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

Jason Milionis[1,2], Jens Ernstberger[3], Joseph Bonneau[4,6], Scott Duke Kominers[5,6], and Tim Roughgarden[1,6]

[1] Columbia University
[2] Category Labs
[3] Technical University of Munich
[4] New York University
[5] Harvard University
[6] a16z Crypto

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 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 some 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.

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 $\hat{\boldsymbol{y}} = \hat{y}_1, \ldots, \hat{y}_n$ of $n \geq 2$ self-interested *observers* that receive signals $\boldsymbol{y}$ 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 elicitation 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 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 a 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 astronomical 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 showing that truthful elicitation is possible in this setting if and only if a natural 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. We make the mild assumption (usual in peer prediction mechanisms) that an observer receiving a different signal induces different posterior distributions of other observers' signals; even if the assumption does not hold, because in this setting we are primarily interested in the source's signal, signal aggregation provides an isomorphic case where the source's signal elicitation is possible.
- 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 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., 25, 26], 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: 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" the possible locations of the object of interest with observers.
- In the bandwidth-proving setting, we consider observers that receive noisy (and possibly throttled) evaluations of a source's bandwidth. Source identifiability does not hold in this setting (intuitively because the source can artificially throttle its bandwidth without the observers ever realizing it); nevertheless, 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 can induce 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 illustrates 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

Decentralized Physical Infrastructure Networks (DePIN) applications have been highlighted for their broad reach and potential in a widespread variety of domains, including telecommunications, broadband network access, energy, compute (including for artificial intelligence applications), storage, and general service provision such as delivery services [1]. The nascent research literature on 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 [18, 25, 26]. Location verification has in particular been highlighted as a critical application for many of web3 DePIN applications [see, e.g., 7, which performs triangulations based on a pre-existing semi-trusted set of challengers]. Our work here crucially differs in that we study the players' rational behavior according to utility functions. Sheng et al. [25, 26] study the respective settings of location and bandwidth capacity elicitation with this in mind. Both Sheng et al. [25] and likewise Maram et al. [18] 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 in a highly practical regime.

Our work relates to the active and expansive body of work on peer prediction mechanisms [15, 16, 19–24, 28–30]. A peer prediction mechanism incentivizes truthful reporting of otherwise unverifiable private signals by paying agents based on how well their report predicts (or agrees with) their peers' reports. A core difference relative to the standard 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.[7] Most mechanisms for the truthful elicitation of unverifiable information are surprisingly brittle (i.e., 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 [24, 30]. 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 [24] and Prelec [21], correspondingly. Prelec [21] analyzes a pairwise stochastically relevant setting. Schoenebeck and Yu [24] require a stronger assumption than stochastic relevance of signals, specifically "second-order stochastic relevance," which is a statement of one's posterior about another player's signal changing conditioned on different values of a third player's (truthful) signal. Generic impossibilities in peer prediction regimes with few assumptions have been given by Waggoner and Chen [28] and Zhang and Chen [30]; 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 [5, 27]. Alternatives to this method include estimating the ground truth [12], 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

[7] 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. [13] 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.

honesty extension bears a semblance to an observation by Gao et al. [10] 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., [8, 14], as well as [17] 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 [2, 9].

Goel et al. [11], 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. [11] 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 [6] 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. [31] 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 introduce the model along with our definitions. Unless otherwise explicitly specified (e.g., when we discuss robustness to coalitions), all agents are assumed rational and risk-neutral. Among all players, there is one agent (the source) who has a distinguished role, in that we are interested in eliciting 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 who 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 $\{p_i\}$.[8]

2. The source chooses any n-dimensional distribution $\mathcal{D}$ either from L_x, where L_x is a feasible set of distributions of signals/observations (according to application-specific modeling), or any other distribution that does not correspond to a feasible distribution for any $x' \in \mathcal{X}$. Formally, the source chooses $\mathcal{D} \in L_x \cup \left\{ \hat{\mathcal{D}} \mid \forall x' \in \mathcal{X} : \hat{\mathcal{D}} \notin L_{x'} \right\}$. 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 $\boldsymbol{y} \sim \mathcal{D}$, and each observer i privately receives their own signal y_i.

