Racionalidad colectiva en la argumentación social. Imposibilidad general y posibilidad restringida

Gustavo Adrián Bodanza[*]

Introducción

Este trabajo trata sobre la toma de decisiones sociales basadas en argumentación. Desde el punto de vista de la Teoría de la Elección Social (TES), las preferencias sociales se establecen por agregación de las preferencias de los individuos, sin tener en cuenta el porqué de tales preferencias. Sin embargo, si se pretenden justificar las decisiones sociales argumentativamente –como proponen los teóricos de la democracia deliberativa– los criterios y ponderaciones individuales sobre los argumentos a favor o en contra de una alternativa deberían ser evaluados de modo tal que permitan hallar los argumentos colectivamente mejor fundados.

Entre los muchos problemas teóricos que suscita una decisión colectiva argumentada, nos enfocaremos en la cuestión de bajo qué condiciones se la puede considerar racional. Siguiendo las intuiciones naturales surgidas en los orígenes de la TES, consideraremos, en términos generales, que las decisiones colectivas deben reflejar, de alguna manera, las mismas condiciones de racionalidad exigibles a las decisiones individuales. En particular, entendemos que las ponderaciones colectivas realizadas sobre los argumentos del caso deben regirse por las mismas condiciones que las ponderaciones individuales. Por ejemplo, si los individuos reconocen en un argumento a a un atacante de otro argumento b, entonces es de esperar que también colectivamente a se reconozca como un atacante de b.

En nuestro estudio idealizaremos una situación de decisión en la que se pondera un conjunto finito de argumentos A, los que interactúan entre sí atacándose unos a otros, interacción que representaremos mediante una relación binaria T sobre A. Así, partiremos de un *marco argumentativo* abstracto $M = \langle A, T\rangle$ (Dung, 1995) que representa idealmente la situación. Asumiremos también que la ponderación de los argumentos de A, tanto en lo individual como en lo colectivo, consiste en determinar, para cada argumento en A, si resulta aceptado, rechazado o no decidido. Las ponderaciones, para ser racionales, deben cumplir ciertos requisitos. Intuitivamente, por ejemplo, es de esperar que si $(a, b) \in T$ (i.e. a ataca a b) entonces la aceptación de a implique el rechazo de b, o el rechazo de a –si no hay otros atacantes de b– implique la aceptación de b. Entonces, si esto es exigible para las ponderaciones individuales, también debería ser exigible para las ponderaciones colectivas.

Rahwan y Tohmé (2010) muestran que no es posible, en general, obtener una agregación racional de ponderaciones individuales, teniendo en cuenta, además, otros requisitos de racionalidad conocidos en la TES. Sin embargo, nos proponemos mostrar que, bajo ciertas restricciones razonables impuestas a las ponderaciones individuales y al número de individuos, se puede garantizar la racionalidad colectiva.

* Universidad Nacional del Sur / CONICET

Marcos argumentativos abstractos y etiquetamientos

Como dijimos, la herramienta para representar una situación de debate será un *marco argumentativo abstracto* (Dung, 1995).

Definición 1. Un marco argumentativo abstracto es un par $\mathrm{M} = \langle \mathrm{A}, \mathrm{T} \rangle$ donde A es un conjunto de argumentos y T es una relación binaria $\mathrm{T} \subseteq \mathrm{A} \times \mathrm{A}$ que representa los ataques que se dan entre elementos de A (tanto la noción de argumento como la de ataque son primitivas en este modelo).

Ejemplo 1. El marco argumentativo $\mathrm{M}_1 = \langle \{a, b, c\}, \{(b, a), (b, c), (c, b)\} \rangle$ representa una situación en la que hay sólo tres argumentos, *a*, *b* y *c*, tales que *b* ataca a *a*, *b* ataca a *c* y *c* ataca a *b*. A su vez, M_1 se puede graficar como un digrafo en el que los nodos representan los argumentos y los arcos, los ataques (fig. 1).

