Better with Byzantine: Manipulation-Optimal Mechanisms

Abraham Othman and Tuomas Sandholm

Computer Science Department
Carnegie Mellon University
{aothman,sandholm}@cs.cmu.edu

Abstract. A mechanism is manipulable if it is in some agents' best interest to misrepresent their private information. The *revelation principle* establishes that, roughly, anything that can be accomplished by a manipulable mechanism can also be accomplished with a truthful mechanism. Yet agents often fail to play their optimal manipulations due to computational limitations or various flavors of incompetence and cognitive biases. Thus, manipulable mechanisms in particular should anticipate byzantine play. We study *manipulation-optimal* mechanisms: mechanisms that are undominated by truthful mechanisms when agents act fully rationally, and do better than any truthful mechanism if *any* agent fails to act rationally *in any way*. This enables the mechanism designer to do better than the revelation principle would suggest, and obviates the need to predict byzantine agents' irrational behavior. We prove a host of possibility and impossibility results for the concept which have the impression of broadly limiting possibility. These results are largely in line with the revelation principle, although the considerations are more subtle and the impossibility not universal.

1 Introduction

Mechanism design is the science of generating rules of interaction—such as auctions and voting protocols—so that desirable outcomes result despite participating agents (humans, companies, software agents, etc.) acting in their own interests. A mechanism receives a set of preferences (i.e. type *reports*) 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 report 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 simply supplementing the manipulable mechanism with a strategy formulator for each agent that acts strategically on the agent's behalf (see, e.g., [1]). Since truthfulness is certainly worth something—simplicity, fairness, and the removal of incentives to invest

in information gathering about others—the revelation principle produces something for nothing, a free lunch. As a result, mechanism design research has largely focused on truthful mechanisms.

In this work, we explore what can happen in manipulable mechanisms when agents do not play optimally. Is it possible to design mechanisms with desirable off-equilibrium properties? There are several reasons why agents may fail to play their optimal manipulations. Humans may play sub-optimally due to cognitive limitations and other forms of incompetence. The field of behavioral game theory studies the gap between game-theoretic rationality and human behavior (an overview is given in [2]). Agents may also be unable to find their optimal manipulations due to computational limits: finding an optimal report is NP-hard in many settings (e.g., [3–6]), and can be #P-hard [4], PSPACE-hard [4], or even uncomputable [7]. One notable caveat 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 its optimal manipulation, an agent may submit its true private type but she could also submit her best guess of what her optimal manipulation might be or, by similar logic, give an arbitrary report. A challenge in manipulable mechanisms is that it is difficult to predict in which specific ways agents will behave if they do not play according to game-theoretic rationality. Byzantine players, who behave arbitrarily, capture this idea.

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 *any* agent fails *in any way* to play her optimal manipulation. This enables the mechanism designer to do better than the revelation principle would suggest, and obviates the need to predict agents' irrational behavior. Conitzer and Sandholm [5] proved the existence of such a mechanism in one constructed game instance, but this work is the first to explore the concept formally and broadly.

2 The general setting

Each agent i has type $\theta_i \in \Theta_i$ and a utility function $u_i^{\theta_i}(o) : O \to \Re$, which depends on the outcome $o \in O$ that the mechanism selects. An agent's type captures all of the agent's private information. For brevity, we sometimes write $u_i(o)$. A mechanism $M : \Theta_1 \times \Theta_2 \times \cdots \times \Theta_n \to O$ selects an outcome based on the agents' type reports.

The mechanism designer has an objective (which can be thought of as mechanism utility) which maps outcomes to real values:

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

where $m(\cdot)$ captures the designer's desires unrelated to the agents' utilities, and $\gamma_i \geq 0$. 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 agents' payments, $\pi_1(o), \ldots, \pi_n(o)$, to the mechanism. Fix $\gamma_i = 0$ and $m(o) = \sum_{i=1}^{n} \pi_i(o)$.

Definition 1. *Agent i has a* manipulable *type θ_i if, for some report of the other agents' types θ_{-i}, there exists $\theta_i' \neq \theta_i$ such that*

$$u_i(M(\theta_i', \theta_{-i})) > u_i(M(\theta_i, \theta_{-i}))$$

Note that a type that is manipulable for some reports of the other agents, but not for other reports of the other agents, is still manipulable.

