Obvious strategyproofness needs monitoring for good approximations (extended abstract)*

Diodato Ferraioli[1] and Carmine Ventre[2]

[1] DIEM, Università degli Studi di Salerno, Italy. dferraioli@unisa.it
[2] CSEE, University of Essex, UK. carmine.ventre@gmail.com

Abstract. Obvious strategyproofness (OSP) is an appealing concept as it allows to maintain incentive compatibility even in the presence of agents that are not fully rational, e.g., those who struggle with contingent reasoning [10]. However, it has been shown to impose some limitations, e.g., no OSP mechanism can return a stable matching [3].
We here deepen the study of the limitations of OSP mechanisms by looking at their approximation guarantees for basic optimization problems paradigmatic of the area, i.e., machine scheduling and facility location. We prove a number of bounds on the approximation guarantee of OSP mechanisms, which show that OSP can come at a significant cost. However, rather surprisingly, we prove that OSP mechanisms can return optimal solutions when they use monitoring — a mechanism design paradigm that introduces a mild level of scrutiny on agents' declarations [9].

1 Introduction

Algorithmic Mechanism Design (AMD) is by now an established research area in computer science that aims at conceiving algorithms resistant to selfish manipulations. As the number of parties (a.k.a., agents) involved in the computation increases, there is, in fact, the need to realign their individual interests with the designer's. Truthfulness is the chief concept to achieve that: in a truthful mechanism, no selfish and *rational* agent has an interest to misguide the mechanism. A question of recent interest is, however, how easy it is for the selfish agents to understand that it is useless to strategize against the truthful mechanism.

Recent research has come up with different approaches to deal with this question. Some authors [14, 4, 6, 1] suggest to focus on "simple" mechanisms; an orthogonal approach is that of verifiably truthful mechanisms [5], wherein agents can run some algorithm to effectively check that the mechanism is incentive compatible. Li [10] has recently formalized the aforementioned idea of simple mechanisms, by introducing the concept of Obviously Strategy-Proof (OSP) mechanisms. This notion stems from the observation that the very same mechanism can be more or less truthful in practice depending on the implementation details. For example, in lab experiments, Vickrey's famous second-price mechanism results to be "less" truthful when implemented via a sealed-bid auction,

* An extended version appeared at AAAI [8]. D. Ferraioli was supported by "GNCS – INdAM". C. Ventre was supported by the EPSRC grant EP/M018113/1.

and "more" truthful when run via an ascending auction. The quite technical definition of OSP formally captures how implementation details matter by looking at a mechanism as an extensive-form game; roughly speaking, OSP demands that strategy-proofness holds among subtrees of the game (see below for a formal definition). An important validation for the 'obviousness' is further provided by Li [10] via a characterization of these mechanisms in terms of agents with limited cognitive abilities (i.e., agents with limited skills in contingent reasoning). Specifically, Li shows that a strategy is obviously dominant if and only if these "limited" agents can recognize it as such. Nevertheless, the notion of OSP mechanisms might be too restrictive. E.g., Ashlagi and Gonczarowski [3] prove that no OSP mechanism can return a stable matching – thus implying that the Gale-Shapley matching algorithm is not OSP despite its apparent simplicity.

Our contribution. We investigate the power of OSP mechanisms in more detail from a theoretical computer science perspective. In particular, we want to understand the quality of approximate solutions that can be output by OSP mechanisms. To answer this question, we focus on two fundamental optimization problems, machine scheduling [2] and facility location [11], arguably (among) the paradigmatic problems in algorithmic mechanism design.

For the former problem, we want to compute a schedule of jobs on selfish related machines so to minimize the makespan. For this single-dimensional problem, it is known that a truthful PTAS is possible [7]. In contrast, we show that there is no better than 2-approximate OSP mechanism for this problem independently from the running time of the mechanism.

For the facility location problem, we want to determine the location of a facility on the real line given the preferred locations of n agents. The objective is to minimize the social cost, defined as the sum of the individual agents' distances between their preferred location and the facility's. Moulin [11] proves that the optimal mechanism, that places the facility on the median of the reported locations, is truthful. OSP mechanisms turn out to be much weaker than that. We prove in fact a tight bound of $n - 1$.

