The Cost and Windfall of Manipulability

Abraham Othman and Tuomas Sandholm

Abstract

A mechanism is manipulable if it is in agents' best interest to misrepresent their private information (lie) to the center. We provide the first formal treatment of the *windfall of manipulability*, the seemingly paradoxical quality by which the failure of any agent to play their best manipulation yields a strictly better result than an optimal truthful mechanism. We dub such mechanisms *manipulation optimal*. We prove that any manipulation-optimal mechanism can have at most one manipulable type per agent. We show the existence of manipulation-optimal multiagent mechanisms with the goal of social welfare maximization, but not in dominant strategies when agents are anonymous and the mechanism is symmetric, the most common setting. For this setting, we show the existence of manipulation-optimal mechanisms when the goal is affine welfare maximization.

1 Introduction

Mechanism design is the science of generating rules of interaction—such as auctions and voting protocols—so that desirable outcomes result despite the participation of self-interested agents. A mechanism receives a set of preferences (i.e. type *revelations*) from the agents, and based on that information imposes an *outcome* (such as a choice of president, an allocation of items, and potentially also payments).

A central concept in mechanism design is *truthfulness*, which means that an agent's best strategy is to reveal its type (private information) truthfully to the mechanism. The *revelation principle*, a foundational result in mechanism design, proves that any social choice function that can be implemented in some equilibrium form, can also be implemented using a mechanism where all the agents are motivated to tell the truth. The proof is based on the idea of supplementing the manipulable mechanism with a strategy formulator for each agent that acts strategically on the agent's behalf (see, e.g., Mas-Colell et al. (1995)). Since truthfulness is certainly worth something—with real people, fairness and simplicity, with virtual agents, the elimination of the need to strategically compute—the revelation principle produces something for nothing, a free lunch. As a result, contemporary research into mechanism design has focused almost exclusively on truthful mechanisms.

But manipulable mechanisms protected by the "shield" of computational hardness are intuitively appealing. Computational complexity could be used to sever the symmetry between manipulable and truthful mechanisms, opening up exciting new possibilities in the outcome space. One notable caveat in this agenda is that an agent's inability to find its optimal manipulation does not imply that the agent will act truthfully. Unable to solve the hard problem of finding their optimal manipulation, an agent may submit their true private type but they could also submit their best guess for what their optimal manipulation might be or, by similar logic, give an arbitrary revelation. A challenge in manipulable mechanisms is that it is difficult to predict in which specific ways agents, particularly human agents, will behave if they do not play according to game-theoretic rationality.

In manipulable mechanisms, there are several reasons why agents may fail to play their optimal manipulations. Humans may play suboptimally due to cognitive limitations and other forms of incompetence. The field of behavioral game theory studies the gap between theoretical optimality and human actions (see Camerer (2003) for a survey). Virtual agents may be unable to find their optimal manipulations due to computational lim-

its: finding an optimal strategy is NP-hard in many settings (e.g., Bartholdi et al. (1989); Conitzer and Sandholm (2003, 2004); Procaccia and Rosenschein (2007)), and can be #P-hard (Conitzer and Sandholm, 2003), PSPACE-hard (Conitzer and Sandholm, 2003), or even uncomputable (Nachbar and Zame, 1996). The idea of using complexity as a shield against manipulations has been most prominent in the field of voting theory (e.g., Bartholdi et al. (1989); Conitzer and Sandholm (2003); Procaccia and Rosenschein (2007)).

In this paper, we explore mechanism design beyond the realm of truthful mechanisms using a concept we call *manipulation optimality*, where a mechanism benefits—and does better than any truthful mechanism—if agents fail to play their optimal manipulations *in any way*. This enables the mechanism designer to do better than the revelation principle would suggest, and obviates the need for predicting agents' irrational behavior. Conitzer and Sandholm (2004) show the existence of such a mechanism in an artificial setting, but leave open the question of how broadly this paradigm applies and whether it applies to any practical settings. These are the questions that we answer in this paper.