Definition 2. *Types θ_i and θ_i' are* distinct *if there exists some report of other agents θ_{-i}, such that the best response for type θ_i is to submit t and the best response for type θ_i' is t', where*

$$M(t, \theta_{-i}) \equiv o \neq o' \equiv M(t', \theta_{-i}), \text{ and}$$

$$u_i^{\theta_i}(o) > u_i^{\theta_i}(o'), \text{ and } u_i^{\theta_i'}(o') > u_i^{\theta_i'}(o)$$

Put another way, types are distinct only if there exists a circumstance under which agents with those types will be motivated to behave distinctly, causing distinct outcomes that provide distinct payoffs.

Definition 3. *A mechanism is (dominant-strategy)* truthful *if no agent has a manipulable type.*

Definition 4. *Let f and g be functions mapping an arbitrary set $S \to \Re$. We say f* Pareto dominates *g (or g* is Pareto dominated by *f) if for all $s \in S$,*

$$f(s) \geq g(s),$$

where the inequality is strict for at least one s.

Definition 5. *A type report of $\theta_i^* \in \Theta_i$ is* optimal *for agent i if, given reports of other agents θ_{-i}, $u_i(M(\theta^*, \theta_{-i})) \geq u_i(M(\theta, \theta_{-i}))$, for all $\theta \in \Theta_i$.*

Now we are ready to introduce the main notion of this paper. We define a manipulable mechanism to be *manipulation optimal* if it does as well as the best truthful mechanism if agents play their optimal manipulations, and strictly better if any agent fails to do so in any way:

Definition 6. *For an arbitrary collection of types, let o represent the outcome that arises when all agents with manipulable types play optimally, and let $\hat{o}$ represent an outcome that can arise when some agents with manipulable types do not play optimally. We call a manipulable mechanism $\hat{M}$ a* strictly manipulation optimal mechanism (strict MOM) *if:*

1. *No truthful mechanism Pareto dominates agents playing optimally in $\hat{M}$. (Here the inputs are the true types of the agents and we measure based on the mechanism designer's objective.)*

2. For all $\hat{o}$, $\mathcal{M}(\hat{o}) > \mathcal{M}(o)$, where $\mathcal{M}(\cdot)$ represents the designer's objective.

If instead of the second condition holding strictly (i.e., for all $\hat{o}$), it holds with equality in some places and with strict inequality in others, we call $\hat{M}$ a *Pareto manipulation optimal mechanism (Pareto MOM).*

We assume that, if an agent's optimal play is to reveal its true type, then it 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. However, our setting and results translate straightforwardly to a fully byzantine setting, where the behavior of every agent (regardless of the truthfulness of their type) is arbitrary. We discuss this setting at the conclusion of this section.

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 require that our MOMs do well for *any* failure to manipulate optimally.

2.1 A broad impossibility result for strict MOMs

While Conitzer and Sandholm [5] showed that manipulation-optimal mechanisms do exist, the following result strongly curtails their existence.

Proposition 1. *No mechanism satisfies Characteristic 2 of Definition 6 if any agent has more than one distinct manipulable type.*

Proof. Suppose for contradiction that $\hat{M}$ is a manipulable mechanism satisfying Characteristic 2 such that agent i has two distinct manipulable types. Let the types be a and b, and let $\mathbf{x}$ represent the reports of other agents where they express its distinction, so that agent i of type a has best response a', and agent i of type b has best response b', and:

$$\hat{M}(a', \mathbf{x}) \neq \hat{M}(b', \mathbf{x})$$

We first define the following shorthand notation:

$$\sum(a') \equiv \sum_{j \neq i} \gamma_j u_j(\hat{M}(a', \mathbf{x})) + m(\hat{M}(a', \mathbf{x}))$$
$$\sum(b') \equiv \sum_{j \neq i} \gamma_j u_j(\hat{M}(b', \mathbf{x})) + m(\hat{M}(b', \mathbf{x}))$$

Because $\hat{M}$ satisfies the strict form of Characteristic 2, we get the following two inequalities on mechanism utilities—for agent i of type b and agent i of type a, respectively.

$$\gamma_i u_i^b(\hat{M}(b', \mathbf{x})) + \sum(b') < \gamma_i u_i^b(\hat{M}(a', \mathbf{x})) + \sum(a')$$
$$\gamma_i u_i^a(\hat{M}(a', \mathbf{x})) + \sum(a') < \gamma_i u_i^a(\hat{M}(b', \mathbf{x})) + \sum(b')$$

But because a' and b' are distinct, $u_i^a(\hat{M}(a', \mathbf{x})) > u_i^a(\hat{M}(b', \mathbf{x}))$ and $u_i^b(\hat{M}(b', \mathbf{x})) > u_i^b(\hat{M}(a', \mathbf{x}))$. Thus since $\gamma_i \geq 0$ we have

$$\gamma_i u_i^b(\hat{M}(a', \mathbf{x})) + \sum(a') \leq \gamma_i u_i^b(\hat{M}(b', \mathbf{x})) + \sum(a')$$
$$\gamma_i u_i^a(\hat{M}(b', \mathbf{x})) + \sum(b') \leq \gamma_i u_i^a(\hat{M}(a', \mathbf{x})) + \sum(b')$$

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. □

This impossibility result is driven by the strict inequality in Characteristic 2 of Definition 6. In the next section, we consider what happens to this result when we loosen the strict inequality.

