

# Incentive-Compatible Recovery from Manipulated Signals, with Applications to Decentralized Physical Infrastructure\*

Jason Milionis<sup>†</sup>    Jens Ernstberger<sup>‡</sup>    Joseph Bonneau<sup>§</sup>  
 Scott Duke Kominers<sup>¶</sup>    Tim Roughgarden<sup>||</sup>

Initial version: February 3, 2025

Current version: March 4, 2025

## Abstract

We introduce the first formal model capturing the elicitation of unverifiable information from a party (the “source”) with implicit signals derived by other players (the “observers”). Our model is motivated in part by applications in decentralized physical infrastructure networks (a.k.a. “DePIN”), an emerging application domain in which physical services (e.g., sensor information, bandwidth, or energy) are provided at least in part by untrusted and self-interested parties. A key challenge in these *signal network* applications is verifying the level of service that was actually provided by network participants.

We first establish a condition called source identifiability, which we show is necessary for the existence of a mechanism for which truthful signal reporting is a strict equilibrium. For a converse, we build on techniques from peer prediction to show that in every signal network that satisfies the source identifiability condition, there is in fact a strictly truthful mechanism, where truthful signal reporting gives strictly higher total expected payoff than any less informative equilibrium. We furthermore show that this truthful equilibrium is in fact the unique equilibrium of the mechanism if there is positive probability that any one observer is unconditionally honest (as would happen, for example, if an observer were run by the network owner). Also, by extending our condition to coalitions, we show that there are generally no collusion-resistant mechanisms in the settings that we consider.

We apply our framework and results to two DePIN applications: proving *location*, and proving *bandwidth*. In the location-proving setting observers learn (potentially enlarged) Euclidean distances to the source. Here, our condition has an appealing geometric interpretation, implying that the source’s location can be truthfully elicited if and only if it is guaranteed to lie inside the convex hull of the observers. In the bandwidth-proving setting, we consider observers that receive noisy (and possibly throttled) evaluations of a source’s bandwidth; we show that our mechanism gives a quasi-strict truthful equilibrium, meaning that the source is disincentivized from reporting a larger bandwidth than they have available.

---

\*Milionis’s research is supported in part by NSF awards CNS-2212745, CCF-2332922, CCF-2212233, DMS-2134059, and CCF-1763970, by an Onassis Foundation Scholarship, and an A.G. Leventis educational grant. Kominers gratefully acknowledges support from the Digital Data Design D<sup>3</sup> Institute at Harvard and the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. Roughgarden’s research at Columbia University is supported in part by NSF awards CCF-2006737 and CNS-2212745. Bonneau, Kominers, and Roughgarden hold positions at a16z crypto (for general a16z disclosures, see <https://www.a16z.com/disclosures/>). Milionis and Ernstberger performed work in part during an internship at a16z crypto. Notwithstanding, the ideas and opinions expressed herein are those of the authors, rather than of a16z or its affiliates. Kominers and Roughgarden also advise companies on marketplace and incentive design.

<sup>†</sup>Columbia University. Email: [jm@cs.columbia.edu](mailto:jm@cs.columbia.edu)

<sup>‡</sup>Technical University of Munich. Email: [jens.ernstberger@gmail.com](mailto:jens.ernstberger@gmail.com)

<sup>§</sup>New York University, and a16z crypto. Email: [jb6395@cs.nyu.edu](mailto:jb6395@cs.nyu.edu)

<sup>¶</sup>Harvard Business School, Harvard University, and a16z crypto. Email: [kominers@fas.harvard.edu](mailto:kominers@fas.harvard.edu)

<sup>||</sup>Columbia University, and a16z crypto. Email: [tim.roughgarden@gmail.com](mailto:tim.roughgarden@gmail.com)

# 1 Introduction

## 1.1 Sources, Observers, and Manipulated Signals

We consider a mechanism designer interested in eliciting information  $x$ , drawn from some abstract set  $\mathcal{X}$ , known to a self-interested agent that we call the *source*. We assume that the designer cannot directly verify the accuracy of a self-report  $\hat{x} \in \mathcal{X}$  by the source, but can instead rely also on the reports  $\hat{\mathbf{y}} = \hat{y}_1, \dots, \hat{y}_n$  of  $n \geq 2$  self-interested *observers* that receive signals  $\mathbf{y}$  related to  $x$ . We allow the source to manipulate the distribution from which observers’ signals are drawn.

For example,  $x$  could represent the true location of an object of interest and  $\hat{x}$  the alleged location of that object (as reported by its owner, for example). Each observer  $i$  could represent a sensor, with  $y_i$  being that sensor’s estimate of its distance from the object, as measured e.g. by the empirical round-trip time of communicating with it. The object may be able to manipulate observers’ distance estimates, for example by deliberately delaying before responding to communication requests.

The primary goal of the paper is to characterize when this mechanism problem—the incentive-compatible recovery of the source’s information from the (possibly manipulated and/or misreported) signals received by the mechanism—is solvable. More precisely, we ask:

1. Under what condition(s) on the allowable source manipulations does there exist a prior-free mechanism for which truthful behavior is a strict Bayesian Nash equilibrium?
2. Under what conditions can the truthful equilibrium be made unique?

And conversely:

3. Under what conditions is such a mechanism impossible?

Our study is motivated in part by applications in decentralized physical infrastructure networks (a.k.a. “DePIN”), an emerging application domain in which physical services are provided at least in part by untrusted and self-interested parties. A key challenge in such applications is how to verify the level of service that was actually provided by participants. The location-elicitation problem outlined above is a canonical DePIN application, which arises, for example, in contexts such as verifying that a resource like server or processing capacity is geographically distributed (which is important for robustness to local shocks such as weather events), as well as for confirming that decentralized data collection entities such as weather trackers are in the right place. Another canonical DePIN application is the elicitation of a source’s available bandwidth, based on noisy measurements taken by observers that may have been manipulated by the source artificially throttling its bandwidth.

We stress, however, that the model introduced in this paper is general and is not overly tailored to DePIN applications. For example, the following problem is isomorphic to the above bandwidth-elicitation problem: elicit the true “quality” of a candidate (student, job applicant, etc.) from noisy measurements by observers (letters of recommendation, references, etc.) that may have been manipulated in certain ways by the candidate (e.g., the candidate misrepresenting their abilities to the observers).

## 1.2 Our Contributions

On the modeling and analysis side, our primary contributions are the following:

- We introduce a novel information elicitation problem, with the key feature that the desired information is known solely to one self-interested agent (the source) who can both misreport that information and manipulate the distribution over the correlated signals observed by other agents.
- We provide a sharp characterization of when truthful elicitation is possible in this setting: if and only if an intuitive condition that we call *source identifiability* holds. Intuitively, source identifiability asserts that the source’s true information could in principle be recovered from an infinite number of samples from the manipulated signal distribution. In concrete examples, source identifiability translates to usable guidelines in practice.
- We prove that whenever source identifiability fails to hold, there is no mechanism for which truthful signal reporting is a strict equilibrium.
- When the source identifiability condition holds, meanwhile, we build on techniques from peer prediction to design a signal elicitation mechanism for which truthful reporting is a strictly optimal equilibrium for network participants, in the sense that any less informative equilibrium has strictly lower total expected payoff than is achieved under truthful signal reporting.
- Our mechanism’s guarantee is even stronger when at least one observer is unconditionally honest with positive probability—in that case, the truthful, value-maximizing equilibrium is unique.
- We extend our characterization through source identifiability to coalitions, and as a consequence show that there are generally no collusion-resistant mechanisms in the settings that we consider.