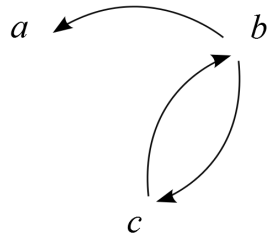

Figura 1

N = {1, 2, …, *n*} será un conjunto de agentes, los miembros del grupo o sociedad que deben tomar la decisión colectiva teniendo en cuenta los argumentos de A. Cada agente realizará sus ponderaciones individuales determinando, para cada argumento $a \in \mathrm{A}$, si lo acepta, lo rechaza o si no decide al respecto. Esto se representará mediante *etiquetamientos* (*labellings* –Caminada, 2006) sobre el marco.

Definición 2. Un *etiquetamiento* sobre un marco argumentativo M = ⟨A, T⟩ es una función total L: $\mathrm{A} \rightarrow$ {IN, OUT, UNDEC} (IN: aceptado; OUT: rechazado; UNDEC: no decidido). Un etiquetamiento es *completo* si y sólo si para todo argumento $x \in$ A:

1. $L(x) = \mathrm{OUT} \Leftrightarrow \exists y \in \mathrm{A}\ ((y, x) \in \mathrm{T} \wedge L(y) = \mathrm{IN})$
2. $L(x) = \mathrm{IN} \Leftrightarrow \forall y \in \mathrm{A}\ ((y, x) \in \mathrm{T} \wedge L(y) = \mathrm{OUT})$

(Además, de 1, 2 y el hecho de que L es una función total, se sigue que $L(x) = \mathrm{UNDEC} \Leftrightarrow L(x) \neq \mathrm{OUT} \wedge L(x) \neq \mathrm{IN}$).

Notación: dada una ordenación (x_1, x_2, …, x_k) de elementos de A, un etiquetamiento L de esos elementos será denotado por $L(x_1, x_2, \ldots, x_k) = (l_1, l_2, \ldots, l_k)$, donde $L(x_j) = l_j$, para todo j, $1 \leq j \leq k$.

Ejemplo 2. El marco del Ejemplo 1 tiene tres etiquetamientos completos posibles para (*a*, *b*, *c*): (IN, OUT, IN), (OUT, IN, OUT) y (UNDEC, UNDEC, UNDEC).

La decisión colectiva como agregación de etiquetamientos

Veamos un simple ejemplo motivador.

Ejemplo 3. Imaginemos un equipo de tres médicos (1, 2 y 3) que deliberan acerca de la terapia adecuada para un paciente. Luego de diversas consideraciones, la discusión se centra en tres argumentos:

a: "Los síntomas x, y, z son indicios de la enfermedad e1, luego debería aplicarse la terapia t1"

b: "Los síntomas x, w, z son indicios de la enfermedad e2, luego debería aplicarse la terapia t2"

c: "Los síntomas x, z son indicios de la enfermedad e3, luego debería aplicarse la terapia t3"

Supongamos además que son incompatibles las terapias t1 con t2 y t2 con t3, de modo que entran en conflicto los argumentos *a* y *b*, por un lado, y *b* y *c*, por otro, atacándose entre sí. Entonces representamos esta situación mediante el marco argumentativo $M_2 = \langle \{a, b, c\}, \{(a, b), (b, a), (b, c), (c, b)\} \rangle$ (fig. 2).

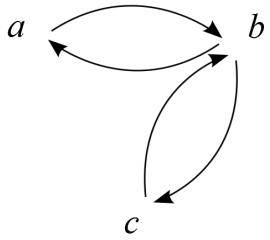

Figura 2

Ahora, cada agente $i \in \{1, 2, 3\}$ hace su ponderación individual, que será representada por un etiquetamiento L_i. Sean tales ponderaciones $L_1(a, b, c) = L_2(a, b, c) =$ (IN, OUT, IN) y $L_3(a, b, c) =$ (OUT, IN, OUT) (fig. 3).