We prove an impossibility result that curtails the windfall of manipulability significantly. Specifically, we show impossibility if any agent has more than one manipulable type. Curtailed by this first impossibility result, we proceed to study settings where each agent has at most one manipulable type. For single-agent settings, we show impossibility under the social welfare maximization objective and possibility under affine welfare maximization. In contrast, in the multiagent setting we get possibility under both of those objectives, but only under the affine version if agents are symmetric.

2 The general setting

Each agent i has type θ_i and a utility function $u_i^{\theta_i}(o)$, which depends on the outcome o that the mechanism selects. An agent's type captures all of the agent's private information. For brevity, we sometimes write $u_i(o)$.

The mechanism designer has an objective (which can be thought of as mechanism utility) that he tries to maximize:

$$\mathcal{M}(o) = \sum_{i=1}^{n} \gamma_i u_i(o) + m(o),$$

where $m(\cdot)$ captures the designer's desires unrelated to the agents' utilities. This formalism has three widely-explored objectives as special cases:

- Social welfare: $\gamma_i = 1$ and $m(\cdot) = 0$.
- Affine welfare: $\gamma_i > 0$ and $m(\cdot) \geq 0$.
- Revenue: Let outcome o correspond to agent payments to the mechanism of $\pi_1(o), \ldots, \pi_n(o)$. Fix $\gamma_i = 0$ and $m(o) = \sum_{i=1}^{n} \pi_i(o)$.

An agent's type is *manipulable* if reporting some other type yields higher utility for the agent. That report is the agents' *best response* and is generally conditional on the reports of the other agents. If a certain report is a best response for every possible report of the other agents, it is known as a *dominant strategy*. A mechanism implements a social choice function, a function from agent reports to outcomes.

Two manipulable types are *distinct* if, for some revelation of the other agents, the types have different optimal manipulations which lead to different outcomes.

A mechanism M is *truthful* if each agent's dominant strategy in the mechanism is to reveal her true type. A mechanism M is an *optimal truthful mechanism* if it is not (weakly) Pareto-dominated by any other truthful mechanism.

Now we are ready to introduce the main notion of this paper. We say a mechanism is *manipulation optimal* if, when agents play their optimal strategies, the mechanism utility equals that of the best truthful mechanism, and *any* failure of agents to perform their optimal manipulations yields greater mechanism utility.

We assume that, if an agent's optimal play is to reveal her true type, then she will do so. The mechanism, for instance, can publish which types are truthful, and it can be expected that those agents will behave rationally. With software agents, such behavior can be hard-coded in. On the other hand, agents with manipulable types may not behave optimally; for instance, finding an optimal manipulation can be computationally intractable. It is important to note that we do *not* assume that an agent necessarily tells the truth if it fails to find its optimal manipulation.

We now proceed to formalize this. Let o be the outcome that would arise from all agents playing strategically optimally in some manipulable mechanism $\hat{M}$. We will denote by $\hat{o}$ an outcome in $\hat{M}$ that arises if one or more agents fail to perform their optimal manipulations. Now (using the revelation principle), transform $\hat{M}$ into the truthful mechanism M which, given the true types of agents, yields outcome o.

Definition 1 *We call* $\hat{M}$ manipulation-optimal *if it meets the following characteristics:*

1. *M is an optimal truthful mechanism.*
2. *$\forall \hat{o} \neq o, \ \mathcal{M}(\hat{o}) > \mathcal{M}(o)$.*

2.1 A general impossibility result

While Conitzer and Sandholm (2004) showed that manipulation-optimal mechanisms do exist, the following result strongly curtails their existence generally.

