Harvard CS 1360, Spring 2026
Instructor: Ariel Procaccia
Scribe: Taig Singh, Jakob de Raaij

Strategic Manipulation in Elections

—

Lecture 7

1 Strategyproofness

We start with a formal model of voting, as we discussed it in the last two lectures:

Definition 1 (Ranked Voting)**.** An instance of voting with *strict ordinal preferences* consists of

- a set of voters $N = \{1, \ldots, n\}$ (we assume that $n \geq 2$),
- a set of alternatives A with $|A| = m$,
- for each voter, a ranking $\sigma_i \in \mathcal{L}$ over the alternatives. $\mathcal{L}$ is the set of all possible rankings of the m alternatives (without ties). We write $x \succ_{\sigma_i} y$ if voter i (with ranking σ_i) prefers x to y.
- We denote by $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n) \in \mathcal{L}^n$ the *preference profile*, the collection of all voters' rankings.

A social choice function $f : \mathcal{L}^n \to A$ picks a single *winner* from the alternatives for each possible preference profile $\boldsymbol{\sigma}$.[1]

So far we have considered voters to be honest by voting with their true ranking. However, this may not be realistic in practice: Sometimes, voters can report a ranking that differs from their true ranking, to achieve a better outcome (under their original ranking).

Example 1 (Manipulation in the Borda Count)**.** Let's consider the following (true) preference profile

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

If we use the Borda Count voting rule, we find that b has 8 points, a has 7 points, c has 3 points, and d has 0 points. Thus, b wins under the Borda Count for the true preferences.

Voter 3 is not fully satisfied with the outcome: They would prefer a to win over b. However, voter 3 now realizes that if they strategically misreport their ranking, in particular making b seem less favorable to them, they can actually change the outcome of the election to a. Consider voter 3 misreporting as shown in the following preference profile:

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

[1] A function $f : \mathcal{L}^n \to \mathcal{L}$ that outputs a ranking over all alternatives is called a *social welfare function*; an example is Kemeny's rule from last lecture.

Under Borda Count, a still has 7 points but b dropped to 6 points. Since c and d have even fewer points, 4 and 1, respectively, a wins the election under Borda Count. Voter 3 is able to strategically manipulate their vote to result in a better outcome for themselves.

Faced with the criticism that his rule was easily manipulatable, Jean-Charles de Borda (1733–1799) replied: "My rule is intended for honest men!" While this may have been an acceptable rebuttal for his intended use case of voting in the Academy of Sciences of 18th century France, the finding that the rule is easy to strategically manipulate seems concerning in many modern-day voting scenarios: How can we claim that a vote accurately aggregates the voters' (true) preferences if the voters may be incentivized to misreport their preferences?

Definition 2 (Strategyproofness)**.** A social choice function f is *strategyproof (SP)* if a voter can never benefit from lying about their preferences:

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

$\boldsymbol{\sigma}_{-i} = (\sigma_1, \dots, \sigma_{i-1}, \sigma_{i+1}, \dots, \sigma_n)$ denotes the preference profile $\boldsymbol{\sigma}$ without the ranking of voter i; $(\sigma_i', \boldsymbol{\sigma}_{-i}) = (\sigma_1, \dots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \dots, \sigma_n)$ denotes the preference profile $\boldsymbol{\sigma}$ where voter i has ranking σ_i' instead of σ_i.

In other words, f is strategyproof if in every preference ranking $\boldsymbol{\sigma}$, for every voter i, for every alternative ranking σ_i' that voter i can misreport, the winning alternative if voter i votes truthfully, i.e., $f(\boldsymbol{\sigma})$, is at least as preferred by voter i (in their true ranking) as the winner that arose from voter i misreporting their preferences, i.e., $f(\sigma_i', \sigma_{-i})$. In short, this says that truthfully reporting the ranking is a dominant strategy for every voter since no other strategy gives a better outcome.

Example 2 (Strategyproofness of plurality)**.** For 2 alternatives, plurality is strategyproof: If a voter's preferred alternative is winning, they are already as satisfied as possible so cannot benefit from misreporting. If a voter's preferred alternative is losing, the only way in which they could misreport is by ranking their less preferred alternative, which lowers the plurality score of their preferred alternative.

To illustrate this, consider the general preference profile below. If $k \geq n/2$ voters prefer a to b, alternative a wins (We assume here that ties are broken in favor of a).

