Implementación y diseño de mecanismos

Una de niños cuarteados

En el capítulo 3 del Libro de los Reyes del Antiguo Testamento se relata el conocido como "Juicio del Rey Salomón". Dos prostitutas, *p* y *q*, se presentan ante el rey. Explica *p* que ambas comparten una casa, en compañía de nadie más, y que ambas han dado recientemente a luz. Continúa explicando *p* que el hijo de *q* ha fallecido y que, mientras dormía, *q* le ha arrebatado a su hijo. Por ello se presenta ante el rey para pedirle la devolución de su hijo. Niega *q* los cargos, y asegura que el bebé fallecido es de la acusadora *p*. Entonces el rey ordena que el bebé sea partido por la mitad con una espada y cada parte entregada a una madre. En ese momento, la madre del niño implora al rey que, antes que muerto, prefiere ver a su bebé en manos de la otra madre. Ésta, por el contrario, acepta la partición del bebé. La decisión final del rey es entregar el bebé a la madre que aceptó ceder el niño, aparentemente basándose en la presunción de que la auténtica madre preferiría que su hijo viviera aunque fuera con otra madre.
http://www.newadvent.org/bible/1ki003.htm http://en.wikipedia.org/wiki/King_Solomon

Esta historia permite ilustrar en qué consiste el diseño de mecanismos: en construir un juego la solución del cual conduzca a un resultado previamente seleccionado y pretendido. Este resultado seleccionado es el resultado a implementar (obtener) mediante el juego. En el juicio, el resultado que pretende obtener el rey es entregar el bebé a la auténtica madre. El problema es que el rey no dispone de la información suficiente para tomar esta decisión: es información privada quién es la madre auténtica.

Un mecanismo (también llamado "forma de juego" o *game form*) consiste en la especificación de estrategias para cada jugador (los mensajes mediante los que éstos pueden transmitir su información privada) junto con una regla que determine qué resultados produce cada posible combinación de estrategias de los jugadores. Un mecanismo no es un juego. Pero los jugadores se suponen dotados de preferencias sobre los resultados, de forma que cuando se combina el mecanismo con esas preferencias lo que se obtiene sí es un juego.

El siguiente paso consiste en escoger qué concepto de solución de un juego se adopta. Tres de los conceptos de referencia son el equilibrio en estrategias dominantes (equilibrio dominante), el equilibrio de Nash y el equilibrio bayesiano. Escogido un concepto de solución, ya puede definirse en qué consiste la implementación del resultado deseado mediante el concepto de solución escogido: el resultado deseado se dice que es implementable mediante el mecanismo diseñado en términos del concepto de solución que se haya adoptado si ese resultado es el único que produce el concepto de solución en el juego obtenido cuando se combina el mecanismo propuesto con las preferencias que tengan los jugadores sobre los resultados.

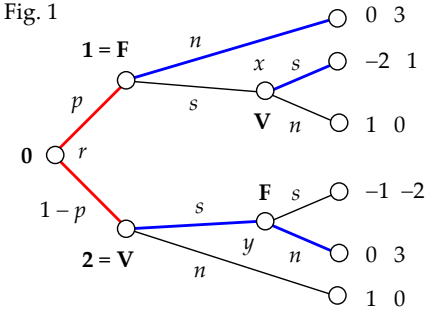
El diseño de mecanismos va en dirección opuesta al análisis de un juego. Cuando se analiza un juego, se parte de las estrategias de que disponen los jugadores y se trata de investigar qué tipo de resultados pueden razonable o justificadamente obtenerse en el juego. Cuando se diseña un mecanismo ocurre lo contrario: se parte de un resultado y se trata de estructurar un mecanismo asignando estrategias a los jugadores para que sus decisiones produzcan el resultado deseado cuando se juega de acuerdo con un cierto tipo de solución.

Un mecanismo para el juicio del Rey Salomón

El juicio descrito anteriormente pretende, aparentemente, evidenciar la sabiduría del rey. Sin embargo, el rey consiguió implementar el resultado deseado por fortuna, no por un diseño adecuado del mecanismo: el rey fue, más que sabio, afortunado. La razón es que la madre falsa no jugó una mejor respuesta a la estrategia seguida por la madre auténtica. ¿Qué habría hecho el rey si la madre falsa hubiese replicado la estrategia que, a primera vista, escogería una madre auténtica? Esto es, ¿qué habría decidido el rey si la madre falsa también hubiese implorado que, antes que arrebatarle la vida, el niño fuese entregado a la otra madre?

