Information Elicitation from Rowdy Crowds

Grant Schoenebeck
University of Michigan
Ann Arbor, USA
schoeneb@umich.edu

Fang-Yi Yu
Harvard University
Boston, USA
fangyiyu@seas.harvard.edu

Yichi Zhang
University of Michigan
Ann Arbor, USA
yichiz@umich.edu

ABSTRACT

We initiate the study of information elicitation mechanisms for a crowd containing both self-interested agents, who respond to incentives, and adversarial agents, who may collude to disrupt the system. Our mechanisms work in the peer prediction setting where ground truth need not be accessible to the mechanism or even exist.

We provide a meta-mechanism that reduces the design of peer prediction mechanisms to a related robust learning problem. The resulting mechanisms are ϵ-informed truthful, which means truth-telling is the highest paid ϵ-Bayesian Nash equilibrium (up to ϵ-error) and pays strictly more than uninformative equilibria. The value of ϵ depends on the properties of robust learning algorithm, and typically limits to 0 as the number of tasks and agents increase.

We show how to use our meta-mechanism to design mechanisms with provable guarantees in two important crowdsourcing settings even when some agents are self-interested and others are adversarial.

CCS CONCEPTS

• **Theory of computation** → **Algorithmic mechanism design**; *Unsupervised learning and clustering*; • **Information systems** → *Incentive schemes*; • **Mathematics of computing** → Probabilistic inference problems.

KEYWORDS

information elicitation, crowdscourcing, robust algorithms

ACM Reference Format:
Grant Schoenebeck, Fang-Yi Yu, and Yichi Zhang. 2021. Information Elicitation from Rowdy Crowds. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3442381.3449840

1 INTRODUCTION

Crowdsourcing, the process of employing workers to complete concise tasks, enables the requester (mechanism designer) to collect valuable information. Image annotation, relevance judgment, sentiment analysis, and language translation are now routinely completed through crowdsourcing on platforms like Amazon Mechanical Turk and CrowdFlower. One major challenge for crowdsourcing is ensuring reliable results from a diverse set of workers.

To effectively elicit reliable information, a crowdsourcing mechanism needs to account for agents' incentives, which may vary between individual agents: agents could be strategic, adversarial, and altruistic [16]. Workers who are strategic can be motivated with monetary payments. However, the values of such payments must be chosen carefully; self-interested strategic agents may manipulate their information or effort level to try to gain additional payments from the mechanism. For example, an agent participating in a mechanism that pays agents based on the total number of tasks they complete may hurry through the tasks without investing the effort necessary to complete each task well. A good mechanism must appropriately reward truthful and effortful work.

Workers who are adversarial may have some external incentive to collude to sabotage the mechanism [10, 11, 13, 15]. Such agents are unlikely to respond to monetary incentives and thus are often outside of the purview of the models of agent behavior traditionally considered in mechanism design. However, a rich history of compromised computer systems serves as a warning of the peril of ignoring the possibility of these non-strategic agents attempting to disrupt the system. Examples include denial-of-service attacks against websites [12], computer viruses [17], Google Bombing [25], Goldfinger attacks [20] against nascent cryptocurrencies [3], and, recently, Zoom Bombing [2].

Finally, agents motivated by altruism or honesty may both exert effort and report truthfully regardless of incentives. If every agent is honest, eliciting information reduces to a statistical inference problem illustrated in Figure 1 (a).

One potential way to handle the possible presence of all these different types of workers is to insert random "gold-standard" questions whose answers are known. However, these questions can be cumbersome to construct (e.g., calibrating examples for peer grading is a costly use of instructors' time). Another solution is to pay agents for agreeing with a trusted reviewer. However, this begs the question of how one can determine which reviewers are trustworthy. Moreover, mechanisms employing either of these solutions necessarily incur additional costs—either paying workers to answer questions with known answers or employing additional trusted workers. Furthermore, neither of these methods applies when there is no accessible ground truth, e.g. on matters of opinion.

Peer prediction (sometimes called information elicitation without verification) literature has introduced new techniques to circumvent these hurdles. However, peer prediction in the presence of rowdy crowds with both strategic and adversarial agents faces several challenges. Adversarial inputs may degrade the quality of the output. Additionally, adversarial reports may also malign the

Grant Schoenebeck, Fang-Yi Yu, and Yichi Zhang are pleased to acknowledge the support of the National Science Foundation under grants NSF 1618187 and 2007256. Fang-Yi Yu is partially supported by the National Science Foundation under grants CCF-1718549 and IIS-2007887.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449840

incentives for strategic agents causing even strategic agents to act unpredictably. For example, strategic agents may answer incorrectly believing this will increase their payments due to the effects of adversarial agents. Removing the effects of the adversarial behavior in crowdsourcing systems is made additionally difficult because crowdsourcing workers are often transient and/or anonymous.

Our Contributions. We design crowdsourcing mechanisms that can elicit information from rowdy crowds, where a constant fraction of the crowd can adversarially collude and the remaining agents are strategic. Our mechanisms are asymptotically informed truthful. This means that for any $\epsilon > 0$, for a sufficient number of tasks and/or agents that: 1) truth-telling is an ϵ-Bayesian Nash equilibrium; 2) in expectation, the truth-telling equilibrium pays each agent within ϵ of their optimal payment under any ϵ-Bayesian Nash equilibrium and strictly more than under any uninformative strategy profile, (i.e. where the agents' strategies do not depend on their information). In particular, this means that the effect of the adversaries goes to zero in limit.

In such mechanisms, truth-telling is essentially the best that each agent can expect to do. Thus, strategic agents should always report truthfully and effortfully, just as an honest agent would.

We present a meta-algorithm, the ***Robust Mutual Information Framework (RMIF)***, for designing asymptotically informed truthful mechanisms for rowdy crowds (Sect. 4). This meta-algorithm reduces the mechanism design problem to a certain robust learning problem. As shown in Figure 1. (d), the key is to use the output of the robust learning algorithm to both compute payments and produce outputs that are robust against adversarial influence. To our knowledge, this is the first work that considers information elicitation from a combination of strategic and adversarial agents.

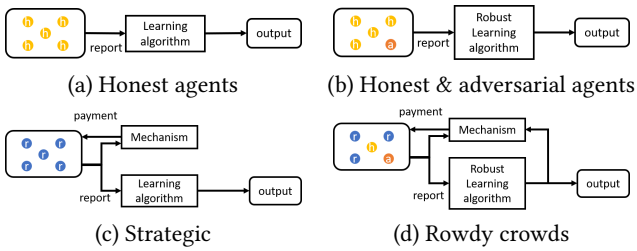
We apply our meta-algorithm to the multi-task setting, where each agent is asked a batch of a priori similar questions, e.g., "is there a bus in the picture?" First, we focus on a general model of peer prediction, where each task need not to have a ground truth, but agents' information for each task is assumed to be correlated (Sect. 5). Second, we consider the Dawid-Skene model [7], where each task has a ground truth, and agents' information is independent conditioned on the ground truth.

For both models, we provide asymptotically informed truthful mechanisms for rowdy crowds that are minimal—i.e. agents need not report any additional information—and detail free—i.e. the mechanism requires no foreknowledge of agents' beliefs or distribution of answers.

1.1 Related Work

The literature on information elicitation without verification focuses on capturing the strategic aspect of human agents (c.f. Figure 1 (c)). In the multi-task setting, Dasgupta and Ghosh [6] proposed a seminal informed truthful[1], minimal, detail-free mechanism. Shnayder et al. [24] and Kong and Schoenebeck [19] independently generalized this beyond binary signals. The former also introduced the concept of informed truthfulness. The latter work proposed a mutual information based meta-algorithm, which our Robust Mutual Information Framework (RMIF) builds upon. The prior mechanism design work does not consider adversarial agents, and offers no guarantees in the presence of adversaries.