Proposition 1 *No mechanism satisfies Characteristic 2 of Definition 1 if any agent has more than one distinct manipulable type.*

Proof. Suppose, for contradiction, that f is a social choice function satisfying Characteristic 2. Let agent i with type a have a best-response revelation a', and let agent i with type b have a best-response revelation b'. Fix the plays of the other agents as $\mathbf{x}$, such that $f(a', \mathbf{x}) \neq f(b', \mathbf{x})$. Since a and b are distinct, there must exist such an $\mathbf{x}$.

We first define the following shorthand notation:

$$\begin{aligned} \sum(a') &\equiv \sum_{j \neq i} \gamma_j u_j(f(a', \mathbf{x})) + m(f(a', \mathbf{x})) \\ \sum(b') &\equiv \sum_{j \neq i} \gamma_j u_j(f(b', \mathbf{x})) + m(f(b', \mathbf{x})) \end{aligned}$$

Because f satisfies Characteristic 2, we get the following two inequalities on mechanism utilities—for agent i of type b and agent i of type a, respectively.

$$\begin{aligned} \gamma_i u_i^b(f(b', \mathbf{x})) + \sum(b') &< \gamma_i u_i^b(f(a', \mathbf{x})) + \sum(a') \\ \gamma_i u_i^a(f(a', \mathbf{x})) + \sum(a') &< \gamma_i u_i^a(f(b', \mathbf{x})) + \sum(b') \end{aligned}$$

Because a' and b' are best-response plays for agents of their respective types, $u_i^a(f(a', \mathbf{x})) \geq u_i^a(f(b', \mathbf{x}))$ and $u_i^b(f(b', \mathbf{x})) \geq u_i^b(f(a', \mathbf{x}))$. Thus since $\gamma_i \geq 0$ we have

$$\begin{aligned} \gamma_i u_i^b(f(a', \mathbf{x})) + \sum(a') &\leq \gamma_i u_i^b(f(b', \mathbf{x})) + \sum(a') \\ \gamma_i u_i^a(f(b', \mathbf{x})) + \sum(b') &\leq \gamma_i u_i^a(f(a', \mathbf{x})) + \sum(b') \end{aligned}$$

Combining the first lines of the above two equation blocks yields $\sum(b') < \sum(a')$, while combining the second lines yields $\sum(a') < \sum(b')$, a contradiction. ■

Note that this impossibility result is driven by the strict inequality in Characteristic 2 of Definition 1. Weakening the strict inequality to loose inequality results in a very different conclusion: that the mechanism should have identical utilities for reports of both a' and b'. Taken more generally, replacing the strict inequality with loose inequality yields the result that the outcome from reporting any type which is an optimal manipulation must deliver to the mechanism exactly the same utility.

We argue that strict inequality, and the impossibility it implies, is more appropriate for this setting. When we talk about the "windfall of manipulability", what we are talking about are beneficial results beyond the scope of what truthful mechanisms can reach. That is, we want the mechanism to do better when agents make mistakes, not to be so indifferent to agent inputs that it does not matter agents are making mistakes! Moreover, strict inequality was used by Conitzer and Sandholm (2004), so our results are in keeping with that work.

In the rest of this section we explore mechanisms where each agent can have at most one manipulable type; the above impossibility result precludes the existence of manipulation-optimal mechanisms if the other types are not dominant-strategy truthful.

2.2 Single-agent settings

In this subsection we study settings where there is only one agent reporting their private information. (If there are other agents, their types are assumed to be known.)

Proposition 2 *There exist no single-agent manipulation-optimal mechanisms with the objective of social welfare maximization.*

Proof. In the single-agent context, social welfare maximization means maximizing the utility of the single agent. For contradiction, let f be a manipulation-optimal mechanism. Let the agent of type a have optimal play a' such that $u(f(a')) \geq u(f(x)) \ \forall x \in \Theta_i$. But by Characteristic 2, $u(f(a')) < u(f(x)) \ \forall x \in \Theta_i \setminus a'$. ■

Proposition 3 *There exist single-agent manipulation-optimal mechanisms with the objective of affine welfare maximization.*

Proof. We can derive this result from the constructive proof of Conitzer and Sandholm (2004). Because the transformation is non-trivial, we restate that result here.