La Fig. 1, construida a partir de Binmore (2008), muestra un juego que sí habría permitido implementar el resultado deseado (que el bebé fuese entregado a su madre) empleando como concepto de solución el equilibrio perfecto en subjuegos. El mecanismo en el que se basa el juego es el siguiente. La naturaleza (jugador 0) determina quién es la madre demandante: con probabilidad $0 < p < 1$, la madre demandante es la madre falsa (jugador 1) y, por tanto, el bebé está inicialmente en manos de la madre verdadera; y con probabilidad $1 - p$, la madre demandante es la madre verdadera (jugador 2) y, por tanto, el bebé está inicialmente en manos de la madre falsa. Una vez que la naturaleza establece quién es la madre demandante, ésta decide si manifestar que ella no es la madre (acción *n*) o afirmar que sí lo es (acción *s*). Si la declaración de la madre demandante es "no soy la madre", el juego acaba. Si la declaración es "soy la madre" entonces la otra madre decide si sostener que es la madre o no.

Fig. 1



Los resultados del mecanismo son los siguientes. Si el bebé está inicialmente en manos de la madre verdadera (lo que hace que la demandante sea la madre falsa y nos encontremos en la parte alta de la Fig. 1) entonces el bebé queda en manos de la madre verdadera a menos que ésta diga que el bebé no es suyo y la madre demandante diga que sí lo es. Además, si ambas declaran ser las madres, cada una pagará una multa. Si el bebé está inicialmente en manos de la madre falsa (lo que hace que la demandante sea la madre verdadera) entonces el bebé seguirá en manos de la madre falsa a menos que ésta diga que el bebé no es suyo y la madre demandante diga que sí lo es. Si ambas declaran ser las madres, cada una pagará una multa. Cuando se añadan pagos al mecanismo se obtendrá un juego.

En la Fig. 1, los pagos se han establecido suponiendo que el valor de tener el bebé para la madre verdadera (jugador 2) es 3 y que el valor de tenerlo para la madre falsa (jugador 1) es 1. El importe de la multa (que se impone si las dos declaran ser la madres) es de 2 para ambas. Por último, no recibir el bebé ni asumir la multa implica un pago de 0.

Resolviendo el juego por inducción hacia atrás, tomemos el nudo de decisión *x*. La mejor respuesta en este nudo para el jugador 2 (la madre verdadera) es *s*. Dada la elección de *s* en *x*, la mejor respuesta del jugador 1 en el nudo que precede inmediatamente a *x* es *n*. Pasando al nudo de decisión *y*, la mejor respuesta en este nudo para el jugador 1 (la madre falsa) es *n*. Dada la

elección de n en y, la mejor respuesta del jugador 2 en el nudo justo antes de y es s. Por tanto, decida lo que decida la naturaleza, el resultado siempre es el mismo: la madre falsa niega ser la madre y la madre auténtica afirma serlo. Mediante el equilibrio perfecto en subjuegos, el mecanismo que da lugar al juego de la Fig. 1 ha permitido implementar el resultado deseado: que el bebé siempre vaya a parar a manos de la madre verdadera (vector de pagos (0, 3)).

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	⟶	a	b	c	c
2 puntos	⟶	b	c	b	b
1 punto	⟶	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 $M = M_1 \times M_2 \times \dots \times M_n$, 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 $m \in M = M_1 \times M_2 \times \dots \times M_n$ 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 $M = M_1 \times M_2 \times \dots \times M_n$. 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.

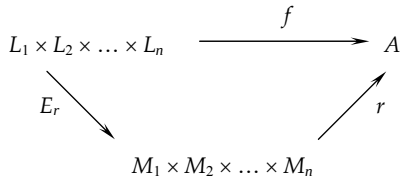
Por ejemplo, los equilibrios dominantes de Nash son aquellos equilibrios de Nash formados por estrategias débilmente dominantes (y, por tanto, son equilibrios en los que ninguna estrategia es débilmente dominada).