The observers and source then participate in a mechanism $\mathscr{M}$, with common knowledge of all information above, including the model specification L. We envision the mechanism as implemented via a smart contract on a blockchain such as Ethereum, or by any non-strategic entity, such as a trusted third party.[9] The mechanism may request reports about any information that the parties either hold or can predict based on their common knowledge about the structure of the model specification and signals each participant individually and privately observes. The mechanism then aggregates the information and computes rewards and penalties.

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.

Note that in this paper we consider discrete signal spaces. Our work can be generalized to continuous signal spaces by using techniques analogous to those of Schoenebeck and Yu [24], Radanovic and Faltings [22], and Richardson and Faltings [23].

We next define our condition (Definition 1) that characterizes the existence of a mechanism in which signal-truthfulness is a strict Bayesian Nash equilibrium.[10]

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

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

[8] Bayesian Nash equilibria of the mechanism are with respect to this prior, but our mechanism will be independent of this distribution (i.e., prior-free). We assume that x can take on at least two different values and that $n \geq 2$. Observers have characteristics that can be used, for example, to define the information structure that gives the measured signals; e.g., they could be the observers' locations as in Environment 1.

[9] We do not model the incentives of the mechanism operator—which we consider non-strategic—in this work.

[10] We use the standard definitions for the Bayesian Nash equilibrium in games with incomplete information.

i.e., there are no two different source signals leading to the same feasible distribution. Equivalently, a source is identifiable if and only if the multi-valued *function defined by $x \rightrightarrows L_x$ is injective.*

We refer to this property as a form of *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 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, the benefit of source identifiability for our elicitation problem is that, if rational observers are honest in their signal reporting, then a strictly proper scoring rule will be able to use those signals to cross-check the source's reported value with the realized signal distribution. 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); it 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: Suppose that both the source and observers are located somewhere on the plane. The observers' locations 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 $x \in \mathbb{R}^2$. Observers gather information from the source—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 signals/observations) 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.[11]

What does Definition 1 translate to in this context? In Section 5.1, we show that in location verification, 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. This convex hull condition is intuitive and—importantly—gives guidance for how observers should be positioned in practice.

[11] The randomization by nature of $\boldsymbol{y} \sim \mathcal{D}$ is vacuous in this example, as every "distribution" is just a point mass.

3 Main results

3.1 Impossibility when the source is not identifiable

We begin with our impossibility result for a strictly signal-truthful mechanism without source identifiability. All proofs are deferred to Appendix A.

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

3.2 Positive results

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*:[12]

Assumption 1 (Technical condition – conditional stochastic relevance) *For any $x \in \mathcal{X}$, distribution $\mathcal{D} \in L_x$, $i \in [n]$, and two $y_i \neq y_i'$,*

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

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 observers' signals.

Assumption 1 effectively means that different observations by observer i always lead to different posterior distributions of the observations of the remaining observers. In most common regimes, such an assumption would hold, for example because the source has non-overlapping sets of y_is thus conditioning on one would preclude another (c.f., Section 5.1), or because each of the observers obtains an independent estimate centered on the source's quality of service, hence the posterior distribution would shift (c.f., Section 5.2).

The sub-mechanism that we use to gather information from the observers about the source belongs to the class of Bayesian Truth Serum (BTS) mechanisms, pioneered by Prelec [20]; we specifically use one of the mechanisms of Prelec [21], although we remark that similar theorems to ours could be proven using many other similar mechanisms, such as those of Schoenebeck and Yu [24].

Our "cross-checking" mechanism $\mathscr{M}^{\mathrm{CC}}$ is presented in Algorithm 1. We denote

[12] Because the elicitation of the source's signal is the final sought-after consequence, our results can be generalized to the case that Assumption 1 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.

