A Computer-Aided Proof to Gibbard-Satterthwaite Theorem

Pingzhong Tang
Fangzhen Lin
Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

May 10, 2008

Abstract

The Gibbard-Satterthwaite Theorem is a landmark result in both social choice theory and implementation theory, as it bridges normative and strategic analysis of voting problems. The theorem states that any social choice function that is strategy-proof and onto is also dictatorial. In this paper, we provide a computer-aided inductive proof for the theorem. We first show that this result holds if and only if it holds for the base case where there are exactly 2 agents and 3 alternatives. We then verify the base case using a computer program. This Proof strategy turns out to be very general. We have used it to prove Arrow's impossibility theorem, and we believe it can be used to prove other theorems as well.
Keyword: Social Choice Theory, Gibbard-Satterthwaite Theorem
JEL Classification: D71

1 Gibbard-Satterthwaite Theorem

A voting model is a tuple (N, O), where N is a finite set of individuals (agents) and O a finite set of outcomes (alternatives). An agent's preference ordering is a linear ordering of O, and a preference profile $>$ of (N, O) is a tuple $(>_1, ..., >_n)$, where $>_i$ is agent i's preference ordering, and n the size of N. In the following, when N is clear from the context, we also call $>$ a preference profile of O. Similarly, when O is clear from the context, we also call it a preference profile of N.

Definition 1 *Given a voting model (N, O), a* social choice function *is a function $W : L^n \rightarrow O$, where L is the set of linear ordering of O, and n the size of N.*

In this way, a social choice function defines a winning outcome for each preference profile. The first assumption about a social choice function is that of *onto*.

Definition 2 *A social choice function C is an* onto *if for each $a \in O$, there exists a $>\in L^n$ such that $C(>) = a$.*

In other words, this assumption guarantees that every alternative has a chance to win. Before we get to the second assumption called *strategy-proof*, we need another concept called *manipulation*.

Definition 3 *A social choice function C is* manipulable *at profile $>$ by individual i via $>'$ if $C(>_{-i}, >'_i) >_i C(>)$, where $(>_{-i}, >'_i)$ is the profile resulting from replacing $>_i$ with $>'_i$ in $>$.*

Manipulability is a typical feature of strategic voting in contrast with normative voting where agents always report the truth. If a social choice function C is manipulable, there is always a state where some agent is better off by lying about his true preference, thus the resulting outcome may not truthfully represent a social choice. Manipulability is not a desirable property and should be precluded from the picture of a rational social choice function.

Definition 4 *A social choice function is* strategy-proof *if it is not manipulable by any individual at any profile.*

Strategy-proof is sometimes said to be *dominant strategy incentive compatible*, because if a choice function is strategy-proof, then every agent always has the incentive to report his true preference in the sense that no matter what the other agents report, he is better off to report his true preference.

Definition 5 *An agent i is a* dictator *in a social choice function C if C always selects i's top choice: for every preference profile $>$, $C(>) = a$ iff for all $a' \in O$ that is different from a, $a >_i a'$. C is* non-dictatorial *if it has no dictator.*

Theorem 1 ***(Gibbard-Satterthwaite Theorem)(1973; 1975)*** *For any voting model (N, O) such that $|O| \geq 3$, any social choice function that is strategy-proof and onto is also dictatorial.*

For the purpose of this paper, we also introduce another property called *unanimity*.

Definition 6 *A social choice function is* unanimous *if for each $a \in O$, if $a >_i a'$ for each i and for each $a' \in O$ that is different from a, then $C(>) = a$.*

Unanimity means, if all the agents consider a the best alternative, a should be chosen.

2 An inductive Proof of Gibbard-Satterthwaite's Theorem

For ease of presentation, we assume the following notations.