1	...	k	$k+1$	...	n
a	...	a	b	...	b
b		b	a		a

For first k voters, their most-preferred alternative a wins, so they clearly cannot benefit from misreporting. At the same time, the remaining $n-k$ voters can only help their less-preferred alternative a and harm their more-preferred alternative b by misreporting

However, for 3 or more alternatives, plurality is not strategyproof. For example, consider the following preference profile.

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

Without loss of generality, assume that plurality breaks ties in favor of candidate b. Thus, b is the winner, since it has (tied) the most first-place votes.

However, the first voter may realize that their most-preferred candidate a has no chance at winning the election, but that their vote can be decisive between b and c. Thus, they can misrepresent their preferences as

1	2	3	4	5
~~a~~ c	c	c	b	b
~~c~~ a	b	b	c	c
b	a	a	a	a

Now, alternative c is the winner since it has the most first-place votes. Voter 1 was able to change the winner to an alternative that is more preferred by them (under their true preferences) by misreporting their preferences.

2 Gibbard-Satterthwaite Theorem

We may hope that Borda Count and plurality failing to be strategyproof with 3 or more alternatives is due to them being rather simple social choice functions, and that more intricate functions will be able to avoid this. Unfortunately, this turns out to be false:

Definition 3. A social choice function f is *dictatorial* if there exists a voter $i \in N$ (the dictator) such that for any preference profile $\boldsymbol{\sigma} \in \mathcal{L}^n$, the winner $f(\boldsymbol{\sigma})$ is the most-preferred alternative of voter i, i.e., the top ranked alternative in σ_i. In other words, that no matter how the other voters rank the alternatives, voter i always gets to dictate the winner.

Theorem 1 (Gibbard-Satterthwaite Theorem)**.** *Let $m \geq 3$. A social choice function f is strategyproof and onto*[2] *A if and only if f is dictatorial.*

This theorem shatters our hopes for a strategyproof social choice function: Any social choice function that meets the arguably absolutely minimal requirements of being non-dictatorial (i.e., more than 1 voter matters) and onto (i.e., any alternative wins in at least one preference profile, for example when everyone ranks it first) is not strategyproof.

To prove the Gibbard-Satterthwaite Theorem, first note that the 'if' direction is straight-forward to show: If f is dictatorial, then f is onto because the dictator can rank any alternative first. f is also SP because the dictator is never better off misreporting because their reported preferences will be the result of the election and all other voters are never better off misreporting because their reported preferences do not matter.

Proving the other direction of the G-S Theorem is far more interesting. While the entire proof is out of scope, we will see a proof of this 'only if' direction under two additional assumptions to illustrate how one can go about proving such a theorem. In particular, we assume that $m \geq n$ and that f is a *neutral* social choice function:

Definition 4 (Neutrality)**.** A social choice function f is *neutral* if for any permutation of alternatives $\pi : A \to A$, it holds that

$$f(\pi(\boldsymbol{\sigma})) = \pi(f(\boldsymbol{\sigma})),$$

where $\pi(\boldsymbol{\sigma})$ is the preference profile obtained by permuting all alternatives in the individual voter rankings by π. In other words, the 'order' or 'labels' of the alternatives do not matter.

Any neutral social choice function is also 'onto': By permuting the alternatives, any alternative can be made the winner. Thus, neutrality is a stricter condition on f. Nonetheless, neutrality is still a very uncontentious property of any social choice function — arguably, no candidate should be preferred by a social choice function, with potentially the exception of tie-breaking in symmetric cases.

The proof makes use of two lemmas which we will just state here and you will prove as a part of your homework.

Lemma 1 (Strong monotonicity)**.** *If a social choice function f is strategyproof then it is also* strongly monotone. *That is, if $\boldsymbol{\sigma}$ is a preference profile so that $f(\boldsymbol{\sigma}) = a$, then $f(\boldsymbol{\sigma'}) = a$ for all profiles $\boldsymbol{\sigma'}$ such that $\forall x \in A, i \in N : a \succ_{\sigma_i} x \implies a \succ_{\sigma'_i} x$.*

[2] A function is 'onto' if for any possible output there exists an input leading to this output. In other words, this means that for any alternative $x \in A$, there exists at least one preference profile in which x is the winner.

Lemma 2 (Unanimity)**.** *If a social choice function f is strategyproof and onto, then it is also* unanimous. *That is, if $\boldsymbol{\sigma}$ is a preference profile where $a \succ_{\sigma_i} b$ for all $i \in N$, then $f(\boldsymbol{\sigma}) \neq b$.*

Let's briefly interpret these two properties. If a social choice function f is strongly monotone and it declares a the winner in some preference profile, then in any altered preference profile where every alternative that was ranked below a in any voter preference stays ranked below a in this voter's preference (regardless of what happens to the alternatives above a), the winner of the election will still be a. If a social choice function f is unanimous and every voter ranks alternative a above alternative b, then b will not be the winner.