On the applied side, our work is—to our knowledge—the first to take DePIN signal elicitation seriously as an incentive design problem. Existing DePIN frameworks have effectively ignored incentive issues by either simply assuming truthful reporting, or through out-of-mechanism procedures for resolving reporting issues through governance or audits. Our model and results offer a number of insights into DePIN applications:

- We use our general results to characterize when truthful signal elicitation is possible in location signal networks and bandwidth signal networks. These two DePIN categories are actively used in practice (see, e.g., Sheng et al. 2024b; Sheng et al. 2024a), and our results imply crucial design considerations for setting them up, as well as how signal elicitation should be conducted once these networks are deployed.
- In the location-proving setting, observers learn (potentially enlarged) Euclidean distances to the source. Here, the source identifiability condition has an appealing geometric interpretation, implying that the source’s location can be truthfully elicited if and only if it is guaranteed to lie inside the convex hull of the observers. In other words, for incentive-compatible location recovery, be sure to “surround” with observers the possible locations of the object of interest.
- In the bandwidth-proving setting, we consider observers that receive noisy (and possibly throttled) evaluations of a source’s bandwidth; we show that our mechanism gives a quasi-strict truthful equilibrium, meaning that the source is disincentivized from reporting a larger bandwidth than they have available.

- Our result on equilibrium uniqueness under a mild unconditional honesty assumption speaks to and reinforces the importance of “decentralization” in DePIN: this assumption seems particularly likely to hold in a large decentralized setting because when there are many independent agents, there is a nontrivial possibility that at least one of them is not compromised; hence, in a well organized, (sufficiently) decentralized physical infrastructure network, the mere threat of being compared against an honest agent induces coordination on a truthful revelation equilibrium.
- Our impossibility result for collusion-resistant mechanisms (e.g., in settings where an agent can, through sybils, act as both a source and an observer) can be interpreted as the first formal treatment of what is known as the “self-dealing” problem in DePIN. Our result implies that self-dealing must be handled through out-of-mechanism means, such as restrictions on permissionless entry, further refined trust assumptions, or both.

More broadly, our work here shows that DePIN networks are some of the largest and most natural applications for peer prediction and related techniques to ever arise “in the field.”

### 1.3 Related work

Our work relates to the active and expansive body of work on peer prediction mechanisms (Prelec 2004; Miller et al. 2005; Witkowski and Parkes 2012; Zhang and Chen 2014; Waggoner and Chen 2013; Prelec 2021; Schoenebeck and Yu 2023; Kong and Schoenebeck 2019; Radanovic and Faltings 2014; Kong et al. 2020; Richardson and Faltings 2024). A core difference relative to the peer prediction setting is that, in our work, the source is allowed to actively manipulate the other players’ observed signals before those signals are elicited.<sup>1</sup> Most mechanisms for the truthful elicitation of unverifiable information are surprisingly brittle (sensitive) to a number of assumptions; restrictive assumptions have been usually placed on the information structure, population size, signal spaces, and whether the mechanism is aware of the setting’s joint distribution (Zhang and Chen 2014; Schoenebeck and Yu 2023). Currently, the peer prediction mechanisms with the most minimal set of assumptions to obtain ex-ante Pareto dominance to any uninformative equilibrium and strong truthfulness respectively have been given by Schoenebeck and Yu (2023) and Prelec (2021) correspondingly. The former uses a stochastically relevant setting about signals received from individuals by the nature, and the latter requires a stronger assumption than stochastic-relevance of signals, specifically second-order stochastic relevance about how one’s posterior distribution about another player’s signal changes, using a third player’s (truthful) signal. Generic impossibilities in peer prediction regimes with few assumptions have been given by Waggoner and Chen (2013) and Zhang and Chen (2014); our technique for proving impossibility in non-source-identifiable model specifications is inspired by their general ideas.

In the multiple-questions peer prediction regime, to obtain truthfulness, agents are asked to report on multiple correlated tasks (Dasgupta and Ghosh 2013; Shnayder et al. 2016). Alternatives

---

<sup>1</sup>While the possibility of the source manipulating observers’ signals has not been considered in the peer prediction literature, it does seem plausible that it would be a concern in some settings in which peer prediction is used in practice. For example, in settings like that of Hussam et al. (2022) where peer prediction is used to elicit the ability of microentrepreneurs from assessments by their neighbors, we might imagine that, prior to participating in the peer prediction mechanism, individuals would invest effort in convincing their neighbors that they are especially effective at innovating and/or making efficient use of capital. In this sense, our work suggests how to augment the traditional goals of peer prediction mechanism design to address a practical robustness concern that is typically left outside the boundaries of that model.

to this method including estimating the ground truth (Han et al. 2023), including with the help of machine learning techniques, thereby almost making the problem one where partial access to the ground truth can be granted. Relatedly, our unconditional honesty extension bears a semblance to an observation by Gao et al. (2020) that in costly information-gathering scenarios (such as peer grading where effort has to be exerted) comparison to the ground truth with low probability is sufficient to yield a truthful elicitation mechanism; in their work, the trusted evaluator that provides an unbiased estimator of the ground truth is known in advance.

The role of the possibility of an unconditional observer in equilibrium selection is reminiscent of the role of “commitment types” in reputation games (see, e.g., Fudenberg and Maskin 1986; Jaramillo and Srikant 2010, as well as Levin 2006 and the references therein), although in our setting, the commitment type disciplines behavior in a single-shot mechanism rather than in a repeated game where a reputation for commitment can be observed over time. Likewise, the need for the signal structure to be refined enough to render different strategies probabilistically distinguishable appears in various forms throughout game theory; for example, such a condition is used in characterizing when cooperation is possible in repeated games with imperfect public monitoring (Fudenberg et al. 2009; Abreu et al. 1990).

The nascent literature on Decentralized Physical Infrastructure Networks (DePIN) has studied Byzantine (i.e., arbitrary adversarial) behavior in information elicitation systems, with a focus on setting limits on the fraction of the population that can be Byzantine, and assuming that the rest are unconditionally honest, without the consideration of any incentives (Sheng et al. 2024a; Maram et al. 2021; Sheng et al. 2024b). Our work here crucially differs in that we study the players’ rational behavior according to utility functions. Sheng et al. (2024a) and Sheng et al. (2024b) study the respective settings of location and bandwidth capacity elicitation with this in mind. Both Sheng et al. (2024a) and likewise Maram et al. (2021) substantiate the practicality of using (possibly enlarged by manipulation) distances as a relevant assumption in the setting of location verification, and treat players as non-strategic; instead, the former is based on the adversarial model and performs Byzantine-resistant triangulation, while the latter considers the servers trustworthy in their timestamping. We formally study how incentives play out with such mechanisms, and thus achieve a great synergy with high practical relevance.

Goel et al. (2021), motivated in part by the design of decentralized oracle networks, give a non-strongly-truthful peer prediction mechanism in a setting with subjective, correlated beliefs when there are binary observations. The key novelty in the model of Goel et al. (2021) is the assumption that agents face some outside incentive to misreport (which depends on the aggregate outcome), and the paper focuses on how to adapt mechanisms for peer consistency (Faltings and Radanovic 2017) and use suitable side payments between agents to overcome these incentives; the paper also derives stronger results under assumptions about the number of agents that are unconditionally honest. Zhao et al. (2024) study the specific homogeneous partially-verifiable setting of proof verification, where the status of a common object (the “proof”) can be obtained by players exerting costly effort, and implement a peer prediction mechanism to address rational verifier apathy (in a blockchain context, the “verifier’s dilemma”); in our setting, the model is built on the presumption of manipulability of signals received by participants.

## 2 Setting

In this section, we will introduce the model along with our definitions. Unless otherwise explicitly specified (e.g., when we will be discussing robustness to coalitions), all agents are assumed rational and risk-neutral. Among all players, there is one agent (the source) which has a distinguished role, in that we are interested to elicit her (unverifiable) private information from her interaction with the rest of the players in the game induced by the mechanism. The rest of the players have the role of observers which interact with the source and the mechanism, as described informally in [Section 1](#).

### 2.1 The basic model

A complete description of the model follows:

1. Nature chooses, from a joint prior distribution, the source’s signal  $x \in \mathcal{X}$  and  $n$  (private) observers’ characteristics  $\{p_i\}$ .<sup>2</sup>
2. The source chooses an  $n$ -dimensional distribution  $\mathcal{D}$  either from  $L_x$ , where  $L_x$  is a feasible set of distributions of reports (according to application-specific modeling), or any other distribution that does not correspond to any feasible distribution if the source were truthful. Formally, the source chooses  $\mathcal{D} \in L_x \cup \{\hat{\mathcal{D}} \mid \forall x \in \mathcal{X} : \hat{\mathcal{D}} \notin L_x\}$ . We denote by  $L$  the multi-valued function defined by  $f(x) \triangleq L_x$  wherever the context is clear, and we term  $L$  the *model specification*.
3. Nature chooses  $\mathbf{y} \sim \mathcal{D}$ , and each  $y_i$  gets sent to every one of the  $n$  observers (each one privately observes their own signal).