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. 2



La Fig. 2 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 $M_1, M_2, \ldots, M_n$ a cada individuo y definir una función de resultados r (esto es, construir un mecanismo) de manera que, para cada perfil de preferencias $P = (P_1, P_2, \ldots, P_n)$, embutir cada perfil de mensajes $m \in M_1 \times M_2 \times \ldots \times M_n$ 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 $P \in L_1 \times L_2 \times \ldots \times L_n$, $f(P) = r[E_r(P)]$: la vía inferior (la vía descentralizada) replica el resultado de la vía superior (la vía centralizada).

Por ejemplo, en el juicio de Salomón, se puede definir a como la alternativa "el bebé se entrega a la madre 1", b como la alternativa "el bebé se entrega a la madre 2" y c como la alternativa "el bebé no se entrega a ninguna madre" (al menos, entero). El conjunto de individuos sería {1, 2}, formado por las dos madres. El rey se enfrentaría a la situación con los dos perfiles de preferencia $P = (P_1, P_2)$ y $Q = (Q_1, Q_2)$ tales que

P_1	P_2
a	b
b	c
c	a

Q_1	Q_2
a	b
c	a
b	c

en donde se interpreta que el perfil P se corresponde con el caso en que la madre 1 es la verdadera y el perfil Q con el caso en que la madre 2 es la verdadera. El rey pretendía implementar una regla de elección social f tal que $f(P) = a$ y $f(Q) = b$.

Implementación honesta

Un mecanismo es directo si el conjunto $M = M_1 \times \ldots \times M_n$ de perfiles de mensajes coincide con el conjunto $L = L_1 \times \ldots \times L_n$ de perfiles de preferencias. En un mecanismo directo, los mensajes de los jugadores consisten en indicar una preferencia (no necesariamente la preferencia auténtica).

Sea $f : L \to A$ una FES. El mecanismo directo (N, L, A, r) implementa honestamente la función de elección social f mediante la solución de equilibrio E si, para todo perfil de preferencias $P \in L$: (H1) $P \in E_r(P)$; y (H2) $r(P) = f(P)$. Cuando existe un mecanismo directo (N, L, A, r) que implementa honestamente f mediante E se dice que f es honestamente implementable mediante la solución E [y el mecanismo directo (N, L, A, r)].

La condición (H1) afirma que revelar la preferencia auténtica por parte de cada jugador es un equilibrio (del tipo requerido) del juego asociado con el mecanismo. La condición (H2) establece que, cuando todos los jugadores revelan sus preferencias auténticas, el resultado del mecanismo coincide con la elección que hace la FES con aquellas preferencias.

La implementación honesta es más débil que la implementación, ya que la honesta requiere el uso de mecanismos directos y permite equilibrios (en el juego asociado con el mecanismo) en los que no se es honesto sobre las preferencias y que generan un resultado distinto al de la FES.

El principio de revelación (para equilibrios de Nash dominantes, Gibbard (1973))

Si una función de elección social f es implementable mediante equilibrios dominantes entonces f es honestamente implementable mediante equilibrios dominantes.

Demostración. Sea f una FES implementable mediante equilibrios dominantes. Esto quiere decir que existe algún mecanismo (N, M, A, r) tal que, para todo perfil de prefencias $P \in L$ y para todo jugador $i \in N = \{1, 2, \ldots, n\}$, existe un mensaje $m_i^*(P) \in M_i$ tal que:

(i) $m_i^*(P)$ es una estrategia (débilmente) dominante para i en el juego asociado con (N, M, A, r) y P; y
(ii) $f(P) = r(m_1^*(P), m_2^*(P), \ldots, m_n^*(P))$.

Puesto que $m_i^*(P)$ es débilmente dominante, no depende de las preferencias de los demás jugadores. Por ello, podemos escribir $m_i^*(P_i)$ en lugar de $m_i^*(P)$.

Definamos el mecanismo directo (N, L, A, h) tal que, para todo $P \in L$, $h(P) = r(m^*(P))$, en donde $m^*(P) = (m_1^*(P_1), m_2^*(P_2), \ldots, m_n^*(P_n))$ es el perfil de estrategias débilmente dominantes que se ha escogido para el juego asociado con (N, M, A, r) y P.

