Game Theory

Lecture Notes By

Y. Narahari

Department of Computer Science and Automation

Indian Institute of Science

Bangalore, India

July 2012

The Gibbard Satterthwaite Theorem and Arrow's Impossibility Theorem

Note: *This is a only a draft version, so there could be flaws. If you find any errors, please do send email to* `hari@csa.iisc.ernet.in`. *A more thorough version would be available soon in this space.*

1 The Gibbard–Satterthwaite Impossibility Theorem

We have seen in the last section that dominant strategy incentive compatibility is an extremely desirable property of social choice functions. However the DSIC property, being a strong one, precludes certain other desirable properties to be satisfied. In this section, we discuss the Gibbard–Satterthwaite impossibility theorem (G–S theorem, for short), which shows that the DSIC property will force an SCF to be dictatorial if the utility environment is an unrestricted one. In fact, in the process, even ex-post efficiency will have to be sacrificed. One can say that the G–S theorem has shaped the course of research in mechanism design during the 1970s and beyond, and is therefore a landmark result in mechanism design theory. The G–S theorem is credited independently to Gibbard in 1973 [1] and Satterthwaite in 1975 [2]. The G–S theorem is a brilliant reinterpretation of the famous Arrow's impossibility theorem (which we discuss in the next section). We start our discussion of the G–S theorem with a motivating example.

Example 1 (Supplier Selection Problem) We have seen this example earlier (Example 2.29). We have $N = \{1, 2\}$, $X = \{x, y, z\}$, $\Theta_1 = \{a_1\}$, and $\Theta_2 = \{a_2, b_2\}$. Consider the following utility functions (note that these are different from the ones considered in Example 2.29):

$$\begin{aligned} u_1(x, a_1) &= 100;\;\; u_1(y, a_1) = 50;\;\; u_1(z, a_1) = 0 \\ u_2(x, a_2) &= 0;\;\; u_2(y, a_2) = 50;\;\; u_2(z, a_2) = 100 \\ u_2(x, b_2) &= 30;\;\; u_2(y, b_2) = 60;\;\; u_2(z, b_2) = 20. \end{aligned}$$

We observe for this example that the DSIC and BIC notions are identical since the type of player 1 is common knowledge and hence player 1 always reports the true type (since the type set is a singleton).

Consider the social choice function f given by $f(a_1, a_2) = x;\quad f(a_1, b_2) = x$. It can be seen that this SCF is ex-post efficient.

To investigate DSIC, suppose the type of player 2 is a_2. If player 2 reports his true type, then the outcome is x. If he misreports his type as b_2, then also the outcome is x. Hence there is no incentive for player 2 to misreport. A similar situation presents itself when the type of player 2 is b_2. Thus f is DSIC.

In both the type profiles, the outcome happens to be the most favorable one for player 1, that is, x. Therefore, player 1 is a dictator and f is dictatorial. Thus the above function is ex-post efficient and DSIC but dictatorial.

Now, let us consider a different SCF h defined by $h(a_1, a_2) = y; h(a_1, b_2) = x$. Following similar arguments as above, h can be shown to be ex-post efficient and nondictatorial but not DSIC. Table 1 lists all the nine possible social choice functions in this scenario and the combination of properties each function satisfies.

i	$f_i(a_1, a_2)$	$f_i(a_1, b_2)$	EPE	DSIC	NON-DICT
1	x	x	√	√	×
2	x	y	√	×	√
3	x	z	×	×	√
4	y	x	√	×	√
5	y	y	√	√	√
6	y	z	×	×	√
7	z	x	√	√	√
8	z	y	√	√	×
9	z	z	×	√	√

Table 1: Social choice functions and properties satisfied by them

Note that the situation is quite desirable with the following SCFs.

$$\begin{aligned} f_5(a_1, a_2) &= y;\quad f_5(a_1, b_2) = y \\ f_7(a_1, a_2) &= z;\quad f_7(a_1, b_2) = x. \end{aligned}$$

The reason is these functions are ex-post efficient, DSIC, and also nondictatorial. Unfortunately however, such desirable situations do not occur in general. In the present case, the desirable situations do occur because of certain reasons that will become clear soon. In a general setting, ex-post efficiency, DSIC, and nondictatorial properties can never be satisfied simultaneously. In fact, even DSIC and nondictatorial properties cannot coexist. This is the implication of the powerful Gibbard–Satterthwaite theorem.

1.1 The G–S Theorem

We will build up some notation before presenting the theorem. We have already seen that the preference of an agent i, over the outcome set X, when its type is θ_i can be described by means of a *utility function* $u_i(\cdot, \theta_i) : X \to \mathbb{R}$, which assigns a real number to each element in X. A utility function

$u_i(\cdot, \theta_i)$ always induces a *unique* preference relation $\succsim$ on X which can be described in the following manner

$$x \succsim y \Leftrightarrow u_i(x, \theta_i) \geq u_i(y, \theta_i).$$

The above preference relation is often called a rational preference relation and it is formally defined as follows.