The observers and source then participate in a mechanism  $M$ , with common knowledge of all information above, including the model specification  $L$ .

This model allows potentially for the source to pick among adversarial values, if the distributions belonging to each  $L_x$  are modeled as point masses. In that special case, the set of distributions is then a set of points, out of which the source may choose their favorite one.

Note that in this paper we will consider discrete signal spaces. Our work can be generalized to continuous signal spaces by using techniques in a similar fashion to Schoenebeck and Yu (2023), Radanovic and Faltings (2014), and Richardson and Faltings (2024).

We move on to define our condition ([Definition 1](#)) that we will tie to the existence of a mechanism where signal-truthfulness is a strict Bayesian Nash equilibrium. We use the standard definitions for the Bayesian Nash equilibrium in games with incomplete information.

**Definition 1** (Source identifiability). A source in a model specification  $L$  is called *identifiable* if for any two different  $x_1 \neq x_2$ ,

$$L_{x_1} \cap L_{x_2} = \emptyset,$$

i.e., there exists no distribution that’s exactly the same for two different source signals. Equivalently, a source is identifiable if and only if the *multi-valued* function defined by  $x \mapsto L_x$  is injective.

We call this property *identifiability*, because in line with statistics, it roughly implies that the model’s parameters can be uniquely determined from the probability distribution of the observed

---

<sup>2</sup>Our mechanism will be independent of this distribution (i.e., prior-free). Bayesian Nash equilibria of the mechanism are with respect to this prior. We assume that  $x$  can take on at least two different values and that  $n \geq 2$ .

data. In other words, if one could somehow *perfectly* observe the true data-generating process—e.g., with an infinite amount of data—they would be able to uniquely deduce the value of the parameter from that distribution. Thus, our mechanism’s intuition is to make use of strictly proper scoring rules to ensure that we can truthfully obtain the source’s value, given the rational observers are honest in their signal reporting. We stress that [Definition 1](#) allows for distributions in two different sets  $L_{x_1}$  and  $L_{x_2}$  that are arbitrarily close to each other (e.g., in total variation distance), and only forbids identical distributions.

## 2.2 Example: proof of location

One example we referred to in [Section 1](#) was *location verification*. We can now see how this maps to the formalism in our model, in the following way: Suppose that both the source and observers are located somewhere on the plane. The observers’ locations on the plane are known and in our model, correspond to vectors  $p_i \in \mathbb{R}^2$ . The mechanism designer’s objective is to estimate the (a priori unknown) location of the source, which is going to be  $x \in \mathbb{R}^2$ . Observers gather information from the source, which consist of positive numbers  $y_i$  that are interpreted as the distances between observer  $i$  and the source.

For this example, we suppose that the source can misrepresent its distance to each observer, but can only artificially *increase* its distance to each one individually (e.g., by delaying communications); it cannot make its distance seem smaller than it actually is. In this sense, this example allows arbitrary “one-sided manipulation” by the source. This constraint would be represented with our model specification as  $L_x$  (the feasible set of reports) being a (possibly uncountable) set of point-mass distributions: the set of all potentially enlarged distances to each observer. The source is therefore able to choose its favorite enlarged distances that each observer individually receives.<sup>3</sup>

What does [Definition 1](#) translate to in this setting? In [Section 5.1](#), we show that source identifiability translates to a convex hull condition: a source’s location is identifiable if and only if all possible locations of the source are contained in the convex hull formed by the observers’ locations.<sup>4</sup> This convex hull condition is intuitive and—importantly—gives guidance for how observers should be positioned in practice.

## 3 Main results

We begin with our impossibility result for a signal-truthful mechanism in the case of a model specification where the source is not identifiable.

**Theorem 3.1** (Impossibility when source is not identifiable). *Given any model specification  $L$  where the source is not identifiable, i.e., does not satisfy [Definition 1](#), there exists no mechanism  $M$  taking as input not only the players’ self-declared  $\hat{x}, \hat{\mathcal{D}}, \hat{\mathbf{y}}$  but also the model specification  $L$ , for which signal truthfulness is a strict Bayesian Nash equilibrium.*

*Proof.* For the sake of contradiction, assume there was such a mechanism  $M$ , and that it assigns a payoff  $u((\hat{x}, \hat{\mathcal{D}}), \hat{\mathbf{y}}, L)$  to the source. Because the source in  $L$  is not identifiable, there exist  $x_1 \neq x_2$  and a joint distribution of manipulated observer signals  $\mathcal{D}$  such that  $\mathcal{D} \in L_{x_1} \cap L_{x_2}$ .

<sup>3</sup>The randomization by nature of  $\mathbf{y} \sim \mathcal{D}$  is meaningless in this example, as every “distribution” is just a point mass.

<sup>4</sup>For the exact formalism and details, we refer the interested reader to [Section 5.1](#).

Since signal truthfulness is a strict Bayesian Nash equilibrium for  $M$ , call the respective strategy profile functions  $(s_0(\cdot), s_1(\cdot), \dots, s_n(\cdot))$ , where  $s_0$  denotes the source's strategy, mapping the private values of each player to their actions in the mechanism (the actions are declaring  $(\hat{x}, \hat{\mathcal{D}})$  for the source and  $\hat{y}_i$  for observer  $i$ ); then, it must be that

$$\mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_1, \mathcal{D}), \mathbf{y}, L)|x_1] > \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_2, \mathcal{D}), \mathbf{y}, L)|x_1] \quad (1)$$

$$\mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_2, \mathcal{D}), \mathbf{y}, L)|x_2] > \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_1, \mathcal{D}), \mathbf{y}, L)|x_2] \quad (2)$$

Build the following rogue (i.e., non-truthful) strategy where the source is truthful when its private value is  $x_1$  but behaves the same for  $x_2$  (obviously the truthful  $\mathcal{D}$ , chosen by the source, is feasible for both signals  $x_1, x_2$  by the model specification), i.e.,  $s'_0(x_2) = (x_1, \mathcal{D})$  and otherwise  $s'_0$  is the same as  $s_0$ . We now prove that, since this gives the same expected payoff to the source (conditioning on  $x_1$ ) as the truthful strategy, the Bayesian Nash equilibrium cannot be strict, which is the contradiction finishing the proof.

Indeed, we have that

$$\begin{aligned} \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_2, \mathcal{D}), \mathbf{y}, L)|x_2] &= \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_2, \mathcal{D}), \mathbf{y}, L)|x_1] < \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_1, \mathcal{D}), \mathbf{y}, L)|x_1] = \\ \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_1, \mathcal{D}), \mathbf{y}, L)|x_2] &< \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [u((x_2, \mathcal{D}), \mathbf{y}, L)|x_2], \end{aligned}$$

which is a contradiction, and where the equalities hold because the conditional distribution  $\mathcal{D}$  is the same and the conditioned random variable is independent of the conditioning random variable, and the inequalities are [Eqs. \(1\) and \(2\)](#) respectively.  $\square$

We move on to the positive results, and give a mechanism to truthfully elicit the unverifiable information of the source and observers, subject to [Definition 1](#). For technical convenience, and without loss of generality, we will also make the following assumption which is roughly stochastic relevance *conditioned on the source's truthfulness*.<sup>5</sup>

**Assumption 1** (Technical Condition). *For any  $x \in \mathcal{X}$ , distribution  $\mathcal{D} \in L_x$ ,  $i \in [n]$ , and two  $y_i \neq y'_i$ ,*

$$\Pr_{\mathcal{D}|y_i}[\mathbf{y}_{-i}|y_i] \neq \Pr_{\mathcal{D}|y'_i}[\mathbf{y}_{-i}|y'_i],$$

*i.e., there do not exist two different  $y_i \neq y'_i$  that induce the same conditional distribution (for the truthful  $x$ ) on the rest of all truthfully-received observers' signals.*

[Assumption 1](#) effectively means that  $y_i$  causes the posterior of any observer  $i$  to change based on the (truthful) value they receive from the source. In most common regimes, such an assumption would hold, for example because the source has non-overlapping sets of  $y_i$ 's (c.f., [Section 5.1](#)), or because each of the observers obtains an independent estimate centered on the source's quality of service (c.f., [Section 5.2](#)).