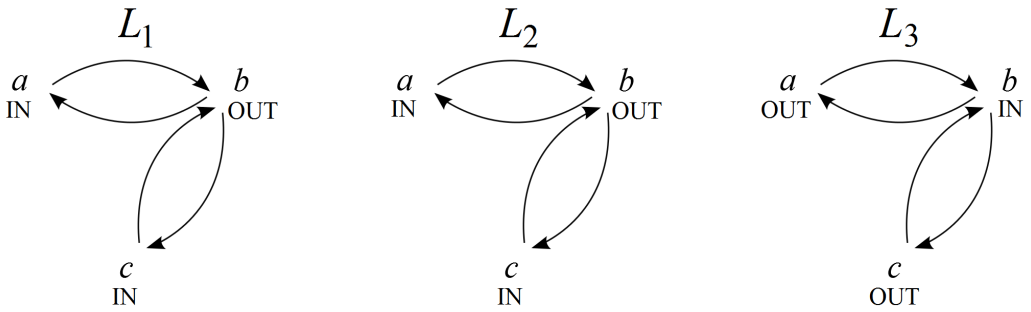

Figura 3

Luego se aplica el mecanismo de agregación. Los mecanismos podrían ser varios. Por ejemplo, podemos pensar en aplicar *mayoría simple*, ya sea contando las etiquetas de cada argumento (*argument-wise plurality voting* –Rahwan & Tohmé, 2010) o contando los etiquetamientos completos. En este caso, ambos procedimientos arrojarían el mismo etiquetamiento colectivo L_c=L_1=L_2, ya que los agentes 1 y 2 hacen el mismo etiquetamiento y logran imponerlo.

Formalmente, cada mecanismo es representado por un *operador de agregación*, i.e. una función parcial

$$F : \boldsymbol{L}(\langle \mathrm{A,T} \rangle)^n \rightarrow \{\mathrm{IN, OUT, UNDEC}\}^{|\mathrm{A}|}$$

donde $\boldsymbol{L}(\langle \mathrm{A,T} \rangle)$ es la clase de etiquetamientos de ⟨A,T⟩ y n es el número de agentes. Es decir, dado un perfil[1] de etiquetamientos individuales, F podrá asignarle a éste un etiquetamiento colectivo (dado que es un función parcial, podría no asignar etiquetamientos colectivos para algunos perfiles).

Ahora bien, ¿podemos asegurar que los etiquetamientos asignados por F serán siempre completos (en el sentido de la Definición 2)? O sea, dado *cualquier* mecanismo de agregación que represente F, y dado *cualquier* perfil de etiquetamientos individuales, ¿será completo el etiquetamiento colectivo L_c? Esta duda, según nuestro criterio, cuestiona la posibilidad de obtener decisiones colectivas argumentadas racionales en general. De hecho –siguiendo a Rahwan y Tomé (2010)– un operador de agregación F cumple *racionalidad colectiva* si y sólo si para cualquier marco argumentativo ⟨A,T⟩ y cualquier perfil (L_1, L_2, …, L_n) de etiquetamientos individuales de ⟨A,T⟩, $L_c = F(L_1, L_2, \ldots, L_n)$ es un etiquetamiento completo.

Es fácil ver que no siempre el etiquetamiento colectivo resultante será completo.

Ejemplo 4. Sean el marco $M_3 = \langle \{a, b, c\}, \{(a, b), (b, a), (b, c), (c, b), (c, a), (a, c)\} \rangle$, los agentes N={1, 2, 3} y las ponderaciones individuales de los argumentos $L_1(a, b, c)$ = (IN, OUT, OUT), $L_2(a, b, c)$ = (OUT, IN, OUT) y $L_3(a, b, c)$ = (OUT, OUT, IN). Entonces, si F

representa el mecanismo de agregación por mayoría simple sobre las etiquetas de cada argumento, obtenemos el etiquetamiento colectivo $L_c = F(L_1, L_2, L_3)$ = (OUT, OUT, OUT) que, claramente, no es completo (fig. 4).