However, a surprising connection of OSP mechanisms with a novel mechanism design paradigm – called *monitoring* – allows us to prove strong positive results. Building upon the notion of mechanisms with verification [12, 13], Kovacs et al. [9] introduce the idea that a mechanism can check the declarations of the agents at running time and guarantee that those who overreported their costs end up paying the exaggerated costs. This can be enforced whenever costs can be easily measured and certified. For example, a mechanism can force a machine that in her declaration has augmented her running time to work that long by keeping her idle for the difference between real and reported running time.

We prove that, for machine scheduling, there is an optimal OSP mechanism with monitoring. The construction of this mechanism can use any algorithm for the problem as a black box, thus implying that there is PTAS that is OSP. Our construction is based upon the first-price (truthful) mechanism (with monitoring) recently designed in [15]. Our results effectively show how it is possible to modify this mechanism to allow OSP implementations.

2 Preliminaries

Mechanisms and strategy-proofness. In this work we consider a classical mechanism design setting, in which we have a set of outcomes $\mathcal{O}$ and n selfish agents. Each agent i has a *type* $t_i \in D_i$, where D_i is defined as the *domain* of i. The type t_i is *private knowledge* of agent i. Moreover, each selfish agent i has a *cost function* $c_i \colon D_i \times \mathcal{O} \to \mathbb{R}$. For $t_i \in D_i$ and $X \in \mathcal{O}$, $c_i(t_i, X)$ is the cost paid by agent i to implement X when her type is t_i.

A *mechanism* consists of a protocol whose goal is to determine an outcome $X \in \mathcal{O}$. To this aim, the mechanism is allowed to interact with agents. During this interaction, agent i is observed to take *actions*; these actions may depend on her presumed type $b_i \in D_i$ that can be different from the real type t_i. We say that agent i takes *actions according to* b_i to stress this. For a mechanism $\mathcal{M}$, we let $\mathcal{M}(\mathbf{b})$ denote the outcome returned by the mechanism when agents take actions according to their presumed types $\mathbf{b}$. Usually, this outcome is given by a pair $(f, \mathbf{p})$, where $f = f(\mathbf{b})$ (termed *social choice function*) maps the actions taken by the agents according to $\mathbf{b}$ to a feasible solution for the problem at the hand, and $\mathbf{p} = (p_1(\mathbf{b}), \ldots, p_n(\mathbf{b})) \in \mathbb{R}^n$ maps the actions taken by the agents according to $\mathbf{b}$ to *payments* from the mechanism to each agent i.

A mechanism $\mathcal{M}$ is *strategy-proof* if for every i, every $\mathbf{b}_{-i}$ and every $b_i \in D_i$, it holds that $c_i(t_i, \mathcal{M}(t_i, \mathbf{b}_{-i})) \leq c_i(t_i, \mathcal{M}(b_i, \mathbf{b}_{-i}))$, where t_i is the true type of i. That is, in a strategy-proof mechanism the actions taken according to the true type are dominant for each agent. Moreover, a mechanism $\mathcal{M}$ satisfies *voluntary participation* if for every i and every $\mathbf{b}_{-i}$, it holds that $c_i(t_i, \mathcal{M}(t_i, \mathbf{b}_{-i})) \leq 0$.

Obvious strategyproofness. An *extensive-form mechanism* $\mathcal{M}$ is defined by a directed tree $\mathcal{T} = (V, E)$ such that: (i) every leaf ℓ of the tree is labeled by a possible outcome $X(\ell) \in \mathcal{O}$ of the mechanism; (ii) every internal vertex $u \in V$ is labeled by a subset $S(u) \subseteq [n]$ of agents; (iii) every edge $e = (u, v) \in E$ is labeled by a subset $T(e) \subseteq D$ of type profiles such that:

- the subsets of profiles that label the edges outgoing from the same vertex u are disjoint, i.e., for every triple of vertices u, v, v' such that $(u, v) \in E$ and $(u, v') \in E$, we have that $T(u, v) \cap T(u, v') = \emptyset$;
- the union of the subsets of profiles that label the edges outgoing from a non-root vertex u is equal to the subset of profiles that label the edge going in u, i.e., $\bigcup_{v \colon (u,v) \in E} T(u, v) = T(\phi(u), u)$, where $\phi(u)$ is the parent of u in $\mathcal{T}$;
- the union of the subsets of profiles that label the edges outgoing from the root vertex r is equal to the set of all profiles, i.e., $\bigcup_{v \colon (r,v) \in E} T(r, v) = D$;
- for every u, v such that $(u, v) \in E$ and every two profiles $\mathbf{b}, \mathbf{b}' \in T(\phi(u), u)$ such that $(b_i)_{i \in S(u)} = (b'_i)_{i \in S(u)}$, if $\mathbf{b}$ belongs to $T(u, v)$, then also $\mathbf{b}'$ must belong to $T(u, v)$.