Definition 1 (Rational Preference Relation) *We say that a relation $\succsim$ on the set X is called a rational preference relation if it possesses the following three properties:*

1. *Reflexivity: $\forall\ x \in X$, we have $x \succsim x$.*

2. *Completeness: $\forall\ x, y \in X$, we have that $x \succsim y$ or $y \succsim x$ (or both).*

3. *Transitivity: $\forall\ x, y, z \in X$, if $x \succsim y$ and $y \succsim z$, then $x \succsim z$.*

The following proposition establishes the relationship between these two ways of expressing the preferences of an agent i over the set X.

Proposition 1

1. *If a preference relation $\succsim$ on X is induced by some utility function $u_i(\cdot, \theta_i)$, then it will be a rational preference relation.*

2. *For every preference relation $\succsim$ on X, there may not exist a utility function that induces it. However, when the set X is finite, given any preference relation, there will exist a utility function that induces it.*

3. *For a given preference relation $\succsim$ on X, there might be several utility functions that induce it. Indeed, if the utility function $u_i(\cdot, \theta_i)$ induces $\succsim$, then $u'_i(x, \theta_i) = f(u_i(x, \theta_i))$ is another utility function that will also induce $\succsim$, where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing function.*

Strict Total Preference Relations

We now define a special class of rational preference relations that satisfy the antisymmetry property also.

Definition 2 (Strict-total Preference Relation) *We say that a rational preference relation $\succsim$ is strict-total if it possesses the antisymmetry property, in addition to reflexivity, completeness, and transitivity. By antisymmetry, we mean that, for any $x, y \in X$ such that $x \neq y$, we have either $x \succsim y$ or $y \succsim x$, but not both.*

The strict-total preference relation is also known as a *linear order relation* because it satisfies the properties of the usual *greater than or equal to* relationship on the real line. Let us denote the set of all rational preference relations and strict-total preference relations on the set X by $\mathscr{R}$ and $\mathscr{P}$, respectively. It is easy to see that $\mathscr{P} \subset \mathscr{R}$.

Ordinal Preference Relations

In a mechanism design problem, for agent i, the preference over the set X is described in the form of a utility function $u_i : X \times \Theta_i \to \mathbb{R}$. That is, for every possible type $\theta_i \in \Theta_i$ of agent i, we can define a utility function $u_i(\cdot, \theta_i)$ over the set X. Let this utility function induce a rational preference relation $\succsim_i (\theta_i)$ over X. The set $\mathscr{R}_i = \{\succsim \,:\, \succsim = \succsim_i (\theta_i) \text{ for some } \theta_i \in \Theta_i\}$ is known as the set of ordinal preference relations for agent i. It is easy to see that $\mathscr{R}_i \subset \mathscr{R} \quad \forall\, i \in N$.

With all the above notions in place, we are now in a position to state the G–S theorem.

Theorem 1 (Gibbard–Satterthwaite Impossibility Theorem) *Consider a social choice function $f : \Theta \to X$. Suppose that*

1. *The outcome set X is finite and contains at least three elements,*
2. *$\mathscr{R}_i = \mathscr{P} \quad \forall\, i \in N$,*
3. *$f(\cdot)$ is an onto mapping, that is, the image of SCF $f(\cdot)$ is the set X.*

Then the social choice function $f(\cdot)$ is dominant strategy incentive compatible iff it is dictatorial.

For a proof of this theorem, the reader is referred to Proposition 23.C.3 of the book by Mas-Colell, Whinston, and Green [3]. We only provide a brief outline of the proof. To prove the necessity, we assume that the social choice function $f(\cdot)$ is dictatorial and it is shown that $f(\cdot)$ is DSIC. This can be shown in a fairly straightforward way. The proof of the sufficiency part of the theorem starts with the assumption that $f(\cdot)$ is DSIC and proceeds in three steps:

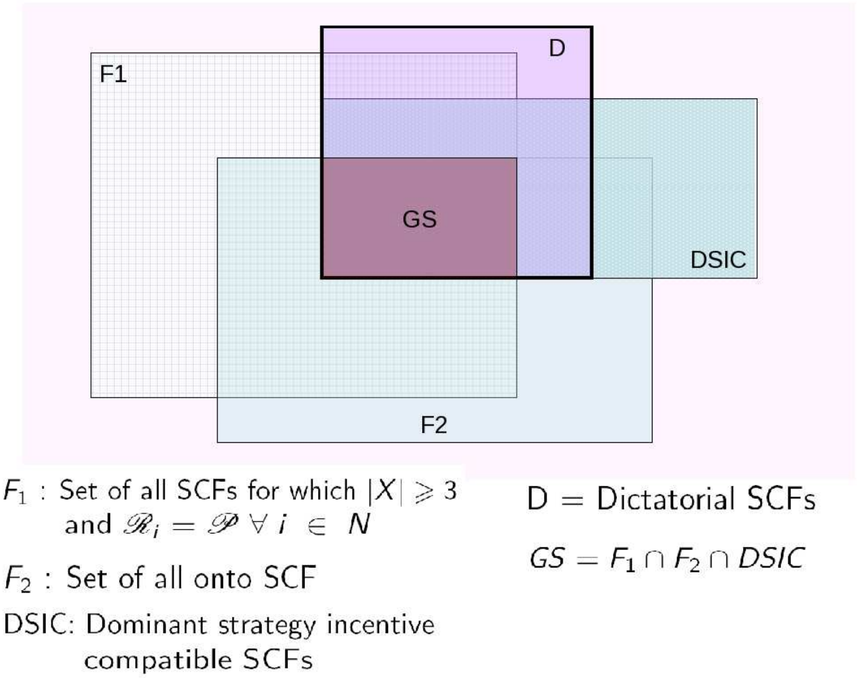
1. It is shown using the second condition of the theorem ($\mathscr{R}_i = \mathscr{P} \quad \forall \; i \in N$) that $f(\cdot)$ is monotonic.
2. Next using conditions (2) and (3) of the theorem, it is shown that monotonicity implies ex-post efficiency.
3. Finally, it is shown that a SCF $f(\cdot)$ that is monotonic and ex-post efficient is necessarily dictatorial.

Figure 1 shows a pictorial representation of the G–S theorem. The figure depicts two classes F_1 and F_2 of social choice functions. The class F_1 is the set of all SCFs that satisfy conditions (1) and (2) of the theorem while the class F_2 is the set of all SCFs that satisfy conditions (1) and (3) of the theorem. The class GS is the set of all SCFs in the intersection of F_1 and F_2 which are DSIC. The functions in the class GS have to be necessarily dictatorial.

1.2 Implications of the G–S Theorem

One way to get around the impossible situation described by the G–S Theorem is to hope that at least one of the conditions (1), (2), and (3) of the theorem does not hold. We discuss each one of these below.

- Condition (1) asserts that $|X| \geq 3$. This condition is violated only if $|X| = 1$ or $|X| = 2$. The case $|X| = 1$ corresponds to a trivial situation and is not of interest. The case $|X| = 2$ is more interesting but is of only limited interest. A public project problem where only a go or no-go decision is involved and no payments by agents are involved corresponds to this situation.



F_1 : Set of all SCFs for which $|X| \geqslant 3$ and $\mathscr{R}_i = \mathscr{P} \; \forall \; i \; \in \; N$

F_2 : Set of all onto SCF

DSIC: Dominant strategy incentive compatible SCFs

D = Dictatorial SCFs

$GS = F_1 \cap F_2 \cap DSIC$

Figure 1: An illustration of the Gibbard–Satterthwaite Theorem

- Condition (2) asserts that $\mathscr{R}_i = \mathscr{P} \; \forall \; i \in N$. This means that the preferences of each agent cover the entire space of strict total preference relations on X. That is, each agent has an extremely rich set of preferences. If we are able to somehow restrict the preferences, we can hope to violate this condition. One can immediately note that this condition was violated in the motivating example (Example 1, the supplier selection problem). The celebrated class of VCG mechanisms has been derived by restricting the preferences to the quasilinear domain. This will be discussed in good detail in a later section.

- Condition (3) asserts that f is an onto function. Note that this condition also was violated in Example 1. This provides one more route for getting around the G–S Theorem.

Another way of escaping from the jaws of the G–S Theorem is to settle for a weaker form of incentive compatibility than DSIC. We have already discussed Bayesian incentive compatibility (BIC) which only guarantees that reporting true types is a best response for each agent whenever all other agents also report their true types. Following this route leads us to Bayesian incentive compatible mechanisms. These are discussed in good detail in a future section.

The G–S Theorem is an influential result that defined the course of mechanism design research in the 1970s and 1980s. As already stated, the theorem happens to be an ingenious reinterpretation, in the context of mechanism design, of the celebrated Arrow's impossibility theorem, which is discussed next.

2 Proof of GS Theorem

Lower Contour Sets

Given an outcome $x \in X$, and agent $i \in N$ and a type of agent i, $\theta_i \ \in \ \Theta_i$, the lower contour set $L_i(x, \theta_i)$ is defined as

$$L_i(x,\theta_i) = \{y \in X \ : \ u_i(x,\theta_i) \ \geq \ u_i(y,\theta_i)\}$$

The lower contour set consists of all outcomes which produce equal or less utility than $u_i(x, \theta_i)$.