Issues of adversarial inputs are broadly studied by the robust learning literature. In particular, the multi-tasks setting corresponds to robust batch learning which is studied in Qiao and Valiant [22] and Chen et al. [5]. In both works, an adversary controls an α fraction of the input samples, while the other $1-\alpha$ fraction of data are i.i.d. sampled from an unknown target distribution. The proposed learning algorithms are shown to be robust such that as the number of samples increases, the error between the output and the target distribution decreases. In addition, empirical works like Goodfellow et al. [14] and Papernot et al. [21] also provide promising approaches to defend against adversary who can alter input data in a separate manner. This approach is illustrated in (b) in Figure 1. Notice that these papers do not consider the mechanism by which the non-adversarial data is procured. Instead, it implicitly assumes that all non-adversarial data is solicited from honest rather than strategic agents.



(a) Honest agents (b) Honest & adversarial agents

(c) Strategic (d) Rowdy crowds

Figure 1: Pipeline models for information elicitation from diverse crowds.

2 MODEL

There are n agents and m tasks. Each agent will report on all the tasks. There is a finite set of possible signals $\mathcal{X}$. As is common in the literature [1, 6, 18], we assume that the tasks are *a priori* similar where the signals for all agents on each tasks are i.i.d. sampled from some prior P on $\mathcal{X}^n$. We use X to denote the random variable of the joint distribution of all agents' signals on all tasks (e.g. with support $\mathcal{X}^{n\times m}$). We use X_i to denote agent i's signals; $X_{i,s}$ to denote agent i's signal on task s; and X_{-i} to denote the signals of all agents except i. Moreover, $\hat{X}$, $\hat{X}_i$, $\hat{X}_{i,s}$ and $\hat{X}_{-i}$ are similar notions for agent's reports. We often consider a set of permissible priors $\mathcal{P}$, and it is a common knowledge among the agents that the actual prior is permissible, i.e., $P \in \mathcal{P}$.

A *multi-task peer prediction mechanism* $\mathcal{M} = (n, m, \mathcal{X}, \mathcal{L})$ collects m reports in the set $\mathcal{X}$ from each of n agents, denoted as $\hat{X}$, and rewards the agents according to the function $\mathcal{L} : \mathcal{X}^{n\times m} \to \mathbb{R}^n$. $\mathcal{L}_i(\hat{X})$ denotes the $i-$th index of the reward function, which is agent i's payoff.

Our model assumes agents have no cost while obtaining signals. However, by scaling the payments, our techniques can be generalized to the setting where agents incur a cost to obtain signals.

[1] Actually, it is strongly truthful, a slightly stronger notion.

2.1 Model for Rowdy Crowds

In this paper, we disregard the existence of the honest agents so that all agents are either *rational*, acting in their selfish interest to maximize their utility, or *adversarial*, acting arbitrarily. We use $\mathcal{A}$ to denote the set of adversarial agents and $\mathcal{R}$ the set of rational agent where $[n] = \mathcal{A} \cup \mathcal{R}$. Honest agents will be discussed in future work (section 7).

Let α be the fraction of adversarial agents so that $|\mathcal{A}| = \alpha n$. Adversarial agents first observe their signals, $\{X_{i,s} : i \in \mathcal{A} \text{ and } s \in [m]\}$. Then they can collude and submit arbitrary reports $\hat{x}_{i,s}$. Since the adversarial agents can collectively decide their reports, we model this as one adversary controlling all adversarial agents behavior. We define the adversary's mapping from their signals to reports as $\sigma_{\mathcal{A}} : \mathcal{X}^{\alpha n \times m} \to \mathcal{X}^{\alpha n \times m}$. We use $\mathcal{S}_{\mathcal{A}}$ to denote the set of strategies available to the adversary.

A rational agent $i \in \mathcal{R}$ wants to maximize her expected payment by reporting strategically. Her strategy $\sigma_i : X_i \to \hat{X}_i$ can be seen as a (random) mapping from her signals X_i to reports $\hat{X}_i$. We make the common assumption from the peer prediction literature [1, 6, 19] that agents' strategies are task-independent. This means agent i chooses a mapping $\sigma_i : \mathcal{X} \to \Delta\mathcal{X}$, which is applied independently to each signal. Thus, $\hat{X}_{i,s}$ is a random variable drawn from $\sigma_i(X_{i,s})$. We use $\mathcal{S}_i$ to denote the set of agent i's possible strategies. We use $\mathcal{S}_{\mathcal{R}} := \prod_{i \in \mathcal{R}} \mathcal{S}_i$ to denote the set of *rational strategy profiles*.

Importantly, note that once all agents' strategies are fixed, $\hat{X}$ is itself a random variable which depends on the randomness of X and the randomness of the strategies.

We call the above setting ***multi-task information elicitation from rowdy crowds*** with parameters $(\mathcal{P}, \alpha)$—RowdyCrowds$(\mathcal{P}, \alpha)$, for short. Given a mechanism $\mathcal{M}$, with prior P, strategies $\boldsymbol{\sigma}_{\mathcal{R}} \in \mathcal{S}_{\mathcal{R}}$, and $\boldsymbol{\sigma}_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$, for agent i we denote agent i's ex-ante payment as

$$u_i^{P,\mathcal{M}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) := \mathop{\mathbb{E}}_{X, \boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}} \left[\mathcal{L}_i(\hat{X})\right].$$

Definition 2.1. In RowdyCrowds$(\mathcal{P}, \alpha)$, a (rational) strategy profile $\boldsymbol{\sigma}_{\mathcal{R}} \in \mathcal{S}_{\mathcal{R}}$ is called an ***ϵ-Bayesian Nash equilibrium (ϵ-BNE)*** in $\mathcal{M}$ if $\boldsymbol{\sigma}$ is an ϵ-BNE regardless of the adversary's strategy or $P \in \mathcal{P}$. Formally, for all $i \in \mathcal{R}$, $\sigma_i' \in \mathcal{S}_i$, and all adversarial strategies $\boldsymbol{\sigma}_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$ with $|\mathcal{A}| = \alpha n$,

$$u_i^{P,\mathcal{M}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \geq u_i^{P,\mathcal{M}}\left(\boldsymbol{\sigma}_{\mathcal{R}\setminus\{i\}}, \sigma_i', \boldsymbol{\sigma}_{\mathcal{A}}\right) - \epsilon.$$

2.2 Mechanism Design

In the literature of information elicitation, there are two particularly important classes of task-independent strategies: the first is the ***truth-telling strategy profile***, $\boldsymbol{\tau}$, where all rational agents' reports are equal to their private signals. The second is an ***uninformed strategy profile***, $\boldsymbol{\theta}$, where all rational agents' strategies are independent of their signals. Ideally, truth-telling should be an equilibrium. It should also be a desirable equilibrium for the agents, so that they play it rather than any other equilibrium.

The below definition of an ϵ-informed truthful mechanism rigorously formulates this goal by adapting the informed truthful definition from Shnayder et al. [24] to our setting.

Definition 2.2. A mechanism $\mathcal{M}$ for the RowdyCrowds$(\mathcal{P}, \alpha)$ setting is ***ϵ-informed-truthful*** if the mechanism is ϵ-informed truthful regardless of adversary's strategy. Formally,

(1) The truth-telling strategy is an ϵ-Bayesian Nash equilibrium;
(2) The truth-telling strategy has the highest payment with ϵ additive error for each agent: for all adversarial strategies $\boldsymbol{\sigma}_{\mathcal{A}}$ and $\boldsymbol{\sigma}_{\mathcal{A}}'$ (which need not be the same), rational strategy profile $\boldsymbol{\sigma}_{\mathcal{R}}$, and $i \in \mathcal{R}$, $u_i^{P,\mathcal{M}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \geq u_i^{P,\mathcal{M}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}') - \epsilon$;
(3) For any uninformed strategy profile $\boldsymbol{\theta}_{\mathcal{R}}$, adversary strategies $\boldsymbol{\sigma}_{\mathcal{A}}$ and $\boldsymbol{\sigma}_{\mathcal{A}}'$ (which need not be the same), and $i \in \mathcal{R}$, $u_i^{P,\mathcal{M}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) > u_i^{P,\mathcal{M}}(\boldsymbol{\theta}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}')$.

