# Probabilistic Verification for Obviously Strategyproof Mechanisms

Extended Abstract

Diodato Ferraioli
University of Salerno
Fisciano (SA), Italy
dferraioli@unisa.it

Carmine Ventre
University of Essex
Colchester, United Kingdom
c.ventre@essex.ac.uk

## ABSTRACT

Obviously strategyproof (OSP) mechanisms maintain the incentive compatibility of agents that are not fully rational. They have been object of a number of studies since their recent definition. We are motivated by the result showing that OSP mechanisms without money cannot return good approximations, even if the designer monitors the agents during the execution of the mechanism [10]. We ask whether there are different (harsher) forms of punishments and novel ways to exert control over the agents that can overcome this impossibility. We define a model of probabilistic verification wherein agents are caught misbehaving with a certain probability and show how OSP mechanisms without money can implement a given social choice function at the cost of either imposing very large fines for lying or verifying a linear number of agents.

**ACM Reference Format:**
Diodato Ferraioli and Carmine Ventre. 2018. Probabilistic Verification for Obviously Strategyproof Mechanisms. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

## 1 INTRODUCTION

Will people strategize against an incentive-compatible mechanism? The answer depends on whether they will *understand* that doing so is against their own interest and, ultimately, on their rationality and cognitive skills. This question has often been raised in literature (see, e.g., [11, 19]) and much of the recent research in (algorithmic) mechanism design is motivated by this question. Several definitions for "simple" mechanisms have been recently given in literature: posted price mechanisms and variants [1, 4, 8], Bulow-Klemperer-like auctions [14], verifiably truthful mechanisms [6]. This quest for the right definition for simple mechanisms culminated with the introduction of *obviously strategyproof* (OSP) mechanisms [16].

OSP mechanisms are the only ones that preserve the incentive-compatibility of agents who lack contingent reasoning skills [16], that is, a class of agents that are not fully rational and prone to strategize uselessly. Consequently, this concept has attracted a considerable amount of recent work [3, 5, 10] that mainly focuses on the limitations of these appealing mechanisms. Of particular interest for our study are the results proved by Ferraioli and Ventre [10] showing that OSP mechanisms cannot have good approximation guarantees for machine scheduling and facility location, two canonical optimization problems studied in the area. This negative result is reinforced for our purposes by the fact that, when the designer can "monitor" the agents at work (meaning that the utility of lying agents depends on their type *and* bid) monetary transfers are sufficient and necessary for the existence of optimal OSP mechanisms. Since money is undesirable in many applications (cf. the vast literature on approximate mechanism design without money initiated by [18]) our main aim here is to understand how we can reconcile approximation and OSP mechanisms without money.

Given the current state of the art, we need to look at novel ways the designer can limit the agents' ability to misbehave. We introduce a model of *probabilistic verification* wherein the mechanism designer has a (potentially faulty) verification device that she can use at runtime to check whether an agent has lied. The device will catch the lie of the checked agent with certainty, or with a certain probability if faulty. E.g., if the type $t$ of an agent is her location on the real line (as in facility location) the designer can use a GPS logger to check where the agent is against her reported type $b$. In our terminology, this tool is faulty if its reading $t'$ of $t$ is subject to some measurement error $\delta$ and the agent would be caught only if $|b - t'| > \delta$; more generally, different tools can make mistakes in their measurements with some probability rather than in range (e.g., it gets better as the difference between reported and real type increases). This notion generalizes and combines the different notions of verification introduced in related literature (see, e.g., [7, 17]).

We begin by studying what we call the *full probabilistic verification model*, wherein every agent is verifiable and therefore there is a non-null probability of catching lies. We prove that, in this setting, it is always possible to obtain an OSP mechanism without money. Since in some contexts it might be impossible that for all the agents to be verifiable (e.g., not all the agents might have been equipped with a GPS logger), we look at the *partial probabilistic verification model*, where for some agents we cannot use any verification. We then prove that there is a problem $P$ for which all $\varepsilon$-OSP mechanisms (i.e., agents will not deviate for small gains $\varepsilon$) that solve $P$ need to verify in expectation a linear number of agents.