Weak Preference Reversal Property

Recall the necessary and sufficient condition for DSIC of a social choice function

$$f(.) : u_i(f,(\theta_i,\theta_{-i}),\ \theta_i) \ \geq \ u_i(f,(\hat{\theta}_i,\theta_{-i}),\ \theta_i) \ \forall\, \theta_i \ \in \ \Theta_i,\ \forall\, \hat{\theta}_i \ \in \ \Theta_i,\ \forall\, \theta_{-i} \ \in \ \Theta_{-i},\ \forall\, i \ \in \ N$$

Consider an agent $i \in N$ and let $\theta_i^{'},\ \theta_i^{''} \ \in \ \Theta_i$ be any two possible types function $f(.)$ is DSIC. Then the above necessary and sufficient condition yields the following two inequalities:

$$u_i(f(\theta_i^{'},\theta_{-i}),\ \theta_i^{'}) \ \geq \ u_i(f({\theta_i}^{''},\theta_{-i}),\ \theta_i^{'}) \ \forall\, \theta_{-i} \ \in \ \Theta_{-i}$$

$$u_i(f(\theta_i^{''},\theta_{-i}),\ \theta_i^{''}) \ \geq \ u_i(f({\theta_i}^{'},\theta_{-i}),\ \theta_i^{''}) \ \forall\, \theta_{-i} \ \in \ \Theta_{-i}$$

clearly, the preference ranking of the outcomes $f(\theta_i^{'},\theta_{-i})$ and $f(\theta_i^{''},\theta_{-i})$ *weakly reverses* when the type changes from $\theta_i^{'}$ to $\theta_i^{''}$.

On the other hand, if a social choice function $f(.)$ is such that the above weak preference reversal property holds for all $\theta_{-i} \ \in \ \Theta_{-i}$ and for all possible pairs $\theta_i^{'},\theta_{-i}^{''} \ \in \ \Theta_i$, it can be shown that $f(.)$ is DSIC. Thus DSIC can also be characterized as being equivalent to the weak preference reversal property. In terms of lower contour sets, the above observations can be summarized as the following proposition.

Proposition : A social choice function $f : X \rightarrow \Theta$ is DSIC iff $\forall\, i \ \in \ N$, $\forall\, \theta_{-i} \ \in \ \Theta_{-i}$ and all pairs $\theta_i^{'},\theta_{-i}^{''} \ \in \ \Theta_i$, the following equalities are satisfied.

$$f(\theta_i^{''},\theta_{-i}) \ \in L_i(f(\theta_i^{'},\theta_{-i}),\ \theta_i^{'}) \text{ and } f(\theta_i^{'},\theta_{-i}) \ \in L_i(f(\theta_i^{''},\theta_{-i}),\ \theta_i^{''})$$

Monotonicity

Monotonicity is an important property of a social choice function and plays a crucial role in mechanism design theory. Suppose $\theta \ \in \ \Theta$ and $f(\theta) = x \in X$. Let the type profile θ change to $\theta^{'} \in \Theta$ and assume that in the new type profile $\theta^{'}$, no agent i finds that some alternative which was weakly worse than x under type θ_i becomes strictly preferred to x. Then monotonicity of $f(.)$ means that x must continue to be the social choice in $\theta^{'}$, that is $f(\theta^{'}) = x$. This is formalized in the following definition.

Definition: A social choice function $f : \Theta \rightarrow X$ is monotonic if $\forall\, \theta \ \in \ \Theta$, $\forall\, \theta^{'} \ \in \ \Theta(\theta^{'} \neq \theta)$,

$$L_i(f(\theta),\theta_{-i}), \subset L_i(f(\theta),\theta_{-i}^{'}) \ \forall\, i \in N \Longrightarrow f(\theta_{-i}^{'}) = f(\theta)$$

Proof of Gibbard Satterthwaite Theorem

The proof is simple in one direction: Suppose all the conditions $(1) - (3)$ are satisfied and $f(.)$ is dictatorial, it can be shown easily that $f(.)$ is DSIC. This is left as an exercise.

In the other direction, we are given that conditions $(1) - (3)$ are satisfied and $f(.)$ is DSIC. We have to show that $f(.)$ is dictatorial. The proof of this proceeds in three steps. We have followed closely the proof given by Mascolell, Whinston, and Green MASCOLELL95]

Step 1 : Showing that $f(.)$ is Monotonic

We are given that $f(.)$ is DSIC. Consider two profiles θ and $\theta^{'}$ such that

$$L_i(f(\theta),\theta_i), \subset L_i(f(\theta),\theta_i^{'}) \ \forall \ i \in N$$