El mecanismo directo (L, h) meramente internaliza las decisiones que los jugadores hacen en el mecanismo original (M, r). En este mecanismo, un jugador i con preferencia P_i escoge el mensaje $m_i^*(P_i)$ y luego la función de resultados r determina el resultado $r(m_1^*(P_1), \ldots, m_n^*(P_n))$, que es justamente el resultado $f(P_1, \ldots, P_n)$ escogido por la FES cuando las preferencias son $(P_1, \ldots, P_n)$. En el nuevo mecanismo directo, cuando el jugador elige la preferencia P_i como mensaje (en principio, P_i no tiene por qué ser la preferencia auténtica de i), el mecanismo mismo escoge la estrategia $m_i^*(P_i)$ y determina el resultado $r(m_1^*(P_1), \ldots, m_n^*(P_n))$. Por tanto, el mecanismo directo es como un agente que replica la decisión que cada jugador tomaría en el mecanismo original. La única diferencia entre ambos mecanismos es que se ha movido la línea que marca lo que queda dentro o fuera del mecanismo, pero todo el proceso es el mismo y, por tanto, genera el mismo resultado. La demostración consiste, precisamente, en verificar este extremo.

La prueba del principio de revelación se reduce a verificar que el mecanismo directo (L, h) implementa honestamente la FES mediante equilibrios dominantes. Supongamos que no es así. El objetivo es llegar a una contradicción: si la negación de una proposición conduce a una contradicción, la proposición es verdadera. Si (L, h) no implementa f honestamente mediante equilibrios dominantes falla (H1) o falla (H2) (o fallan ambas condiciones).

Supongamos que (H1) no se cumple. Entonces existen $i \in N$, $P \in L$ y $Q_i \in L_i$ tal que transmitir el mensaje Q_i es mejor para i que transmitir P_i cuando las preferencias auténticas de todos los jugadores vienen dadas por P. Esto implica que el resultado $h(Q_i, P_{-i})$ si i revela la preferencia falsa Q_i y los demás revelan sus preferencias auténticas es preferido por i al resultado $h(P_i, P_{-i})$ que se obtiene cuando todos son honestos. Formalmente, $h(Q_i, P_{-i})\ P_i\ h(P_i, P_{-i})$.

Por definición de h, $h(Q_i, P_{-i})\ P_i\ h(P_i, P_{-i})$ equivale a $r(m_i^*(Q_i), m_{-i}^*(P_{-i}))\ P_i\ r(m_i^*(P_i), m_{-i}^*(P_{-i}))$. Lo anterior dice lo siguiente: si i prefiere declarar la preferencia falsa Q_i a la preferencia verdadera P_i en el mecanismo directo entonces i también preferirá declarar, en el mecanismo original, el mensaje $m_i^*(Q_i)$ correspondiente a Q_i al mensaje $m_i^*(P_i)$ correspondiente a P_i. Pero entonces, contrariamente a lo que se había asumido, el mensaje $m_i^*(P_i)$ no es una estrategia dominante cuando la preferencia auténtica es P_i. De esta contradicción se deduce que, para todo perfil de preferencias P y todo jugador i, revelar P_i en el mecanismo directo cuando P es el perfil de preferencias auténticas es una estrategia dominante para i. Con ello se verifica el cumplimiento de la condición (H1) de implementación honesta: declarar las preferencias auténticas constituye un equilibrio dominante.

La prueba concluye verificando el cumplimiento de (H2): para todo $P \in L$, $h(P) = f(P)$. Esta parte es más fácil. Por definición, $h(P) = r(m_1^*(P_1), \dots, m_n^*(P_n))$. Por la hipótesis que (M, r) implementa f, se tiene que $r(m_1^*(P_1), \dots, m_n^*(P_n)) = f(P)$. Y ya está: $h(P) = f(P)$.■

Lo que significa el principio de revelación

La implementación mediante equilibrios dominantes es, a priori, la opción más atractiva y satisfactoria. La razón es que, en el juego inducido por el mecanismo implementador, cada jugador tiene algún mensaje (débilmente) dominante, de modo que es razonable esperar que cada jugador lo escoja. El atractivo de ese concepto de equilibrio es que el problema estratégico de los jugadores se puede considerar, de hecho, un problema de decisión individual: al jugador i no le importa qué eligen los demás jugadores si él dispone de alguna estrategia que siempre le da el máximo pago hagan lo que hagan los demás. Este hecho hace más robusto el funcionamiento del mecanismo.