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

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

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

Algorithm 1: Mechanism $\mathscr{M}^{\mathrm{CC}}$ with inputs $(x, \mathcal{D}), \boldsymbol{y} \sim \mathcal{D}$ after model

1. Each observer i submits $\hat{y}_i$ 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. Each observer i submits $\pi_i \in (0, 1]$ 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.[13] 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 receives reward
$$\log\left(\frac{\hat{q}_i(\hat{y}_j)}{\pi_j}\right) - \left|\log\left(\frac{\hat{q}_j(\hat{y}_i)\pi_j}{\hat{q}_i(\hat{y}_j)\pi_i}\right)\right| + \mathbf{1}\{\hat{x}_1 = \cdots = \hat{x}_n\}.$$
6. The source receives reward
$$P(\hat{\mathcal{D}}, \hat{\boldsymbol{y}}) + \mathbf{1}\{\hat{x} = \hat{x}_1 = \cdots = \hat{x}_n\}, \tag{1}$$
where $P(\cdot, \cdot)$ is any strictly proper scoring rule.

This mechanism is prior-free. Furthermore, 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:*

– *The source, given x, chooses $\mathcal{D} \in L_x$, and then submits $(\hat{x}, \hat{\mathcal{D}}) = (x, \mathcal{D})$.*

[13] We note that, per standard procedure in peer prediction mechanisms [see, e.g., 24], 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.

– *Observer i, given y_i, the source's strategy $(\hat{x}, \hat{\mathcal{D}})$, and their pairing with observer 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 some $0 < c \leq 1$ 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 $\hat{q}_i(\cdot) = \hat{\mathcal{D}}_{j|y_i}$ to be the posterior on j's signal conditional on y_i as per $\hat{\mathcal{D}}$.*

Theorem 2 (Truthful equilibrium). *For any $x \in \mathcal{X}$ (and for any prior) 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 $\mathscr{M}^{\mathrm{CC}}$, where strictness is defined disregarding (i.e., aggregating over) any distribution $\mathcal{D} \in L_x$ for the truthful x, for all $x \in \mathcal{X}$.*[14] *Additionally, for any less informative equilibrium of the mechanism (in the sense of Blackwell [3], i.e., a strict garbling of signals), there exists a signal-truthful equilibrium with strictly higher total expected payoff.*

Note that, due to revenue maximization by the source, it is possible that only some $\mathcal{D} \in L_x$ will result in signal-truthful strict Bayesian Nash equilibria. The theorem states that there *exists* a $\mathcal{D}$ that yields a signal-truthful strategy profile, and an equilibrium does not need to argue about *how* the players arrive at such. A common observation with some peer prediction mechanisms [24] is that it can be hard to imagine that players would arrive at non-truthful equilibria that require unnatural coordination in their play. In mechanism $\mathscr{M}^{\mathrm{CC}}$ 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 a strictly lower 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 2 can be sharpened further whenever there is positive probability that at least one observer is unconditionally honest through the following lemma:

Lemma 1 (Positive 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 $\mathscr{M}^{\mathrm{CC}}$.*

Unconditional honesty of any 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

[14] Recall that this does not detract from signal truthfulness by source identifiability.

equilibrium; this is why they must account for 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 arising in the game. This—along with application of the implications of Theorem 2—is the reason why the only feasible (unique) equilibrium is the signal-truthful one.

The intuition and formal argument for Lemma 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 themself in a way that is common knowledge.

4.2 Individual rationality

In mechanism $\mathscr{M}^{\mathrm{CC}}$, 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 2. More specifically, according to straightforward calculations from the mechanism's payoffs, we note that the expected payoff of $\mathscr{M}^{\mathrm{CC}}$ to any observer (conditional on their signal) if all observers behave truthfully is non-negative, because it is [see, e.g., 21] 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 $\mathcal{D}$.[15]

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 of Brier [4] as P in Eq. (1) for the source's reward, then indeed, by adding $1/2$ for a transformed scoring rule, the payoff to the source is always non-negative, since a strictly monotone affine transformation of a strictly proper scoring rule is also a strictly proper scoring rule.

[15] Recall that for distributions P and Q on the same support, the Kullback–Leibler (KL) divergence is $D_{\mathrm{KL}}(P\|Q) \triangleq \mathbb{E}_{z\sim P}\left[\log \frac{P(z)}{Q(z)}\right]$, and it always satisfies $D_{\mathrm{KL}}(P\|Q) \geq 0$ (with equality iff $P = Q$). Here $P = \mathbf{Pr}_{\mathcal{D}}[y_j = \cdot \mid y_i]$ and $Q = \mathbf{Pr}_{\mathcal{D}}[y_j = \cdot]$, so the non-negativity follows directly and the divergence measures how much y_i shifts the beliefs of observer i about y_j.