It is required that truth-telling pay more (up to additive error ϵ) than any other equilibrium and strictly more than any uninformative equilibrium. This implicitly accounts for the cost to the agents of observing the signal. By scaling up the payments, the payment gap between the truthful and uninformed strategies can be made arbitrary large to overcome the cost of observing the signal. Note that only uninformative strategies can be played without incurring the cost of observing the signal.

Furthermore, we say a mechanism $\mathcal{M}$ is ***asymptotically informed-truthful***, if for all $P \in \mathcal{P}$ and $\epsilon > 0$, the mechanism is ϵ-informed-truthful against an α adversary when n and m are large enough.[2]

As shown by Kong and Schoenebeck [19], f-mutual information can serve as an important tool for truthful mechanism design. Here, we consider a special case of this tool which is used throughout this paper— total variation distance mutual information.

Definition 2.3. Let $P_{X,Y}$ be the joint distribution of random variables X and Y, and P_X, P_Y be the marginal distributions of X, Y respectively. The *total variation distance mutual information* is the total variation distance between $P_{X,Y}$ and $P_X P_Y$, i.e.

$$\|P_{X,Y} - P_X P_Y\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} |P_{X,Y}(x,y) - P_X(x) P_Y(y)|.$$

We will simply use $\mathrm{MI}(X;Y) := \|P_{X,Y} - P_X P_Y\|_{\mathrm{TV}}$.

In this paper, we usually deal with the mutual information between random vectors with i.i.d. entries. For simplicity, we introduce the *termwise mutual information* (twMI). Formally, if X, Y are two random vectors with length of m and i.i.d. entries, $\mathrm{MI}(X_s, Y_s) = \mathrm{MI}(X_{s'}, Y_{s'})$ for all $s, s' \in [m]$. The *termwise mutual information* of the vectors denotes the mutual information of any pair of entries, i.e. $\mathrm{twMI}(X,Y) = \mathrm{MI}(X_s, Y_s)$ for all $s \in [m]$.

The empirical estimator of termwise mutual information, $\widetilde{\mathrm{twMI}}(X,Y)$, uses a realized version of two random vectors X and Y of length m with i.i.d entries with support $\mathcal{X}$ and $\mathcal{Y}$ respectively to estimate their termwise mutual information:

$$\frac{1}{2} \sum_{x,y} \left| \frac{|\{s : X_s = x, Y_s = y\}|}{m} - \frac{|\{s : X_s = x\}||\{s : Y_s = y\}|}{m^2} \right| \quad (1)$$

The ***data-processing inequality*** (DPI) is a well-known (and very useful for our purposes) property of mutual information: suppose the X and Y are two random variables and $M(X)$ is a (random)

[2]Here we assume there exist $\rho > 0$ such that in permissible priors P, the termwise mutual information between any pair of agents i and j is great than ρ.

function applied to X so that $M(X)$ and Y are independent conditioned on X, then

$$\mathrm{MI}(M(X);Y) \le \mathrm{MI}(X;Y). \tag{DPI}$$

Because MI is symmetric, the analogous statement holds if M is applied to Y.

The data-processing inequality implies that applying any strategy on agents' signals can only decrease the mutual information. This property plays an important role in mechanism design. Note that (DPI) also applies to termwise mutual information.

3 PRELIMINARY: ROBUST LEARNING

The problem of learning a discrete distribution given access to independent samples has been intensely studied in the statistics community. In this section, we introduce two different settings for density estimation for a distribution P on a finite space Ω.

In the first setting, estimation is from i.i.d. samples (with no corruption). If we have m_L samples $w_1, \dots, w_{m_L}$ from P, the empirical distribution $\tilde{P}$ from those m_L samples is defined as

$$\tilde{P}(w) := \frac{1}{m_L} \sum_{l \le m_L} \mathbf{1}[w_l = w] \text{ for all } w \in \Omega.$$

The following show that the empirical distribution $\tilde{P}$ has a small total variation distance from the real distribution P

Lemma 3.1 (Theorem 3.1 in [8]). *For any $\epsilon, \delta > 0$, finite domain Ω, and distribution P on Ω, there exists $M = O\left(\frac{1}{\epsilon^2}\max(|\Omega|, (1/\delta))\right)$ such that for all $m_L \ge M$ the empirical distribution with m_L i.i.d. samples, $\tilde{P}_{m_L}$, satisfies $\Pr[\|P - \tilde{P}_{m_L}\|_{\mathrm{TV}} \le \epsilon] \ge 1 - \delta$.*

Now, we introduce the setting of density estimation with an α fraction of corrupted batches. Specifically, the input consists of n_L batches and each batch has k_L data $W \in \Omega^{n_L \times k_L}$. At lease a $(1-\alpha)$ fraction of the batches draw their samples from the distribution P i.i.d. which are called honest batches. The remaining α fraction of batches can be arbitrarily corrupted. The following result shows if the corruption has this batched structure. We can accurately recover the density function P such that the error approaches zero as the number of batches and the size of batches are large enough. Formally,

Theorem 3.2 (Qiao and Valiant [22]). *Let $\alpha \le \alpha_{\mathrm{batch}} = 1/900$, $\delta \in (0,1)$, and $k_L \ge 1$. Given $n_L = O((|\Omega| + k_L + \log(1/\delta))/\alpha^2)$ batches of samples of which a $(1-\alpha)$ fraction of batches consists of k_L iid draws from a distribution P with support Ω, there is an algorithm $\mathcal{L}_{\mathrm{batch}}$ that runs in time $\mathrm{poly}(2^{|\Omega|}, k_L, 1/\alpha, \log(1/\delta))$ and returns a distribution $\tilde{P}$ such that $\|P - \tilde{P}\|_{\mathrm{TV}} = O(\alpha/\sqrt{k_L})$ with probability at least $1-\delta$.*

Note that we can estimate the distribution P with vanishing error as k_L increases. That is because the adversary in the batch model can only corrupt α rows of the data in $W \in \Omega^{n_L \times k_L}$ instead of an arbitrary α fraction of data. This structure allows the algorithm to better detect corrupted data. (See [5] and [22] for more discussion.)

4 META ALGORITHM

Now we provide a framework for designing information elicitation mechanisms for rowdy crowds. We pay each agent i a robust estimation of the termwise mutual information between i's reports $\hat{X}_i$ and a (potentially randomly chosen) *projection function* of the other agents' reports $f(\hat{X}_{-i})$. A projection function $f : \mathcal{X}^{n'} \to \mathcal{Z}$ maps a collection of n' agents' reports to a finite signal space $\mathcal{Z}$. We can extend this to $\boldsymbol{f} : \mathcal{X}^{n' \times m} \to \mathcal{Z}^m$ by applying f to each task independently.

Definition 4.1. In the RowdyCrowds$(\mathcal{P}, \alpha)$ setting, let $\mathcal{L} : \mathcal{X}^{n \times m} \to \mathbb{R}^n$ be a payment function, and let F be a distribution over projection functions, f. Then $(\mathcal{L}, F)$ is a *robust mutual information estimation pair* with ϵ_1, ϵ_2 error if:

For any prior P, truth-telling strategy profile $\boldsymbol{\tau}$, any rational strategy profile $\boldsymbol{\sigma}_{\mathcal{R}} \in \mathcal{S}_{\mathcal{R}}$, and any adversary strategy $\boldsymbol{\sigma}_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$, the expected payment of agent $i \in \mathcal{R}$ satisfies:

$$\begin{aligned} u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) &= \mathop{\mathbb{E}}_{X, \boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}} \left[\mathcal{L}_i(\hat{X})\right] \\ &\ge \mathbb{E}_{f \sim F}[\mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i}))] - \epsilon_1, \text{ and} \qquad (2) \\ u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) &= \mathop{\mathbb{E}}_{X, \boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}} \left[\mathcal{L}_i(\hat{X})\right] \\ &\le \mathbb{E}_{f \sim F}[\mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i}))] + \epsilon_2. \qquad (3) \end{aligned}$$

Moreover, if additionally, we insist the prior is permissible, $P \in \mathcal{P}$, we require:

$$\mathop{\mathbb{E}}_{f \sim F}[\mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i}))] > \epsilon_1 + \epsilon_2 \text{ for all } i \in \mathcal{R}. \tag{4}$$

Under this definition, when the (non-adversarial) agents report truthfully, the expected payment $\mathcal{L}_i$ of each agent $i \in \mathcal{R}$ is approximately lower bounded by a certain mutual information (Eq. 2). Moreover, under any reports, the expected $\mathcal{L}_i$ is approximately upper bounded by this same mutual information (Eq. (3)).