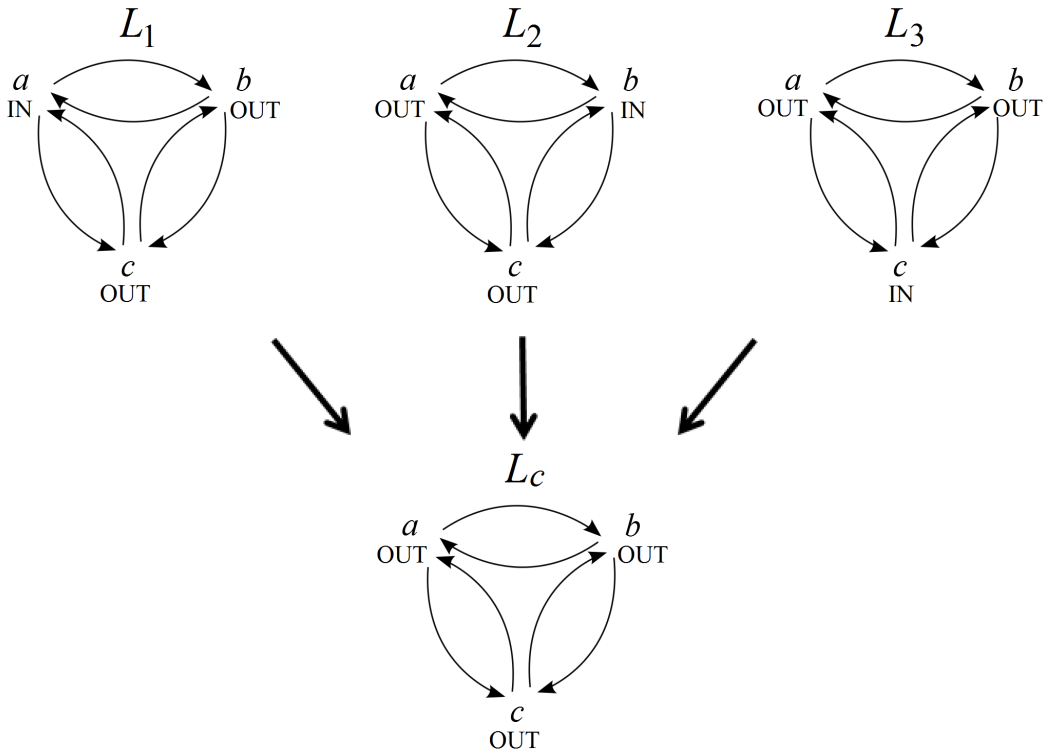

Figura 4

Un resultado negativo y una propuesta de escape

a. El resultado negativo

Habiendo visto que el mecanismo de mayoría simple no garantiza la racionalidad colectiva, surge la pregunta de si no será posible para otros mecanismos. Rahwan y Tohmé (2010) muestran que *ningún* mecanismo de agregación que cumpla racionalidad colectiva puede a la vez satisfacer un conjunto de otras propiedades deseables. Tales propiedades, familiares en el ámbito de los teóricos de la elección social, son las siguientes:

Dominio Irrestricto: todo posible perfil de etiquetamientos individuales ($L_1, \ldots, L_n$) está en el

dominio de F.

Unanimidad: Si $L_i = L$ para $i = 1, \ldots, n$, entonces

$$F(L_1, \ldots, L_n) = L.$$

Anonimidad: dada cualquier permutación

$$p : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\},$$

$F(L_1, \ldots, L_i, \ldots, L_n) = F(L_{p(1)}, \ldots, L_{p(i)}, \ldots, L_{p(n)})$.

Sistematicidad: para cualesquiera argumentos $a, b \in$ A y dos perfiles $(L_1, \ldots, L_n)$ y $(L'_1, \ldots, L'_n)$, si $\forall i\ L_i(a) = L'_i(b)$ entonces $F(L_1, \ldots, L_n)[a] = F(L'_1, \ldots, L'_n)[b]$.

La propiedad de *dominio irrestricto* dice que F debe arrojar un resultado, cualesquiera sean las ponderaciones individuales. *Unanimidad* –como es obvio– dice que si todos los agentes hacen exactamente la misma ponderación, ésta debe ser la resultante en lo colectivo. Por *anonimidad* se entiende que la ponderación colectiva no puede cambiar si los agentes intercambian entre sí sus ponderaciones –dicho de otro modo, todo depende de las ponderaciones mismas y no de quiénes las hacen. Finalmente, por *sistematicidad* se entiende que si todos los individuos ponderan un argumento a en un perfil determinado, exactamente del mismo modo que ponderan otro argumento b en otro perfil, entonces la ponderación colectiva de a en el primer perfil debe ser exactamente la misma que la ponderación colectiva de b en el otro perfil.