Consider the outcome $f(\theta_1^{'},\theta_2,\ldots,\theta_n)$. Then by the weak preference reversal property, we have

$$f(\theta_1^{'},\theta_2,\ldots,\theta_n) \ \in L_1(f(\theta),\theta_1)$$

By the assumption we have made, we have

$$f(\theta_1^{'},\theta_2,\ldots,\theta_n) \ \in L_1(f(\theta),\theta_1^{'})$$

By the weak preference reversal property, we again have

$$f(\theta_1,\theta_2,\ldots,\theta_n) \ \in L_1(f(\theta_1^{'},\theta_2,\ldots,\theta_n),\theta_1^{'})$$

Since $\mathbf{R}_i = \mathbf{P} \ \forall i \in N$, no two alternatives can be indifferent in the preference relation $\tau/1(\theta_1^{'})$. therefore it must be that

$$f(\theta_1^{'},\theta_2,\ldots,\theta_n) = f(\theta)$$

On similar lines, it can be shown that

$$f(\theta_1^{'},\theta_2^{'},\theta_3\ldots,\theta_n) = f(\theta)$$

Extending the above argument iteratively we get what we need for monotonicity of $f(.)$:

$$f(\theta_1^{'},\theta_2^{'},\ldots,\theta_n^{'}) = f(\theta)$$

Step 2 : Showing that $f(.)$ is Ex-Post Efficient

Here we show that $f(.)$ is ex-post efficient if $\mathbf{R}_i = \mathbf{P} \ \forall \ i \in N, \ f(\Theta) = X$, and f is monotonic. We prove this by contradiction. Suppose $f(.)$ is not ex-post efficient. Then $\ni$ a type profile $\theta \in \Theta$ and an outcome $y \in X$ such that

$$u_i(y,\theta_i) > u_i(f(\theta),\theta_i) \ \forall \ i \in N$$

The above involves only strict inequality because no two alternatives can be indifferent for any agent as $\mathbf{R}_i = \mathbf{P} \ \forall \ i \in N$.

Since $f(\Theta) = X$, there exists a type profile $\theta^{'} \in \Theta$ such that $f(\theta^{'}) = y$. Choose $\theta^{''} \in \Theta$ such that $\forall \ i \in N$,

$$u_i(y,\theta_i^{''}) > u_i(f(\theta),\theta_i^{''}) > u_i(z,\theta_i^{''}) \ \forall \ z \neq f(\theta), y$$

The above choice is certainly possible since all preferences in $\mathbf{P}$ are allowed. We now invoke monotonicity and note that

$$L_i(y, \theta_1^{'}) \subset L_i(y, \theta_i^{''}) \;\forall\; i \in N \Rightarrow f(\theta_i^{''}) = y$$

However, since $L_i(f(\theta), \theta_i) \subset L_i(f(\theta), \theta_i^{''}) \;\forall\; i \in N$, monotonicity again implies that

$$f(\theta_i^{''}) = f(\theta)$$

The above is contradiction, since $y \neq f(\theta)$. This in turn implies that f must be ex-post efficient.

Step 3 : Showing that $f(.)$ is Dictatorial

We are given that $f(.)$ is DISC and EPE and we are supposed to show that $f(.)$ is dictatorial. This result can be obtained as a Corollary of the Arrow's impossibility result (see next section). We direct the reader to chapter 21 of the book by Mascolell, Whinston, and Green [MASCOLELL95].

Some Notes and Observations

- It may be noted that the finiteness of X is not required for GS Theorem. However, if X is not finite, the assumption about agents being expected utility maximizers may not be compatible with the condition $\mathbf{R}_i = \mathbf{P} \;\forall\; i \in N$ [MASCOLELL95]. If X is not finite, the GS Theorem will still hold $\mathbf{R}_i$ for each $\forall\; i \in N$ is the set of all continuous preferences on X [MASCOLELL95].

- If $|X| = 2$, the GS theorem is not true. We have already seen this earlier in this section while discussing the example.

- When $\mathbf{R}_i = \mathbf{P} \;\forall\; i \in N$, it may be noted that any ex-post efficient social choice function must have $f(\Theta) = X$.

- The GS theorem holds even if the assumption (2) is relaxed to

$$\mathbf{P} \subset \mathbf{R}_i \;\forall\; i \in N$$

- We define $f : \Theta \to X$ as dictatorial on $Y \subset X$ if there exists an agent $d \in N$ such that $\forall\; \theta \in \Theta$,

$$u_d(f(\theta), \theta_d) \geq u_d(y, \theta_d) \;\forall\; y \in Y$$

  The GS theorem holds good under the following special setting also: suppose X is finite, $|f(\Theta)| \geq 3$, and $\mathbf{P} \subset \mathbf{R}_i \;\forall\; i \in N$. then $f(.)$ is DISC iff $f(.)$ is dictatorial on $f(\Theta)$.