- For any set S, we use S_{-a} to denote $S \setminus \{a\}$, i.e. the result of deleting a in S.
- We extend the above notation to tuples as well: if $t = (t_1, ..., t_n)$, then we use t_{-i} denotes the tuple $(t_1, ..., t_{i-1}, t_{i+1}, ..., t_n)$. Furthermore, we use (t_{-i}, s) to denote the result of replacing ith item in t by s: $(t_{-i}, s) = (t_1, ..., t_{i-1}, s, t_{i+1}, ..., t_n)$. We use $t_{-\{i,j\}}$ to denote $(t_{-i})_{-j}$.
- If $>$ is a linear ordering of O, and $a \in O$, then we let $>_{-a}$ be the restriction of $>$ on O_{-a}: for any $a', a'' \in O_{-a}$, $a' >_{-a} a''$ iff $a' > a''$. On the other hand, if $>$ is a linear ordering of O_{-a} for some $a \in O$, then we let $>^{+a}$ to be the extension of $>$ to O such that for any $a' \in O_{-a}$, $a' >^{+a} a$. Similarly, we let $>^{a+}$ to be the extension of $>$ to O such that for any $a' \in O_{-a}$, $a >^{a+} a'$. Thus if $>$ is a linear ordering of O, and $a \in O$, then $>^{+a}_{-a}$ is $(>_{-a})^{+a}$, i.e. the result of moving a to the bottom of the ordering. These notations extend to tuples of orderings. Thus if $>$ is a preference profile of (N, O_{-a}), then
$$>^{+a} = (>_1, ..., >_n)^{+a} = (>_1^{+a}, ..., >_n^{+a}),$$
which will be a preference profile of (N, O). Similarly for $>^{a+}$

2.1 The Inductive Step

The inductive step consists of the following two lemmas.

Lemma 1 *If there is a social choice function for n individuals and $m+1$ alternatives that is onto, strategy-proof and non-dictatorial, then there is a social choice function for n individuals and m alternatives that satisfies these three conditions as well, for all $n \geq 2, m \geq 3$.*

Proof: Let (N, O) be a voting model such that $|N| = n$ and $|O| = m + 1$, and C a social choice function that satisfies the three conditions in the lemma. For any $a \in O$, we define C_a to be a social choice function that is the "restriction" of C on O_{-a}: for any preference profile $>$ of O_{-a}, $C_a(>) = C(>_{+a})$.

- We first show that C_a is well defined. That is, $C_a(>) \neq a$ for any $>$ on O_{-a}. Suppose $C_a(>) = C(>_{+a}) = a$ for some $>$. We have $C(>_{-i}, >'_i) = a$ for any agent i and any ordering $>'_i$, otherwise i can manipulate at $>$ via $>'_i$. Similarly, we have $C(>_{-\{i,j\}}, >'_i, >'_j) = a$, otherwise j can manipulate at $(>_{-i}, >'_i)$ via $>'_j$. We continue the above argument until we get $C(>') = a$ for any $>'$, which leads to the contradiction that C is an onto.

- We now show that C_a is strategy-proof. Suppose otherwise: some agent i can manipulate at $>$ via $>'_i$ in C_a. This is equivalent of saying i can manipulate at $>^{+a}$ via $>'^{+a}$, which leads to the contradiction that C is not strategy-proof.

- Instead of showing C_a is an onto, we prove C and C_a are unanimous, which implies onto by letting each alternative on top of everyone's preference. Since C is an onto, we can find some $>$ for each b such that $C(>) = b$. We have $C(>_{-i}, (>'_i)^{b+}_{-b}) = b$. In words, by moving b to the top of any $>'_i$, the resulting choice will still be b. Otherwise, i can manipulate at $(>_{-i}, (>'_i)^{b+}_{-b})$ via $>_i$. We continue the argument until we get $C((>')^{b+}_{-b}) = b$. Therefore C is unanimous and so is C_a by definition.

- We finally show that an alternative a can be chosen properly so that C_a is non-dictatorial. We proceed in two steps:

  1. We now prove that for any two distinct alternative c, d, C_c and C_d have the same dictator, if they have one. Suppose otherwise, that is, C_c has dictator i, C_d has dictator j and $i \neq j$. We consider the following profile $>$, where we have
     - $a >_s b >_s ... >_s d >_s c, for\ \ all\ \ s \neq j$
     - $b >_j a >_j ... >_j d >_j c, for >_j$

