Gibbard-Satterthwaite theorem

September 17, 2018

Author's affiliations:

Pierre Bernhard
Biocore team, Université Côte d'Azur-INRIA, BP 93, 06902 Sophia Antipolis Cedex, France
pierre.bernhard@inria.fr

Marc Deschamps
CRESE EA3190, Univ. Bourgogne Franche-Comté, F-25000 Besançon, France
marc.deschamps@univ-fcomte.fr

Definition paragraph

One seminal question in social choice theory was: is it possible to find a social choice function such that each agent is always better off when telling the truth concerning his preferences no matter what the others report? In other words, can we find a strategy-proof voting rule? With at least three alternatives and two voters the answer is clearly no under a very general framework, as was proved independently by Allan Gibbard and Mark Satterthwaite. Since then, the Gibbard-Satterthwaite theorem is at the core of social choice theory, game theory and mechanism design.

1 Introduction

Since K. Arrow's 1951 analysis, which marks the revival of the theory of social choice, economists investigate from an axiomatic point of view the aggregation of individual preferences in order to obtain a social welfare function (i.e a complete and transitive ranking based on the individual preferences) or a social choice function (i.e one alternative from the individual preferences). Such questions concern huge domains of human beings, for example, the family's choice of the walls color in the living room, the choice of the Palme d'Or winner by the jury of the Cannes International Film Festival, or the vote for the President of the European Commission. Thus, in addition to his (im)possibility theorem, K. Arrow has initiated a new domain in economic analysis.

Since then, the question of the strategic behavior of individuals during an election was raised again on several levels. Indeed for a very long time, since we find for example a letter from Pline the Younger in Roman Antiquity, the question of manipulation had attracted attention. Whether playing on who will be a voter, who can be a candidate, the choice of the voting rule, abstention, beliefs about preferences or preferences expressed during the vote, the idea that at least one individual can do a manipulation to his advantage has interested and concerned many thinkers.

Concerning this last form of manipulation, it was historically mentioned at least seven times before the Gibbard-Satterthwaite theorem (an expression first used by [Schmeidler and Sonnenschein, 1978]). First, a story probably apocryphal tells that Borda responded to one of his critics who pointed out that his method was manipulable that it was intended for use by honest people. There is also a reference to this question in Ch. Dodgson (alias Lewis Carroll), who said in a specific voting system, "This voting principle makes an election more of a skill game than real test of voters' wishes" (see [Black, 1958]). In the modern period, [Black, 1948] discusses the link between unimodal preferences and strategic votes, while [Arrow, 1951, p. 7] explicitly states that he will not deal with this issue even though he later returns to it in a footnote [footnote 8, p. 80-81]. Arrow's general analysis, however, will lead [Vickrey, 1960, pp. 517-519] to conjecture that immunity to strategic manipulation is logically equivalent to the association of the axiom of independence of irrelevant alternatives and that of positive association. Finally, it will be R. Farquharson in his 1958 doctoral dissertation, published in 1969, [Farquharson, 1969], who will introduce the distinction between "sophisticated strategy" and "sincere strategy", and then in an article with M. Dummett in 1961 when they make the conjecture: "It seems unlikely that there is any voting procedure in which it can never be advantageous for any voter to vote 'strategically', i.e non sincerely" [Dummett and Farquharson, 1961, p. 34]. In an interview given in 2006 to R. Fara and M. Salles, M. Dummett confesses that he felt at this time that proving this conjecture would be extremely difficult and that is why they did not try the demonstration. We refer the reader to [Barberà, 2010] for more details on this historical part.

Thus, for more than twenty years after [Arrow, 1951], all scholars of social choice theory seem to have been convinced that the question of the manipulation of preferences by an individual (that is, the fact that he does not express his true preferences in order to lead to a social choice that satisfies him better than would have been obtained if he had been honest) was an important question, easy to explain, but very difficult to prove.

As an illustration of this question of the manipulation of preferences by an individual we can give the following example (from [Feldman, 1979, p. 459]):

Imagine a committee made up of 21 members each having one vote and whose actual preferences presented in descending order can be broken down into three groups.