4.3 Collusion of the source with observers

A significant concern in decentralized systems is collusion, especially since it 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; we show that it is essentially impossible for a mechanism to be collusion-resistant and strictly truthful in this context.

Definition 3 (Source↔observer collusion–resistance). *Consider a (specific) subset of observers $C \subseteq [n]$ that collude with the source. In our setting, we call a mechanism C-*collusion-resistant *if for any joint (coordinated) reports of the source and subset C, strict truthfulness holds for the source's value (in the same sense as in Theorem 2), 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. The following lemma formalizes this impossibility.

Lemma 2. *Assume that $C \subseteq [n]$ is common knowledge to all players and the mechanism. For any model specification L, consider the following refining as a multi-valued function $x \rightrightarrows L_x|_C \triangleq \{\mathcal{D}_{\overline{C}} \mid \mathcal{D} \in L_x\}$, i.e., every distribution is a marginal of the original model specification over all observers not in the colluding set C. Unless source identifiability holds for the model specification defined by $L|_C$, there is no mechanism that can be C-collusion-resistant.*

In the context of decentralized physical infrastructure networks (DePIN), whereby participation in 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 those "observers" are Sybils of the source itself, rather than just coordinating with it. This is referred to as *self-dealing* in the context of DePIN. One particularly worrying special case is if *all* observers are in fact the source itself. Our Lemma 2 above (setting $C = [n]$ shows that it is not possible under any circumstances to satisfy source identifiability for $L|_{[n]}$ regardless of L) 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 outside of the 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.[16]

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*

$$L_x = \{\{(y_1, y_2, \ldots, y_n)\} \mid y_i \geq \mathrm{dist}(p_i, x)\ \forall i \text{ and } \exists x' \in \mathcal{X} : \forall i : y_i = \mathrm{dist}(p_i, x')\},$$

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 the model specification of Environment 1 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., $L_x \cup \left\{\hat{\mathcal{D}} \mid \forall x \in \mathcal{X} : \hat{\mathcal{D}} \notin L_x\right\}$, includes the full set of strategic choices $\{\{(y_1, y_2, \ldots, y_n)\} \mid y_i \geq \mathrm{dist}(p_i, x)\ \forall i\}$.[17] Therefore, the source may report any individual values that are larger than the actual distances; of course, by the guarantees of Theorem 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.

Common alternative noisy models also fall into our framework, e.g., a setting with small local errors such as

$$L_x = \left\{ \left\{ \begin{matrix} (y_1 + \epsilon_1, \ldots, y_n + \epsilon_n) \text{ w.p. } 1/2, \\ (y_1 - \epsilon_1, \ldots, y_n - \epsilon_n) \text{ w.p. } 1/2 \end{matrix} \right\} \middle| \begin{matrix} y_i \geq \mathrm{dist}(p_i, x)\ \forall i, \\ \exists x' \in \mathcal{X} : \forall i : y_i = \mathrm{dist}(p_i, x') \end{matrix} \right\}.$$

Similar noisy models can represent observers who ping the source regularly and can thus provide posterior distributions given their knowledge of how the noise relates to the links they have with the source.

Proposition 1 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 implemented 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

[16] Everybody knows that the source is somewhere on $\mathcal{X}$ by the common knowledge property. The model specification below includes all possible enlarged distances that have some corresponding point in $\mathcal{X}$, i.e., $x' \in \mathcal{X}$ in the language of Environment 1. As we will see later, if L is source identifiable, then for all x, L_x should not have any such other $x' \neq x$.

[17] 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.

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 1 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 $P(\hat{\mathcal{D}}, \hat{\boldsymbol{y}})$ rewards consistency at the signal-truthful equilibrium; either the vectors obtained by the observers (which according to Proposition 1 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 $\mathrm{Conv}(\{p_1, \ldots, p_n\})$ the convex hull defined by the points $\{p_1, \ldots, p_n\}$. Our convex hull characterization depends on two helpful lemmata about Euclidean spaces that may be of independent interest.