2.2 A characterization of Pareto MOMs

Recall that the difference between the two MOM concepts was that strict MOMs require that the mechanism always do strictly better when an agent plays sub-optimally, while Pareto MOMs only require that the mechanism does not do worse and does strictly better at some point when an agent plays sub-optimally. It follows that possibility results for strict MOMs implies possibility for Pareto MOMs, and impossibility results for Pareto MOMs imply impossibility for strict MOMs.

We now revisit the impossibility of Proposition 1, but with the Pareto MOM notion. Instead of obtaining impossibility, we derive the following result:

Proposition 2. *In any mechanism that satisfies the Pareto version of Characteristic 2 of Definition 6, the report of every type that is a manipulable best-response for any other type results in identical mechanism utility. Furthermore, for any agent i with more than one distinct manipulable type, $\gamma_i = 0$.*

Proof. This proof follows the same guidelines as the one for strict MOMs. Define a, b, a', b', $\mathbf{x}$, $\sum(a')$ and $\sum(b')$ as before. The difference is that we now have only:

$$\gamma_i u_i^b(\hat{M}(b', \mathbf{x})) + \sum(b') \leq \gamma_i u_i^b(\hat{M}(a', \mathbf{x})) + \sum(a')$$
$$\gamma_i u_i^a(\hat{M}(a', \mathbf{x})) + \sum(a') \leq \gamma_i u_i^a(\hat{M}(b', \mathbf{x})) + \sum(b')$$

because there is no guarantee that the strict relation is expressed at $\mathbf{x}$. But because a and b are distinct, $u_i^a(\hat{M}(a', \mathbf{x})) > u_i^a(\hat{M}(b', \mathbf{x}))$ and $u_i^b(\hat{M}(b', \mathbf{x})) > u_i^b(\hat{M}(a', \mathbf{x}))$. Now if $\gamma_i > 0$, we have

$$\gamma_i u_i^b(\hat{M}(a', \mathbf{x})) + \sum(a') < \gamma_i u_i^b(\hat{M}(b', \mathbf{x})) + \sum(a')$$
$$\gamma_i u_i^a(\hat{M}(b', \mathbf{x})) + \sum(b') < \gamma_i u_i^a(\hat{M}(a', \mathbf{x})) + \sum(b')$$

which yields a contradiction. However, if $\gamma_i = 0$, we get

$$\gamma_i u_i^b(\hat{M}(a', \mathbf{x})) + \sum(a') = \gamma_i u_i^b(\hat{M}(b', \mathbf{x})) + \sum(a')$$
$$\gamma_i u_i^a(\hat{M}(b', \mathbf{x})) + \sum(b') = \gamma_i u_i^a(\hat{M}(a', \mathbf{x})) + \sum(b')$$

which, when combined with the MOM characterization above, yields possibility only when $\sum(a') = \sum(b')$, which, because $\gamma_i = 0$, indicates that mechanism utility is identical for reports of a' and b'. □

Corollary 1. *There exist no mechanisms with the social welfare maximization objective that satisfy the Pareto version of Characteristic 2 of Definition 6 if any agent has more than one distinct manipulable type.*

Corollary 2. *In any mechanism that satisfies the Pareto version of Characteristic 2 of Definition 6, the mechanism utility corresponding to reports of types that are not the best responses of some manipulable type must be at least the mechanism utility obtained from the best-response type reports, with at least one report inducing strictly greater mechanism utility.*

Proposition 3. *If every type is a manipulable best response for some other type, then there exist no mechanisms that satisfy the Pareto version of Characteristic 2 of Definition 6.*

Proof. Since every type is a manipulable best response for some other type, we have that every outcome must have identical mechanism utility. But then the "at least one strict" condition of Pareto dominance fails. □

From Proposition 3 we see that we get almost as broad impossibility for Pareto MOMs as we did for strict MOMs (Proposition 1).

