Harvard CS 1360, Spring 2025
Instructor: Ariel Procaccia
Scribe: Taig Singh

Strategic Manipulation in Elections

Lecture 7

We will begin with a reminder of the voting model.

Definition 1 (The Voting Model)**.** The voting model includes

- Set of voters $N = \{1, \dots, n\}$ (assume $n \geq 2$).
- Set of alternatives A with $|A| = m$.
- Each voter has a ranking $\sigma_i \in L$ over the alternatives, where $x \succ_{\sigma_i} y$ means that voter i prefers x to y.
- A preference profile $\sigma \in L^n$ is a collection of all voters' rankings.
- A social choice function $f : L^n \to A$.

So far we have assumed voters are honest and represent their preferences as they truly are. However, voters can manipulate their votes strategically to achieve better outcomes.

Example 1 (Manipulation in the Borda Count)**.** Consider the following true preference profile

1	2	3
b	b	a
a	a	b
c	c	c
d	d	d

Here, b has 8 points, a has 7 points, and c and d are both weaker, and so b wins under the Borda Count. Now, what if voter 3 changes their vote, resulting in the following preference profile:

1	2	3
b	b	a
a	a	c
c	c	d
d	d	b

Now, a still has 7 points, but b has 6 points. Since c and d have even fewer points, a would be the winner of this election. Thus, voter 3 manipulated their vote and misrepresented their true rankings to result in a better outcome for themselves, because, in reality, they wanted a to win.

Seeing that his rule was easily manipulatable, Jean-Charles de Borda (1733–1799) said, "My rule is intended for honest men!"

Definition 2 (Strategyproofness)**.** Denote $\sigma_{-i} = (\sigma_1, \dots, \sigma_{i-1}, \sigma_{i+1}, \dots, \sigma_n)$. A social choice function f is strategyproof (SP) if a voter can never benefit from lying about their preferences:

$$\forall \sigma \in L^n, \forall i \in N, \forall \sigma_i' \in L, f(\sigma) \succeq_{\sigma_i} f(\sigma_i', \sigma_{-i}).$$

In other words, f is strategyproof if truthful reporting is a dominant strategy because no other strategy can guarantee a voter a better outcome.

Example 2 (When is Plurality Strategyproof?)**.** It turns out that plurality is only strategyproof for a maximum of 2 alternatives. For 2 alternatives, the only deviation a voter can do is flip their second favorite alternative with their top choice, which directly contributes to the second-favorite alternative's count, resulting in the same or a worse outcome for the voter. For 3 alternatives, however, if a voter's top choice has no chance of winning but their second and third choices are tied in plurality counts, that voter is better off misreporting their true preference by reporting their second choice as their top choice to ensure their second choice wins over their third choice.

Theorem 1 (Gibbard-Satterthwaite Theorem)**.** *Let $m \geq 3$. A social choice function f is SP and onto A if and only if f is dictatorial.*

By “onto,” we mean that for any alternative, a strategy profile exists that will result in the alternative getting selective. By “dictatorial,” we mean that there is a voter that always get their top choice, no matter what other voters want. This theorem implies that any voting rule that is onto and nondictatorial is manipulable.

For $m \geq 3$, all common rules are onto and nondictatorial. What if we enforce SP and nondictatorial but relax the onto constraint? The rule that says “regardless of anyone’s votes, pick alternative a” is SP and nondictatorial, because no one’s votes matter. Further, the rule that says “in advance, we decide that either a or b wins, and we’ll take the majority between them” is also SP and nondictatorial.

Proof Sketch of G-S. If f is dictatorial, then trivially f is onto because the dictator can rank any alternative first, and f is also SP because the dictator is best-off telling their true preference because it will be the result of the election, and the reported preferences of all other voters do not matter. Proving the other direction of the G-S Theorem is far more interesting.