El principio de revelación facilita la determinación de qué tipo de FES es implementable o no: si una FES no es honestamente implementable en equilibrios dominantes entonces no es implementable en equilibrios dominantes. Por tanto, no es preciso recurrir a la infinita variedad de mecanismos imaginables que se podrían considerar para averiguar si una FES es implementable: basta con restringir la atención a mecanismos directos (lo cual no hace necesariamente más sencillo al mecanismo). Así que si todos los mecanismos directos fallan para implementar honestamente una FES no hace falta romperse el cráneo imaginando otros mecanismos para ver si funcionan: no lo harán.

Sin embargo, *per se*, el principio de revelación no hace equivalentes la implementación (en equilibrios dominantes) y la implementación honesta (en equilibrios dominantes). Con todo, para el caso que se está considerando en el que las preferencias son estrictas y la regla de elección social es una función (selecciona sólo un resultado) y no una correspondencia, ambos tipos de implementación son equivalentes.

Equivalencia de la implementación y la implementación honesta

Una función de elección social es implementable mediante equilibrios dominantes si, y sólo si, es implementable honestamente mediante equilibrios dominantes.

En resumen, para determinar si una FES es implementable mediante equilibrios dominantes basta con considerar mecanismos directos y verificar que la revelación honesta es dominante. ¿Y 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-Satterthwaite (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 αβγ la manera de expresar la preferencia del individuo *i* tal que α P_i β P_i γ (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. 3, en donde cada casilla se corresponde con un perfil de preferencias.

Fig. 3

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

La flecha "→" apunta a la alternativa que la FES selecciona cuando el perfil de preferencias es el indicado en la casilla. La Fig. 3 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. 3. 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. 3. 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. 4.

Fig. 4

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. 4. 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. 4) 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. 5.

Fig. 5

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. 5. 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. 5. Por la hipótesis inicial, *f*(*abc*, *bac*) = *a*, tal y como indica la Fig. 5. 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. 6.

Fig. 6

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. 4, 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. 7

a c *b a* → *a* *c b*	*a c* *c a* → *a* *b b*	*b c* *a a* → *c b*	*b c* *c a* → *a b*	*c c* *a a* → *c* *b b*	*c c* *b a* → *c* *a b*
a c *b b* → *a* *c a*	*a c* *c b* → *a* *b a*	*b c* *a b* → *c a*	*b c* *c b* → *a a*	*c c* *a b* → *c* *b a*	*c c* *b b* → *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. 3, la FES escoge precisamente la alternativa más preferida por R1: *a* (ejercicio 2 de la lista). 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. 7. Por Paretoeficiencia, no puede seleccionarse *a*. Así que *f*(*bac*, *cba*) ∈ {*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. 7, 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. 7 (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.

Implementación de funciones de elección social mediante equilibrios de Nash

Una función de elección social es implementable honestamente mediante equilibrios de Nash si, y sólo si, es implementable honestamente mediante equilibrios dominantes.

Por el principio de revelación, toda FES implementable mediante equilibrios dominantes empleando el más complejo de los mecanismos imaginable es también implementable mediante mecanismos directos. Debido a ello, el estudio de la implementación mediante equilibrios dominantes puede restringirse sin pérdida de generalidad al uso de mecanismos directos.

El mecanisme de Groves-Clarke

Una de celebracions

Els estudiants de Microeconomia Superior 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-Clarke (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) $T_i = \sum_{j\neq i} \hat{u}_j(b) - \sum_{j\neq i} \hat{u}_j(a)$.

El mecanisme de Groves-Clarke no incentiva revelacions falses a l'etapa 1

Triem *i* i suposem que, per a tot $j \neq i$, *j* declara, a l'etapa 1, els valors $\hat{u}_j(a)$ i $\hat{u}_j(b)$, que poden no coincidir amb els valors reals $u_j(a)$ i $u_j(b)$. Es tracta de verificar que declarar els valors autèntics $u_i(a)$ i $u_i(b)$ a la primera etapa constitueix una millor resposta d'*i* a les declaracions dels altres. Hi ha dos casos: $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$; i $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) \leq \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$. El primer cas és aquell on el mecanisme selecciona l'opció *a* quan, donat el que manifesten els altres, *i* revela la seva autèntica utilitat. El segon cas és aquell on el mecanisme tria *b* quan *i* també revela la seva autèntica utilitat, donat el que declaren els altres. El cas 1 s'analitza a continuació. El cas 2 es deixa com a exercici. El cas 1 es pot dividir en quatre subcasos.