Type 1	Type 2	Type 3
A	B	C
B	C	B
C	A	A
10 voters	9 voters	2 voters

In a majority election where voters indicate their real preferences, A gets 10 votes, B gets 9 votes and C gets 2 votes; this leads to the election of A. However, if anticipating this result, the voters in group 3 manipulate their preferences and vote for B, this will lead to 10 voters for A and 11 voters for B and so B will be elected. This strategic choice of the voters of the group 3 allows them to obtain a more favorable result than they would have obtained if they had voted sincerely. ∎

Therefore a question immediately comes to mind: is it possible to imagine a social choice function such that each individual, regardless of the others's choices, is always better off by expressing his true preferences? In other words, is there a social choice function such that always telling the truth is a strictly dominant strategy for each individual? The answer to this question was independently given by the philosopher Allan Gibbard [Gibbard, 1973] and the economist Mark Satterthwaite [Satterthwaite, 1975] (work summarizing part of his doctoral dissertation defended in 1973) and it is unfortunately negative. What we call since the Gibbard-Satterthwaite theorem can be broadly stated as follows: *it is not possible to find a social choice function that is both non-manipulable and non-trivial.*[1]

This last term deserves an explanation. By a trivial function we mean one of the following solutions (or a solution equivalent to it): 1/ a dictatorship (i.e all the power of decision resides with a single individual and he has no indifferent choice), 2/ the permanent choice of an alternative whatever the preferences expressed by the individuals are (which could, for example, correspond to a tradition which would be imposed in all circumstances), 3/ the perfect unanimity of all the

[1] This statement implies that impossibility holds as soon as the social choice function has at least three alternatives in its range. We will only prove a slightly weaker version, namely under the stronger assumption that no alternative is out of its range.

voters (which in fact means having perfect clones and therefore no diversity in the preferences), or 4/ the majority between two alternatives only, whatever the number of other alternatives existing and the votes that are expressed for them. Thus, if all individuals know their preferences and those of others and that there are at least three possible alternatives and at least two voters, there is no social choice function that implies that each individual always has an interest in expressing his or her real preferences. We thus find here the conditions of Arrow's theorem and the same conclusion since when there are only two alternatives, majority voting is at the same time a non-dictatorial and non-manipulable social choice function.

Since then, the Gibbard-Satterthwaite theorem has been considered, along with Arrow's theorem, as one of the two most famous results of social choice theory and has led to a very large literature in this field. It also plays a crucial role in public economics and in the theory of incentive mechanisms, which can be broadly seen as a social engineering approach of finding the rules to achieve a specific outcome from agents interacting strategically and having private information (see [Börger, 2015]). Indeed, because of this theorem, the incentives of individuals must be considered as relevant constraints in the design of any mechanism.

2 Gibbard-Satterthwaite theorem

Many proofs of this theorem have been proposed and it is possible to consider that they take one of the following four paths: 1/ that used by A. Gibbard and which uses Arrow's theorem, 2/ that used by M. Satterthwaite thanks to a combinatorial argument and recurrences on the number of individuals and alternatives, 3/ that considering this theorem as the consequence of the fact that non-manipulability requires strong monotony (see[Moulin, 1988]), and 4/ that developed by S. Barberá and his coauthors using the concept of pivotal agents. For a first presentation of the question of manipulation we refer to [Feldman, 1979] and for a review of this literature we refer to [Sprumont, 1995] and [Barberà, 2010].

Following Gibbard's approach, we will prove the Gibbard-Satterthwaite theorem as a corollary of Arrow's (im)possibility theorem. The presentation we retain will distinguish the formal setup, the links between the two theorems, the proof of the Gibbard-Satterthwaite theorem, and that Arrow's theorem in turn can be seen as a corollary if we directly prove the Gibbard-Satterthwaite theorem.

2.1 Formal set-up

Let $\mathcal{A} = \{a, b, \ldots\}$ be a set of three or more *alternatives*. Let P be the set of linear orders over $\mathcal{A}$, and $\mathcal{P} = P^n$. An element $\Pi = (\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_n) \in \mathcal{P}$ is called a *profile*, and the P_i the *individual preferences*. The order in P_i is denoted $a \succ_i b$ for "player i prefers a to b". We further define

Definition 1 *The* domination set *of $a \in \mathcal{A}$ in P_i*

$$\mathcal{D}(a, \mathsf{P}_i) = \{x \in \mathcal{A} \mid a \succ_i x\}$$

Moreover, for $\Pi = (\mathsf{P}_1, \dots, \mathsf{P}_n)$, $\mathcal{D}(a, \Pi)$ stands for the product set of the $\mathcal{D}(a, \mathsf{P}_i)$. Thus, let $\Pi' = (\mathsf{P}'_1, \dots, \mathsf{P}'_n)$ be another profile, the notation $\mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)$ means: $\forall i \in \{1, \dots, n\}\,,\ \mathcal{D}(a, \mathsf{P}'i) \supseteq \mathcal{D}(a, \mathsf{P}_i)$.