The sub-mechanism that we will use to gather information from the observers about the source belongs to the class of Bayesian Truth Serum (BTS) mechanisms, pioneered by Prelec ([2004](#)); we

---

<sup>5</sup>Because the elicitation of the source's signal is the final sought-after consequence, our results can be generalized to the case that the technical condition does not hold, and the optimal strategy is a quasi-strict equilibrium where observer  $i$  submits any  $y_i$  that—conditioned on the truthful  $x$ —yields the exact same marginal distribution for the rest of all observers, i.e., the strategy groups the non-distinct (in terms of the joint probability distribution)  $y_i$ 's.



specifically use one of the mechanisms in Prelec (2021), although we remark that similar theorems to ours could be proven using many other similar mechanisms, developed by Prelec (2021) and Schoenebeck and Yu (2023).

The mechanism  $M$  is presented in Algorithm 1. We denote by  $\mathbf{1}\{A\}$  the indicator function that is 1 if  $A$  happens, otherwise 0. Recall that a strictly proper scoring rule is a (potentially extended) real-valued function  $P(\mathcal{D}, \mathbf{y})$  that takes as input a probability measure  $\mathcal{D}$  and a realized outcome  $\mathbf{y}$ , and outputs a real number (reward) such that

$$\mathbb{E}_{\mathbf{y} \sim \mathcal{D}'} [P(\mathcal{D}', \mathbf{y})] \leq \mathbb{E}_{\mathbf{y} \sim \mathcal{D}} [P(\mathcal{D}, \mathbf{y})] \text{ for all distributions } \mathcal{D}, \mathcal{D}',$$

with equality if and only if  $\mathcal{D}' = \mathcal{D}$ .

---

**Algorithm 1:** Mechanism  $M$  run after model with inputs  $(x, \mathcal{D}), \mathbf{y} \sim \mathcal{D}$

---

1. Observers submit  $\hat{y}_i$  to the mechanism.
2. The source submits  $(\hat{x}, \hat{\mathcal{D}})$ , where  $\hat{x} \in \mathcal{X}$ , to the mechanism and to the observers.
3. Observers submit  $\pi_i \in (0, 1]$  and  $\hat{x}_i \in \mathcal{X} \cup \{\emptyset\}$  to the mechanism.
4. Each observer  $i$  is paired (by the mechanism) with a random observer  $j$ , and submits a probability distribution for  $j$ 's signal to the mechanism.<sup>6</sup> The probability distribution is defined by non-negative numbers  $\hat{q}_i(\cdot)$  that sum to 1 across  $j$ 's support of signals.
5. Each observer  $i$  obtains reward

$$\log \left( \frac{\hat{q}_i(\hat{y}_j)}{\pi_j} \right) - \left| \log \left( \frac{\hat{q}_j(\hat{y}_i) \pi_j}{\hat{q}_i(\hat{y}_j) \pi_i} \right) \right| + \mathbf{1}\{\hat{x}_1 = \dots = \hat{x}_n\}.$$

6. The source obtains reward

$$P(\hat{\mathcal{D}}, \hat{\mathbf{y}}) + \mathbf{1}\{\hat{x} = \hat{x}_1 = \dots = \hat{x}_n\}, \tag{3}$$

where  $P(\cdot, \cdot)$  is any strictly proper scoring rule.

---

This mechanism is prior-free. Further, the mechanism does not require that  $\hat{\mathcal{D}} \in L_{\hat{x}}$ , and for this reason is also free of the model specification  $L$ . In other words, the mechanism need not know the model specification at all, and our analysis of the mechanism holds so long as the true (private) signals of the observers indeed come from that model.

We next prove a number of desirable properties of this generic mechanism. To state it, we first define the signal-truthful strategy profiles:

**Definition 2.** We call a strategy profile *signal-truthful* if:

---

<sup>6</sup>We note that, per standard procedure in peer prediction mechanisms (see, e.g., Schoenebeck and Yu 2023), one need not ask for an entire probability distribution, but just a single probability (at least in the discrete signals case) by the mechanism choosing a random value as a virtual signal and asking  $i$  for the probability that  $j$ 's signal is that virtual signal;  $i$ 's reward is then to be modified such that if the randomly chosen signal value matches the actually submitted value from  $j$  then the normal reward function is followed, otherwise a (maximal) reward of 0 is given.

- The source, given  $x$ , chooses  $\mathcal{D} \in L_x$ ,<sup>7</sup> and then submits  $(\hat{x}, \hat{\mathcal{D}}) = (x, \mathcal{D})$ .
- Observer  $i$ , given  $y_i$ , the source’s strategy  $(\hat{x}, \hat{\mathcal{D}})$ , and pairing  $j$ , submits  $\hat{y}_i = y_i, \pi_i = c \cdot \hat{\mathcal{D}}_i(y_i)$  (the probability of the marginal on  $i$  to get  $y_i$  as per  $\hat{\mathcal{D}}$ , rescaled by any  $0 < c \leq 1$  which is fixed across observers),  $\hat{x}_i$  such that  $\hat{\mathcal{D}} \in L_{\hat{x}_i}$  (unique by source identifiability) or  $\hat{x}_i = \emptyset$  if none exists, and  $\hat{q}_i(\cdot)$  to be the posterior on  $j$ ’s signal conditional on  $y_i$  as per  $\hat{\mathcal{D}}$ .

**Theorem 3.2** (Truthful equilibrium). *For any  $x \in \mathcal{X}$  (i.e., any prior on  $\mathcal{X}$ ) and any model specification  $L$  where the source is identifiable as per Definition 1, and subject to Assumption 1, there exists  $\mathcal{D} \in L_x$  such that for any  $0 < c \leq 1$ , the signal-truthful strategy profiles as defined in Definition 2 with the choice of  $\mathcal{D}$  are strict Bayesian Nash equilibria of the game induced by the model and the mechanism  $M$ , where strictness is defined disregarding (i.e., aggregating over) any distribution  $\mathcal{D} \in L_x$  for the truthful  $x$ , for all  $x \in \mathcal{X}$ .<sup>8</sup> Additionally, for any less informative equilibrium of the mechanism, there exists a signal-truthful equilibrium with strictly higher total expected payoff.*

*Proof.* First, we prove strict truthfulness. Consider the source and observers separately.<sup>9</sup>

- For the source, assuming all observers are truthful ( $\hat{\mathbf{y}} = \mathbf{y}$ ): the source selects some  $\mathcal{D} \in L_x \cup \left\{ \hat{\mathcal{D}} \mid \forall x \in \mathcal{X} : \hat{\mathcal{D}} \notin L_x \right\}$ . By the strict properness of scoring rule  $P$ , the unique best-response is to submit  $\hat{\mathcal{D}} = \mathcal{D}$ . Because  $\hat{\mathcal{D}} = \mathcal{D}$ , observers will choose  $\hat{x}_i = x$  for all  $i$  by Definition 1 if the chosen  $\mathcal{D} \in L_x$ , otherwise they will choose  $\emptyset$  (which is infeasible for the source to report, as it’s the special signal of the observers that the source was not truthful), since (again by source identifiability) there is no other  $x' \neq x$  that has the same distribution  $\mathcal{D}$ . Strictness for  $\hat{x} = x$  follows.
- For observer  $i$ , assuming all other observers and the source are honest (in particular, this means  $\hat{\mathcal{D}} = \mathcal{D} \in L_x$ ): first,  $\hat{x}_i = \hat{x}$  is the unique best-response by Definition 1. Second, by the stochastic relevance of Assumption 1 conditioned on the source’s truthfulness hence a distribution  $\mathcal{D} \in L_x$ , the submechanism among the observers operates as a strictly truthful peer prediction mechanism (Prelec 2021). Strict truthfulness for the rest of the strategic choices of observer  $i$  follows by the basic mechanism’s strict truthfulness.