• Subcas 1: $u_i(a) > u_i(b)$ i $\sum_{j\neq i} \hat{u}_j(a) > \sum_{j\neq i} \hat{u}_j(b)$. El fet que $u_i(a) > u_i(b)$ significa que *i* prefereix l'opció *a* a la *b*. A més, el cas que s'està considerant (cas 1: $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$) fa que l'opció escollida sigui, precisament, *a*. Això fa que *i* no necessiti mentir sobre les seves utilitats per a què *a* sigui escollida: revelant les utilitats reals, l'opció més preferida (*a*) ja és seleccionada.

• Subcas 2: $u_i(a) > u_i(b)$ i $\sum_{j\neq i} \hat{u}_j(a) \leq \sum_{j\neq i} \hat{u}_j(b)$. Si *i* diu la veritat a l'etapa 1 i revela els valors $u_i(a)$ i $u_i(b)$ es garanteix que *a* (la opció més preferida per *i*) serà l'opció escollida (perquè el cas 1 que s'està analitzat suposa que a l'etapa 2 se selecciona *a* quan *i* diu la veritat). Però si *i* no participés en el mecanisme, *b* seria l'opció seleccionada, ja que $\sum_{j\neq i} \hat{u}_j(a) \leq \sum_{j\neq i} \hat{u}_j(b)$. Com a resultat d'això, *i* ha de pagar, en aquest subcas, $T_i = \sum_{j\neq i} \hat{u}_j(b) - \sum_{j\neq i} \hat{u}_j(a)$. Atès que T_i no depèn del que *i* digui a l'etapa 1, no hi ha manera de reduir aquest pagament si no és alterant l'opció que selecciona el mecanisme a l'etapa 2. Per tant, per a determinar si a *i* li convé revelar informació certa a l'etapa 1, cal comparar la utilitat neta d'*i* quan el mecanisme tria *a* amb la utilitat neta d'*i* quan el mecanisme tria *b*. Quan es tria *a*, la utilitat neta d'*i* és $u_i(a) - T_i$. Quan es tria *b*, la utilitat neta d'*i* és $u_i(b)$ i, en aquest cas, *i* s'estalvia pagaments addicionals.

Així que, per a determinar si a *i* li convé mentir a la primera etapa, només cal comparar el resultat $u_i(a) - T_i$ de dir la veritat amb l'únic resultat alternatiu $u_i(b)$ que es pot aconseguir mentint. Per la hipòtesi del cas 1, $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$. D'aquí, $\sum_{j\neq i} \hat{u}_j(a) - \sum_{j\neq i} \hat{u}_j(b) + u_i(a) > u_i(b)$. De manera equivalent, $u_i(a) - [\sum_{j\neq i} \hat{u}_j(b) - \sum_{j\neq i} \hat{u}_j(a)] > u_i(b)$. Per definició, $T_i = \sum_{j\neq i} \hat{u}_j(b) - \sum_{j\neq i} \hat{u}_j(a)$. En conseqüència, $u_i(a) - T_i > u_i(b)$. Conclusió: *i* no guanya res no revelant els valors reals d'utilitat de les opcions *a* i *b* al subcas 2.

• Subcas 3: $u_i(a) < u_i(b)$. Quan $u_i(a) < u_i(b)$, *i* prefereix *b* a *a*. Al cas 1 que s'està tractant, *a* és l'opció escollida si *i* declara els valors reals $u_i(a)$ i $u_i(b)$. Es tracta d'esbrinar si surt a compte a *i* mentir i forçar el canvi d'*a* a *b*. Si *i* revela els valors reals $u_i(a)$ i $u_i(b)$, *a* és l'opció triada. Això vol dir que $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$. Si $\sum_{j\neq i} \hat{u}_j(a) \leq \sum_{j\neq i} \hat{u}_j(b)$, aleshores $u_i(a) < u_i(b)$ implicaria $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) < \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$, contradient la condició que defineix el cas 1 que s'està analitzant. Per tant, $\sum_{j\neq i} \hat{u}_j(a) > \sum_{j\neq i} \hat{u}_j(b)$. Se segueix de $\sum_{j\neq i} \hat{u}_j(a) > \sum_{j\neq i} \hat{u}_j(b)$ que el mecanisme seleccionaria *a* si *i* no participés. Així doncs, si *i* declarés valors $\hat{u}_i(a)$ i $\hat{u}_i(b)$, on almenys un dels dos no fos el valor real, de manera que *b* fos l'opció seleccionada, *i* hauria de pagar una transferència positiva $T_i = \sum_{j\neq i} \hat{u}_j(a) - \sum_{j\neq i} \hat{u}_j(b)$, perquè la seva declaració provocaria el canvi d'*a* a *b*. Ara comparem les dues alternatives d'*i*: dir la veritat i revelar els autèntics valors $u_i(a)$ i $u_i(b)$, o mentir i declarar valors $\hat{u}_i(a)$ i $\hat{u}_i(b)$ que provoquessin l'elecció de *b*.