En lo que sigue, intentaremos hallar un resultado positivo para, al menos, el mecanismo de agregación por mayoría simple sobre los etiquetamientos de argumentos. Para escapar de la imposibilidad, como es usual en TES, se puede abandonar la condición de dominio irrestricto, i.e., aceptar que para algunos perfiles de ponderaciones individuales no habrá una ponderación colectiva racional. Entonces, ¿cómo pueden caracterizarse los perfiles para los que el resultado positivo es posible? Básicamente, necesitamos tres cosas: 1) evitar empates, 2) que siempre que una mayoría rechace un argumento, acepte a su vez a un atacante de éste, y 3) que siempre que una mayoría acepte un argumento, entonces rechace cualquier posible atacante de éste. Siguiendo a Rahwan y Tohmé (2010) llamamos a estas condiciones 'No empate', 'Derrota Condorcet', y 'No Indecisión Condorcet', respectivamente. Formalmente, se definen como sigue:

1) *No empate:* Un perfil de etiquetamientos $(L_1, \ldots, L_n)$ satisface la condición *no empate* si para cualquier $a \in$ A, existe una etiqueta l tal que

$$|\{i : L_i(a) = l\}| > \max_{l' \neq l} |\{i : L_i(a) = l'\}|$$

Esta condición exige que todo argumento reciba siempre una etiqueta mayoritaria.

Para definir las siguientes dos propiedades necesitamos la noción de 'ganador Condorcet'. Diremos que una etiqueta $l_a \in$ {IN, OUT, UNDEC} de un argumento $a \in$ A es un *ganador*

Condorcet con respecto a un perfil de etiquetamientos (L_1, …, L_n), en símbolos $GC(a, l_a, (L_1, \ldots, L_n))$, si y sólo si $|\{i: L_i(a) = l_a\}| > |\{i: L_i = l'_a\}|$ para toda etiqueta $l'_a \neq l_a$. O sea, la etiqueta l_a es un ganador Condorcet si la cantidad de individuos que ponen esa etiqueta sobre *a* es superior a la de individuos que ponen cualquier otra etiqueta. Entonces,

2) *Derrota Condorcet:* Un perfil (L_1, . . . , L_n) satisface *derrota Condorcet* si y sólo si se cumple:

$GC(a, \text{OUT}, (L_1, \ldots, L_n)) \Leftrightarrow \exists b \in \text{A}$, tal que $(b, a) \in \text{T}$ y $GC(b, \text{IN}, (L_1, \ldots, L_n))$

Es decir, todo perfil debe ser tal que si para un argumento *a* resulta ganadora Condorcet la etiqueta OUT, entonces debe resultar ganadora Condorcet la etiqueta IN para algún argumento atacante de *a*.

3) *No Indecisión Condorcet:* Un perfil de etiquetamientos satisface *no indecisión Condorcet* si y sólo si se cumple:

$GC(a, \text{IN}, (L_1, \ldots, L_n)) \Leftrightarrow \nexists b \in \text{A}$, tal que $(b, a) \in \text{T}$ y, o bien $GC(b, \text{UNDEC}, (L_1, \ldots, L_n))$, o bien $GC(b, \text{IN}, (L_1, \ldots, L_n))$.

Según esta condición, la etiqueta IN no puede resultar ganadora Condorcet para un argumento *a*, si para algún atacante de *a* resulta ganadora Condorcet o bien la etiqueta IN o bien la etiqueta UNDEC.

Es bastante claro que las condiciones de Derrota Condorcet y No Indecisión Condorcet apuntan directamente a obtener la racionalidad colectiva, de acuerdo a los requisitos exigidos para un etiquetamiento completo de argumentos. En efecto, Rahwan y Tohmé (2010) demuestran que estas condiciones son necesarias y suficientes para obtener la racionalidad colectiva.

Por otra parte, nuestro propósito es encontrar algunas condiciones específicas y, sobre todo, razonables, bajo las cuales se cumplan estas tres propiedades.

b. La propuesta de escape

Las condiciones que hallamos no son necesarias sino suficientes. En primer lugar, el modo más obvio de evitar empates para el mecanismo de mayoría simple es el de limitar la cantidad de individuos a un número impar. En segundo lugar, la Derrota Condorcet se puede asegurar si el marco argumentativo en cuestión tiene como máximo tres etiquetamientos completos posibles por cada componente fuertemente conectado del grafo[2]. En tercer lugar, No Indecisión Condorcet se cumplirá si todos los individuos minimizan la etiqueta UNDEC en sus ponderaciones.