Observe that, according to the definition above, for every profile $\mathbf{b}$ there is only one leaf $\ell = \ell(\mathbf{b})$ such that $\mathbf{b}$ belongs to $T(\phi(\ell), \ell)$. For this reason we say that $\mathcal{M}(\mathbf{b}) = X(\ell)$. Moreover, for every type profile $\mathbf{b}$ and every node $u \in V$, we say that $\mathbf{b}$ is *compatible* with u if $\mathbf{b} \in T(\phi(u), u)$. Finally, two profiles $\mathbf{b}$, $\mathbf{b}'$

are said to *diverge* at vertex u if there are two vertices v, v' such that $(u, v) \in E$, $(u, v') \in E$ and $\mathbf{b} \in T(u, v)$, whereas $\mathbf{b}' \in T(u, v')$.

An extensive-form mechanism $\mathcal{M}$ is *obviously strategy-proof (OSP)* if for every agent i, for every vertex u such that $i \in S(u)$, for every $\mathbf{b}_{-i}, \mathbf{b}'_{-i}$, and for every $b_i \in D_i$ such that $(t_i, \mathbf{b}_{-i})$ and $(b_i, \mathbf{b}'_{-i})$ are compatible with u, but diverge at u, it holds that $c_i(t_i, \mathcal{M}(t_i, \mathbf{b}_{-i})) \leq c_i(t_i, \mathcal{M}(b_i, \mathbf{b}'_{-i}))$.

Monitoring. Let $\mathcal{M}(\mathbf{b})$ denote the outcome returned by mechanism $\mathcal{M} = (f, \mathbf{p})$ when agents take actions according to $\mathbf{b}$. Commonly, the cost paid by agent i to implement $\mathcal{M}(\mathbf{b})$ is defined as a quasi-linear combination of agent's true cost $t_i(f(\mathbf{b}))$ and payment $p_i(\mathbf{b})$, i.e., $c_i(t_i, \mathcal{M}(\mathbf{b})) = t_i(f(\mathbf{b})) - p_i(\mathbf{b})$. This approach disregards the agent's declaration for evaluating her cost.

In mechanisms with monitoring the usual quasi-linear definition is maintained but costs paid by the agents are more strictly tied to their declarations [9]. Specifically, in a mechanism with monitoring $\mathcal{M}$, the bid b_i is a lower bound on agent i's cost for $f(b_i, \mathbf{b}_{-i})$, so an agent is allowed to have a real cost higher than $b_i(f(\mathbf{b}))$ but not lower. Formally, we have $c_i(t_i, \mathcal{M}(\mathbf{b})) = \max\{t_i(f(\mathbf{b})), b_i(f(\mathbf{b}))\} - p_i(\mathbf{b})$.

3 Machine scheduling

We are given a set of m different jobs to execute and the n agents control related machines. That is, agent i has a job-independent processing time t_i per unit of job (equivalently, an execution speed $1/t_i$ that is independent from the actual jobs). Therefore, the social choice function f must choose a possible schedule $f(\mathbf{b}) = (f_1(\mathbf{b}), \ldots, f_n(\mathbf{b}))$ of jobs to the machines, where $f_i(\mathbf{b})$ denotes the job load assigned to machine i when agents take actions according to $\mathbf{b}$. The cost that agent i faces for the schedule $f(\mathbf{b})$ is $t_i(f(\mathbf{b})) = t_i \cdot f_i(\mathbf{b})$. Note that our mechanisms for machine scheduling will always pay the agents.

For this setting, monitoring means that those agents who have exaggerated their unitary processing time, i.e., they take actions according to $b_i > t_i$, can be made to process up to time b_i instead of the true processing time t_i. E.g., we could not allow any operation in the time interval $[t_i, b_i]$ or charge $b_i - t_i$.

We focus on social choice functions f^* optimizing the *makespan*, i.e., $f^*(\mathbf{b}) \in \arg\min_{\mathbf{x}} \mathtt{mc}(\mathbf{x}, \mathbf{b})$, where $\mathtt{mc}(\mathbf{x}, \mathbf{b}) = \max_{i=1}^{n} b_i(\mathbf{x})$. We have the following results.