## 2 PRELIMINARIES

A mechanism design setting is defined by a set of $n$ *selfish agents* and a set of allowed *outcomes* $\mathcal{S}$. Each agent $i$ has a *type* $t_i \in D_i$, where $D_i$ is called the *domain* of $i$. The type $t_i$ is assumed to be *private knowledge* of agent $i$. Each selfish agent $i$ has a *valuation function* $v_i : D_i \times \mathcal{S} \to \mathbb{R}$. For $t_i \in D_i$ and $X \in \mathcal{S}$, $v_i(t_i, X)$ is the valuation that agent $i$ has for outcome $X$ when her type is $t_i$. We

*Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.* 

will often use $t_i(X)$ as a shorthand for $v_i(t_i, X)$. The domain $D_i$ of agent $i$ is *bounded* if $t_i(X) \in [t_{\inf}, t_{\sup}]$ for all $i$, $t \in D_i$, $X \in \mathcal{S}$.

A *mechanism* $\mathcal{M}$ is a process for selecting an outcome $X \in \mathcal{S}$, defined by a directed tree $\mathcal{T} = (V, E)$ such that:

- every leaf $\ell$ of the tree is labeled by a possible outcome $X(\ell) \in \mathcal{S}$;
- every internal vertex $u \in V$ either is labeled by an agent $S(u) \in [n]$, or is a *chance vertex* labeled by character $c$;
- every edge $e = (u, v) \in E$ going out from a non-chance vertex is labeled by a set $T(e) \subseteq D = D_1 \times \cdots \times D_n$ of type profiles s.t.:
  - the sets of profiles that label the edges outgoing from the same vertex $u$ are disjoint;,
  - the union of the sets of profiles labeling the edges outgoing from non-root vertex $u$ is equal to the set of profiles labeling the edge going in $u$;
  - the union of the sets of profiles that label the edges outgoing from the root vertex $r$ is equal to the set of all profiles;
  - for every $u, v$ such that $(u, v) \in E$ and every $\mathbf{b}, \mathbf{b}' \in T(\phi(u), u)$ such that $b_{S(u)} = b'_{S(u)}$, if $\mathbf{b} \in T(u, v)$, then also $\mathbf{b}' \in T(u, v)$;
- every $e = (u, v) \in E$, $u$ being a chance vertex, has label $T(e) = D$ if $u$ is a root, and $T(e) = T(\phi(u), u)$ otherwise;
- every non-chance vertex $u$ is associated to an information set $I(u) \subseteq D$, where $I(r) = D$, and, for $u \neq r$, either $I(u) = D$ or $I(u) = T(\phi(v), v)$ for some $v$ in the path from $r$ to $u$.

Observe that 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 $\mathbf{b} \in I(u)$. 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 $\mathbf{b}' \in T(u, v')$.

Now we define obvious strategyproofness. An extensive-form mechanism $\mathcal{M}$ is *$\varepsilon$-obviously strategy-proof ($\varepsilon$-OSP)* if for every agent $i$ with real type $t_i$, for every vertex $u$ such that $i = S(u)$, for every $\mathbf{b}_{-i}, \mathbf{b}'_{-i}$ (with $\mathbf{b}'_{-i}$ not necessarily different from $\mathbf{b}_{-i}$), and for every $b_i \in D_i$, with $b_i \neq t_i$, such that $(t_i, \mathbf{b}_{-i})$ and $(b_i, \mathbf{b}'_{-i})$ are compatible with $u$, but diverge at $u$, it holds that $v_i(t_i, \mathcal{M}(t_i, \mathbf{b}_{-i})) \geq v_i(t_i, \mathcal{M}(b_i, \mathbf{b}'_{-i})) - \varepsilon$. $\mathcal{M}$ is obviously strategy proof (OSP) if $\varepsilon = 0$.

Given a *social choice function* $f : D \to \mathcal{S}$, a mechanism $\mathcal{M}$ *implements* $f$ if $\mathcal{M}(\mathbf{b}) = f(\mathbf{b})$ for every $\mathbf{b}$.