Our **Robust Mutual Information Framework** uses the following theorem to create an $(\epsilon_1+\epsilon_2)$-Informed Truthful Mechanism from a $(\mathcal{L}, F)$ robust mutual information estimation pair.

Theorem 4.2. *For the RowdyCrowds$(\mathcal{P}, \alpha)$ setting, for any $n, m, \mathcal{X}$, if $(\mathcal{L}, F)$ is a robust mutual information estimation pair with (ϵ_1, ϵ_2) error, then $\mathcal{M} = (n, m, \mathcal{X}, \mathcal{L})$ is $(\epsilon_1 + \epsilon_2)$-informed truthful.*

Intuitively, the framework can provide (approximate) informed truthful mechanisms because, first, for strategic agents, truth-telling is an approximately optimal BNE by (2) and (3). Second, in any uninformed equilibrium with information structure P, since $\hat{X}_{\mathcal{R}\setminus\{i\}} = \{\hat{X}_{j,s} : j \in \mathcal{R} \setminus \{i\}, s \in [m]\}$ does not depend on $X_{\mathcal{R}\setminus\{i\}}$, in (3), the rational agents' report is unchanged if we replace the signals to all always be zeros. This renders the right hand side of Eq. (3) equal to zero. Thus, as long as the information structure P and the projection function F satisfy (4), the truth-telling payment exceeds that of any uninformed equilibrium. We leave the complete proof of Theorem 4.2 to Appendix A.

We illustrate two ways to deploy our framework. First, we can design f to be very simple, i.e. projecting onto one variable, and then $\mathcal{L}$ will use robust learning to estimate the termwise mutual information (Sect. 5). Second, we can make f itself robust to adversarial noise (Sect. 6).

5 PEER PREDICTION IN THE GENERAL SETTING

The general setting of peer prediction considers the case where agents' signals are correlated while ground truth need not exist [6,

19, 24]. In this section, we focus on designing $\mathcal{L}$ to be robust while considering F that uniformly outputs a random rational agents' report, i.e. $f(\hat{X}_{-i}) = \hat{X}_j$, where j is selected from $\mathcal{R}\backslash\{i\}$ uniformly at random. We can rewrite Theorem 4.2 as follows.

Corollary 5.1. *In general RowdyCrowds$(\mathcal{P}, \alpha)$ setting, if $\forall P$:*

$$\left| u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) - \mathbb{E}_{j\in\mathcal{R}\backslash\{i\}}\left[\mathbb{E}_{\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_j}[\text{twMI}(\boldsymbol{\sigma}_i(X_i);\boldsymbol{\sigma}_j(X_j))]\right]\right| \le \epsilon \tag{5}$$

for all $i \in \mathcal{R}$ and $\mathbb{E}_{j\in\mathcal{R}\backslash\{i\}}[\text{twMI}(X_i, X_j)] > 2\epsilon$, then RMIF$(\mathcal{L}, F)$ is 2ϵ-informed truthful.

Corollary 5.1 follows directly from Theorem 4.2. On one hand, if all rational agents report truthfully, Eq. (5) is sufficient to show Eq. (2). On the other, Eq. (5) is sufficient to show Eq. (3) since the data processing inequality implies that $E_{\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_j}[\text{twMI}(\boldsymbol{\sigma}_i(X_i), \boldsymbol{\sigma}_j(X_j)] \le \mathbb{E}[\text{twMI}(X_i, X_j)]$ for all $j \in \mathcal{R}\backslash\{i\}$. Thus, we transform the original problem into a problem of estimating $\text{twMI}(\hat{X}_i, \hat{X}_j)$ robustly.

Now, we provide two ϵ-informed truthful mechanisms as applications of our framework. The first is a naive mechanism with $\epsilon = \Theta(\alpha)$; the second is a mechanism based on a robust learning algorithm that is asymptotically informed truthful.

5.1 Naive mechanism

The idea of the naive mechanism (Mechanism 1) is straightforward. We randomly select a peer j and pay i the empirical termwise mutual information between i and j's reports, i.e.

$$\mathcal{L}_i(\hat{X}) = \widetilde{\text{twMI}}\left(\hat{X}_i, \hat{X}_j\right) \text{ for a random } j \in [n] \setminus \{i\}, \tag{6}$$

MECHANISM 1: Naive mechanism for average mutual information between i and other rational agents

Input: Agents' report profile $\hat{X}$ on m tasks, and an index $i \in [n]$
Result: Payment for agent i
Pick $j \in [n]\backslash\{i\}$ uniformly at random.
Compute the empirical joint distribution from agent i's and j's reports, for $x, y \in \mathcal{X}$

$$\tilde{P}_{ij}(x, y) = \frac{|\{s : \hat{X}_{i,s} = x, \hat{X}_{j,s} = y\}|}{m},$$

and compute the empirical marginal distribution, for $x, y \in \mathcal{X}$

$$\tilde{P}_i(x) = \frac{|\{s : \hat{X}_{i,s} = x\}|}{m} \text{ and } \tilde{P}_j(y) = \frac{|\{s : \hat{X}_{j,s} = y\}|}{m}$$

Output

$$\mathcal{L}_i(\hat{X}) := \widetilde{\text{twMI}}(\boldsymbol{\sigma}_i(X_i), \boldsymbol{\sigma}_j(X_j)) = \frac{1}{2}\sum_{x,y\in\mathcal{X}} |\tilde{P}_{ij}(x, y) - \tilde{P}_i(x)\tilde{P}_j(y)| \tag{7}$$

The following theorem shows that this indeed yields an ϵ-informed truthful mechanism. However, the error of the naive algorithm is $\epsilon = \Theta(\alpha)$, where α is the fraction of adversary, which implies that Mechanism 1 is not asymptotically informed truthful.

Theorem 5.2. *In the general RowdyCrowds$(\mathcal{P}, \alpha)$ setting with m tasks and n agents, the naive mechanism is $\left(4(1 + \frac{1}{n-1})\alpha + O_n(\sqrt{\frac{\log m}{m}})\right)$-informed truthful.*

We leave the proof in Appendix B.1, while we sketch how to employ Corollary 5.1 to prove the theorem here.

We must bound the error between $\widetilde{\text{twMI}}$ (payment in Eq. (6)) and the ground truth twMI. On one hand, a rational agent i has probability $1-\alpha$ to be paired with a rational peer j. In this case, $\widetilde{\text{twMI}}$ is close to twMI with an error of $O(\sqrt{\log m/m})$ when the number of tasks is m. In addition, there is an extra error bounded by α which is caused by taking the average over different sets ($j \in [n]\backslash\{i\}$ and $j \in \mathcal{R}\backslash\{i\}$). Furthermore, with probability α the selected peer j is adversarial. In this case, since $\widetilde{\text{twMI}}$ is bounded between 0 and 1, the error is bounded by α.

Theorem 5.2 shows the naive mechanism is $\Theta(\alpha)$-informed truthful. Thus, with a large α, the naive mechanism has a poor truthfulness guarantee. In the next section, we will show that in a general symmetric setting, we can obtain $\epsilon \ll \alpha$.

5.2 Mechanism for Symmetric Priors

We denote $\mathcal{P}_{symm}$ as the set of symmetric priors such that the joint distributions between any pair of agents are identical. Formally, for all $P \in \mathcal{P}_{symm}$ there is a distribution Q on $\mathcal{X}^2$, such that for all $i, j \in [n]$ $P_{i,j} = Q$. Furthermore, we say a joint distribution P is *ε-informative* if the mutual information between any pair of agents' signals is greater than ε. Let $\mathcal{P}^{\varepsilon}_{symm} \subset \mathcal{P}_{symm}$ be the set of symmetric prior such that all $P \in \mathcal{P}^{\varepsilon}_{symm}$ are ε-informative.