We will start with the following lemmas:

- Strong monotonicity: If f is a SP function and σ is some strategy profile where $f(\sigma) = a$, then $f(\sigma') = a$ for all profiles σ' such that $\forall x \in A, i \in N : a \succ_{\sigma_i} x \implies a \succ_{\sigma'_i} x$.
- Unanimity: If f is a SP and onto function, and σ is a strategy profile where $a \succ_{\sigma_i} b$ for all $i \in N$, then $f(\sigma) \neq b$.

Strong monotonicity can be interpreted as the following: if f is strategyproof then if the outcome of some strategy profile is a, then if you make a new strategy profile where everything that was ranked below a in voter preferences stays ranked below a (regardless of where the alternatives above a are ranked), then the outcome of the election will still be a. Unanimity can be interpreted as the following: If f is strategyproof and onto, then if everyone ranks a above b then b will not win the election.

Let us also assume that $m \geq n$ and neutrality, that states that for any permutation of alternatives $\pi : A \to A$, then

$$f(\pi(\sigma)) = \pi(f(\sigma))$$

That is to say, if a was the winner of an election and we permute the alternatives in the strategy profile such that some alternative b takes a’s place in every voter’s ranking, then the winner from this new strategy profile will be b. Note that neutrality and $m \geq n$ are assumptions that are not necessary for the G-S Theorem but they make the proof more simple.

We will even further simplify the proof for $m = 5$ and $n = 4$. You should be able to generalize for $m \geq n$ from here. Assume that f is SP and onto. Consider the profile:

$$\sigma = \begin{array}{c|c|c|c} 1 & 2 & 3 & 4 \\ \hline a & b & c & d \\ b & c & d & a \\ c & d & a & b \\ d & a & b & c \\ e & e & e & e \end{array}$$

Note here that all voters rank e last, and alternatives a through d are ranked in a cyclic fashion across the voters. By unanimity, it is clear that e is not the winner. Since alternatives a through d are all symmetric across the preferences, WLOG assume $f(\sigma) = a$. Now, consider the following strategy profile, created matching the rankings of voters 2 and 3 to the ranking of voter 4:

$$\sigma^1 = \begin{array}{c|c|c|c} 1 & 2 & 3 & 4 \\ \hline a & d & d & d \\ d & a & a & a \\ b & b & b & b \\ c & c & c & c \\ e & e & e & e \end{array}$$

By strong monotonicity, it follows that $f(\sigma^1) = a$. Strong monotonicity applies because the preferences of voters 2 and 3 are the only preferences that changed, and for each of them the alternatives ranked below a in σ are still below a in σ^1. Now, consider the following strategy profile constructed by changing voter 2's ranking by dropping a to their least favorite preference:

$$\sigma^2 = \begin{array}{c|c|c|c} 1 & 2 & 3 & 4 \\ \hline a & d & d & d \\ d & b & a & a \\ b & c & b & b \\ c & e & c & c \\ e & a & e & e \end{array}$$

Note that in this profile, alternative d dominates alternatives b, c, and e, so the winner cannot be either of these three alternatives. Further, since f is SP, the winner cannot be d because otherwise voter 2 would have had a useful misrepresentation if σ_2^1 was their true preference, as they would prefer d over a and by misrepresenting to σ_2^2 they could change the result from a to d. Thus, from the process of elimination, we conclude that $f(\sigma^2) = a$. Similarly, if we construct σ^3 by changing voter 3's ranking by dropping a to their least favorite preference, we will get $f(\sigma^3) = a$ by the same rationale. If we construct σ^4 by further editing voter 4's preferences, we will get $f(\sigma^4) = a$ where

$$\sigma^4 = \begin{array}{c|c|c|c} 1 & 2 & 3 & 4 \\ \hline a & d & d & d \\ d & b & b & b \\ b & c & c & c \\ c & e & e & e \\ e & a & a & a \end{array}$$