*Probabilistic Verification.* Fix $i$ and $\mathbf{b}_{-i}$. Let $t$ and $t'$ denote the true and reported type of agent $i$, respectively. A *mechanism with probabilistic verification* $\mathcal{M}$ catches agent $i$ lying with probability $(1 - p^i_{t',t}(\mathbf{b}_{-i}))$ and punishes the agent caught lying with a fine $F^i_{t',t}(\mathbf{b}_{-i}) > 0$. Except for the fines, the mechanism does not use any other form of transfers: following previous works, we then say that our mechanisms are without money. When misreporting her type to a mechanism with probabilistic verification, agent $i$ will then have a valuation $t(\mathcal{M}(t', \mathbf{b}_{-i})) - (1 - p^i_{t',t}(\mathbf{b}_{-i}))F^i_{t',t}(\mathbf{b}_{-i})$.

Our interest will be in understanding the expected number of agents verified by a mechanism with probabilistic verification. Therefore, we will say that an agent is *verified* with probability $1 - p^i_{t',t}(\mathbf{b}_{-i})$ so that the number of verified agents in a mechanism with probabilistic verification is a random variable $V = \sum_{i=1}^{n} V_i$, where $V_i = 1$ if agent $i$ is caught lying, and 0 otherwise.

We will consider two different categories of mechanisms with probabilistic verification: the *full* model wherein all the agents are verifiable, so that we can define $p^i_{t',t}(\mathbf{b}_{-i}) \in [0, 1]$ for every tuple $(i, t, t', \mathbf{b}_{-i})$, and the *partial* model wherein there exists at least one agent $i$ that is not verifiable, that is, for which we require $p^i_{t',t}(\mathbf{b}_{-i}) = 1$ for every $\mathbf{b}_{-i}$ and every $t, t'$ with $t \neq t'$.

## 3 OUR RESULTS

Theorem 3.1 focuses on the full probabilistic verification model.

Theorem 3.1. *If the domains of agents are bounded, then for every social choice function $f$ there is an OSP mechanism with full probabilistic verification that implements $f$ and verifies in expectation only a constant number of agents.*

Unfortunately, the mechanism that proves Theorem 3.1 needs very large fines. However, we prove that full probabilistic verification still turns out to be a powerful tool even if large fines are not available. In particular, we observe a trade-off between fines and the number of verified agents. Hence, one may be able to work with lower fines, by having more accurate verification (in a sense, we can reduce fines only if we spend more for our verification tools).

Next we focus on the partial probabilistic verification model. Specifically, we investigate whether it is possible to obtain OSP mechanisms that verify few agents.

To this aim, suppose there is a subset $U$ of verifiable agents. Such a subset, just like the outcome, can be chosen randomly and can depend on the declaration of the agents. For bounded domains we can guarantee through fines that, no matter the quality of the verification device, truthtelling will be obviously dominant for all the agents in $U$. Therefore, the mechanism "only" needs to obviously incentivize the agents that are not in $U$. Unfortunately, we show that there is a social choice function for which this is possible only if $|U| = n - o(n)$, that is the number of unverifiable agents is sublinear.

Consider the *public project* problem: we need to decide whether to implement or not a public project (e.g., building a bridge) whose cost is $c$. The society is comprised of $n$ agents. The valuation of agent $i$ if the project is implemented may be either $v_i(1) = 1$ or $v_i(1) = \delta > 0$, where, $\delta \ll 1$ (e.g., $\delta = \frac{1}{n^2}$). We say that the type of $i$ is 1 in the first case, and $\delta$ in the second. Moreover, each agent has valuation $v_i(0) = 0$ if the project is not implemented. This problem has been introduced by [15] and it is a basic and very well studied problem in economics and computer science (see, e.g., [2] and references therein). The designer would like to implement the project only if at least $c$ agents have type 1. I.e., the designer would like to implement the *public project function* $f$ that returns 1 if $\sum_i v_i(1) \geq c$, and 0 otherwise. We have the following theorem.