Using these two lemmas, we now prove the Gibbard-Satterthwaite Theorem for neutral social choice functions with $m = 5$ and $n = 4$. You will generalize this proof to neutral social choice functions with any $m \geq n$ and $m \geq 3$ on the homework.

Proof of Gibbard-Satterthwaite Theorem for neutral social choice functions with $m = 5$ and $n = 4$. Assume that f is a social choice function that is strategyproof and onto. By Lemmas 1 and 2, we thus also know that f is strongly monotone and unanimous.

Consider the preference profile

$\boldsymbol{\sigma} =$

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

In $\boldsymbol{\sigma}$, all voters rank e last, and alternatives a through d are ranked in a cyclic fashion across the voters. Since f is unanimous, it is clear that e cannot be the winner. Alternatives a through d are all symmetric across the preferences, so w.l.o.g. we can assume that $f(\boldsymbol{\sigma}) = a$.

Now, consider the following preference profile, obtained from $\boldsymbol{\sigma}$ by matching the rankings of voters 2 and 3 to the ranking of voter 4:

$\boldsymbol{\sigma^1} =$

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

In $\boldsymbol{\sigma^1}$, any alternative ranked below a in any individual voter's ranking is still below a —we just pushed b and c down in the rankings of voters 2 and 3. Thus, by strong monotonicity, it follows that the winner is still a, i.e., $f(\boldsymbol{\sigma^1}) = a$.

Now, consider the following preference profile obtained from $\boldsymbol{\sigma^1}$ by pushing alternative a to the bottom of the ranking of voter 2:

$\boldsymbol{\sigma^2} =$

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

In profile $\boldsymbol{\sigma^2}$, alternative d is preferred to alternatives b, c, and e by any voter, so by unanimity, the winner cannot be any of these three alternatives. Now, assume towards a contradiction that the winner was d. Then, under preference profile $\boldsymbol{\sigma^1}$, voter 2 could misreport their ranking as their ranking in $\boldsymbol{\sigma^2}$ to lead to a winner, d, that they prefer over the winner a in $\boldsymbol{\sigma^1}$. However, this is a contradiction to f being strategyproof. Thus, by elimination, we know that $f(\boldsymbol{\sigma^2}) = a$.

Similarly, we construct $\boldsymbol{\sigma^3}$ from $\boldsymbol{\sigma^2}$ by pushing a to the bottom of the ranking of player 3 and we construct $\boldsymbol{\sigma^4}$ from $\boldsymbol{\sigma^3}$ by pushing a to the bottom of the ranking of player 4; we get $f(\boldsymbol{\sigma^3}) = a$ and $f(\boldsymbol{\sigma^4}) = a$ by the same as right above.

$\boldsymbol{\sigma^4} =$

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

Thus, we arrive at a preference profile, $\boldsymbol{\sigma^4}$, where the winner is alternative a even though all voters but voter 1 rank a last. From here, strong monotonicity implies that $f(\boldsymbol{\sigma'}) = a$ in every preference profile $\boldsymbol{\sigma'}$ where voter 1 ranks a first: If voter 1 ranks a first in $\boldsymbol{\sigma'}$, then all alternatives that were ranked below a in $\boldsymbol{\sigma^4}$ are ranked below a in $\boldsymbol{\sigma'}$ (because no alternatives are below a in the rankings of voters 2, 3, or 4 in $\boldsymbol{\sigma^4}$). Since $f(\boldsymbol{\sigma^4}) = a$, strong monotonicity implies that $f(\boldsymbol{\sigma'}) = a$.

We now claim that neutrality implies that voter 1 is a dictator. Consider any preference profile $\boldsymbol{\sigma'}$ and denote the alternative that voter 1 ranks first as x. Let π be the permutation that swaps x with a. We know from above that $f(\pi(\boldsymbol{\sigma'})) = a$ since voter 1 ranks a first in $\pi(\boldsymbol{\sigma'})$. By neutrality, it follows that $\pi(f(\boldsymbol{\sigma'})) = a$, so that $f(\boldsymbol{\sigma'}) = x$. Thus, in any preference profile $\boldsymbol{\sigma'}$, the alternative x that voter 1 ranked first is the winner — voter 1 is a dictator. □

To conclude the discussion of the Gibbard-Satterthwaite Theorem, let's briefly consider what happens if we relax the 'onto' condition: If the set of possible winners of f is 1, i.e., regardless of anyone's votes f always picks some alternative x, then f is SP and non-dictatorial, because no one's votes matter. If the set of possible winners of f is 2, this corresponds to in advance deciding that only either one of two alternatives x, y can win. If we only consider comparisons between x and y and take the majority winner between them, this social choice function is SP and non-dictatorial (since plurality for $m = 2$ satisfies both). If the set of possible winners of f is at least 3, we can disregard all other alternatives and apply the Gibbard-Satterthwaite Theorem to those three alternatives to get that f is either not strategyproof or dictatorial.