     Obviously, $C(>) = a$ because C_c has dictator i. Now we change $>_i$ into $a >'_i b >'_i ... >'_i c >'_i d$, then we still have $C(>_{-i}, >'_i) = a$. Otherwise i can manipulate at $(>_{-i}, >'_i)$ via $>_i$. We continue the same argument until we get $C(>'_{-j}, >_j) = a$. However, there is a manipulate for j at $(>'_{-j}, >_j)$ via $>'_j$: $\ b >'_j a >'_j ... >'_j c >'_j d$. So contradiction occurs.
  2. Now suppose C_a has a dictator for each $a \in O$, by (1) above, $\forall a \in O\ \ C_a$ has the same dictator i. Since C is non-dictatorial, there exists some $>$ so that $C(>) = a_1$ is different from the topmost alternative a_2 of $>_i$. Now consider $>' = (>^{a_1+}_{-a_1})^{+a_3}_{-a_3}$, where a_3 is distinct from a_1 and a_3. We can continue to change $>_s$ to $>'_s$ for all $s \neq i$ without changing the social value until we get $C(>'_{-i}, >_i) = a_1$. Now we have a manipulation for agent i at $(>'_{-i}, >_i)$ via $(>_i)^{+a_3}_{-a_3}$, which contradicts to the strategy-proof property of C.

  Therefore, there exists some a such that C_a is non-dictatorial.

■

Lemma 2 *If there is a social choice function for $n+1$ individuals and m alternatives that is onto, strategy-proof and non-dictatorial, then there is a social choice function for n individuals and m alternatives that satisfies these three conditions as well, for all $n \geq 2, m \geq 3$.*

Proof: Let (N, O) be a voting model such that $|N| = n+1$ and $|O| = m$, and C a social choice function that satisfies the three conditions in the lemma. For any pair of agents $i \neq j \in N$, we define $C_{i,j}$ to be the following social welfare function for (N_{-i}, O): for any preference profile $>$ of (N, O), $C_{i,j}(>_{-i}) = C(>_{-i}, >_j)$.

- Again the onto property of $C_{i,j}$ follows from the unanimity of C as shown in the proof of lemma 1.
- For the strategy-proofness of $C_{i,j}$, suppose otherwise that $C_{i,j}(>_{-k}, >'_k) >_k C_{i,j}(>)$ for some $>$, $>'_k$ and k. There are two cases:
  1. If $k \neq j$, then C is not strategy-proof since $C_{i,j}(>_{-k}, >'_k) = C(>_{-k}, >'_k, >_j) >_k C(>, >_j) = C_{i,j}(>)$[1]. A contradiction.
  2. If $k = j$, by definition of $C_{i,j}$, we have $a_1 = C(>_{-j}, >'_j, >'_j) >_j C(>_{-j}, >_j, >_j) = a_2$. Now consider $C(>_{-j}, >_j, >'_j) = a_3$.
     - If $a_1 >_j a_3$, then j can manipulate at $(>_{-j}, >_j, >'_j)$ via $>'_j$
     - If $a_3 \geq_j a_1$, then i can manipulate at $(>_{-j}, >_j, >_j)$ via $>'_j$ (since i, j have the same preference in this profile).

     Either way, there is a contradiction to the strategy-proofness of C.

  So $C_{i,j}$ is strategy-proof.
- We prove in the following that we can find two distinct agents i, j such that $C_{i,j}$ is non-dictatorial. Suppose not, then for every pair of agents i, j, there is an agent $d_{i,j}$ that is a dictator of $C_{i,j}$. We first show that $d_{i,j} = j$ for any i, j. Suppose otherwise, $d_{i,j} = k \neq j$. Since C is non-dictatorial, we can find a profile $>$ such that $a = C(>) \neq b$ where b is on top of $>_k$. We then still have $C(>_{-i}, (>_i)^{a+}_{-a}) = a$, otherwise there would be a manipulation for i at $>' = (>_{-i}, (>_i)^{a+}_{-a})$ via $>_i$. Similarly, we have $C(>'_{-j}, >'_i) = a$. Now we have a contradiction, since according to $d_{i,j} = k$, we would have $C(>'_{-j}, >'_i) = b$.