There exists a manager with three possible true types for a team of workers that needs to be assembled:

1. "Team with no friends", which we abbreviate TNF.
2. "Team with friends", which we abbreviate TF.
3. "No team preference", which we abbreviate NT.

The mechanism implements one of two outcomes: picking a team with friends (TF), or picking a team without friends (TNF). The manager gets a base utility 1 if TNF is chosen, and 0 if TF is chosen. If a manager has a team preference, implementing that team preference (either with or without friends) gives the manager an additional utility of 3.

In addition to the manager, the other agent in the game is the HR director, who has utility 2 if a team with friends is chosen. Even though there are two agents in the game, because the HR director does not report a type, this is not a multiagent setting. In fact, the HR director's utilities are equivalent to the payoffs from the outcome-specific mechanism utility map m.

The optimal truthful mechanism maps reports of NT and TNF to TNF and TF to TF. Now consider the manipulable mechanism which maps reports of TNF to TNF and NT and TF to TF. Note that in this mechanism there is only one manipulable type, NT, and that its optimal strategic play is to report TNF. This mechanism is manipulation-optimal: if the manager has type NT and reports NT or TF instead of TNF, the mechanism generates affine welfare of 2, whereas the optimal truthful mechanism generates affine welfare of 1. ■

Conitzer and Sandholm (2004) showed that, for an NT agent, reporting TNF is NP-hard because actually constructing a team of size k without friends requires solving the independent set problem in a graph of people where the edges are friend relationships. Computational complexity is a strong justification for why an agent may not be able to find its optimal manipulation.

2.3 Multi-agent settings

Though we proved above that there do not exist single-agent social welfare maximizing manipulation-optimal mechanisms, they do exist in multi-agent settings.

Proposition 4 *There exist multi-agent manipulation-optimal mechanisms with the objective of social welfare maximization.*

Proof. Consider a game in which two agents, the row agent and the column agent, can have one of two types, a or a'. Our mechanism maps reports to one of four different outcomes:

Report	a'	a
a'	o_1	o_2
a	o_3	o_4

The following two payoff matrices over the four outcomes constitute a manipulation-optimal mechanism. Payoffs for type a are on the left and for type a' on the right:

Report	a'	a
a'	1,1	4,0
a	0,3	3,0

Report	a'	a
a'	3,4	5,0
a	0,6	0,0

Here, playing a' is a strictly dominant strategy for agents of both types. By the revelation principle, we can "box" this mechanism into a truthful mechanism, M_1, that always chooses o_1. However, when an agent of type a plays a rather than a', social welfare is strictly higher than with o_1. (This holds regardless of how others play.) We have now proven Characteristic 2.

What remains to be proven is Characteristic 1; we must demonstrate that M_1 is optimal among truthful mechanisms. We begin by examining the following table, which lists the payoffs for agents over outcomes given the four possible true type combinations:

True types	o_1	o_2	o_3	o_4
a, a	2	4	3	3
a, a'	5	4	6	3
a', a	4	5	3	0
a', a'	7	5	6	0

Now, for contradiction, let M^D be a truthful mechanism that Pareto dominates M_1. Note that M_1 delivers the highest payoff when both agents are of type a'. Thus, $M^D(a', a') = o_1$. But this implies $M^D(a, a')$ and $M^D(a', a)$ must also equal o_1: mapping them to the outcome that gives higher social welfare (in the former case, o_3 and in the latter, o_2) is

not truthful because the agent of type a has incentive to report a' and force o_1. At the same time, mapping to an outcome that is not o_1 delivers less social welfare than M_1. So, $M^D(a', a') = M^D(a', a) = M^D(a, a') = o_1$. But if these three inputs map to o_1, M^D cannot truthfully map revelations of (a, a) to any outcome other than o_1, because some agent will always want to deviate, report type a', and force outcome o_1. Thus, our construction of M^D fails because $M^D = M_1$. ■

The result above uses dominant strategy equilibrium as the solution concept. Therefore, the result implies possibility for weaker equilibrium notions as well.