Lemma 3. *The map* $x \mapsto (\mathrm{dist}(p_1, x), \ldots, \mathrm{dist}(p_n, x))$ *is injective in any domain* $\mathcal{X}$ *that is a subset of the convex hull* $\mathrm{Conv}(\{p_1, \ldots, p_n\})$.

Lemma 4. *Consider two* $x', x \in \mathrm{Conv}(\{p_1, \ldots, p_n\})$. *If it holds that* $\mathrm{dist}(p_i, x') \geq \mathrm{dist}(p_i, x)\ \forall i$, *then* $x' = x$. *(The converse is trivial, since all distances are the same.)*

Lemma 3 concerns the injectivity of exact distances on any domain that is a subset of the convex hull, and Lemma 4 is about their distance vectors being coordinate-wise incomparable. These lemmata prove Proposition 1.

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

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 of 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,*[18] *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* $\exists i : x \in \text{support}(\mathcal{D}_i^x)$.[19]

Unfortunately, most settings following Environment 2 are not source-identifiable, as Proposition 2 shows.

Proposition 2. *In the model of Environment 2, there is at least one model specification where the source is not identifiable.*

That said, we can derive a modification of the desired guarantees; specifically, we first relax the strictness requirement, as follows, and conclude with Proposition 3.

Definition 4. *A signal-truthful strategy profile of mechanism $\mathscr{M}^{\text{CC}}$ 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.*

Proposition 3. *For any prior on $\mathcal{X}$, signal-truthfulness defined by Theorem 2 in mechanism $\mathscr{M}^{\text{CC}}$, where*[20] *we additionally refine the strategy of every observer by reporting $\hat{x}_i = \max_i \left\{\text{support}(\hat{\mathcal{D}}_i)\right\}$ from the received $\hat{\mathcal{D}}$, in the setting defined by Environment 2, is quasi-strict for the source, as defined by Definition 4.*

5.3 Modeling Remarks

We remark that in deriving a model specification that is both realistic in practice, as well as being able to reflect to the maximal extent the possible misreports of the source, there is a dual mandate: a mechanism designer would like to specify L_x as narrowly as possible (for the sake of satisfying source identifiability more easily), while the feasible set of misreports themselves should be as unrestricted/wide as possible. The former needs to closely reflect the possible states that *nature* can *actually* be put in, while the latter should reflect observers' view of what

[18] We note that this model specification is realistic in the case that bandwidth measurements are performed independently by observers, e.g., sequentially; if they were performed simultaneously, then a different modeling restriction akin to the sum of the bandwidths would need to be appropriately taken into consideration. For the purposes of this incentive design work, we do not model the systems-level procedure by which signals are gathered.

[19] 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., $\mathcal{D}_i^x(x) > 0$.

[20] 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.

source's information looks like. In total, L_x as a complete model specification needs to induce an information structure over all possible observers' signals (hence it should ideally be narrow), but from the viewpoint of each individual observer the set of potential misreports of the source conditioned on the observer's signal—including implied misreports on the information of others—is a superset of the actual feasible misreports if all observers' signals were taken into account (i.e., conditioned on the source's private information).

Acknowledgments The authors thank Yiling Chen, Pranav Garimidi, Yuqing Kong, Bo Waggoner, Guy Wuollet, and seminar audiences at a16z crypto for helpful comments, as well as the the referees of the Financial Cryptography conference, and especially the shepherd.

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^3 Institute at Harvard and the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. Part of this work was conducted while Kominers was visiting the Technological Innovation, Entrepreneurship, and Strategic Management (TIES) Group at the MIT Sloan School of Management; he greatly appreciates their hospitality. Roughgarden's research at Columbia University is supported in part by NSF awards CCF-2006737 and CNS-2212745.

Disclosures Kominers and Roughgarden hold positions at a16z crypto. Milionis and Ernstberger performed work in part during an internship at a16z crypto. Part of this work was conducted while Bonneau held a position at a16z crypto. a16z crypto reviewed a draft of this article for compliance prior to publication and is an investor in various crypto projects, including DePIN protocols (for general a16z disclosures, see https://www.a16z.com/disclosures/). Notwithstanding, the ideas and opinions expressed herein are those of the authors, rather than of a16z or its affiliates. Kominers and Roughgarden advise a number of companies on marketplace and incentive design, and serve as experts on related matters.