3 Arrow's Impossibility Theorem

This famous impossibility theorem is due to Kenneth Arrow (1951), Nobel laureate in Economic Sciences in 1972. This result has shaped the discipline of social choice theory in many significant ways.

Before discussing this result, we first set up some relevant notation. Consider a set of agents $N = \{1, 2, \ldots, n\}$ and a set of outcomes X. Let $\succsim_i$ be a rational preference relation of agent i ($i \in N$). Subscript i in $\succsim_i$ indicates that the relation corresponds to agent i. For example, $\succsim_i$ could be induced by $u_i(., \theta_i)$ where θ_i is a certain type of agent i. Each agent is thus naturally associated with a set $\mathscr{R}_i$ of rational preference relations derived from the utility functions $u_i(., \theta_i)$ where $\theta_i \in \Theta_i$.

Given a rational preference relation $\succsim_i$, let us denote by $\succ_i$ the relation defined by

$$(x, y) \in \succ_i \ \text{ iff } \ (x, y) \in \succsim_i \ \text{ and } \ (y, x) \notin \succsim_i \ .$$

The relation $\succ_i$ is said to be the *strict total preference relation* derived from $\succsim_i$. Note that $\succ_i \ = \ \succsim_i$ if $\succsim_i$ itself is a strict total preference relation. Given an outcome set X, a strict total preference relation can be simply represented as an ordered tuple of elements of X. Given $\succsim_i$, let us denote by $\sim_i$ the relation defined by

$$(x, y) \in \sim_i \ \text{ iff } \ (x, y) \in \succsim_i \ \text{ and } \ (y, x) \in \succsim_i \ .$$

The relation $\sim_i$ is said to be the *indifference relation* derived from $\succsim_i$.

As usual $\mathscr{R}$ and $\mathscr{P}$ denote, respectively, the set of all rational preference relations and strict total preference relations on the set X. Let $\mathscr{A}$ be any nonempty subset of $\mathscr{R}^n$. We define a social welfare functional as a mapping from $\mathscr{A}$ to $\mathscr{R}$.

Definition 3 (Social Welfare Functional) *Given a set of agents $N = \{1, 2, \ldots, n\}$, an outcome set X, and a set of profiles $\mathscr{A}$ of rational preference relations of the agents, $\mathscr{A} \subset \mathscr{R}^n$, a social welfare functional is a mapping $W : \mathscr{A} \longrightarrow \mathscr{R}$.*

Note that a social welfare functional W assigns a rational preference relation $W(\succsim_1, \ldots, \succsim_n)$ to a given profile of rational preference relations $(\succsim_1, \ldots, \succsim_n) \in \mathscr{A}$.

Example 2 (Social Welfare Functional) Consider the example of the supplier selection problem discussed in Example 2.45, where $N = \{1, 2\}$, $X = \{x, y, z\}$, $\Theta_1 = \{a_1\}$, and $\Theta_2 = \{a_2, b_2\}$. Recall the utility functions:

$$\begin{aligned} u_1(x, a_1) &= 100; \ \ u_1(y, a_1) = 50; \ \ u_1(z, a_1) = 0 \\ u_2(x, a_2) &= 0; \ \ u_2(y, a_2) = 50; \ \ u_2(z, a_2) = 100 \\ u_2(x, b_2) &= 30; \ \ u_2(y, b_2) = 60; \ \ u_2(z, b_2) = 20. \end{aligned}$$

The utility function u_1 leads to the following strict preference relation:

$$\succsim_{a_1} = (x, y, z).$$

The utility function u_2 leads to the strict total preference relations:

$$\succsim_{a_2} = (z, y, x); \ \ \succsim_{b_2} = (y, x, z).$$

Let the set $\mathscr{A}$ be defined as

$$\mathscr{A} = \{(\succsim_{a_1}, \succsim_{a_2}), (\succsim_{a_1}, \succsim_{b_2})\}.$$

An example of a social welfare functional here would be the mapping W_1 given by

$$W_1(\succsim_{a_1}, \succsim_{a_2}) = (x, y, z); \quad W_1(\succsim_{a_1}, \succsim_{b_2}) = (y, x, z).$$

Another example would be the mapping W_2 given by

$$W_2(\succsim_{a_1}, \succsim_{a_2}) = (x, y, z); \quad W_2(\succsim_{a_1}, \succsim_{b_2}) = (z, y, x).$$

Note the difference between a social choice function and a social welfare functional. In the case of a social choice function, the preferences are summarized in terms of types and each type profile is mapped to a social outcome. On the other hand, a social welfare functional maps a profile of individual preferences to a social preference relation. Recall that the type of an agent determines a preference relation on the set X through the utility function.

We now define three properties of a social welfare functional: *unanimity* (also called *Paretian property*); *pairwise independence* (also called *independence of irrelevant alternatives* (IIA)), and *dictatorship*.