We consider the strict MOM notion more compelling than the Pareto MOM notion for two reasons:

- Strict inequality is in line with prior work. It was the MOM notion used in the original paper by Conitzer and Sandholm [5] that proved that MOMs exist (although they did not call the mechanisms MOMs).
- The motivation of MOMs is to have a mechanism that does better when agents make mistakes—not to impose artificial caveats on the mechanism designer's utility function. Thus we consider the blanket impossibility result that we obtained for strict MOMs more relevant than the somewhat contrived, barely broader possibility we obtained for Pareto MOMs.

The results in this section extend straightforwardly to a fully byzantine setting, where all agents (including those with truthful types) behave arbitrarily. It is easy to see that no strict MOMs exist in this setting, because participating agents must have more than one type (or else the setting would not require the report of private information), and so the impossibility result of Proposition 1 holds. Furthermore, while there can exist Pareto MOMs for the fully byzantine

setting, because truthful types are their own best response the results of Proposition 2 and its corollaries hold, in the sense that the report of any truthful type must also result in identical mechanism utility. Finally, for the fully byzantine setting, Proposition 3 adjusts so that if every type is the best response for *any* other type (rather than only just manipulable types) we get impossibility.

For the remainder of the paper, we return to our original setting, in which only players with manipulable types are byzantine. We feel the argument that truthful behavior for certain types can be hard-coded into computational agents, and publicly published and verified for human agents, to be the most convincing reason why we should expect players with truthful types to actually behave truthfully.

2.3 Single-agent settings

In this subsection we study settings where there is only one agent reporting its private information. If there are other agents, their types are assumed to be known, so there is only one *type-reporting* agent.

Proposition 4. *There exist no single-agent Pareto MOM with the objective of social welfare maximization.*

Proof. In the single-agent context, social welfare maximization indicates that the utility of the mechanism is equivalent to the utility of the single agent. Let the agent have manipulable type a, which has optimal report a'. Denote $\hat{a}$ as the report satisfying the strict Pareto MOM criterion (we could have $\hat{a} = a$, but both $a \neq a'$ (because a is manipulable) and $\hat{a} \neq a'$ hold). In particular:

$$u^a(\hat{M}(\hat{a})) > u^a(\hat{M}(a'))$$

but a' was an optimal report, so:

$$u^a(\hat{M}(a')) \geq u^a(\hat{M}(\hat{a}))$$

which is a contradiction. □

The impossibility for Pareto MOMs directly implies impossibility for strict MOMs.

Proposition 5. *There exist single-agent strict MOMs with the objective of affine welfare maximization.*

Proof. We can derive this result from the constructive proof of Conitzer and Sandholm [5] by recasting parts of their construction within our framework.

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

- "Team with no friends", which we abbreviate TNF.
- "Team with friends", which we abbreviate TF.
- "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 $m(\cdot)$ (as we defined earlier in this paper).

The optimal truthful mechanism maps reports of NT and TNF to TNF and TF to TF. Now consider the manipulable mechanism that 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. □

This possibility of strict MOMs implies possibility of Pareto MOMs.

In this example, it is NP-hard for an NT agent to report TNF because 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 [5]. Computational complexity is a strong justification for why an agent may not be able to find its optimal manipulation.

2.4 Multi-agent settings

Though we proved above that there do not exist single-agent social welfare maximizing MOMs, they do exist in multi-agent settings!

Proposition 6. *There exist strict multi-agent MOMs with the objective of social welfare maximization.*

Proof. Consider a mechanism in which two agents, the row agent and the column agent, can have one of two types each, 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 payoffs for type a' are 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

Another way to view these payoffs is the following table:

Outcome	θ	u_{row}	u_{column}
o_1	a	1	1
	a'	3	4
o_2	a	4	0
	a'	5	0
o_3	a	0	3
	a'	0	6
o_4	a	3	0
	a'	0	0

In the mechanism, reporting 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 property holds regardless of how the other agent behaves). We have now proven (the strict form of) Characteristic 2.

What remains to be proven is Characteristic 1: we need to prove that M_1 is Pareto undominated among truthful mechanisms. We begin by examining the following table, which shows the social welfare (sum of agents' utilities) for the four possible true type combinations (listed as $\theta_{row}, \theta_{column}$).

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

Suppose that there exists a truthful mechanism, M^D, 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 that $M^D(a, a')$ 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 by reporting type a', and force outcome o_1. Therefore $M^D = M^1$ and so M^1 is undominated among truthful mechanisms.

□

The result above uses dominant strategies as the solution concept. Therefore, the result implies possibility for weaker equilibrium notions as well, such as Bayes-Nash equilibrium. Furthermore, this possibility for strict MOMs implies possibility for Pareto MOMs.

