns pagaments addicionals $\sum_i T_i$. Què es fa amb aquest superàvit? Es podria pensar que no hi hauria cap problema per a distribuir el superàvit $\sum_i T_i$ entre els individus. Malauradament, la distribució de l'excedent $\sum_i T_i$ altera els incentius a dir la veritat. L'exemple de la Fig. 1 il·lustra el problema. Suposem que tothom sap que l'excedent $\sum_i T_i$ es reparteix igualitàriament entre els individus. Aleshores 2 augmentaria la seva utilitat neta declarant $\hat{u}_2(b) = ½$ en comptes d'$u_2(b) = 1$. Declarant $u_2(b) = 1$, la utilitat neta de 2 seria $u_2(b) + T_1/3 = 1 + 6/3 = 1 + 2 = 3$. Declarant $\hat{u}_2(b) = ½$, b és encara l'opció escollida però ara es tindria $T_1' = 13/2$, de manera que la nova utilitat neta de 2 seria superior: $u_2(b) + T_1'/3 = 1 + 13/6 > 3$.

Més inconvenients del mecanisme de Groves-Clarke

La impossibilitat general de repartir l'excedent entre els individus provoca que el resultat de l'MGC no sigui Paretoeficient. Això condueix al següent dilema: per a què el mecanisme no sigui manipulable (i, per tant, ningú no tingui incentiu a mentir), els excedents en general no es podran distribuir; però si aquests excedents no es distribueixen, el resultat del mecanisme no

serà Paretoeficient perquè, donada l'elecció feta pel mecanisme, tothom estaria millor amb una part de l'excedent que genera el mecanisme.

L'MGC no necessàriament satisfà la restricció de participació, que diu que participar en el mecanisme no pot produir un resultat pitjor per a algun individu que no participar. Per exemple, en el cas de la Fig. 1, suposem que, sense el mecanisme, la decisió presa seria a. La utilitat de l'individu 2 seria $u_2(a) = 4$. Si el mecanisme s'aplica, la decisió presa seria b i la utilitat neta de 2 seria $u_2(b) - T_2 = 1 - 0 = 1$. Conclusió: 2 estaria millor si el mecanisme no s'apliqués.

A més a més, el mecanisme de Groves-Clarke no és immune a manipulació per part de coalicions. Per exemple, en la situació representada per la Fig. 1, suposem que els individus 2 i 3 declaren $\hat{u}_2(a) = 7$ en comptes d'$u_2(a) = 4$ i $\hat{u}_3(a) = 9$ en comptes d'$u_3(a) = 6$ (la resta de valors declarats són els reals). En aquest cas, a és l'opció seleccionada, amb $T_1 = 0$ i $T_2 = T_3 = 1$. Dient la veritat, b és l'opció seleccionada, la utilitat neta de 2 és $u_2(b) = 1$ i la utilitat neta de 3 és $u_3(b) = 3$. Declarant els valors falsos $\hat{u}_2(a) = 7$ i $\hat{u}_3(a) = 9$, la utilitat neta de 2 és $u_2(a) - T_2 = 4 - 1 = 3$ i la utilitat neta de 3 és $u_3(a) - T_3 = 6 - 1 = 5$. Així doncs, 2 i 3 augmenten la seva utilitat neta revelant, conjuntament, utilitats falses.

Els inconvenients previs del mecanisme de Groves-Clarke provenen d'un teorema d'impossibilitat de Leonid Hurwicz[1] (1972), que demostrà el següent: en una economia d'intercanvi estàndar, no hi ha cap mecanisme que sigui compatible amb els incentius (no manipulable), que satisfaci la restricció de participació i que generi resultats Paretoeficients. Una interpretació d'aquest teorema és que la informació privada exclou la plena eficiència.

Referencias

Clarke, Edward H. (1971): "Multipart pricing of public goods", *Public Choice* 11, 17–33.

Groves, Theodore (1973): "Incentives in teams", *Econometrica* 41, 617–631.

Hurwicz, Leonid (1972): "On informationally decentralized systems", a McGuire, C. B. i Radner, Roy (ed.): *Decision and Organization: A Volume in Honor of Jacob Marshak*. North-Holland: Amstedam, p. 297–336.

[1] Leonid Hurwicz (1917–2008) va ser un dels tres premis Nobel d'Economia del 2007. Amb 90 anys, és la persona amb més edat que ha rebut un premi Nobel (http://en.wikipedia.org/wiki/Hurwicz). Hurwicz va ser el creador de la teoria del disseny de mecanismes.