Veamos ahora que tan "razonables" son estas condiciones. Primero, está claro que si tenemos un número par de individuos, entonces en casos de dos facciones con $n/2$ votos cada una es imposible evitar los empates. De modo que no hay muchas opciones para evitar esto.

En la vida real, vemos que es usual en algunos organismos resolutivos que en casos de empates algún miembro cuente con voto calificado para desequilibrar la igualdad, procedimiento equivalente a contar doble el voto de ese individuo, llevando el total de votos a un número impar.

Segundo, limitar la cantidad de etiquetamientos posibles del marco argumentativo a un máximo de tres equivale a evitar ataques mutuos entre más de dos argumentos, todos entre sí. Básicamente, esta restricción evita los ciclos de ataques de longitud tres, situación en la que un argumento puede tener etiqueta mayoritaria OUT sin que ninguno de sus atacantes tenga etiqueta mayoritaria IN. Por ejemplo, el marco de la figura 4 tiene cuatro etiquetamientos posibles –además de L_1, L_2 y L_3 también es posible (UNDEC, UNDEC, UNDEC)– dando lugar a un etiquetamiento colectivo no completo: el perfil de etiquetamientos individuales no cumple Derrota Condorcet. Para estos casos quizá resulte más apropiada una decisión basada en el azar antes que en el mecanismo de mayorías, lo que, después de todo, significa reconocer que los argumentos disponibles no son suficientes para la decisión.

Tercero, que todos los individuos minimicen la etiqueta UNDEC representa el hecho de exigir el mayor compromiso posible para aceptar o rechazar argumentos. Siempre que sea lógicamente posible (i.e. manteniendo un etiquetamiento completo), cada individuo *i* debe ser capaz de decidir, para cada argumento *a*, si su etiqueta será $L_i(a)$=IN o $L_i(a)$=OUT. Un dato técnico: esto significa que los etiquetamientos de los individuos deben ser *semi-estables* (Caminada, 2006), caracterizándose por el hecho de que todo argumento que no tiene etiqueta IN es atacado por algún argumento con etiqueta IN (exceptuando el extraño caso de argumentos que se auto-ataquen).

Veamos ahora un bosquejo de la prueba. Asumiendo, entonces, que *n* es impar, que el marco argumentativo tiene a lo sumo tres etiquetamientos completos posibles y que los individuos aplican etiquetamientos en los que UNDEC es mínimo, debemos probar que se cumplen las propiedades de No Empate, Derrota Condorcet y No Indecisión Condorcet.

Comencemos por ver que se cumple Derrota Condorcet. Supongamos que para un argumento *a* la etiqueta OUT resulta ganadora Condorcet. Está claro que en ese marco argumentativo debe existir un argumento *b* que ataca a *a*. Más aún, podemos tener secuencias $a_1, a_2, \ldots, a_n$ de argumentos donde $a_1 = a$ y para cada i, $1 < i \leq n, (a_{i+1}, a_i) \in \mathrm{T}$. En algún etiquetamiento en el que *a* tiene etiqueta OUT, a_2 tendrá etiqueta IN, a_3 tendrá etiqueta OUT, etc. Nótese que un etiquetamiento que asigna UNDEC a todos estos argumentos no puede ser elegida por ningún individuo, puesto que no cumplirían con el compromiso de minimizar esa etiqueta. De modo que en cualquier etiquetamiento individual todos los argumentos de esa secuencia tienen o bien etiqueta IN o bien etiqueta OUT. Esto implicará que algún argumento atacante de *a* recibirá la etiqueta IN, que resultará ganadora Condorcet.