References

1. 0xSmac. DePIN's imperfect present & promising future: A deep dive, 2024. URL https://www.compound.vc/writing/depin.
2. Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58 (5):1041–1063, 1990. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2938299.
3. David Blackwell. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2):265–272, 1953.
4. Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.
5. Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 319–330, 2013.
6. Boi Faltings and Goran Radanovic. *Game Theory for Data Science: Eliciting Truthful Information*, volume 11 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2017.
7. Foamspace Corp. FOAM, 2018. URL https://foam.space/publicAssets/FOAM_Whitepaper.pdf.
8. Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1911307.
9. Drew Fudenberg, David Levine, and Eric Maskin. The folk theorem with imperfect public information. In *A Long-Run Collaboration On Long-Run Games*, pages 231–273. World Scientific, 2009.
10. Xi Alice Gao, James R. Wright, and Kevin Leyton-Brown. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence*, 275:618–638, 2019.
11. Naman Goel, Aris Filos-Ratsikas, and Boi Faltings. Peer-prediction in the presence of outcome dependent lying incentives. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, page Art. 18, 2021. ISBN 9780999241165.
12. Yong Han, Wenjun Wu, Yu Liang, and Lijun Zhang. Peer grading eliciting truthfulness based on autograder. *IEEE Transactions on Learning Technologies*, 16(3):353–363, 2023.
13. Reshmaan Hussam, Natalia Rigol, and Benjamin N. Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, 112(3):861–898, 2022.
14. Juan José Jaramillo and R. Srikant. A game theory based reputation mechanism to incentivize cooperation in wireless ad hoc networks. *Ad Hoc Networks*, 8(4):416–429, 2010. ISSN 1570-8705. https://doi.org/https://doi.org/10.1016/j.adhoc.2009.10.002. URL https://www.sciencedirect.com/science/article/pii/S1570870509001103.

15. Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation*, 7(1):Art. 2, 2019.
16. Yuqing Kong, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu. Information elicitation mechanisms for statistical estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2095–2102, 2020.
17. Jonathan Levin. Reputation in repeated interaction. 2006. URL https://web.stanford.edu/~jdlevin/Econ%20286/Reputation.pdf. Lecture Notes.
18. Deepak Maram, Mahimna Kelkar, Iddo Bentov, and Ari Juels. GoAT: File geolocation via anchor timestamping. In *Proceedings of the International Conference on Financial Cryptography and Data Security*, volume 14745 of *Lecture Notes in Computer Science*, pages 35–72. Springer, 2024.
19. Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.
20. Drazen Prelec. A Bayesian truth serum for subjective data. *Science*, 306 (5695):462–466, 2004. https://doi.org/10.1126/science.1102081. URL https://www.science.org/doi/abs/10.1126/science.1102081.
21. Drazen Prelec. Bilateral Bayesian truth serum: The $n \times m$ signals case. *SSRN 3908446*, 2021.
22. Goran Radanovic and Boi Faltings. Incentives for truthful information elicitation of continuous signals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 2014. https://doi.org/10.1609/aaai.v28i1.8797. URL https://ojs.aaai.org/index.php/AAAI/article/view/8797.
23. Adam Richardson and Boi Faltings. Peer neighborhood mechanisms: A framework for mechanism generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9883–9890, 2024.
24. Grant Schoenebeck and Fang-Yi Yu. Two strongly truthful mechanisms for three heterogeneous agents answering one question. *ACM Transactions on Economics and Computation*, 10(4):Art. 14, 2023.
25. Peiyao Sheng, Vishal Sevani, Ranvir Rana, Himanshu Tyagi, and Pramod Viswanath. BFT-PoLoc: A byzantine fortified trigonometric proof of location protocol using internet delays. *arXiv preprint arXiv:2403.13230*, 2024.
26. Peiyao Sheng, Nikita Yadav, Vishal Sevani, Arun Babu, Anand Svr, Himanshu Tyagi, and Pramod Viswanath. Proof of Backhaul: Trustfree measurement of broadband bandwidth. In *Proceedings of the 2024 Network and Distributed System Security Symposium*, San Diego, CA, USA, 2024. Internet Society. ISBN 9781891562938. https://doi.org/10.14722/ndss.2024.24764. URL https://www.ndss-symposium.org/wp-content/uploads/2024-764-paper.pdf.
27. Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196, 2016.
28. Bo Waggoner and Yiling Chen. Information elicitation sans verification. In *Proceedings of The 3rd Workshop on Social Computing and User-Generated Content (SCUGC-13)*, 2013.