Definition 2

- *A* social welfare function *(or SWF) is an application $f : \mathcal{P} \to P$.*
- *A* social choice function *(or SCF) is an application $F : \mathcal{P} \to \mathcal{A}$.*

The order relation in $f(\Pi)$ will be denoted $a > b$ or, if needed $a >_{f(\Pi)} b$. We need to define the following properties of SWF or SCF:

Definition 3

- *The SWF f is* Pareto efficient *(or* satisfies the unanimity rule*) if*

$$[\forall i \le n\,,\ a \succ_i b] \Rightarrow a > b\,.$$

- *The SCF F is* Pareto efficient, *(or* satisfies the unanimity rule*) if whenever an alternative $a \in \mathcal{A}$ is the most preferred alternative in all individual preferences, it results that $F(\Pi) = a$.*
- *The SWF f is* independent of irrelevant alternatives *(IIA) if, for any a and b in $\mathcal{A}$, the relative ranking of a and b in $f(\Pi)$ only depends on their relative rankings in the P_i, irrespective of the rankings of other alternatives.*
- *The SCF F is* monotonic *if, given two profiles Π and Π',*

$$[F(\Pi) = a \text{ and } \mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)] \Rightarrow F(\Pi') = a\,.$$

- *The SCF F is* strategy-proof *if, when Π and Π' differ only in changing P_i into P'_i, it results that either $F(\Pi) = F(\Pi')$ or $F(\Pi) \succ_i F(\Pi')$.*
- *The SWF f is called* dictatorial *if there exists a player k such that, $\forall \Pi \in \mathcal{P}$, $f(\Pi) = \mathsf{P}_k$. (The social preferences are always player k' preferences.)*
- *The SCF F is called* dictatorial *if there exists a player k such that, $\forall \Pi \in \mathcal{P}$, $\forall x \neq F(\Pi)$, $F(\Pi) \succ_k x$. ($F(\Pi)$ is player k's most preferred alternative.)*

2.2 Arrow and Gibbard-Satterthwaite theorems

Theorem 1 (Arrow) *If a SWF is Pareto efficient and IIA, then it is dictatorial.*

Theorem 2 (Gibbard-Satterthwaite) *If a SCF is strategy-proof and onto (i.e. its range is all of $\mathcal{A}$: $\forall a \in \mathcal{A}$, $\exists \Pi \in \mathcal{P}$ such that $F(\Pi) = a$), it is dictatorial.*

Lemma 1 (Muller and Satterthwaite) *If a SCF is strategy-proof and onto, it is monotonic and Pareto efficient.*

Proof Let Π and Π' be two profiles, $F(\Pi) = a$, $\mathcal{D}(a, \Pi') \supseteq \mathcal{D}(a, \Pi)$, but assume that $F(\Pi') \neq a$. Make the change from Π to Π' one player at a time in numeric order. Denote by Π_k the profile obtained after changing P_k to P'_k. (And $\Pi_0 = \Pi$). At some point, we have that $F(\Pi_{i-1}) = a \neq b = F(\Pi_i)$. By strategy-proofness, it follows that $a \succ_i b$ in P_i while $b \succ_i a$ in P'_i. But by hypothesis, if $a \succ_i b$ in P_i it is a fortiori true in Π'. A contradiction. Therefore the SCF is monotonic.

Because F is assumed to be an onto function, for any given $a \in \mathcal{A}$, there exists a profile Π such that $F(\Pi) = a$. Build Π' by moving a at the top of the preferences of all players. By monotonicity, it still holds that $F(\Pi') = a$. Now, get Π'' by shuffling at will the preferences of all players *below* a, leaving a at their top position. Because of the definition of monotonicity, and specifically its IIA-like character, it still holds that $F(\Pi'') = a$. Therefore the SCF is Pareto efficient. ∎

2.3 Proof of the Gibbard-Satterthwaite theorem

This note is devoted to the proof of the Gibbard-Satterthwaite theorem viewed as a corollary of Arrow's theorem. We assume therefore that the latter is known. Given the above lemma 1, we need to prove

Lemma 2 *If a SCF is Pareto efficient and monotonic, it is dictatorial.*

Or method of proof will be as follows: we assume that a SCF F is known which is Pareto efficient and monotonic. From it we construct a SWF f which we show to be Pareto efficient and IIA, therefore dictatorial, which will imply that F is dictatorial.