The main idea of our mechanism for symmetric priors (Mechanism 2 in Appendix B.2) is to learn this underlying Q robustly and use it to compute the MI as the payments, and then to appeal to Corollary 5.1. We employ the batch learning algorithm $\mathcal{L}_{\text{batch}}$ [22] which can robustly learn an unknown distribution with finite samples. The input of $\mathcal{L}_{\text{batch}}$ is an $N \times K$ matrix where a $1-\alpha$ fraction of the rows contain K i.i.d. sample from the distribution Q, and the remaining α fraction of rows are adversarially chosen. Then, $\mathcal{L}_{\text{batch}}$ returns an estimate of Q with an error asymptotically decreasing with K. (Theorem 3.2)

Theorem 5.3. *Given any $\varepsilon > 0$, for RowdyCrowds$(\mathcal{P}^{\varepsilon}_{symm}, \alpha)$, if and $\alpha < \alpha_{\text{batch}}$, Mechanism 2 is asymptotically informed truthful for symmetric strategy profile.* [3]

Compared with the naive mechanism, in Mechanism 2, we are able to make the error approach 0 by increasing the number of tasks. Details of the proof are included in Appendix B.2. Here we provide a sketch.

Our algorithm for symmetric agents has two phases. First, for a given rational agent i, the joint distributions between $\sigma_i(X_{i,s})$ and $\sigma_j(X_{j,s})$ are the same for all $j \in \mathcal{R}\backslash\{i\}$ and $s \in [m]$, and we denote this as simply $Q_{i,*}$. We create an $(n-1) \times K$ matrix as follows: for each $j \in [n] \setminus \{i\}$ samples (about) K new tasks to obtain a row of K fresh samples from $Q_{i,j}$, where $K = \lfloor m/(n-1) \rfloor$. Because an α fraction of rows are corrupted and the rest are i.i.d. samples from $Q_{i,*}$, we can apply $\mathcal{L}_{\text{batch}}$ to learn an estimate $\tilde{Q}_{i,*}$ of $Q_{i,*}$. Once we have $\tilde{Q}_{i,*}$, we can explicitly estimate the estimated termwise mutual information and pay each agent i accordingly.

Remark. *In order to guarantee all $K(n-1)$ samples are independent, agent i is paired with each $j \in [n]\backslash\{i\}$ K times and each task is used to*

[3] We cannot rule out the possibility that there exists an asymmetric equilibrium in which some rational agent is paid more than in the truth-telling equilibrium.

MECHANISM 2: Algorithm for symm prior on agent i

Input: Agents' report profile $\hat{X}$ on m tasks, and an index $i \in [n]$
Result: Payment for agent i
Set $W := [n] \setminus \{i\}$, $K := \lfloor m/|W| \rfloor$ and $\mathbf{Z}_j = \emptyset$ for all $j \in W$;
// the set of K pairs of reports of agent i and j
for $s \in [m]$ **do**
  while *pick* $j \in W$ *randomly* **do**
    if $|\mathbf{Z}_j| < K$ **then**
      $\mathbf{Z}_j = \mathbf{Z}_j \cup (\hat{X}_{i,s}, \hat{X}_{j,s})$;
      break;
Run $\mathcal{L}_{\text{batch}}$ on $\{\mathbf{Z}_j\}_{j \in W}$ and get a distribution $\tilde{Q}_{i,*}$ on $\mathcal{X}^2$;
// $\{\mathbf{Z}_j\}_{j \in W}$ consists of $|W| = n-1$ batches of i.i.d. samples, and only $O(\alpha)$ fraction of batches are corrupted.
Compute the product of the marginal distribution denoted as $\tilde{R_{i,*}}$ on $\mathcal{X}^2$ where for all $x, y \in \mathcal{X}$

$$\tilde{R}_{i,*}(x, y) = \sum_{z \in \mathcal{X}} \tilde{Q}_{i,*}(x, z) \cdot \sum_{w \in \mathcal{X}} \tilde{Q}_{i,*}(w, y).$$

Output

$$\mathcal{L}_i(\hat{X}) := \frac{1}{2} \sum_{x, y \in \mathcal{X}} |\tilde{Q}_{i,*}(x, y) - \tilde{R}_{i,*}(x, y)| \tag{8}$$

generate one sample from $Q_{i,j}$. Thus, we require the number of tasks to be at least $n-1$, i.e. $K = \lfloor m/(n-1) \rfloor \geq 1$. However, to accommodate small values of m, we can make n small by random selecting a small number the workers and pay them according to our mechanism, and pay the rest of agents zero. Specifically, given α_{batch}, α and ϵ, there exists $n_0 = O\left(\max\left\{\epsilon^{-2}, (\log 1/\epsilon)\max\{(\alpha_{\text{batch}} - \alpha)^{-2}, (\alpha)^{-2}\}\right\}\right)$, such that our mechanism is ϵ-informed truthful mechanism when $n \geq n_0$ and $m \geq \epsilon^{-2}\alpha^2 n_0$. See Appendix B.3.

Furthermore, we can use the batch learning algorithm as a black box, such that the truthfulness guarantee of our mechanism can be improved with any improvement in the batch learning algorithms (see in Appendix B.3.) Our reduction does not lose anything in α, so our mechanism can handle the same fraction of adversaries that the best batch learning algorithm can.

6 PEER PREDICTION ON DAWID AND SKENE MODEL

The RMIF can be particularly powerful if the prior P on agents' signals is a latent variable model where agents' signals are mutually independent conditioned on the latent variables. Examples include Dawid Skene models, Gaussian mixture models, hidden Markov models, and latent Dirichlet allocations.

The key insight is that if P is a latent variable model, using our RMIF, it is sufficient to design a *robust latent label recovery algorithm* r that, for each task $s \in [m]$, can robustly recover the latent variables $Y_s \in \mathcal{Y}$ from the signals on the task, where $\mathcal{Y}$ is the space of possible latent variables. Armed with such a robust latent label recovery algorithm $r : \mathcal{X}^{n'} \to \mathcal{Y}$, we can set the projection function F to deterministically be r. We define the payment function of agent $i \in [n]$ to be $\mathcal{L}(\hat{X}) = \widetilde{\text{twMI}}(\hat{X}_i; \boldsymbol{r}(\hat{X}_{-i}))$, the empirical estimator of the termwise mutual informational between agent i's reports and the recovered latent variables. The pair $(\mathcal{L}, F)$ is a robust mutual information estimation pair.

Here we sketch the general idea while showing a rigorous instantiation in Sect. 6.1. Eq. (2) holds by our definition of $\mathcal{L}$ and F and because the empirical estimate of the mutual information is close to the real value. For Eq. (3), there are three parts. First, by (DPI), the payment at the non-truthful strategy profile is weakly less than the termwise mutual information between agents' signals. In other words, $u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) = \mathbb{E}[\mathcal{L}_i(\hat{X})] = \widetilde{\text{twMI}}(\hat{X}_i; F(\hat{X}_{-i})) \approx \text{twMI}(\hat{X}_i; F(\hat{X}_{-i})) \leq \text{twMI}(X_i; X_{-i})$. Then, because for any $s \in [m]$, $X_{i,s}$ and $X_{j,s}$ are independent conditioned on the latent variable Y_s and $r(X_{-i,s}) \approx Y_s$, we have $\text{twMI}(X_i; X_{-i}) \leq \text{twMI}(X_i; Y) \approx \text{MI}(X_i; F(X_{-i}))$. Finally, because r is a function on the signals on each task $s \in [m]$, $X_{-i,s}$, by (DPI), $\text{twMI}(X_i; F(X_{-i})) \leq \text{twMI}(X_i; X_{-i})$. Combining these, we have Eq. (3), since:

$$\begin{aligned} u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) &\approx \text{twMI}(\hat{X}_i; F(\hat{X}_{-i})) \\ &\leq \text{twMI}(X_i; X_{-i}) \\ &\approx \text{twMI}(X_i; F(X_{-i})) \approx u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}). \end{aligned}$$

In order to satisfy the above conditions, we desire that the robust latent label recovery algorithm r only requires signals on one task as input, instead of all the signals on all tasks. However, we may use other signals to learn the latent label recovery algorithm and use this fixed function to recover latent variable of each tasks.