3 The complexity of manipulation

We have just seen that in any reasonable social choice function, there can exist voters incentivized to misreport their true ranking. However, one fix to this issue may be the computational complexity of manipulation. In particular, if it is computationally hard to find a beneficial manipulation, we may hope that voters report their true preferences.

Definition 5 (Manipulation problem)**.** In an instance of the f-Manipulation problem, given a preference profile $\boldsymbol{\sigma}_{-i}$ (with the votes of voter i missing) and a *preferred* alternative p, the question is whether there exists a ranking σ_i of player i that makes p the winner under f.

In the manipulation problem, it is usually assumed that ties are broken against alternative p. Thus, for social choice functions where ties can occur with no specified tie-breaking rule, such as Borda Count or plurality, the problem asks whether p can be made the 'unique' winner.

Example 3 (The Borda-Manipulation problem)**.** Given is the following preference profile, with the votes of voter 3 missing.

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

The preferred alternative is $p = a$. Thus, the Borda-Manipulation problem asks whether voter 3 can report a ranking that makes a the winner under the Borda Count social choice function.

In this case, the answer is yes, as we saw in Example 1. Voter 3 can report $a \succ c \succ d \succ b$, resulting in a being the winner with 7 points, while b, c, and d only get 6, 4, and 1 points, respectively.

We now consider a greedy algorithm for the general f-Manipulation problem:

Algorithm 1 Greedy algorithm for f-Manipulation

```
Put p in the first place of the ranking σ_i.
for position j ∈ [2, ..., m] in the ranking do
    if there exists an alternative that can be placed at position j in ranking σ_i without preventing p from being the unique winner then
        place it there.
    else
        Return False.
    end if
end for
Return True.
```

Example 4 (Greedy algorithm for Borda-Manipulation)**.** Let's reconsider the Borda-Manipulation instance from Example 3. Following the greedy algorithm, we place a in the first place of the ranking of voter 3:

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

Next, the algorithm considers the second place of the ranking of voter 3. We cannot put alternative b there, since this will give b 8 points, making it the winner over a with 7 points. However, we can place c there, since c will only have 4 points so that a is still the winner:

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

The algorithm now considers which alternative to place in the third position of the ranking of voter 3. Again, we cannot put b in the third place, since this would give alternative b a total of 7 points, making it tied with a (recall that we want a to be the unique winner). If we place d in the third place, d will have a total of 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	?

Finally, we attempt to place b in the fourth place of voter 3's ranking. This gives b a total of 6 points, less than the 7 points of a, making a the winner. Thus, we get the preference profile

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

and the algorithm (correctly) returns TRUE — there exists a ranking σ_i that makes p the unique winner.

We next look at this algorithm for manipulating Llull's social choice function. Recall that under Llull's rule, each alternative receives a point for each pairwise comparison won.

Example 5 (Greedy algorithm for LLULL-MANIPULATION)**.** Given is the following preference profile with the ranking of voter 5 missing. The preferred alternative is $p = a$, which following the greedy algorithm we already placed in the top spot of the ranking of voter 5. Next to the preference profile is the pairwise comparison matrix to compute the Llull scores.

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	–

Alternative a has a Llull score 3, as it has a majority against c, d, e. Let's consider the second spot of the ranking of voter 5. If we try to place b there, b will achieve a score of 4, as it would beat every other alternative, making it the winner. If we place c as the second place, c will achieve a score of 2, which does not prevent a from being the winner. Thus, we make this placement and get the following preference 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 does not prevent a from being the winner. Thus, we make this placement and get the following preference 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, giving it a Llull score of 2, so not preventing a from being the winner

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	3	3	2	–

Finally, we note that placing b last will not change any of the pairwise comparisons. b has a Llull score of 2, so a is the unique winner of the completed preference profile. The algorithm correctly returned TRUE.

We have just seen that this algorithm seems to work well for some social choice functions. We will now show that this holds for a whole class of social choice functions.