It is left to prove the second part of the theorem. Any less informative equilibrium in  $M$  exhibits either pooling on  $\hat{x} = \hat{x}_i \neq x$  or is a less informative equilibrium of the sub-mechanism with observers. In the latter case, first consider the associated payoffs of the observers based only on their reports except for  $\hat{x}_i$ ’s. Applying the data processing inequality twice (see, e.g., Prelec 2021), any signal garbling equilibrium that is less informative (by either randomizing or pooling over a strategy) has strictly less expected payoffs for every observer than the corresponding signal-truthful equilibrium. Therefore, there exists a  $c < 1$  such that the total expected payoff (including the source) of the corresponding signal-truthful equilibrium is strictly higher. For the former case, we repeat the latter argument, because the equilibrium with  $\hat{x} = \hat{x}_i = x$  is a tie in the individual expected payoffs conditional on each player’s signals. This proves the second part of the theorem.  $\square$

<sup>7</sup>Note that our theorem will state that there exists some  $\mathcal{D}$  for a signal-truthful strategy profile; not all  $\mathcal{D}$ ’s might correspond to signal-truthful profiles that are strict Bayesian equilibria.

<sup>8</sup>Recall that this does not detract from signal truthfulness by source identifiability.

<sup>9</sup>In what follows, because of the aggregating notion of strictness explained in the theorem’s statement, we show that, in the extensive form game, strictness is satisfied disregarding (i.e., conditioning on) the choice  $\mathcal{D} \in L_x$  that the source makes in the first step of the game before mechanism  $M$ .

A common observation with some peer prediction mechanisms (Schoenebeck and Yu 2023) is that it is sometimes hard to imagine that players would arrive at non-truthful equilibria that require unnatural coordination in their play. In mechanism  $M$  and any signal-truthful equilibrium,  $\hat{\mathcal{D}} = \mathcal{D}$  provides a natural point for reports on  $\hat{x}, \hat{x}_i$  to pool on; any other choice of  $\hat{\mathcal{D}}$  by the source would in expectation provide them with strictly less payoff, therefore the aforementioned equilibrium could be a natural coordinating strategy profile in the extensive-form game.

## 4 Extensions

### 4.1 Unconditional honesty

The guarantee of [Theorem 3.2](#) can be sharpened further whenever there is positive probability that at least one observer is unconditionally honest:

**Lemma 4.1** (Any probability of observer unconditional honesty yields unique truthful equilibrium). *If there is a positive probability that any one observer is unconditionally honest, then the truthful equilibrium is the only equilibrium of mechanism  $M$ .*

Unconditional honesty of any (random) observer in the game turns the extensive-form game into one where any (other) observer’s information set cannot feasibly have an implicit guarantee around their pair’s behavior given by a Bayesian Nash equilibrium; this is why they must randomize over the (non-trivial) possibility that they get paired with the unconditionally honest observer. It turns out that the mere threat of being matched up to such an observer is enough to deter non-truthful, less-informative equilibria from forming in the game. This—along with application of the implications of [Theorem 3.2](#)—is the reason why the only feasible (unique) equilibrium is the signal-truthful one.

*Proof of [Lemma 4.1](#).* Name the probability  $p_0 > 0$ , and say observer  $j$  is unconditionally honest with probability  $p_0$ . Then, by the strictness of the truthful equilibrium in  $M$ , if observer  $i \neq j$  played any strategy other than the truthful one, then with probability  $p_0/n$  they would obtain strictly less than the maximal payoff achieved with the truthful strategy (because they got paired with a truthful observer), and with probability  $1 - p_0/n$  they will obtain a payoff that (by the second part of [Theorem 3.2](#)) is in expectation less than the truthful one. Thus, by  $p_0/n > 0$ , an equilibrium is only possible if all observers report truthfully, therefore by strictness of the truthful equilibrium of  $M$ , the source will also be truthful. The lemma follows.  $\square$

The intuition and formal argument for [Lemma 4.1](#) make it clear that the role that (the possibility of) unconditional honesty plays here reflects a general idea, which we suspect may be useful more broadly: In peer prediction-based mechanisms, agents’ reports are cross-examined against each other—and the possibility that at least some agents may be unconditionally honest means that any putative non-truthful equilibrium behavior has some risk of being identified, and punished, through cross-examination with an unconditionally honest agent (who always reports truthfully). Thus, even a small positive probability of an unconditionally honest agent helps isolate the truthful equilibrium.

We note also that the assumption that at least one observer might be unconditionally honest is particularly natural in the context of large signal networks with many independent participants—like in the DePIN applications we examine in [Section 5](#). Indeed, with many independent observers, it becomes increasingly reasonable to assume that each observer believes that at least one

observer may not be compromised. (Moreover, in many applications, it may be possible for the signal network’s organizer to directly guarantee that at least one observer is unconditionally truthful—perhaps by managing that observer themselves in a way that is common knowledge.)

## 4.2 Individual rationality

In mechanism  $M$ , with respect to an arbitrary source, the observers can guarantee non-negative expected payoffs if they behave according to a signal-truthful equilibrium as per [Theorem 3.2](#). More specifically, according to straightforward calculations from the mechanism’s payoffs, we note that the expected payoff of  $M$  to any observer (conditional on their signal) if all observers behave truthfully is non-negative, because it is exactly the Kullback-Leibler divergence between the posterior probability distribution of  $j$  conditional on  $i$ ’s signal and the marginal distribution of  $j$ ’s signal according to  $\mathcal{D}$ . This divergence is guaranteed to be non-negative.

For the source, the usual comments applicable to affine transformations of scoring rules to guarantee individual rationality hold: for example, if we choose the quadratic scoring rule, then indeed, by adding  $1/2$  for a transformed scoring rule, the payoff to the source is always non-negative.

## 4.3 Collusion of source with observers

A significant concern in decentralized systems is collusion. A commonly cited reason is that collusion can be readily facilitated with smart contracts that provide the mechanism for parties to coordinate and credibly commit to prescribed behavior. In this section, we will be particularly concerned with collusion of (a subset of) observers with the source, and show that it is essentially impossible for a mechanism to be collusion-resistant and strictly truthful.

**Definition 3** (Source-observers collusion-resistance). Consider a (specific) subset of observers  $\mathcal{C} \subseteq [n]$  that collude with the source. In our setting, we will call a mechanism  $\mathcal{C}$ -collusion-resistant, if and only if for any joint (coordinated) reports of the source and subset  $\mathcal{C}$ , strict truthfulness holds for the source’s value, i.e., it is a strict best-response for the source to report its true value to the mechanism.

We note that this definition is akin to a quasi-strictness definition, because it aggregates over the actions of the other colluding players in the game induced by the mechanism and the model.

**Lemma 4.2.** *Assume that  $\mathcal{C} \subseteq [n]$  is common knowledge to all players and the mechanism. For any model specification  $L$ , consider the following refining as a multi-valued function  $x \rightrightarrows L_x|_{\mathcal{C}} \triangleq \{\mathcal{D}_{\bar{\mathcal{C}}} \mid \mathcal{D} \in L_x\}$ , i.e., every distribution is a marginal of the original model specification over all observers not in the colluding set  $\mathcal{C}$ . Unless source identifiability holds for the model specification defined by  $L|_{\mathcal{C}}$ , there is no mechanism that can be  $\mathcal{C}$ -collusion-resistant where signal-truthfulness is a Bayesian Nash equilibrium (in the same sense as in [Theorem 3.2](#)).*

*Proof.* Forward: Construct an instantiation of mechanism  $M$  ([Algorithm 1](#)), where the mechanism only operates over the subset  $\bar{\mathcal{C}}$  of all observers that do not collude; the mechanism otherwise ignores (does not request) input from observers in  $\mathcal{C}$ . By source identifiability on  $L|_{\mathcal{C}}$  and [Theorem 3.2](#), the desired properties hold.