6.1 Crowdsourcing on Symmetric Dawid and Skene Model

We give an example of this approach by considering the symmetric Dawid and Skene model [7].

Definition 6.1. The ***symmetric Dawid Skene model*** (symm DS model) has parameters $(\mathcal{X}, \mathcal{Y}, w, \Gamma)$ where $\mathcal{X}$ is a set of signals, $\mathcal{Y}$ is the set of latent labels, e.g., $\{\text{good}, \text{bad}\}$, $w \in \Delta_{\mathcal{Y}}$ is the prior distribution of latent labels, and $\Gamma \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ which encodes a distribution of $\mathcal{X}$ conditional on each $y \in \mathcal{Y}$, i.e. $\Gamma_{y,x} = \Pr[X = x | Y = y]$. Formally, for the multitask setting, $P(x_1, \ldots, x_n) = \sum_{y \in \mathcal{Y}} w_y \prod_{i \in [n]} \Gamma_{y,x_i}$. We use Γ_y to denote the row vector of Γ.

Now we use the RMIF from Sect. 4, and design an asymptotically informed-truthful mechanism for rowdy crowds. Note that in the DS model, for any agent i and task s, her signal $X_{i,s}$ is independent of other agents' signals $X_{-i,s}$ on the tasks conditioned on the latent label of task s, Y_s. Suppose we have a robust latent label recovery algorithm $r(X_{-i,s})$ which outputs the latent label Y_s for each s and set the projection function $F \equiv r$. Then Eq. (2) and (3) are satisfied by the above derivation. Therefore, by Theorem 4.2, we can design an approximate informed-truthful mechanism for rowdy crowds by designing an accurate an robust latent label recovery algorithm r.

To successful recover the latent labels, the DS model cannot be "singular". A common assumption requires each row of Γ to be independent. Additionally, we require the fraction of adversary to be smaller than the distance between each row of Γ, $\gamma_{DS} := \frac{\min_{y,y':y \neq y'} D_{\text{KL}}(\Gamma_y; \Gamma_{y'})}{4 \max_{y,x} |\log \Gamma_{y,x}|}$ where $D_{\text{KL}}(\Gamma_y; \Gamma_{y'}) := \sum_x \Gamma_{y',x} \log(\Gamma_{y',x}/\Gamma_{y,x})$ is the KL-divergence from $\Gamma_{y'}$ to Γ_y.

Theorem 6.2. *Let $\alpha^* = \min\{\alpha_{\text{batch}}/3, \gamma_{DS}\}$, where α_{batch} is defined in Theorem 3.2 and γ_{DS} is defined above. For*

RowdyCrowds($\mathcal{P}_{DS}, \alpha$) setting on the symmetric DS model such that $\alpha < \alpha^*$, Mechanism 3 is asymptotically informed-truthful.

We first state our mechanism formally, and then provide a proof of Theorem 6.2 in Appendix C.

To minimize the required number of agents and tasks for our mechanism, we can use the subsampling idea in the remark after Theorem 5.3. Formally, given small enough $\epsilon > 0$ and $\alpha \le \alpha^*$, there exists $n_0 = O\left(\max\left\{(\alpha^* - \alpha)^{-2}\log 1/\epsilon, \left(d_r^{-2} + \alpha^{-2}\log 1/\epsilon\right)^{\frac{1}{2}}\right\}\right)$ where $d_r = \frac{1}{8}\min_{x,y}\Gamma_{y,x}\min\left\{4, \min_{y\neq y'} D_{\mathsf{KL}}(\Gamma_y;\Gamma_{y'})\right\}$, such that if $n > n_0$, $m > d_r^{-2}\alpha^2 n_0^3$, we have an ϵ-informed truthful mechanism.

Furthermore, we also provide a black-box reduction for Mechanism 3. The main difference is that in addition to the black-box batch learning algorithm, we write the latent label recovery algorithm in a black-box form. The details of the subsampling trick and the black-box reduction are provided in Appendix C.4.

6.2 Mechanism details

Our mechanism (Mechanism 3) has three stages.

(1) Estimate the parameters of the DS model (w, Γ) from agents' report profile $\hat{X}$.
(2) For each task $s \in [m]$ infer the latent label $\tilde{y}_s$ based on the reports $\{\hat{X}_{i,s}\}_{i\in[n]}$.
(3) Finally, pay each agent with the empirical estimation of termwise mutual information between her reports and the estimated latent labels.

In the first stage, we estimate the parameters w and Γ robustly. First, we observe that to recover the parameters w and Γ, it suffices to estimate the moments of the distribution P. [26] The first moment M_1 is the marginal distribution of one agent's signal where $(M_1)_x = \sum_y w_y\Gamma_{y,x}$ for all $x \in \mathcal{X}$, and the second and third moments are

$$(M_2)_{x_1,x_2} := \sum_y w_y\Gamma_{y,x_1}\Gamma_{y,x_2} \text{ for all } x_1, x_2 \in \mathcal{X} \tag{9}$$

$$(M_3)_{x_1,x_2,x_3} := \sum_y w_y\Gamma_{y,x_1}\Gamma_{y,x_2}\Gamma_{y,x_3} \text{ for all } x_1, x_2, x_3 \in \mathcal{X}, \tag{10}$$

where M_2 is the probability of two agents' signals and M_3 is the probability of three agents' signals. Because M_2 and M_3 are distributions on a finite space ($\mathcal{X}^2$ and $\mathcal{X}^3$ respectively), we can use a robust batch learner (Theorem 3.2) for density estimation to derive estimations $\tilde{M}_2$ and $\tilde{M}_3$ for second and third moments respectively. Algorithm 1 shows how to use the $\mathcal{L}_{\text{batch}}$ algorithm to estimate the second moment M_2. The idea is very similar to Mechanism 2 in section 5. The algorithm of M_3 can be defined similarly. These moments are indeed density functions on finite domains. Therefore, with a careful decomposition of tasks, we can use a robust density estimation algorithm for batch learning setting [22], to derive estimations of the moments and offset the adversary's attack. Thus, as the number of tasks m increases, the error between $(\tilde{w}, \tilde{\Gamma})$ and the real parameter (w, Γ) vanishes.

In the second stage, with accurate parameters $(\tilde{w}, \tilde{\Gamma})$, we can use maximum likelihood estimator to infer the latent label for each tasks $\hat{y}_s$ for each s in $[m]$. If the fraction of adversaries α is smaller than some constant α^*, which depends on the parameters of DS model (w, Γ), we can recover latent labels for all tasks with high probability when the number of agents n is large enough.

Finally, in the third stage, for each agent $i \in [n]$ we use the empirical estimation of the termwise mutual information between her reports $\hat{X}_i$ and the estimated latent labels $\hat{Y}$.