29. Jens Witkowski and David C. Parkes. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 964–981, New York, NY, USA, 2012. ISBN 9781450314152. https://doi.org/10.1145/2229012.2229085. URL https://doi.org/10.1145/2229012.2229085.
30. Peter Zhang and Yiling Chen. Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pages 245–252, 2014.
31. Zishuo Zhao, Xi Chen, and Yuan Zhou. It takes two: A peer-prediction solution for blockchain verifier's dilemma. *arXiv preprint arXiv:2406.01794*, 2024.

A Proofs Omitted from the Main Text

A.1 Proof of Theorem 1

For the sake of contradiction, assume there were such a mechanism $\mathscr{M}$, and that it assigns a payoff $u((\hat{x}, \hat{\mathcal{D}}), \hat{\boldsymbol{y}}, L)$ 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 $\mathcal{D} \in L_{x_1} \cap L_{x_2}$.

Since signal truthfulness is a strict Bayesian Nash equilibrium for $\mathscr{M}$ (by hypothesis), call the respective strategy profile functions $(s_0(\cdot), s_1(\cdot), \ldots, s_n(\cdot))$, where s_0 denotes the source's strategy, and s_i denotes observer i'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

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_1, \mathcal{D}), \boldsymbol{y}, L) \mid x_1\right] > \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_2, \mathcal{D}), \boldsymbol{y}, L) \mid x_1\right] \tag{2}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_2, \mathcal{D}), \boldsymbol{y}, L) \mid x_2\right] > \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_1, \mathcal{D}), \boldsymbol{y}, L) \mid x_2\right] , \tag{3}$$

where the expectations are conditional on the source's signal.

Build the following rogue (i.e., non-truthful) strategy where the source is truthful when its private value is x_1 but reports x_1 and chooses the same distribution of observers' signals when its private value is x_2 (allowable, since the above $\mathcal{D}$, 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

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_2, \mathcal{D}), \boldsymbol{y}, L) \mid x_2\right] &= \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_2, \mathcal{D}), \boldsymbol{y}, L) \mid x_1\right] \\
&< \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_1, \mathcal{D}), \boldsymbol{y}, L) \mid x_1\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_1, \mathcal{D}), \boldsymbol{y}, L) \mid x_2\right] \\
&< \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{D}}\left[u((x_2, \mathcal{D}), \boldsymbol{y}, L) \mid x_2\right] ,
\end{aligned}$$

which is impossible (because the first and last term are the same), 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. (2) and (3) respectively.

A.2 Proof of Theorem 2

First, we prove strict truthfulness. Consider the source and observers separately.[21]

[21] 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 $\mathscr{M}^{\mathrm{CC}}$.

– For the source, assuming all observers are truthful ($\hat{\boldsymbol{y}} = \boldsymbol{y}$): 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 sub-mechanism among the observers operates as a strictly truthful peer prediction mechanism [21]. 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 $\mathscr{M}^{\mathrm{CC}}$ 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., 21], any signal garbling equilibrium that is less informative (by either randomizing or pooling over a strategy) has strictly lower 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 can 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.

A.3 Proof of Lemma 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 $\mathscr{M}^{\mathrm{CC}}$, if observer $i \neq j$ played any strategy other than the truthful one, then with probability p_0/n they would obtain strictly lower 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 is in expectation over the observer they get matched to less than or equal to the truthful one (by the proof of the second part of Theorem 2, this holds for every observer). Thus, by $p_0/n > 0$, an equilibrium is only possible if all observers report truthfully, therefore by strictness of the truthful equilibrium of $\mathscr{M}^{\mathrm{CC}}$, the source will also be truthful. The lemma follows.