Reverse: In the framework of our impossibility proof in [Theorem 3.1](#), by the coordination of source and observers’ actions, effectively the actions/reports of players in  $\mathcal{C}$  are dictated by the source. Therefore, the source’s expected payoff ranges only over  $\mathbf{y}_{\bar{\mathcal{C}}} \sim \mathcal{D}_{\bar{\mathcal{C}}}$  for some  $\mathcal{D}_{\bar{\mathcal{C}}} \in L_x|_{\mathcal{C}}$ .

Now, as in [Theorem 3.1](#), by the source’s non-identifiability in  $L|_{\mathcal{C}}$ , consider two different  $x_1 \neq x_2$  for which the conditional distribution can be chosen by the source to be the same, i.e., it holds that  $\mathcal{D}_{\bar{c}} \in L_{x_1|_{\mathcal{C}}} \cap L_{x_2|_{\mathcal{C}}} \neq \emptyset$ . Following the rest of the expected payoffs gives rise to a similar contradiction like in [Theorem 3.1](#).  $\square$

In the context of decentralized physical infrastructure networks (DePIN), whereby participation to a mechanism on the blockchain is generally permissionless and unconstrained, i.e., new players are free to join the mechanism, one special case of such collusion of a source with a subset of observers is when these ”observers” are the source itself. This is referred to as *self-dealing* in the context of DePIN, and our [Lemma 4.2](#) above essentially proves that it is impossible to handle, at least in a prior-free mechanism. Thus, we formally prove that self-dealing *must* be handled out of mechanism, either via restrictions to permissionless entry, further refined trust assumptions, or both.

## 5 Applications

### 5.1 Location signal networks

Continuing the discussion of location verification we began in [Section 2.2](#), we have that the mechanism designer wants to estimate the source’s location, and use the observers’ information gathering to properly incentivize them to conclude the actual source’s location.

More specifically,  $x \in \mathbb{R}^d$  is a vector of a Euclidean space, and each observer’s location is fixed as  $p_i \in \mathbb{R}^d$ . The model specification consists of (possibly enlarged) distances to the source so long as these are plausibly feasible by some other  $x' \in \mathcal{X}$  in the model, and is given in [Environment 1](#).<sup>10</sup>

**Environment 1** (Location signal network). *The source’s location is a point  $x \in \mathcal{X} \subset \mathbb{R}^d$ . Observers are represented by points  $p_i \in \mathbb{R}^d$ , and  $L_x = \{ \{(y_1, y_2, \dots, y_n)\} \mid y_i \geq \text{dist}(p_i, x) \forall i \text{ and } \exists x' \in \mathcal{X} : \forall i : y_i = \text{dist}(p_i, x')\}$ , i.e., every distribution that belongs to  $L_x$  is just a point mass, and feasible reports of the source include all individual values greater than its (minimum) distance to observer  $p_i$  that are consistent with some feasible  $x' \in \mathcal{X}$ .*

As a matter of fact, the definition of this model specification means that the source can claim *any* potentially enlarged distances to observers, not just the plausibly feasible ones. This is because, according to the model description in [Section 2](#), the full set of potential source choices to be revealed to observers, i.e.,  $L_x \cup \{ \hat{\mathcal{D}} \mid \forall x \in \mathcal{X} : \hat{\mathcal{D}} \notin L_x \}$ , includes the full set of strategic choices  $\{ \{(y_1, y_2, \dots, y_n)\} \mid y_i \geq \text{dist}(p_i, x) \forall i \}$ .<sup>11</sup> Therefore, the source may report any individual values that are larger than the actual distance; of course, by the guarantees of [Theorem 3.2](#), they will only be strictly worse off if they do choose to do so and observers are signal-truthful, if the source is identifiable according to [Definition 1](#).

We remark that common alternative noisy models also fall into our framework, e.g.,

$$L_x = \left\{ \left\{ \begin{array}{l} (y_1 + \epsilon_1, \dots, y_n + \epsilon_n) \text{ w.p. } 1/2, \\ (y_1 - \epsilon_1, \dots, y_n - \epsilon_n) \text{ w.p. } 1/2 \end{array} \right\} \mid y_i \geq \text{dist}(p_i, x) \forall i \text{ and } \exists x' \in \mathcal{X} : \forall i : y_i = \text{dist}(p_i, x') \right\}.$$

<sup>10</sup>Everybody knows that the source is somewhere on  $\mathcal{X}$  by the common knowledge property. It should not be possible for the source to enlarge their distances such that they claim some  $x' \neq x$  in order for it to be identifiable, but we include it in the fully general model specification.

<sup>11</sup>It includes many other possible lies of the source as well, but the particular ones of enlarged distances are of interest, as described in [Sections 1](#) and [2.2](#).

Similar noisy models can represent observers which ping the source and are well-suited to participate in our mechanism, since they can readily provide posterior distributions by virtue of noise estimates from such links they have with the source.

**Proposition 5.3** gives a sufficient and roughly necessary condition that characterizes the truthful elicitation of the source’s location: a mechanism with a strictly truthful Bayesian Nash equilibrium can be given if and only if the source is guaranteed to lie inside the convex hull of the observers. Note that for the necessity, we have to exclude trivially distinguishable cases, such as  $\mathcal{X}$  being just two points outside the convex hull on opposite sides of it. To overcome these, since such trivial cases do not add value to the characterization, we require (to prove that the source is not identifiable in these cases) that a non-measure zero (in  $\mathbb{R}^d$ ) mass outside of the convex hull is included in  $\mathcal{X}$ .

In practice, the condition of **Proposition 5.3** is very actionable: it indicates that one should think about where the source  $x$  might be (in  $\mathbb{R}^d$ ), and make sure to ”surround” it on the perimeter with sensors.

We note that in this setting, the source’s reward attains a particularly satisfying format: any scoring rule  $P(\hat{D}, \hat{y})$  rewards consistency at the signal-truthful equilibrium; either the vectors obtained by the observers (which according to **Proposition 5.3** cannot be manipulated) match exactly the claimed ones by the source (which may be arbitrary, since they don’t need to conform to any guidelines according to the model specification) in which case this component of the source’s reward is maximized, or the source does not obtain the maximum reward.

In what follows, we denote by  $\text{Conv}(\{p_1, \dots, p_n\})$  the convex hull defined by the points  $\{p_1, \dots, p_n\}$ . We move on with two helpful lemmas about Euclidean spaces, whose proofs we include in **Appendix A** (**Lemma 5.1** concerns the injectivity of exact distances on any domain that is a subset of the convex hull, and **Lemma 5.2** is about their distance vectors being coordinate-wise incomparable) that will be used to prove **Proposition 5.3**.

**Lemma 5.1.** *The map  $x \mapsto (\text{dist}(p_1, x), \dots, \text{dist}(p_n, x))$  is injective in any domain  $\mathcal{X}$  that is a subset of the convex hull  $\text{Conv}(\{p_1, \dots, p_n\})$ .*

**Lemma 5.2.** *Consider two  $x', x \in \text{Conv}(\{p_1, \dots, p_n\})$ . If it holds that  $\text{dist}(p_i, x') \geq \text{dist}(p_i, x) \forall i$ , then  $x' = x$ . (The converse is trivial, since all distances are the same.)*

**Proposition 5.3** (Convex hull characterization). *In the model defined by **Environment 1**, if  $\mathcal{X} \subseteq \text{Conv}(\{p_1, \dots, p_n\})$ , then the source is identifiable. Conversely, if  $\mathcal{X}$  is a superset of a non-measure zero mass of points outside  $\text{Conv}(\{p_1, \dots, p_n\})$ , then the source is not identifiable.*

*Proof.* Forward: Recall that we need to show that the multi-valued function  $x \rightrightarrows L_x$  is injective. As a result of **Lemma 5.2**, any enlarged distances fall outside of the (truthful) model specification  $L_x$ , because they are not plausibly feasible by any other truthful  $x' \in \mathcal{X} \subseteq \text{Conv}(\{p_1, \dots, p_n\})$ . Therefore,  $x \rightrightarrows L_x$  corresponds exactly to the map that **Lemma 5.1** proves is injective, and this direction is complete.