No Indecisión Condorcet: Supongamos que *a* tiene etiqueta IN ganadora Condorcet, i.e. $GC(a, \text{IN}, (L_1, \ldots, L_n))$. Está claro que para cualquier argumento *b* que ataca a *a* existe otro argumento *c* que ataca a *b*. Por hipótesis, existen algunos etiquetamientos en los que *c* no recibe etiqueta OUT; por otra parte, si *c* recibe etiqueta UNDEC, tal etiquetamiento no minimizará el uso de UNDEC, por lo cual no podrá ser elegido por ningún individuo. Luego, en todo etiquetamiento elegido *c* tendrá etiqueta IN, lo que implica que *b* no tendrá etiqueta IN ni UNDEC.

Finalmente, la condición de No Empate se verifica como sigue. Descartando la posibilidad de que *a* pueda ser etiquetado sólo con UNDEC (caso en el que obviamente no habrá empates), *a* sólo podrá ser etiquetado por los individuos o bien con IN o bien con OUT. Puesto que el número de individuos es impar, no podrá haber un empate entre esos etiquetamientos.

Conclusión

En este trabajo hemos tratado sobre ciertas limitaciones formales para la toma de decisiones colectivas argumentadas y hemos propuesto algunas condiciones suficientes para obtener resultados racionales en base al mecanismo de agregación de mayoría simple. En concreto, hemos comentado el resultado de Rahwan y Tohmé (2010) acerca de la imposibilidad de obtener una evaluación colectiva de argumentos racional, siempre que se pretendan salvar las condiciones de Dominio Irrestricto, Unanimidad, Anonimidad y Sistematicidad. Nuestra vía de escape consistió en identificar tres condiciones suficientes para su posibilidad, consistentes en 1) restringir el dominio de perfiles individuales exigiendo el máximo compromiso posible de parte de los individuos en la aceptación o rechazo de los argumentos, 2) acotar la clase de marcos argumentativos sobre los que es posible establecer una decisión racional, evitando que tres o más argumentos se ataquen entre sí a la vez, y 3) evitando la posibilidad de empates exigiendo un número impar de individuos. Por supuesto, aunque suficientes, estas condiciones no son necesarias.

El problema que hemos tratado tiene claras similitudes con el de *agregación de juicios* (List y Pettit, 2002, etc.). En ese caso, se estudia la posibilidad de agregar conjuntos de fórmulas. Dado un conjunto de fórmulas F que representan proposiciones sobre las que cada individuo juzgará su verdad o falsedad, el propósito es encontrar un conjunto de fórmulas que represente la opinión colectiva sobre esas proposiciones. Las evaluaciones de cada individuo *i* son representadas por un conjunto F_i que contendrá, por cada fórmula *a* en F, o bien la fórmula *a* (representando la evaluación de *a* como verdadera) o bien la fórmula $\neg a$ (*a* evaluada como falsa). La fórmulas pueden estar lógicamente relacionadas, por lo que se exigen ciertas condiciones de racionalidad tanto para los conjuntos individuales como para el colectivo. Básicamente, cada conjunto de fórmulas debe ser consistente, completo (en el sentido de que

por cada fórmula *a,* o bien contiene *a* o bien contiene ¬*a*) y deductivamente clausurado. El conjunto colectivo resultante de la agregación es racional si y sólo si el resultado es también un conjunto consistente, completo y deductivamente clausurado. List y Pettit demuestran la imposibilidad de obtener este este resultado en caso de cumplirse otras condiciones, como Dominio Irrestricto, Sistematicidad y Anonimidad. Luego estudian algunas restricciones que lo posibilitan.

Notas

1. Por 'perfil' se entiende el conjunto de todas las ponderaciones individuales.

2. Un *componente fuertemente conectado* de un grafo es un subgrafo máximo tal que para cualesquiera de sus nodos *a* y *b*, existe tanto un camino de *a* hacia *b* como uno de *b* hacia *a*.

Bibliografía

DUNG, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2) (pp. 321–357).

CAMINADA, M. (2006). Semi-stable semantics. *Proc. Computational Models of Agents (COMMA 2006)* (pp. 121-130).

LIST, C., PETTIT, P. (2002). Aggregating sets of judgements: An impossibility result, *Economics and Philosophy*, 18 (pp. 89-110).

RAHWAN, I., TOHMÉ, F. (2010). "Collective argument evaluation as judgement aggregation". En van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May, 10–14, 2010, Toronto, Canada (pp. 417-424).