Definition 4 (Unanimity) *A social welfare functional $W : \mathscr{A} \longrightarrow \mathscr{R}$ is said to be unanimous if $\forall \ (\succsim_1, \ \ldots, \ \succsim_n) \in \mathscr{A}$ and $\forall x, y \in X$,*

$$(x, y) \in \ \succsim_i \ \ \forall i \in N \Longrightarrow (x, y) \in W_p(\succsim_1 \ldots, \succsim_n)$$

where $W_p(\succsim_1 \ldots, \succsim_n)$ is the strict preference relation derived from $W(\succsim_1 \ldots, \succsim_n)$.

The above definition means that, for all pairs $x, y \in X$, whenever x is preferred to y for every agent, then x is also socially preferred to y.

Example 3 (Unanimity) For the problem being discussed, let

$$W_1(\succsim_{a_1}, \succsim_{a_2}) = W_1((x, y, z), (z, y, x)) = (x, y, z)$$

$$W_1(\succsim_{a_1}, \succsim_{b_2}) = W_1((x, y, z), (y, x, z)) = (y, x, z).$$

This is unanimous because

- $(y, z) \in \ \succsim_{a_1}, (y, z) \in \ \succsim_{b_2}$, and $(y, z) \in W_1(\succsim_{a_1}, \succsim_{b_2})$; and
- $(x, z) \in \ \succsim_{a_1}, (x, z) \in \ \succsim_{b_2}$, and $(x, z) \in W_1(\succsim_{a_1}, \succsim_{b_2})$.

On the other hand, let

$$W_2((x, y, z), (z, y, x)) = (x, y, z); \ \ W_2((x, y, z), (y, x, z)) = (z, y, x)$$

Here $(y, z) \in \succsim_{a_1}$ and $(y, z) \in \succsim_{b_2}$ but $(y, z) \notin W_2(\succsim_{a_1}, \succsim_{b_2})$. So W_2 is not unanimous.

Definition 5 (Pairwise Independence) *The social welfare functional $W : \mathscr{A} \longrightarrow \mathscr{R}$ is said to satisfy pairwise independence if $\forall x, y \in X$, the social preference between x and y will depend only on the individual preferences between x and y. That is, $\forall x, y \in X$, $\forall(\succsim_1 \ldots, \succsim_n) \in \mathscr{A}$, $\forall(\succsim_1' \ldots, \succsim_n') \in \mathscr{A}$, with the property that*

$$(x, y) \in \succsim_i \ \ \Leftrightarrow (x, y) \in \ \succsim_i' \ \ \text{and} \ \ (y, x) \in \ \succsim_i \Leftrightarrow (y, x) \in \ \succsim_i' \ \ \forall i \in N,$$

we have that

$$(x, y) \in \ W(\succsim_1, \ldots, \succsim_n) \Leftrightarrow (x, y) \in W(\succsim_1', \ldots, \succsim_n'); \ \ \text{and}$$

$$(y, x) \in W(\succsim_1, \ldots, \succsim_n) \Leftrightarrow (y, x) \in W(\succsim_1', \ldots, \succsim_n').$$

Example 4 (Pairwise Independence) Consider the example as before and let

$$W_3(\succsim_{a_1}, \succsim_{a_2}) = W_3((x,y,z),(z,y,x)) = (x,y,z)$$

$$W_3(\succsim_{a_1}, \succsim_{b_2}) = W_3((x,y,z),(y,x,z)) = (y,z,x).$$

Here agent 1 prefers x to y in both the profiles while agent 2 prefers y to x in both the profiles. However in the first case, x is socially preferred to y while in the second case y is socially preferred to x. Thus the social preference between x and y is not exclusively dependent on the individual preferences between x and y. This shows that W_1 is not pairwise independent. On the other hand, consider W_3 given by

$$W_4((x,y,z),(z,y,x)) = (x,y,z)$$

$$W_4((x,y,z),(y,x,z)) = (z,x,y).$$

Now this social welfare functional satisfies pairwise independence.

The pairwise independence property is a very appealing property since it ensures that the social ranking between any pair of alternatives x and y does not in any way depend on other alternatives or the relative positions of these other alternatives in the individual preferences. Secondly, the pairwise independence property has a close connection to a property called the weak preference reversal property, which is quite crucial for ensuring dominant strategy incentive compatibility of social choice functions. Further, this property leads to a nice decomposition of the problem of social ranking. For instance, if we wish to determine a social ranking on the outcomes of a subset Y of X, we do not need to worry about individual preferences on the set $X \backslash Y$.