The agents in our construction are not symmetric. (Symmetry means that all agents have the same payoffs for their reports relative to the reports of the other agents.) We may ask whether manipulation-optimal mechanisms exist for what can be considered the most common setting: where agents are symmetric, the mechanism is anonymous, and the objective is welfare maximization. (By anonymous we mean that the mechanism selects an outcome based only on the distribution of reported types, rather than the agents who reported those types.)

Proposition 5 *There exist no dominant-strategy anonymous multi-agent manipulation-optimal mechanisms with the objective of social welfare maximization for symmetric agents.*

While the impossibility results earlier in this paper were based on a violation of Characteristic 2 of manipulation-optimal mechanisms alone, here the impossibility comes from not being able to satisfy Characteristics 1 and 2 together.

Proof. By Proposition 1, we can focus on mechanisms with a single manipulable type. Call the type a, with dominant strategy a'. Suppose mechanism $\hat{M}$ satisfies Characteristic 2. By the revelation principle it has a corresponding truthful mechanism M. We show that we can construct a truthful mechanism M that Pareto dominates M'.

First, if a set of reports includes a type other than a or a', we set M^D to simply mirror the action taken by M. Strategic implications for agents other than types a and a' are unaffected because for agents of those types revealing their true type was a dominant strategy under $\hat{M}$.

Let o be the outcome implemented by M when all agents reveal a, and let o' be the outcome implemented by M when all agents reveal a'. Denote by $\tilde{a}$ any combination of revelations a and a'; note that $M'(\tilde{a}) = o'$.

By Characteristic 2 we know that we get higher social welfare if agents of type a—whose best manipulation is to report a'—cannot find the manipulation and report a instead. Since agents are symmetric, this implies $u^a(o') < u^a(o)$. This is akin to the Prisoner's Dilemma: the dominant strategy of type a is to report a' , but the outcome is worse for agents if they all report a' rather than a.

Now we construct M^D based on the payoff structure of agents of type a'.

- **Case I:** $u^{a'}(o') < u^{a'}(o)$. In this case we let M^D map each $\tilde{a}$ to o. M^D Pareto dominates M.
- **Case II:** $u^{a'}(o') \geq u^{a'}(o)$. In this case we let M^D select o if all agents report a and o' for any other $\tilde{a}$. M^D Pareto dominates M. Note that M^D is identical to M for all reports except the one where all agents report a. ■

Note that both our possibility results and this impossibility result have used the dominant strategy solution concept. This implies the strongest possibility, but the weakest impossibility. Here, our requirement for dominant strategy manipulability avoids issues with degenerate special cases.

We can get around this impossibility by moving to the affine welfare objective. Note that for an anonymous mechanism, the outcome-specific mechanism utility function $m(\cdot)$

can depend only on the distribution of types, rather than the identities associated with those types.

Proposition 6 *There exist dominant-strategy anonymous multi-agent manipulation-optimal mechanisms with the objective of affine welfare maximization for symmetric agents.*

Proof. We provide a constructive proof with the same framework as Proposition 4. But now let the payoff matrices be as follows (left matrix for type a and right matrix for type a').

Report	a'	a
a'	2,2	1,1
a	1,1	0,0

Report	a'	a
a'	4,4	1,3
a	3,1	0,0

Let $\gamma_i = 1$ for all i, and let the mechanism's additional payoff, $m(\cdot)$, be $\{0, 3, 3, 5\}$ for outcomes o_1 through o_4, respectively. Note that the row and column agents are symmetric (the payoff matrices are symmetric) and that $m(o_2) = m(o_3)$. The dominant strategy equilibrium for this mechanism is for every agent to report type a'. Therefore this mechanism has truthful analogue M_1, the mechanism that always chooses o_1.