Definition 7. *An* anonymous *mechanism selects an outcome based only on the distribution of reported types, rather than based on the identities of the agents who reported those types.*

Definition 8. *Let i and j be any two* symmetric *agents, θ be a true type, $\widehat{\theta}$ be a report, and $\mathbf{x}$ be some report of the $n-1$ other agents. Then $u_i^{\theta}(M(\widehat{\theta}, \mathbf{x})) = u_j^{\theta}(M(\widehat{\theta}, \mathbf{x}))$ for all true types θ, all reports $\widehat{\theta}$ and all other report vectors $\mathbf{x}$.*

The agents in our construction in the proof above are not symmetric. We may ask whether MOMs exist for what can be considered the most common setting: where agents are symmetric, the equilibrium concept is dominant strategies, the mechanism is anonymous, and the objective is welfare maximization.

Proposition 7. *There exist no dominant-strategy anonymous strict multi-agent MOMs with the objective of social welfare maximization for symmetric agents.*

Proof. By Proposition 1, we can restrict attention to settings with a single manipulable type. Call the type a, and let the best report of that type of an agent be 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^D 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, reporting the true type was a dominant strategy under $\hat{M}$.

Let o be the outcome implemented by M when all agents report a, and let o' be the outcome implemented by M when all agents report a'. Denote by $\tilde{a}$ any combination of reports a and a'; observe 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.

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

We use the strict MOM concept here rather than the Pareto MOM concept, because we cannot assert that $u^a(o') < u^a(o)$ necessarily in the Pareto context. 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 circumvent the above 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 of the agents reporting those types.

Proposition 8. *There exist dominant-strategy anonymous multi-agent strict MOMs with the objective of affine welfare maximization, even for symmetric agents.*

Proof. We provide a constructive proof with the same structure as Proposition 6, but now let the payoff matrices be as follows (the left matrix is for type a and the 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 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 Pareto undominated among truthful mechanisms. 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 6) that due to agent incentives to deviate, any truthful mechanism that would dominate M_1 must map all reports to o_1. Thus M_1 is Pareto undominated among truthful mechanisms.

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 work

The strategic equivalence of manipulable and non-manipulable mechanisms—captured by the revelation principle—does not mean that every manipulable mechanism is automatically flawed. It is well-known that agents often fail to play their optimal manipulations in mechanisms due to computational limitations or various flavors of incompetence and cognitive biases. Yet it is difficult to predict how such game-theoretically irrational agents will act (or which particular equilibrium, among many, each agent will play). We studied the notion of *manipulation-optimal mechanisms*: mechanisms that are undominated by truthful mechanisms when agents play fully rationally, and do better than any truthful mechanism if *any* agent fails to play rationally *in any way*. This enables the mechanism designer to do better than the revelation principle would suggest, and obviates the need to predict agents' irrational behavior.

For the general setting, we showed that manipulation optimality is limited to mechanisms that have at most one manipulable type per agent. We also proved a host of other impossibility and possibility results for the existence of manipulation-optimal mechanisms for a variety of settings and mechanism design objectives. In particular, the possibility result for strict MOMs in the multi-agent social welfare maximization setting was very surprising. However, the overall impression was one of broad impossibility. Thus, our results suggest that in many settings 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.

Manipulation-optimal mechanisms open an avenue for numerous forms of future research. For one, 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* [8] 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.

Acknowledgements

This material is based upon work supported by the National Science Foundation under ITR grant IIS-0427858. An earlier version of this work appeared in COMSOC-08.

References

1. Mas-Colell, A., Whinston, M., Green, J.R.: Microeconomic Theory. Oxford University Press, New York, NY (1995)
2. Camerer, C.: Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press (2003)
3. Bartholdi, III, J., Tovey, C., Trick, M.: The computational difficulty of manipulating an election. Social Choice and Welfare **6**(3) (1989) 227–241
4. Conitzer, V., Sandholm, T.: Universal voting protocol tweaks to make manipulation hard. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI), Acapulco, Mexico (2003) 781–788
5. Conitzer, V., Sandholm, T.: 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.
6. Procaccia, A.D., Rosenschein, J.S.: Junta Distributions and the Average-Case Complexity of Manipulating Elections. Journal of Artificial Intelligence Research (JAIR) **28** (2007) 157–181
7. Nachbar, J.H., Zame, W.R.: Non–computable strategies and discounted repeated games. Economic Theory **8**(1) (June 1996) 103–122
8. Conitzer, V., Sandholm, T.: Complexity of mechanism design. In: Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Edmonton, Canada (2002) 103–110