Theorem 2. *Fix any voter $i \in N$ and any preference profile $\boldsymbol{\sigma}_{-i}$ for all voters but i. Let f be a social choice function such that there exists a function $s : \mathcal{L} \times A \to \mathbb{R}$ such that*

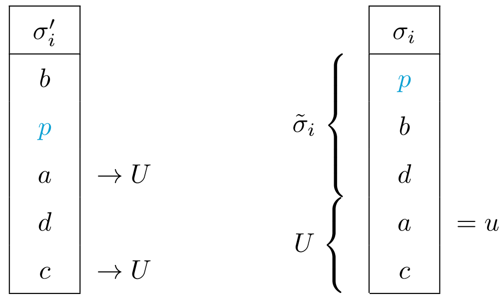
1. *for every σ_i, f chooses the alternative that uniquely maximizes $s(\sigma_i, x)$, and*
2. *if $\{y : y \prec_{\sigma_i} x\} \subseteq \{y : y \prec_{\sigma_i'} x\}$, then $s(\sigma_i, x) \leq s(\sigma_i', x)$.*

Then, the greedy algorithm decides the f-MANIPULATION problem correctly for i and $\boldsymbol{\sigma}_{-i}$.

Here, voter i is the manipulator. s is a 'score' function of an alternative under a preference profile. The first criterion states that there is a unique alternative maximizing the score and that this is the winner under f. The second criterion states that if you have two possible rankings for the manipulator σ_i and σ_i' and all alternatives ranked below x in σ_i are also 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. The greedy algorithm only returns TRUE when it found a ranking making p the unique winner. Thus, the only source of error of the algorithm is if there does exist a ranking making p the unique winner but the algorithm incorrectly returned FALSE.

Thus, assume towards a contradiction that the algorithm returned FALSE even though there exists some ranking σ_i' that makes p be the unique winner. Let's denote the position for which the algorithm failed to find an alternative that does not prevent p from winning as j, and the ranking up until this point as $\tilde{\sigma}_i$. Let U denote the set of alternatives not ranked in $\tilde{\sigma}_i$, and let u be the alternative in U that is ranked the highest in σ_i'. Now, obtain a complete ranking σ_i by appending u to $\tilde{\sigma}_i$, followed by the other alternatives in U in arbitrary order:



We will show that the score of p in σ_i is at least as high as the score of u in σ_i. By property 2, we know that $s(\sigma_i, p) \geq s(\sigma'_i, p)$ because σ_i ranks p first. By property 1 and the fact that σ'_i makes p the winner, it follows that $s(\sigma'_i, p) > s(\sigma'_i, u)$. By property 2, we know that $s(\sigma'_i, u) \geq s(\sigma_i, u)$ because by definition u is the highest ranked alternative in U within σ', so all alternatives ranked below u in σ_i, i.e. U, are also ranked below u in σ'_i. Putting the inequalities together, we get that $s(\sigma_i, p) > s(\sigma_i, u)$.

Thus, the greedy algorithm could have inserted u next, a contradiction to it returning FALSE at this step. □

This theorem applies to many social choice functions that rely on some sort of score, like Borda Count, plurality, or Llull's rule. For example, for Borda Count, one can set $s(\sigma_i, x)$ to be the Borda score of x in the profile $(\sigma_i, \boldsymbol{\sigma}_{-i})$, with an added $\frac{j}{m+1}$ when x is the jth alternative in A to handle tie-breaking so that there is a unique maximum of $s(\sigma_i, x)$ for any σ_i.

This is bad news in our search for a social choice function that is hard to manipulate: A simple greedy algorithm does well on a large class of voting rules. However, there are a couple of rules that are known to be computationally difficult to manipulate. In particular, the popular Instant Runoff Voting (IRV) and Llull's rule with more intricate tie-breaking rules are hard to manipulate.

Nonetheless, this may not imply that IRV or this variant of Llull are safe from manipulation. Computational hardness is a worst-case property, saying that there exists a (potentially small) fraction of elections that are hard to manipulate. However, it does not immediately tell us how hard manipulation will be in an average case — it could be the case that in almost every preference profile one would ever encounter in practice, manipulation is easy.

Lastly, note while computational hardness of manipulation in itself is not a sufficient obstacle to manipulation, computational hardness of computing the winner of a voting rule (such as the Kemeny rule, as we saw last lecture) seems to be an almost insurmountable obstacle in practice. Even if we can calculate the winner of a high-stakes political election quickly in 99% of the cases, the small chance of not being able to name a winner of an election even after months or years would make the social choice function unusable.