Theorem 3.2. *For every $\varepsilon$-OSP mechanism implementing the public project function, with $\varepsilon \in [0, 1)$, there is an instance for which the mechanism verifies in expectation $n - o(n)$ agents.*

*Future Directions.* Li [16] proved that OSP is the "right" definition of truthfulness for "bounded rational" agents, where the kind of bounded rationality (i.e., limited contingent reasoning) is exactly the one observed in many experimental settings. Still, it would be interesting to investigate mechanism design for other (possibly, less stringent) notions of bounded rationality.

It would be also interesting to find settings in which an OSP mechanism with partial probabilistic verification exists that verifies only few agents.

## REFERENCES


[1] Marek Adamczyk, Allan Borodin, Diodato Ferraioli, Bart de Keijzer, and Stefano Leonardi. 2015. Sequential posted price mechanisms with correlated valuations. In *International Conference on Web and Internet Economics*. Springer, 1–15.

[2] Krzysztof R. Apt and Arantza Estévez-Fernández. 2009. Sequential Pivotal Mechanisms for Public Project Problems. In *Algorithmic Game Theory, Second International Symposium, SAGT 2009*. Springer, 85–96.

[3] Itai Ashlagi and Yannai A Gonczarowski. 2015. No stable matching mechanism is obviously strategy-proof. *arXiv preprint arXiv:1511.00452* (2015).

[4] Moshe Babaioff, Nicole Immorlica, Brendan Lucier, and S Matthew Weinberg. 2014. A simple and approximately optimal mechanism for an additive buyer. In *Foundations of Computer Science (FOCS)*. IEEE, 21–30.

[5] Sophie Bade and Yannai A. Gonczarowski. 2017. Gibbard-Satterthwaite Success Stories and Obvious Strategyproofness. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17*. ACM, 565.

[6] Simina Brânzei and Ariel D Procaccia. 2015. Verifiably truthful mechanisms. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 297–306.

[7] Ioannis Caragiannis, Edith Elkind, Mario Szegedy, and Lan Yu. 2012. Mechanism design: from partial to probabilistic verification. In *ACM Conference on Electronic Commerce, EC '12*. ACM, 266–283.

[8] Shuchi Chawla, Jason D Hartline, David L Malec, and Balasubramanian Sivan. 2010. Multi-parameter mechanism design and sequential posted pricing. In *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 311–320.

[9] Diodato Ferraioli, Paolo Serafino, and Carmine Ventre. 2016. What to Verify for Optimal Truthful Mechanisms without Money. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. IFAAMAS, 68–76.

[10] Diodato Ferraioli and Carmine Ventre. 2017. Obvious Strategyproofness Needs Monitoring for Good Approximations. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 516–522.

[11] Diodato Ferraioli, Carmine Ventre, and Gabor Aranyi. 2015. A Mechanism Design Approach to Measure Awareness. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI, 886–892.

[12] Dimitris Fotakis, Piotr Krysta, and Carmine Ventre. 2017. Combinatorial Auctions Without Money. *Algorithmica* 77, 3 (2017), 756–785.

[13] Dimitris Fotakis, Christos Tzamos, and Manolis Zampetakis. 2016. Mechanism Design with Selective Verification. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16*. ACM, 771–788.

[14] Jason D Hartline and Tim Roughgarden. 2009. Simple versus optimal mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*. ACM, 225–234.

[15] Matthew Jackson and Herve Moulin. 1992. Implementing a public project and distributing its cost. *Journal of economic Theory* 57, 1 (1992), 125–140.

[16] Shengwu Li. 2015. Obviously strategy-proof mechanisms. *Available at SSRN 2560028* (2015).

[17] Paolo Penna and Carmine Ventre. 2014. Optimal Collusion-Resistant Mechanisms with Verification. *Games and Economic Behavior* 86 (2014), 491–509.

[18] Ariel D. Procaccia and Moshe Tennenholtz. 2013. Approximate Mechanism Design without Money. *ACM Trans. Economics and Comput.* 1, 4 (2013), 18:1–18:26.

[19] Tuomas Sandholm and Andrew Gilpin. 2003. Sequences of take-it-or-leave-it offers: Near-optimal auctions without full valuation revelation. In *International Workshop on Agent-Mediated Electronic Commerce*. Springer, 73–91.