Reverse: If  $\mathcal{X}$  is a superset of a non-measure zero set of points outside of  $\text{Conv}(\{p_1, \dots, p_n\})$ , then there are two different  $x_1 \neq x_2 \in \mathbb{R}^d$  and a separating hyperplane from the convex hull (represented by its unit normal vector  $u \in \mathbb{R}^d$ ) such that  $\forall i : \langle p_i, u \rangle \geq 0$  and  $x_1 = -\alpha u, x_2 = -\beta u$  for some  $\alpha, \beta > 0$ . Without loss of generality, order  $x_1, x_2$  such that  $\beta > \alpha$ . We show that  $\{(\text{dist}(p_1, x_2), \dots, \text{dist}(p_n, x_2))\} \in L_{x_1} \cap L_{x_2} \neq \emptyset$ , therefore the source is not identifiable. Indeed, it

suffices to prove that  $\forall i : \text{dist}(p_i, x_2) > \text{dist}(p_i, x_1)$ .<sup>12</sup> This is true by computation, since for any  $i$ :

$$\|p_i - x_2\|^2 - \|p_i - x_1\|^2 = (\beta - \alpha)(\beta + \alpha + 2\langle p_i, u \rangle) > 0. \quad \square$$

## 5.2 Bandwidth signal networks

In this setting, the mechanism designer wants to elicit the source’s (ideally maximum available) bandwidth. Observers obtain noisy and possibly throttled estimates about the source’s bandwidth; the model specification is given in [Environment 2](#). A primary rationale for this model specification is the observation that internet connections between two nodes might be throttled, and internet links can operate over multiple hops, therefore even though an observer might have the capacity to notice the full declared bandwidth of the source if connected through a direct peer-to-peer link, they may in fact be connected via a set of intermediate nodes that cannot support this bandwidth. The model, then, would reasonably be expected to be unable to certify a high connection speed, if no observer can witness it. Thus, the model specification below also bakes in the assumption that there is at least one observer capable of probabilistically observing the actual source’s bandwidth.

**Environment 2** (Bandwidth signal network). *The source’s bandwidth is  $x \in \mathbb{R}^+$ . Given  $x$ , every observer obtains independent estimates of the source’s bandwidth, coming from distributions whose support is upper bounded (or truncated) at some value that’s at most  $x$ , i.e.,  $L_x = \{\mathcal{D}_1^x \times \cdots \times \mathcal{D}_n^x \mid 0 \leq \text{support}(\mathcal{D}_i^x) \leq x \forall i\}$ , and  $\exists \mathcal{D}_1^x \times \cdots \times \mathcal{D}_n^x \in L_x$  such that  $\exists i : x \in \text{support}(\mathcal{D}_i^x)$ .*<sup>13</sup>

Unfortunately, most settings following [Environment 2](#) are not source-identifiable, as [Proposition 5.4](#) proves.

**Proposition 5.4.** *In the model of [Environment 2](#), there is at least one model specification where the source is not identifiable.*

*Proof.* There are many example instantiations of the generic model given by [Environment 2](#) that do not satisfy source identifiability.

For example, consider the further refined model, where some of the included distributions (let’s denote them by  $\mathcal{D}_i$ ) in the product distributions contained in  $L_x$  (among others) are distributions upper bounded at some fixed value  $p_i$ , i.e.,  $\text{support}(\mathcal{D}_i) \leq \min\{p_i, x\}$ . We can model this way the source’s choice to *artificially throttle* the bandwidth that it appears that it has to each of the observers; note that in most realistic regimes, this option is practically available to the source. The source can then (strategically) choose these throttled distributions—perhaps to its detriment in a system where high bandwidth is incentivized.

Formally, for any two different  $x_1, x_2$  such that  $x_1 > x_2 > \max_{i \in [n]} \{p_i\}$ , it is clear that

$$\{\mathcal{D}_1 \times \cdots \times \mathcal{D}_n\} \subseteq L_{x_1} \cap L_{x_2} \neq \emptyset;$$

hence, the source is not identifiable according to [Definition 1](#). □

We can now derive a modification of the given guarantees; specifically, we first relax the strictness requirement, as follows.

<sup>12</sup>Notice that here, the quantifier “for all  $i$ ” is the non-trivial part, and why we use the co-linear vectors  $x_1, x_2$  with the hyperplane’s normal vector  $u$ .

<sup>13</sup>This is the condition we impose, because we remind that we consider discrete distributions. Otherwise, we need to impose non-zero measure in a continuous distribution, i.e.,  $\mathcal{D}_i^x(x) > 0$ .

**Definition 4.** A signal-truthful strategy profile of mechanism  $M$  will be called *quasi-strict for the source*, if any  $\hat{x} > x$  attains strictly less payoff for the source when the observers are following the specified signal-truthful strategies.

The relevant lemma follows in [Lemma 5.5](#).

**Lemma 5.5.** *For any prior on  $\mathcal{X}$ , signal-truthfulness defined by [Theorem 3.2](#) in mechanism  $M$ , where<sup>14</sup> we additionally refine the strategy of every observer by reporting  $\hat{x}_i = \max_i \{\text{support}(\hat{\mathcal{D}}_i)\}$  from the received  $\hat{\mathcal{D}}$ , in the setting defined by [Environment 2](#), is quasi-strict for the source, as defined by [Definition 4](#).*

*Proof.* Modifying the proof of [Theorem 3.2](#), for the source’s strategy only, by strict properness of the scoring rule, it’s still going to be that  $\hat{\mathcal{D}} = \mathcal{D}$  for some  $\mathcal{D} \in L_x$  that the source chooses. The source can attain the additional reward of 1 from the indicator function *and* with every challenger reporting  $\hat{x}_i = x$  according to the signal-truthful Bayesian Nash equilibrium, by choosing  $\mathcal{D}$  appropriately, since by [Environment 2](#),  $\exists \mathcal{D} \triangleq \mathcal{D}_1^x \times \dots \times \mathcal{D}_n^x \in L_x$  such that  $\exists i : x \in \text{support}(\mathcal{D}_i^x)$ .<sup>15</sup> Thus, any  $\hat{x} > x$  will give strictly lower payoff to the source than  $x$ , because the indicator will be 0 for any  $\hat{x} > x$ , while at  $x$ , it will be 1.  $\square$

## Acknowledgments

The authors thank Pranav Garimidi, Guy Wuollet, and seminar audiences at a16z crypto for helpful comments.

---

<sup>14</sup>We need to specify the strategy, because for any given  $\hat{\mathcal{D}}$ , there is no longer a unique  $\hat{x}_i$  such that  $\hat{\mathcal{D}} \in L_{\hat{x}_i}$ , due to the source not being identifiable.

<sup>15</sup>Note that the source might also attain 1 from the indicator function if it chooses some other appropriate  $\mathcal{D} \in L_x$ , but no such  $\mathcal{D} \in L_x$  will result in challengers choosing  $\hat{x}_i > x$  at the signal-truthful equilibrium. Rather, challengers might all agree on  $\hat{x}_i < x$ .