We now show that M_1 is an optimal truthful mechanism. First, note that M_1 maximizes the objective when both agents have type a'. It can be shown that (using a construction akin to the last table in the proof of Proposition 4) that due to agent incentives to deviate, any truthful mechanism that would dominate M_1 must map all reports to o_1. Therefore M_1 is an optimal truthful mechanism.

The manipulation-optimality of the mechanism defined by the payoff matrices above comes from noting that whenever agents of type a fail to report a', affine welfare is strictly higher. ■

3 Conclusions and Future Directions

The strategic equivalence of manipulable and truthful mechanisms—captured by the revelation principle—does not mean that any manipulable mechanism is automatically flawed. The failure of agents to perform their best response (or play a particular equilibrium among many), either due to computational constraints or any flavor of incompetence, can actually increase mechanism utility. When the equivalent truthful mechanism to such a manipulable mechanism is optimal among truthful mechanisms, then the manipulable mechanism is truly a better solution. We call such a mechanism *manipulation optimal*.

For a completely general setting, we show that manipulation optimality is limited to mechanisms that have at most one manipulable type per agent. Thus there is a "cost of manipulability" — implementing a manipulable mechanism inherently exposes the designer to achieving an unnecessarily poor result when agents do not perform optimally. This result is, in large part, in line with the revelation principle, although here the considerations are more subtle and the impossibility not universal. This is an inauspicious finding for the concept of using computational complexity as a "trick" to get around the revelation principle: if our mechanism utility function is sufficiently non-trivial (i.e. so that agents' reports with manipulable types can affect it), then the mechanism designer is exposed to the risk of bad outcomes.

It is worth noting how our results apply to the now-copious literature on the complexity of voting schemes. Here they are less disheartening, because non-trivial voting schemes are inherently manipulable by the Gibbard-Satterthwaite impossibility result. So in the voting setting, there is no "truthful analogue" that our manipulable mechanism is performing worse

than. Note also that most voting schemes use cardinal utility (ranked preference lists) as opposed to the ordinal utilities employed here.

It would be interesting to study manipulation optimality under other objectives, such as notions of fairness. As another direction, we plan to explore whether automated mechanism design (Conitzer and Sandholm, 2002) can be used to design manipulation-optimal mechanisms. Given priors over types (and perhaps also over behaviors), it may be possible to ignore incentive compatibility constraints and design manipulable mechanisms that yield higher mechanism utility.

Acknowledgments

This material is based upon work supported by the National Science Foundation under ITR grant IIS-0427858.

References

John Bartholdi, III, Craig Tovey, and Michael Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989. ISSN 0176-1714.

Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 103–110, Edmonton, Canada, 2002.

Vincent Conitzer and Tuomas Sandholm. Universal voting protocol tweaks to make manipulation hard. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 781–788, Acapulco, Mexico, 2003.

Vincent Conitzer and Tuomas Sandholm. Computational criticisms of the revelation principle. In *The Conference on Logic and the Foundations of Game and Decision Theory (LOFT)*, Leipzig, Germany, 2004. Earlier versions: AMEC-03, EC-04.

Andreu Mas-Colell, Michael Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York, NY, 1995.

John H Nachbar and William R Zame. Non–computable strategies and discounted repeated games. *Economic Theory*, 8(1):103–122, June 1996.

Ariel D. Procaccia and Jeffrey S. Rosenschein. Junta Distributions and the Average-Case Complexity of Manipulating Elections. *Journal of Artificial Intelligence Research (JAIR)*, 28:157–181, 2007.

Abraham Othman
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: `aothman@cs.cmu.edu`

Tuomas Sandholm
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
Email: `sandholm@cs.cmu.edu`