Theorem 1. *For every $\varepsilon > 0$, there is no $(2 - \varepsilon)$-approximate mechanism for the machine scheduling problem that is OSP without monitoring and satisfies voluntary participation.*

Theorem 2. *For every α-approximate algorithm f for the machine scheduling problem on related machines there is an α-approximate mechanism for the same problem that is OSP with monitoring and satisfies voluntary participation.*

4 Facility location

In the facility location problem, the type t_i of each agent consists of her position on the real line. The social choice function f must choose a position $f(\mathbf{b}) \in \mathbb{R}$ for

the facility. The cost that agent i pays for a chosen position $f(\mathbf{b})$ is $t_i(f(\mathbf{b})) = d(t_i, f(\mathbf{b})) = |t_i - f(\mathbf{b})|$. So, $t_i(f(\mathbf{b}))$ denotes the distance between t_i and the location of the facility computed by f when agents take actions according to $\mathbf{b}$.

We can implement monitoring also in this setting whenever evidences of the distance can be provided (and cannot be counterfeited). In fact, in this context, monitoring means that $t_i(f(\mathbf{b})) = \max\{d(t_i, f(\mathbf{b})), d(b_i, f(\mathbf{b}))\}$. Therefore, once the evidence is provided, the mechanism can check whether $t_i(f(\mathbf{b})) < b_i(f(\mathbf{b}))$ and charge the agent the difference for cheating.

We focus on optimizing the *social cost*, i.e., $f^*(\mathbf{b}) \in \arg\min_{x \in \mathbb{R}} \texttt{cost}(x, \mathbf{b})$, where $\texttt{cost}(x, \mathbf{b}) = \sum_{i=1}^{n} b_i(x)$. We have the following results.

Theorem 3. *For every $\varepsilon > 0$, there is no $(n - 1 - \varepsilon)$-approximate mechanism for the facility location problem that is OSP without monitoring.*

Theorem 4. *There is a $(n-1)$-approximate mechanism for the facility location problem that is OSP, even without monitoring.*

References

1. Adamczyk, M., Borodin, A., Ferraioli, D., de Keijzer, B., Leonardi, S.: Sequential posted price mechanisms with correlated valuations. In: WINE '15. pp. 1–15 (2015)
2. Archer, A., Tardos, É.: Truthful mechanisms for one-parameter agents. In: FOCS '01. pp. 482–491 (2001)
3. Ashlagi, I., Gonczarowski, Y.A.: No stable matching mechanism is obviously strategy-proof. arXiv preprint arXiv:1511.00452 (2015)
4. Babaioff, M., Immorlica, N., Lucier, B., Weinberg, S.M.: A simple and approximately optimal mechanism for an additive buyer. In: FOCS '14. pp. 21–30 (2014)
5. Brânzei, S., Procaccia, A.D.: Verifiably truthful mechanisms. In: ITCS '15. pp. 297–306 (2015)
6. Chawla, S., Hartline, J.D., Malec, D.L., Sivan, B.: Multi-parameter mechanism design and sequential posted pricing. In: STOC '10. pp. 311–320 (2010)
7. Christodoulou, G., Kovács, A.: A deterministic truthful PTAS for scheduling related machines. SIAM J. Comput. 42(4), 1572–1595 (2013)
8. Ferraioli, D., Ventre, C.: Obvious strategyproofness needs monitoring for good approximations. In: AAAI '17. pp. 516–522 (2017)
9. Kovács, A., Meyer, U., Ventre, C.: Mechanisms with monitoring for truthful ram allocation. In: WINE '15. pp. 398–412 (2015)
10. Li, S.: Obviously strategy-proof mechanisms. Available at SSRN 2560028 (2015)
11. Moulin, H.: On strategy-proofness and single-peakedness. Public Choice 35, 437–455 (1980)
12. Nisan, N., Ronen, A.: Algorithmic Mechanism Design. Games and Economic Behavior 35, 166–196 (2001)
13. Penna, P., Ventre, C.: Optimal collusion-resistant mechanisms with verification. Games and Economic Behavior 86, 491–509 (2014)
14. Sandholm, T., Gilpin, A.: Sequences of take-it-or-leave-it offers: Near-optimal auctions without full valuation revelation. In: International Workshop on Agent-Mediated Electronic Commerce. pp. 73–91 (2003)
15. Serafino, P., Ventre, C., Vidali, A.: Towards a characterization of budget-feasible mechanisms with monitoring (2017), submitted