Si *i* declara el valors reals d'utilitat $u_i(a)$ i $u_i(b)$, *a* se selecciona i la seva utilitat neta és $u_i(a)$, perquè $T_i = 0$. De fet, com ja s'ha demostrat, $\sum_{j\neq i} \hat{u}_j(a) > \sum_{j\neq i} \hat{u}_j(b)$. Això faria que *a* també se seleccionés si *i* no participés. En no provocar la declaració d'*i* un canvi en l'elecció d'opció, $T_i = 0$. D'altra banda, si *i* declara valors $\hat{u}_i(a)$ i $\hat{u}_i(b)$ que provoquen l'elecció de *b*, la utilitat neta d'*i* és $u_i(b) - T_i = u_i(b) - [\sum_{j\neq i} \hat{u}_j(a) - \sum_{j\neq i} \hat{u}_j(b)]$. Atès que (per tractar-se del cas 1) $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$, se segueix que $u_i(a) > [\sum_{j\neq i} \hat{u}_j(b) - \sum_{j\neq i} \hat{u}_j(a)] + u_i(b) = u_i(b) - [\sum_{j\neq i} \hat{u}_j(a) - \sum_{j\neq i} \hat{u}_j(b)] = u_i(b) - T_i$. En resum, *i* no augmenta la seva utilitat neta mentint i forçant el canvi d'*a* a *b*.

• Subcas 4: $u_i(a) = u_i(b)$. Ara *i* és indiferent entre *a* i *b*. Si revela els valors reals, $u_i(a)$ i $u_i(b)$, el mecanisme selecciona *a*, pel fet que s'està analitzant el cas 1. Justament pel cas 1, $\sum_{j\neq i} \hat{u}_j(a) + u_i(a) > \sum_{j\neq i} \hat{u}_j(b) + u_i(b)$. Si $u_i(a) = u_i(b)$, aleshores $\sum_{j\neq i} \hat{u}_j(a) > \sum_{j\neq i} \hat{u}_j(b)$. Això implica que, si *i* no participés, *a* també seria l'opció seleccionada. Per tant, $T_i = 0$. En resum, dient la veritat a l'etapa 1, la utilitat neta d'*i* és $u_i(a)$. Si *i* és plantegés declarar $\hat{u}_i(a)$ i $\hat{u}_i(b)$ que provoquessin l'elecció de *b*, $T_i = \sum_{j\neq i} \hat{u}_j(a) - \sum_{j\neq i} \hat{u}_j(b) > 0$. En aquest cas, la utilitat neta seria $u_i(b) - T_i(a) < u_i(b) = u_i(a)$. Així doncs, *i* no augmenta la seva utilitat neta mentint i forçant el canvi d'*a* a *b*.

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 $T_1' = 13½$, de manera que la nova utilitat neta de 2 seria superior: $u_2(b) + T_1'/3 = 1 + 13⅙ > 3$.

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.

Bibliografia

- Binmore, Ken (2008): L*a teoría de juegos. Una breve introducción*. Alianza Editorial: Madrid.
- Campbell, Donald E. (1995): *Incentives: Motivation and the Economics of Information*, Cambridge University Press: Cambridge, pp. 12–13, 262–264, 283–290 i 294–297.
- Clarke, Edward H. (1971): "Multipart pricing of public goods", *Public Choice* 11, 17–33.
- Groves, Theodore (1973): "Incentives in teams", *Econometrica* 41, 617–631.
- 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.
- Osborne, Martin y Rubinstein, Arial (1994): *A Course in Game Theory*. The MIT Press: Cambridge, Massachussetts, capítulo 10.
- 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.