  Now consider a profile any $>$ on $(N+1, O)$, any triple of agents i, j, k and any triple of alternatives (this is possible since $|N+1| \geq 3$, $|O| \geq 3$) where
  - $a >_i c >_i b >_i$ ... for $>_i$
  - $c >_j b >_j a >_j$ ... for $>_j$
  - $b >_k a >_k c >_k$ ... for $>_k$

  Notice that $>_i, >_j, >_k$ only differ in $\{a, b, c\}$. There are the following cases:
  - $C(>) = a$, then we change $>_j$ to $>_k$ and denote the new profile $>'$. Then,

[1] According to the definition of $C_{i,j}$, $(>_{-k}, >'_k, >_j)$ is a preference profile on $|N| = n+1$ by adding $>_j$ as agent i's preference to $(>_{-k}, >'_k)$, which is on $N \setminus \{i\}$; similarly for $(>, >_j)$ and others.

1. If $C(>') \neq b$, then $d_{j,k} \neq k$
2. If $C(>') = b$, then j can manipulate at $>$ via $>'_j$

– Other cases where $C(>) = b, c$ or other alternatives are similar to the case above.

Therefore, we conclude that there are two distinct agents i, j such that $C_{i,j}$ is non-dictatorial.

■

2.2 The Base Step

We use computer programs to verify this part. The base case says that when $|N| = 2$ and $|O| = 3$, there is no social choice function on (N, O) that is onto, strategy-proof, and non-dictatorial. A straightforward way is to generate all possible social choice functions in (N, O) and check all of them one by one for these three conditions. However, there are too many such functions for this to be feasible on current computers: there are $3! = 6$ number of linear orderings of O, resulting in $6 \times 6 = 36$ total number of preference profiles of (N, O), and 3^{36} possible social choice functions.

Therefore, one should generate all the strategy-proof social choice functions instead of all the social choice functions. Notice that strategy-proof is really a inter-profile property. Therefore, we use the so-called *depth first search* strategy: we incrementally choose a social choice for each preference profile and backtrack when the current choice violates the strategy-proof property with the social choices that have already been chosen for other preference profiles. In this way, we finish generating a strategy-proof function when we have chosen a social choice for each preference profile.

We implement the above idea in a computer program which returns in less 1 second on a laptop with 2.0GHZ CPU and 512MB RAM. There are 17 strategy-proof social choice functions when $|N| = 2$ and $|O| = 3$. Among these 17 functions, 3 of them are constant functions which choose a single outcome for every preference profile, 12 of them have the ranges of size 2, the remaining 2 of them are onto which correspond to the dictatorships of two agents. In fact, all these 17 functions are dictatorial on their images (p256, (Arrow, Sen, & Suzumura 2002)).

This verifies the base case, thus completes our inductive proof of Gibbard-Satterthwaite Theorem.

3 Concluding Remarks

There are several existing proofs of the Gibbard-Satterthwaite theorem. Some of these proofs (Gibbard 1973) finds a connection between a social welfare function that satisfies Arrow's condition and a social choice function, thus by proving the strategy-proofness of the underlying social choice function, dictatorship follows from that of Arrow's impossibility theorem. Others (Barbera 1983; Reny 2001) prove them directly without using Arrow's theorem. Particularly, Sen (2001) also provides a semi-inductive proof in which he reduces the theorem to a base case where there are n agents and 3 outcomes. It turns out that it needs nearly as much effort to prove the base case as to prove the theorem itself. In our proof, we completely reduce the theorem to the base case where there are exactly 2 agents and 3 outcomes so that we can verify it using a simple computer program. Our proof follows the same structure of proof that we gave earlier for Arrow's theorem (Lin & Tang 2008). We believe this proof methodology provides new insights for using computers to uniformly prove other impossibility theorems as well as discover new theorems.

References

Arrow, K. J.; Sen, A. K.; and Suzumura, K., eds. 2002. *Handbook of Social Choice and Welfare*, volume 1 of *Handbook of Social Choice and Welfare*. Elsevier.

Barbera, S. 1983. Strategy-proofness and pivotal voters: A direct proof of the gibbard-satterthwaite theorem. *International Economic Review* 24(2):413–17.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41(4):587–601.

Lin, F., and Tang, P. 2008. Computer-aided proofs of arrow's and other impossibility theorems. In *In Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08)*, to appear.

Reny, P. J. 2001. Arrow's theorem and the gibbard-satterthwaite theorem: a unified approach. *Economics Letters* 70(1):99–105.

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

Sen, A. 2001. Another direct proof of the gibbard-satterthwaite theorem. *Economics Letters* 70(3):381–385.