MECHANISM 3: Mechanism for symmetric DS model

Input: Agents' reports $\hat{X}$ on m tasks
Randomly partition agents $[n]$ into three groups $\{G_0, G_1, G_2\}$ with size at least $\lfloor n/3 \rfloor$ and tasks $[m]$ into $\{T_L, T_R\}$ with size at least $\lfloor m/2 \rfloor$ which partition the reports into six blocks: $\hat{X}^{(g,h)} := \{\hat{X}_{i,s} : i \in G_g, s \in T_h\}$ for $g = 0, 1, 2$ and $h = L, R$.
for $g \in \{0, 1, 2\}$ **do** // Estimate the parameters
  Estimate the second and the third order moments $\tilde{M}_2^{(g)} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ and $\tilde{M}_3^{(g)} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|\times|\mathcal{X}|}$ defined in (9) and (10) by running the robust batch learning algorithm $\mathcal{L}_{\text{batch}}$ (in Theorem 3.2) on $X^{(g,L)}$.
  Compute the whitening matrix $Q \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ where $\tilde{M}_2^{(g)}(Q, Q) := Q^\top \tilde{M}_2^{(g)} Q = \boldsymbol{I}_{|\mathcal{Y}|} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$.
  Use the robust tensor power method to compute eigenvalue-eignevector pairs $\{(\lambda_y, v_y) : y \in \mathcal{Y}\}$ of the whitened tensor $\tilde{M}_3^{(g)}(Q, Q, Q)$. Then compute $\tilde{w}_y^{(g)} = \lambda_y^{-2}$ and $\tilde{\Gamma}_y^{(g)} = \lambda_y(Q^\top)^{-1}v_y \in \mathbb{R}^{|\mathcal{X}|}$.
  Set $\tilde{w}^{(g)} \in \Delta_{\mathcal{Y}}$ and $\tilde{\Gamma}^{(g)} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ naturally combine the $\tilde{w}_y^{(g)}$ and the $\tilde{\Gamma}_y^{(g)}$ respectively.
for $g \in \{0, 1, 2\}$, $s \in T_R$ **do** // Estimate the latent labels
  Estimate the latent label of task s, $\tilde{Y}_s^{(g+1 \pmod 3)}$ with a maximum likelihood estimator using the parameters from group G_g, $(\tilde{w}^{(g)}, \tilde{\Gamma}^{(g)})$, and reports from group $G_{g+1 \pmod 3}$, $\{\hat{X}_{j,s} : j \in G_{g+1 \pmod 3}\}$,

$$\tilde{Y}_s^{(g+1 \pmod 3)} = \operatorname*{argmax}_{y\in\mathcal{Y}}\left\{\log \tilde{w}_y^{(g)} + \sum_{j\in G_{g+1 \pmod 3}} \log \tilde{\Gamma}_{y,\hat{X}_{j,s}}^{(g)}\right\}. \tag{11}$$

for $i \in [n]$ **do** // Compute the payment for each agent
  Set $\tilde{Y} = Y^{(g)}$ when $i \in G_{g+1 \pmod 3}$
  Compute and pay agent i with the empirical total variational distance mutual information $\widetilde{\text{twMI}}$ from the at least $\lfloor m/2 \rfloor$ samples $\{(\hat{X}_{i,s}, \tilde{Y}_s) : s \in T_R\}$ in $\mathcal{X} \times \mathcal{Y}$.

Recall that for our RIMF, we desire that 1) the robust latent label recovery algorithm r only requires signals on one task as input, but 2) we may use other signals to learn the latent label recovery algorithm and use this fixed function to recover the latent variable of each tasks.

In order to achieve these, we need to decompose the agents into three groups G_0, G_1, and G_2. We also decompose the tasks into two blocks: T_L and T_R. We use T_L to estimate the parameters (w, Γ), and recover the latent labels for tasks in T_R. Thus, agents' reports $\hat{X} \in \mathcal{X}^{n\times m}$ are decomposed into six blocks. In Figure 2, we use an example to show how to use this decomposition in our mechanism.

7 CONCLUSION AND FUTURE WORK

We provide a framework (RMIF) for the design of informed truthful mechanisms that uses robust learning algorithms to thwart adversarial attacks. This can be used to understand which properties of

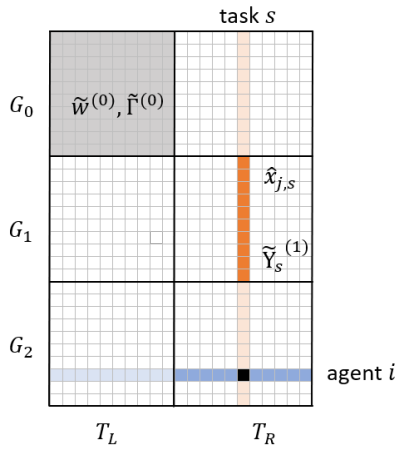


Figure 2: Suppose agent i is in G_2 and we want to recover a latent label of task $s \in T_R$ to pay agent i: The mechanism consists of three stages: 1) We use reports in *gray area* (tasks in T_L from agents in G_0) to obtain estimate $(w^{(0)}, \Gamma^{(0)})$. 2) To infer the latent label $\tilde{Y}_s^{(1)}$ for task $s \in T_R$, we apply maximum likelihood estimator on reports in *dark orange area* (group G_1), with estimated parameters $(w^{(0)}, \Gamma^{(0)})$ from stage 1. 3) Finally, we pay $i \in G_2$ the empirical estimate of the termwise mutual information between her reports in *dark blue area* $(\{\hat{X}_{i,s} : s \in T_R\})$ and the estimated latent labels $\{\tilde{Y}_s^{(1)} : s \in T_R\}$.

ALGORITHM 1: Algorithm for moments M_2

Input: An $\tilde{\alpha}$-corrupted reports $\hat{X} \in \Omega^{n_L \times m_L}$
Let $k_2 = m_L / \binom{n_L}{2}$ and $B = \{\boldsymbol{\iota} = (i, j) : i < j \in [n_L]\}$. ;
// Generate $\binom{n_L}{2}$ batches B and each batch has k_2 samples
for $\boldsymbol{\iota} \in B$ **do**
  Set $\mathbf{X}_{\boldsymbol{\iota}} = \emptyset$ and $l_{\boldsymbol{\iota}} = |\mathbf{X}_{\boldsymbol{\iota}}|$.
for $s \in [m_L]$ **do**
  while *pick* $\boldsymbol{\iota} = (i, j) \in B$ *randomly* **do**
    if $l_{\boldsymbol{\iota}} < k_2$ **then**
      $l_{\boldsymbol{\iota}} = l_{\boldsymbol{\iota}} + 1$;
      $\mathbf{X}_{(i,j), l_{\boldsymbol{\iota}}} = (\hat{X}_{i,s}, \hat{X}_{j,s})$;
      break;
Run $\mathcal{L}_{\text{batch}}$ on $\{\mathbf{X}_{\boldsymbol{\iota}} : \boldsymbol{\iota} \in B\}$ and output a distribution $\tilde{M}_2$ on Ω^2;

robust learning algorithms are useful for information elicitation from rowdy crowds. In particular, under two commonly used settings, we provide three mechanisms based on our framework to show that both robust recovery of joint distributions and robust recovery of latent variables lead to asymptotically informed truthful mechanisms for rowdy crowds.

The current paper focuses on handling the setting with strategic and adversarial agents and the truthful guarantee is the informed truthfulness. In future work, we believe it is possible to achieve even stronger truthfulness guarantees by considering the existence of honest agents or mechanisms in Schoenebeck and Yu [23]. Moreover, in this paper, our truthfulness guarantee only considers payments to strategic agents. An interesting future direction is to study how to detect and punish adversarial agents.

REFERENCES

[1] Arpit Agarwal, Debmalya Mandal, David C Parkes, and Nisarg Shah. 2017. Peer Prediction with Heterogeneous Users. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 81–98.

[2] Shannon Bond. 2020. A Must For Millions, Zoom Has A Dark Side — And An FBI Warning. *NPR* (Apr 2020). https://www.npr.org/2020/04/03/826129520/a-must-for-millions-zoom-has-a-dark-side-and-an-fbi-warning

[3] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Narayanan, Joshua A Kroll, and Edward W Felten. 2015. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In *2015 IEEE Symposium on Security and Privacy*. IEEE, 104–121.

[4] Arun Tejasvi Chaganty and Percy Liang. 2013. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*. 1040–1048.

[5] Sitan Chen, Jerry Li, and Ankur Moitra. 2019. Efficiently Learning Structured Distributions from Untrusted Batches. *arXiv preprint arXiv:1911.02035* (2019).

[6] Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 319–330.

[7] A. Philip Dawid and Allan Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm.

[8] Luc Devroye and Gábor Lugosi. 2012. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media.