Definition 6 (Dictatorship) *A social welfare functional $W : \mathscr{A} \longrightarrow \mathscr{R}$ is called a dictatorship if there exists an agent, $d \in N$, called the dictator such that $\forall x, y \in X$ and $\forall (\succsim_1, \ldots, \succsim_n) \in \mathscr{A}$, we have*

$$(x,y) \in \succsim_d \Rightarrow (x,y) \in W_p(\succsim_1, \ldots, \succsim_n).$$

This means that whenever the dictator prefers x to y, then x is also socially preferred to y, irrespective of the preferences of the other agents. A social welfare functional that does not have a dictator is said to be nondictatorial.

Example 5 (Dictatorship) Consider the social welfare functional

$$W_5((x,y,z),(z,y,x)) = (x,y,z)$$

$$W_5((x,y,z),(y,x,z)) = (x,y,z).$$

It is clear that agent 1 is a dictator here. On the other hand, the social welfare functional

$$W_3((x,y,z),(z,y,x)) = (x,y,z)$$

$$W_3((x,y,z),(y,x,z)) = (y,z,x)$$

is not dictatorial.

Ideally, a social planner would like to implement a social welfare functional that is unanimous, satisfies the pairwise independence property, and is nondictatorial. Unfortunately, this belongs to the realm of impossible situations when the preference profiles of the agents are *rich*. This is the essence of the Arrow's Impossibility Theorem, which is stated next.

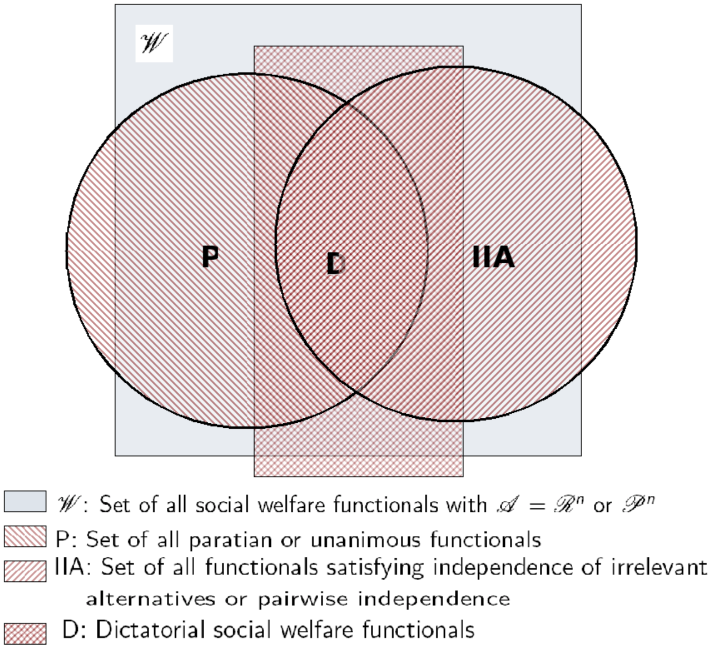

$\mathscr{W}$
P
D
IIA
$\mathscr{W}$: Set of all social welfare functionals with $\mathscr{A} = \mathscr{R}^n$ or $\mathscr{P}^n$
P: Set of all paratian or unanimous functionals
IIA: Set of all functionals satisfying independence of irrelevant alternatives or pairwise independence
D: Dictatorial social welfare functionals

Figure 2: An illustration of the Arrow's impossibility theorem

Theorem 2 (Arrow's Impossibility Theorem) *Suppose*

1. $|X| \geq 3$,
2. $\mathscr{A} = \mathscr{R}^n$ or $\mathscr{A} = \mathscr{P}^n$.

Then every social welfare functional $W : \mathscr{A} \longrightarrow \mathscr{R}$ *that is unanimous and satisfies pairwise independence is dictatorial.*

For a proof of this theorem, we refer the reader to proposition 21.C.1 of Mas-Colell, Whinston, and Green [3]. Arrow's Impossibility Theorem is pictorially depicted in Figure 2. The set P denotes the set of all Paretian or unanimous social welfare functionals. The set IIA denotes the set of all social welfare functionals that satisfy independence of irrelevant alternatives (or pairwise independence). The diagram shows that the intersection of P and IIA is necessarily a subset of D, the class of all dictatorial social welfare functionals.

The Gibbard–Satterthwaite theorem has close connections to Arrow's Impossibility Theorem. The property of unanimity of social welfare functionals is related to ex-post efficiency of social choice functions. The notions of dictatorship of social welfare functionals and social choice functions are closely related. The pairwise independence property of social welfare functionals has intimate connections with the DSIC property of social choice functions through the weak preference reversal property and monotonicity. We do not delve deep into this here; interested readers are referred to the book of Mas-Colell, Whinston, and Green [3] (Chapters 21 and 23).

References

[1] A. Gibbard. Manipulation of voting schemes. *Econometrica*, 41:587–601, 1973.

[2] M.A. Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorem for voting procedure and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[3] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Micoreconomic Theory*. Oxford University Press, 1995.