From here, we claim that strong monotonicity implies that $f(\sigma') = a$ for every σ' where voter 1 ranks a first. That is because if voter 1 ranks a first in σ', then all alternatives that were ranked below a in σ^4 are ranked below a in σ' (because no alternatives are ranked below a in the preferences of voters 2, 3, or 4), and so strong monotonicity says that $f(\sigma') = a$ as well. We now claim that neutrality implies that voter 1 is a dictator. This is because for any strategy profile σ', if voter 1 ranks alternative x first in σ', then we can construct a permutation π that swaps x with a. Since we have just shown that $f(\pi(\sigma')) = a$ (voter 1 ranks a first in $\pi(\sigma')$), it follows from neutrality that $\pi(f(\sigma')) = a$ and so $f(\sigma') = x$, implying that voter 1 is a dictator, because whatever they ranked first won the election! Note that if originally $f(\sigma)$ was b, c, or d, we would have shown that voter 2, 3, or 4 was the dictator, respectively. □

Manipulation may be unavoidable, but can we design "reasonable" voting rules where manipulation is computationally hard? If it is too computationally hard to find a useful manipulation, we can get voters to just represent their true preferences and circumvent the G-S Theorem!

Definition 3 (The Computational Problem)**.** The f-MANIPULATION problem is defined as follows: Given votes of non-manipulators and a preferred alternative p, can a manipulator cast a vote that makes p uniquely win under f?

Example 3 (The Computational Problem using Borda)**.** If the manipulator is the third voter, the preferred alternative is $p = a$, and the votes of the first two (non-manipulating) voters are given by

1	2	3
b	b	
a	a	
c	c	
d	d	

can voter 3 make a uniquely win under the Borda Count? In this case, the answer to the f-MANIPULATION is yes. Voter 3 can report $a \succ c \succ d \succ b$, resulting in a winning the election with 8 points while b, c, and d lose with 6, 4, and 1 points, respectively.

Consider the following greedy algorithm for the f-MANIPULATION problem:

- Rank p in first place.
- While there are unranked alternatives:
  - If an alternative can be placed without preventing p from winning, place it.
  - Otherwise, return *false*.

Example 4 (Greedy Algorithm with Borda)**.** Consider the same problem as before, where voter 3 wants a to win given the other two voters' rankings. Following the greedy algorithm, we place a first to get

1	2	3
b	b	a
a	a	
c	c	
d	d	

Now, the algorithm attempts to put b as the second place, but making this placement will give b 8 points while a only has 7, so we cannot do that. However, if we place c as the second place, c will only have 4 points, so a will still be the winner. Thus, we have Step-by-step ranking process:

1	2	3
b	b	a
a	a	c
c	c	
d	d	

Now, we attempt to place b as the third choice. This would give b 7 points, but since we want a to be the unique winner, we cannot place b here. If we place d in the third place, d will have 1 point and so a will still be winning. Thus, we make this placement and get

1	2	3
b	b	a
a	a	c
c	c	d
d	d	

as our working preference profile. Finally, we attempt to place b in the fourth place of voter 3's rankings, giving b 6 points which still allows a to win with 7 points. Thus, we make this placement to get the preference profile

1	2	3
b	b	a
a	a	c
c	c	b
d	d	d

The algorithm concludes after successfully finding a ranking that satisfies the f-MANIPULATION problem.

We now look at this algorithm in the context of Llull's Rule. Remember that this is the rule that assigns points according to the number of pairwise comparisons won by a given alternative.

Example 5 (Greedy Algorithm with Llull)**.** Consider the following preference profile that results in the pairwise comparison matrix on the right

1	2	3	4	5
a	b	e	e	a
b	a	c	c	
c	d	b	b	
d	e	a	a	
e	c	d	d	

→

	a	b	c	d	e
a	–	2	3	5	3
b	3	–	2	4	2
c	2	2	–	3	1
d	0	0	1	–	2
e	2	2	3	2	–

Here, a is winning with a score of 3, as it has a majority against c, d, e. Now, if we try to place b as the second place, b will achieve a score of 4, as it would beat every other alternative, so we can't make this placement. If we place c as the second place, c will achieve have a score of 2, which still results in a as the winner. Thus, we make this placement and get the following strategy profile with updated pairwise comparisons:

1	2	3	4	5
a	b	e	e	a
b	a	c	c	c
c	d	b	b	
d	e	a	a	
e	c	d	d	

→

	a	b	c	d	e
a	–	2	3	5	3
b	3	–	2	4	2
c	2	3	–	4	2
d	0	0	1	–	2
e	2	2	3	2	–

If we try to place b as the third place, b will achieve a score of 3 and tie with a, so we can't make this placement. If we place d as the third place, d will achieve a score of 1, which still results in a as the winner. Thus, we make this placement and get the following strategy profile with updated pairwise comparisons:

1	2	3	4	5
a	b	e	e	a
b	a	c	c	c
c	d	b	b	d
d	e	a	a	
e	c	d	d	

→

	a	b	c	d	e
a	–	2	3	5	3
b	3	–	2	4	2
c	2	3	–	4	2
d	0	1	1	–	3
e	2	2	3	2	–

If we try to place b as the fourth place, b will still achieve 3 points and tie with a. But we can place e next, as it will only have a score of 2, leading to the profile:

1	2	3	4	5
a	b	e	e	a
b	a	c	c	c
c	d	b	b	d
d	e	a	a	e
e	c	d	d	

→

	a	b	c	d	e
a	–	2	3	5	3
b	3	–	2	4	2
c	2	3	–	4	2
d	0	1	1	–	3
e	2	3	3	2	–

Finally, we note that placing b last will not change any of the pairwise comparisons and b will have a score of 2. Thus, we can place b and a will remain the winner. We now conclude that the ranking $a \succ c \succ d \succ b \succ e$ satisfies the f-MANIPULATION problem, and we are done.

So when does this greedy algorithm work?

Theorem 2. *Fix $i \in N$ and the votes of other voters. Let f be a rule such that a function $s(\sigma_i, x)$ satisfies:*

1. *For every σ_i, f chooses an alternative that uniquely maximizes $s(\sigma_i, x)$.*
2. *If $\{y : y \prec_{\sigma_i} x\} \subseteq \{y : y \prec_{\sigma'_i} x\}$, then $s(\sigma_i, x) \leq s(\sigma'_i, x)$.*

Then, the greedy algorithm decides the f-MANIPULATION problem correctly.

Here, voter i is the manipulator. s can be thought of as the score function of an alternative under a preference profile. The second criterion just states that if you have two rankings for the manipulator σ_i and σ'_i, and the alternatives ranked below x in σ_i is a subset of the alternatives ranked below x in σ'_i, then the score of x under σ'_i is at least as large as the score of x under σ_i.

Proof. Suppose the algorithm failed, producing a partial ranking σ_i. Assume for contradiction that some σ'_i makes p win. Let U denote the set of alternatives not ranked in σ_i. Let u denote the highest-ranked alternative in U according to σ'_i. Now. complete σ_i by adding u next, and then the other alternatives in U arbitrarily. By Property 2, $s(\sigma_i, p) \geq s(\sigma'_i, p)$ because σ_i ranks p first. Additionally, by Property 1 and the fact that σ'_i makes p the winner, it follows that $s(\sigma'_i, p) > s(\sigma'_i, u)$. Finally, by Property 2 $s(\sigma'_i, u) \geq s(\sigma_i, u)$ because u is the highest ranked alternative in U within σ' and the only alternatives ranked below u in σ are the other alternatives in U. Thus, by transitivity, we get that $s(\sigma_i, p) > s(\sigma_i, u)$, and so the algorithm could have inserted u next, contradicting its failure. □

This theorem captures rules like Borda Count, plurality, Llull's rule, and many other rules that are based on simple scores, as long as you have simple tie-breaking rules. This is bad news, as we set out to find rules that are computationally hard to manipulate, but a simple greedy algorithm allows these rules to be manipulated.

However, there are a couple of rules that are known to be computationally difficult to manipulate. These include IRV and Llull's rule with more complicated tie-breaking rules (like second-order Llull scores). But worst-case hardness isn't necessarily an obstacle to manipulation in the average case!