A.4 Proof of Lemma 2

Forward (if source identifiability holds for the model specification defined by $L|_C$, then there is a C-collusion-resistant mechanism for which signal-truthfulness is a Bayesian Nash equilibrium): Construct an instantiation of mechanism $\mathscr{M}^{\mathrm{CC}}$ (Algorithm 1) in which the mechanism only operates over the subset $\overline{C}$ of all observers that do not collude; the mechanism otherwise ignores (does not request) input from observers in C. By source identifiability on $L|_C$ and Theorem 2, the desired properties hold.

Reverse (if source identifiability does not hold for the model specification defined by $L|_C$, then there is no mechanism that can be C-collusion-resistant where signal-truthfulness is a Bayesian Nash equilibrium): In the framework of our impossibility proof in Theorem 1, one possible method of coordination of source and observers' actions would be for the source to dictate the actions/reports of players in C. Therefore, the source's expected payoff ranges only over $\boldsymbol{y}_{\overline{C}} \sim \mathcal{D}_{\overline{C}}$ for some $\mathcal{D}_{\overline{C}} \in L_x|_C$. Now, as in Theorem 1, by the source's non-identifiability in $L|_C$, 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 $\mathcal{D}_{\overline{C}} \in L_{x_1}|_C \cap L_{x_2}|_C \neq \emptyset$. Following the rest of the expected payoffs gives rise to a similar contradiction like in Theorem 1.

A.5 Proof of Lemma 3

Assume the contrary, i.e., that there are two $x, y \in \mathrm{Conv}(\{p_1, \ldots, p_n\})$ such that $x \neq y$ and $\forall i : \mathrm{dist}(p_i, x) = \mathrm{dist}(p_i, y)$. 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.

A.6 Proof of Lemma 4

By $x' \in \mathrm{Conv}(\{p_1, \ldots, p_n\})$, 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 in the lemma's statement, we obtain that

$$\begin{aligned} 0 &\leq \sum_i \lambda_i \left(\|p_i - x'\|^2 - \|p_i - x\|^2\right) \\ &= \|x - x'\|^2 - 2\left\langle x - \sum_i \lambda_i p_i, x - x' \right\rangle \\ &= -\|x - x'\|^2\,; \end{aligned}$$

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

A.7 Proof of Proposition 1

Forward: Recall that we need to show that the multi-valued function $x \rightrightarrows L_x$ is injective. As a result of Lemma 4, any enlarged distances fall outside of the (truthful) model specification L_x, because they are not plausibly feasible by any other truthful $x' \in \mathcal{X} \subseteq \mathrm{Conv}(\{p_1, \ldots, p_n\})$. Therefore, $x \rightrightarrows L_x$ corresponds exactly to the map that Lemma 3 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 $\mathrm{Conv}(\{p_1, \ldots, p_n\})$, 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 $\{(\mathrm{dist}(p_1, x_2), \ldots, \mathrm{dist}(p_n, x_2))\} \in L_{x_1} \cap L_{x_2} \neq \emptyset$, therefore the source is not identifiable. Indeed, via Environment 1, it suffices to prove that $\forall i : \mathrm{dist}(p_i, x_2) > \mathrm{dist}(p_i, x_1)$.[22] This is true by computation, since for any i:

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

A.8 Proof of Proposition 2

There are many example instantiations of the generic model given by Environment 2 that do not satisfy the source identifiability condition.

For example, consider the further refined model in which some of the included distributions (we 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., $\mathrm{support}(\mathcal{D}_i) \leq \min\{p_i, x\}$. 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 $x_1 > x_2 > \max\limits_{i \in [n]}\{p_i\}$, 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.

A.9 Proof of Proposition 3

Modifying the proof of Theorem 2, for the source's strategy only, by strict properness of the scoring rule, it is 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

[22] 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.

in Environment 2, $\exists \mathcal{D} \triangleq \mathcal{D}_1^x \times \cdots \times \mathcal{D}_n^x \in L_x$ such that $\exists i : x \in \text{support}(\mathcal{D}_i^x)$.[23] 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.

[23] 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 $\hat{x}_i < x$.