## References

- Abreu, Dilip, David Pearce, and Ennio Stacchetti (1990). “Toward a Theory of Discounted Repeated Games with Imperfect Monitoring”. In: *Econometrica* 58.5, pp. 1041–1063. ISSN: 00129682, 14680262. URL: <http://www.jstor.org/stable/2938299> (visited on 02/11/2025).
- Dasgupta, Anirban and Arpita Ghosh (2013). “Crowdsourced judgement elicitation with endogenous proficiency”. In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 319–330.
- Faltings, Boi and Goran Radanovic (2017). “Game theory for data science: Eliciting truthful information”. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.2, pp. 1–151.
- Fudenberg, Drew, David Levine, and Eric Maskin (2009). “The folk theorem with imperfect public information”. In: *A Long-Run Collaboration On Long-Run Games*. World Scientific, pp. 231–273.
- Fudenberg, Drew and Eric Maskin (1986). “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information”. In: *Econometrica* 54.3, pp. 533–554. ISSN: 00129682, 14680262. URL: <http://www.jstor.org/stable/1911307> (visited on 02/11/2025).
- Gao, Alice, James Wright, and Kevin Leyton-Brown (July 2020). “Incentivizing Evaluation with Peer Prediction and Limited Access to Ground Truth (Extended Abstract)”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Journal track. International Joint Conferences on Artificial Intelligence Organization, pp. 5140–5144. DOI: [10.24963/ijcai.2020/723](https://doi.org/10.24963/ijcai.2020/723). URL: <https://doi.org/10.24963/ijcai.2020/723>.
- Goel, Naman, Aris Filos-Ratsikas, and Boi Faltings (2021). “Peer-prediction in the presence of outcome dependent lying incentives”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI’20. Yokohama, Yokohama, Japan. ISBN: 9780999241165.
- Han, Yong, Wenjun Wu, Yu Liang, and Lijun Zhang (June 2023). “Peer Grading Eliciting Truthfulness Based on Autograder”. In: *IEEE Trans. Learn. Technol.* 16, pp. 353–363. ISSN: 1939-1382. DOI: [10.1109/TLT.2022.3216946](https://doi.org/10.1109/TLT.2022.3216946). URL: <https://doi.org/10.1109/TLT.2022.3216946>.
- Hussam, Reshmaan, Natalia Rigol, and Benjamin N. Roth (2022). “Targeting high ability entrepreneurs using community information: Mechanism design in the field”. In: *American Economic Review* 112.3, pp. 861–898.
- Jaramillo, Juan José and R. Srikant (2010). “A game theory based reputation mechanism to incentivize cooperation in wireless ad hoc networks”. In: *Ad Hoc Networks* 8.4, pp. 416–429. ISSN: 1570-8705. DOI: <https://doi.org/10.1016/j.adhoc.2009.10.002>. URL: <https://www.sciencedirect.com/science/article/pii/S1570870509001103>.
- Kong, Yuqing and Grant Schoenebeck (2019). “An information theoretic framework for designing information elicitation mechanisms that reward truth-telling”. In: *ACM Transactions on Economics and Computation (TEAC)* 7.1, pp. 1–33.
- Kong, Yuqing, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu (2020). “Information elicitation mechanisms for statistical estimation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02, pp. 2095–2102.
- Levin, Jonathan (2006). “Reputation in Repeated Interaction”. In.
- Maram, Deepak, Iddo Bentov, Mahimna Kelkar, and Ari Juels (2021). “GoAT: File geolocation via anchor timestamping”. In: *Cryptology ePrint Archive*.

- Miller, Nolan, Paul Resnick, and Richard Zeckhauser (2005). “Eliciting informative feedback: The peer-prediction method”. In: *Management Science* 51.9, pp. 1359–1373.
- Prelec, Dražen (2004). “A Bayesian Truth Serum for Subjective Data”. In: *Science* 306.5695, pp. 462–466. DOI: [10.1126/science.1102081](https://doi.org/10.1126/science.1102081). eprint: <https://www.science.org/doi/pdf/10.1126/science.1102081>. URL: <https://www.science.org/doi/abs/10.1126/science.1102081>.
- Prelec, Drazen (2021). “Bilateral Bayesian truth serum: The nxm signals case”. In: *Available at SSRN 3908446*.
- Radanovic, Goran and Boi Faltings (June 2014). “Incentives for Truthful Information Elicitation of Continuous Signals”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 28.1. DOI: [10.1609/aaai.v28i1.8797](https://doi.org/10.1609/aaai.v28i1.8797). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/8797>.
- Richardson, Adam and Boi Faltings (Mar. 2024). “Peer Neighborhood Mechanisms: A Framework for Mechanism Generalization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.9, pp. 9883–9890. DOI: [10.1609/aaai.v38i9.28849](https://doi.org/10.1609/aaai.v38i9.28849). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/28849>.
- Schoenebeck, Grant and Fang-Yi Yu (2023). “Two strongly truthful mechanisms for three heterogeneous agents answering one question”. In: *ACM Transactions on Economics and Computation* 10.4, pp. 1–26.
- Sheng, Peiyao, Vishal Sevani, Ranvir Rana, Himanshu Tyagi, and Pramod Viswanath (2024a). “BFT-PoLoc: A Byzantine Fortified Trigonometric Proof of Location Protocol using Internet Delays”. In: *arXiv preprint arXiv:2403.13230*.
- Sheng, Peiyao, Nikita Yadav, Vishal Sevani, Arun Babu, Anand Svr, Himanshu Tyagi, and Pramod Viswanath (2024b). “Proof of backhaul: trustfree measurement of broadband bandwidth”. In: *Proceedings 2024 Network and Distributed System Security Symposium*. San Diego, CA, USA: Internet Society. ISBN: 9781891562938. DOI: [10.14722/ndss.2024.24764](https://doi.org/10.14722/ndss.2024.24764). URL: <https://www.ndss-symposium.org/wp-content/uploads/2024-764-paper.pdf> (visited on 02/11/2025).
- Shnayder, Victor, Arpit Agarwal, Rafael Frongillo, and David C Parkes (2016). “Informed truthfulness in multi-task peer prediction”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179–196.
- Waggoner, Bo and Yiling Chen (2013). “Information elicitation sans verification”. In: *Proceedings of the 3rd workshop on social computing and user generated content (SC13)*.
- Witkowski, Jens and David C. Parkes (2012). “Peer prediction without a common prior”. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC ’12. Valencia, Spain: Association for Computing Machinery, pp. 964–981. ISBN: 9781450314152. DOI: [10.1145/2229012.2229085](https://doi.org/10.1145/2229012.2229085). URL: <https://doi.org/10.1145/2229012.2229085>.
- Zhang, Peter and Yiling Chen (2014). “Elicitability and knowledge-free elicitation with peer prediction”. In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 245–252.
- Zhao, Zishuo, Xi Chen, and Yuan Zhou (2024). “It Takes Two: A Peer-Prediction Solution for Blockchain Verifier’s Dilemma”. In: *arXiv preprint arXiv:2406.01794*.

## A Proofs Omitted from the Main Text

### A.1 Proof of Lemma 5.1

*Proof.* Assume the contrary, i.e., that there are two  $x, y \in \text{Conv}(\{p_1, \dots, p_n\})$  such that  $x \neq y$  and  $\forall i : \text{dist}(p_i, x) = \text{dist}(p_i, y)$ . Rearranging, we obtain

$$\langle x - y, p_i \rangle = \frac{\|x\|^2 - \|y\|^2}{2} = c,$$

which is a constant independent of  $i$ . Since  $x, y$  are points that belong to the convex hull, there exist  $\lambda_i, \mu_i \geq 0$  such that  $\sum_i \lambda_i = \sum_i \mu_i = 1$  and  $x = \sum_i \lambda_i p_i$ ,  $y = \sum_i \mu_i p_i$ . We compute

$$\langle x - y, x \rangle = \sum_i \lambda_i \langle x - y, p_i \rangle = c \sum_i \lambda_i = c,$$

and similarly  $\langle x - y, y \rangle = c$ . Thus,  $\|x - y\|^2 = \langle x - y, x - y \rangle = 0$ , therefore  $x = y$ . This is a contradiction.  $\square$

### A.2 Proof of Lemma 5.2

*Proof.* By  $x' \in \text{Conv}(\{p_1, \dots, p_n\})$ , there exist  $\lambda_i \geq 0$  such that  $\sum_i \lambda_i = 1$  and  $x' = \sum_i \lambda_i p_i$ . We calculate

$$\|p_i - x'\|^2 - \|p_i - x\|^2 = \|x - x'\|^2 - 2\langle x - p_i, x - x' \rangle,$$

and then by weighing and summing the square of all inequalities of the lemma's statement, we obtain that

$$0 \leq \sum_i \lambda_i \left( \|p_i - x'\|^2 - \|p_i - x\|^2 \right) = \|x - x'\|^2 - 2 \left\langle x - \sum_i \lambda_i p_i, x - x' \right\rangle = -\|x - x'\|^2,$$

therefore it has to be that  $x' = x$ , since the square norm is non-negative.  $\square$