Let F be a SCF. Given a profile Π, our social ordering $f(\Pi)$ is built via the following algorithm:

Algorithm

- Let $\Pi_1 = \Pi$.
- For i ranging from 1 to n, do:

  - define $a_i = F(\Pi_i)$,
  - define Π_{i+1} by moving a_i at the bottom of all individual preferences in Π_i.

- Define $f(\Pi)$ as: for all $i \in \{1, \ldots, n-1\}$, $a_i > a_{i+1}$.

Proposition 1 *If the SWF F is Pareto efficient and monotonic, the above algorithm produces a linear order on $\mathcal{A}$.*

Proof What we need to prove is that all elements of $\mathcal{A}$ will be ranked, i.e. that for all $i \in \{1, \ldots, n\}$, and for all $j < i$, $a_i \neq a_j$.

Let first $i < n$. Therefore, not all alternatives have been numbered as one of the a_k. Define Π'_i as follows: Take an alternative b which has not yet been ranked. Raise it at the top of all individual preferences. This does not change the domination sets: $\mathcal{D}(a_j, \Pi'_i) = \mathcal{D}(a_j, \Pi_i)$ since in Π_i, all the alternatives that have already been selected in the algorithm, including a_j (remember that $j < i$), are stacked at the bottom of all individual preferences. Therefore, by monotonicity, if $F(\Pi_i) = a_j$, it also holds that $F(\Pi'_i) = a_j$. But by Pareto efficiency, $F(\Pi'_i) = b$, a contradiction.

Finally, in Π_n, $n-1$ different alternatives have been placed at the bottom of all individual preferences, thus the same and last alternative is alone at the top of all, and is therefore selected by F by Pareto efficiency.

The following propositions end our proof:

Proposition 2 *If the SCF F is Pareto efficient and monotonic, the SWF defined by the algorithm is*

1. *Pareto efficient,*

2. *independent of irrelevant alternatives (IIA),*

and therefore dictatorial by Arrow's theorem.

Proof

1. Let $a, b \in \mathcal{A}$, and assume that in Π, $\forall i \in \{1, \ldots, n\}$, $a \succ_i b$. Assume that at some step of our algorithm, $F(\Pi_i) = b$, but a has not yet been chosen. In Π_i, raise alternative a at the top of all individual preferences. This does not change the domination sets of b in any individual preferences. Therefore, F should still select b. But by Pareto optimality, it should select a. A contradiction. ∎

2. Let $i < j$ and therefore by our algorithm, $a_i >_{f(\Pi)} a_j$. In Π, move another alternative b in some individual preferences, call the new profile Π'. We claim that in the order created by our algorithm applied to Π', it still holds that $a_i > a_j$.

   Assume that $F(\Pi') = b$. Then at step 2, b is brought at the bottom of all individual preferences, and, as compared to profile Π, the domination sets of a_1 have been enlarged or kept unchanged for all individual preferences. Therefore $F(\Pi'_2) = a_1$. If $b \neq a_2$, at step 2, for the same reason a_2 will be selected, and so forth until the end of the algorithm. Therefore, we will still select a_i before a_j.

   Assume $F(\Pi') = c \neq b$, and assume $c \neq a_1$. This is possible only if in one individual preferences at least, b has been moved from below a_1 to above it. In these individual preferences only, bring b back below a_1. Call this profile Π''. $\mathcal{D}(a_1, \Pi'') \supseteq \mathcal{D}(a_1, \Pi)$. Therefore $F(\Pi'') = a_1$. But in going from Π' to Π'', b has only been moved down in some individual preferences. Therefore $\mathcal{D}(c, \Pi'') \supseteq \mathcal{D}(c, \Pi')$. Therefore $F(\Pi'') = c$. Hence $a_1 = c$ contrary to the hypothesis. Hence $F(\Pi') = a_1$.

   Repeat this argument at each step before b is selected, and once this happens, repeat the previous argument. It follows that, except for b, all other alternatives are chosen in the same order as for Π. ∎

It follows that, by Arrow's theorem, f is dictatorial. There is a player k such that for all Π, $f(\Pi) = \mathsf{P}_k$, and in particular $F(\Pi)$ is player k's preferred alternative. And this proves lemma 2, and consequently the Gibbard-Satterthwaite theorem. ∎

2.4 Complement: Arrow's theorem as a corollary of Gibbard Satterthwaite

It can also be shown that, if the Gibbard-Satterthwaite theorem has been proved directly, Arrow's theorem is an easy corollary, thus establishing the equivalence between the two results. The process is symmetrical from the above one, deriving a SCF from a Pareto efficient and IIA SWF by simply picking the top alternative in the social preferences, and showing that it is onto and strategy-proof.

We might also mention that Lemma 1 has an easy reciprocal.

3 Conclusion

Since its demonstration, the Gibbard-Satterthwaite theorem has generated a huge literature on the question of the manipulation of preferences, distinguishing, in particular, the cases where the social choice function is manipulable by an individual from the case where it is manipulable by a coalition of individuals. It has also been extended to take into account set-valued ([Duggan and Schwartz, 2000]) and non-deterministic choice functions ([Gibbard, 1977]), that is those where the result depends on the votes of individuals but also by chance. (Often both set-valued and non-deterministic.)

In our opinion, three main ways of circumventing this theorem have been followed. The first is to restrict the preference domain (with functions defined on restricted sets of preference profiles, it is possible to find social choice functions that are both non-dictatorial and non-manipulable, e.g unimodal preferences *à la* Black). The second is to change the goal. Indeed, the framework in which the Gibbard-Satterthwaite theorem is situated is very strong since it seeks a social choice function for which telling the truth is a dominant strategy for each individual. The path followed by implementation theory is to simply ask that it be a Nash strategy, or a perfect subgame strategy, or a Bayesian Nash strategy, depending on the informational context of the individuals. Finally, more recently, a third way explain that this problem is real but may not be very important empirically. Indeed, besides the integrity, ignorance or stupidity of individuals that can prevent them from performing manipulations, the fact that a social choice function is manipulable does not imply that it will be manipulated. And since [Bartholdi et al., 1989], economists consider that it may be empirically impossible for individuals to decide how to manipulate even when they have all the information to do so, as the problem may be NP-hard.

Cross References

Arrow (im)possibility theorem

Electoral systems

Heterarchy

Liberty

Political competition

Political economy

Politicians

Public goods

Public interest

Rationality

Simple majority

References

[Arrow, 1951] Arrow, K. J. (1951). *Social Choice and Individual Values*. Wiley.

[Barberà, 2010] Barberà, S. (2010). Strategy-proof social choice. Barcelona Economic Working Paper 420, University of Barcelona.

[Bartholdi et al., 1989] Bartholdi, J., Tovey, C., and Trick, M. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241.

[Black, 1948] Black, D. (1948). On the rationale of group decision making. *Journal of Political Economy*, 56:23–34.

[Black, 1958] Black, D. (1958). *The Theory of Committees and Elections*. Cambridge University Press.

[Börger, 2015] Börger, T. (2015). *An Introduction to the Theory of Mechanism Design*. Oxford University Press.

[Duggan and Schwartz, 2000] Duggan, J. and Schwartz, T. (2000). Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized. *Social Choice and Welfare*, 17:86–93.

[Dummett and Farquharson, 1961] Dummett, M. and Farquharson, R. (1961). Stability in voting. *Econometrica*, 29:33–44.

[Farquharson, 1969] Farquharson, R. (1969). *Theory of voting*. Yale university Press.

[Feldman, 1979] Feldman, A. M. (1979). Manipulating voting procedures. *Economic Inquiry*, 17:452–474.

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

[Gibbard, 1977] Gibbard, A. (1977). Manipulation of schemes that mix voting with chance. *Econometrica*, 45:665–682.

[Moulin, 1988] Moulin, H. (1988). *Axioms of cooperative decision making*. Economic Society Monographs. Cambridge University Press.

[Satterthwaite, 1975] Satterthwaite, M. (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.

[Schmeidler and Sonnenschein, 1978] Schmeidler, D. and Sonnenschein, H. (1978). Two proofs of the gibbard-satterthwaite theorem on the possibility of a strategy-proof social choice function. In Gottinger, H. W. and Leinfellner, W., editors, *Decision Theory and Social Ethics*, pages 227–234. Reidel Publishing Company. Published version of a 1974 working paper.

[Sprumont, 1995] Sprumont, Y. (1995). Strategy-proof collective choice in economic and political environments. *Canadian Journal of Economics*, 28:68–107.

[Vickrey, 1960] Vickrey, W. (1960). Utility, strategy and social decision rules. *Quarterly Journal of Economics*, 74:507–535.