[9] Gabor Lugosi Devroye Luc. 2001. *Combinatorial Methods in Density Estimation*. Springer New York.

[10] Carsten Eickhoff and Arjen de Vries. 2011. How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*. 11–14.

[11] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.

[12] Lee Garber. 2000. Denial-of-service attacks rip the Internet. *Computer* 4 (2000), 12–17.

[13] Rosario Gennaro, Craig Gentry, and Bryan Parno. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Annual Cryptology Conference*. Springer, 465–482.

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML]

[15] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 64–67.

[16] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. Detroit, Michigan, USA, 1–11.

[17] J. O. Kephart and S. R. White. 1993. Measuring and modeling computer virus prevalence. In *Proceedings 1993 IEEE Computer Society Symposium on Research in Security and Privacy*. 2–15.

[18] Yuqing Kong. 2020. Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, Shuchi Chawla (Ed.). SIAM, 2398–2411. https://doi.org/10.1137/1.9781611975994.147

[19] Yuqing Kong and Grant Schoenebeck. 2019. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)* 7, 1 (2019), 2.

[20] Joshua A Kroll, Ian C Davey, and Edward W Felten. 2013. The economics of Bitcoin mining, or Bitcoin in the presence of adversaries. In *Proceedings of WEIS*, Vol. 2013. 11.

[21] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2015. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. arXiv:1511.04508 [cs.CR]

[22] Mingda Qiao and Gregory Valiant. 2017. Learning discrete distributions from untrusted batches. *arXiv preprint arXiv:1711.08113* (2017).

[23] Grant Schoenebeck and Fang-Yi Yu. 2021. Learning and Strongly Truthful Multi-Task Peer Prediction: A Variational Approach. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 185)*. 78:1–78:20.

[24] Victor Shnayder, 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* (Maastricht, The Netherlands) *(EC '16)*. ACM, New York, NY, USA, 179–196.

[25] Tom Zeller Jr. 2006. A New Campaign Tactic: Manipulating Google Data. *The New York Times* (Oct 2006). https://www.nytimes.com/2006/10/26/us/politics/26googlebomb.html

[26] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. 2014. Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing. (June

2014). arXiv:1406.3824 [stat.ML]

A PROOF OF ROBUST MUTUAL INFORMATION FRAMEWORK

By Definition 2.2, there are three steps to prove ϵ-informed truthfulness. We first show that truth-telling is an approximate BNE, where any agent who deviates from truth-telling cannot achieve an extra payment larger than ϵ. Suppose all other rational agents report truthfully. Agent i's expected payment under strategy σ_i' is

$$\begin{aligned} u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}\setminus\{i\}}, \sigma_i', \boldsymbol{\sigma}_{\mathcal{A}}) \leq & \mathbb{E}_{\boldsymbol{f}\sim F}[\text{twMI}(X_i; \boldsymbol{f}(X_{-i}))] + \epsilon_2 \quad \text{(Eq. (3))} \\ \leq & u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) + \epsilon_1 + \epsilon_2 \quad \text{(Eq. (2))} \end{aligned}$$

Therefore, the truth-telling strategy profile is an $(\epsilon_1 + \epsilon_2)$-BNE.

Next, we show that the truth-telling strategy profile is paid approximately the highest by similar derivations.

$$\begin{aligned} u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \leq & \mathbb{E}_{\boldsymbol{f}\sim F}[\text{twMI}(X_i; \boldsymbol{f}(X_{-i}))] + \epsilon_2 \quad \text{(Eq. (3))} \\ \leq & u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) + \epsilon_1 + \epsilon_2 \quad \text{(Eq. (2))} \end{aligned}$$

Finally, we show that the truthful equilibrium is paid strictly higher than any uninformed strategy profile. Note that any uninformed strategy profile is equivalent to the situation that agents receive uninformative signals but report truthfully. Thus, for an arbitrary uninformed strategy profile $\boldsymbol{\theta}_{\mathcal{R}}$ (for rational agents), it's equivalent to say the agents receive the signals that are all zeros while reporting truthfully, i.e. $\theta_i(X_i) = \tau_i(\mathbf{0})$. Thus,

$$\begin{aligned} u_i^{P,\mathcal{L}}(\boldsymbol{\theta}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \leq & \mathbb{E}_{\boldsymbol{f}\sim F}[\text{twMI}(\mathbf{0}; \boldsymbol{f}(X_{-i}))] + \epsilon_2 = \epsilon_2 \\ & \text{(by Eq. (2) and uninformed strategy)} \end{aligned}$$

Moreover, we require the prior is permissible (Eq. (4)), so that $\epsilon_1 + \epsilon_2 < \mathbb{E}_{\boldsymbol{f}\sim F}[\text{twMI}(X_i; \boldsymbol{f}(X_{-i}))]$, and thus,

$$\epsilon_2 < \mathbb{E}_{\boldsymbol{f}\sim F}[\text{twMI}(X_i; \boldsymbol{f}(X_{-i}))] - \epsilon_1 \leq u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}})$$

where the second inequality is by Eq. (2). Putting this together with Eq. (4), we have that $u_i^{P,\mathcal{L}}(\boldsymbol{\theta}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \leq \epsilon_2 < u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}})$.

B PROOF OF THEOREMS IN GENERAL SETTING

B.1 Proof of Theorem 5.2

By Eq. (5), we aim to upper bound the expected difference between the estimated termwise mutual information $\widetilde{\text{twMI}}$ and the ground truth twMI. First, we break this difference into three terms. Then, we derive the upper bound of the three terms separately.

To simplify notation, we define $\widetilde{\text{twMI}}_{ij} := \widetilde{\text{twMI}}(\boldsymbol{\sigma}_i(X_i); \boldsymbol{\sigma}_j(X_j))$, $\mathbb{E}\left[\widetilde{\text{twMI}}_{ij}\right] = \mathbb{E}_{X, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j}\left[\widetilde{\text{twMI}}(\boldsymbol{\sigma}_i(X_i), \boldsymbol{\sigma}_j(X_j))\right]$, and $\mathbb{E}\left[\text{twMI}_{ij}\right] = \mathbb{E}_{X, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j}\left[\text{twMI}(\boldsymbol{\sigma}_i(X_i), \boldsymbol{\sigma}_j(X_j))\right]$.

Suppose i is rational, we can bound the left hand of Eq. (5) as

$$\begin{aligned} & \left|\mathbb{E}_{j\in[n]\setminus\{i\}}\left[\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right]\right] - \mathbb{E}_{j\in\mathcal{R}\setminus\{i\}}\left[\mathbb{E}\left[\text{twMI}_{i,j}\right]\right]\right| \\ \leq & \frac{1}{n-1}\left|\sum_{j\in\mathcal{A}}\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right]\right| + \frac{1}{n-\alpha n-1}\left|\sum_{j\in\mathcal{R}\setminus\{i\}}\left(\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right] - \mathbb{E}\left[\text{twMI}_{i,j}\right]\right)\right| \\ & + \frac{\alpha n}{(n-1)(n-\alpha n-1)}\left|\sum_{j\in\mathcal{R}\setminus\{i\}}\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right]\right|. \end{aligned} \tag{12}$$

This inequality follows directly from separating the summations and use triangle inequality.

We derive the upper bound of the above three terms separately. As for the first, though we have no knowledge of what the adversarial agents can do, we know that the total variation distance of two distributions is always between 0 and 1. Therefore, we know that the first term in Ineq. 12 is upper bounded by $\frac{n}{n-1}\alpha$. Similarly, because there are $n-\alpha n-1$ rational agents, we can derive an upper bound for the third term which is also $\frac{n}{n-1}\alpha$.

To bound the second term, we have to derive the upper bound of $\left|\widetilde{\text{twMI}}_{i,j} - \text{twMI}_{i,j}\right|$ for rational agents.

We first note that the difference $\left|\widetilde{\text{twMI}}_{i,j} - \text{twMI}_{i,j}\right|$ is less than

$$\frac{1}{2}\sum_{(x,y)\in\mathcal{X}^2}\left|\tilde{P}_{ij}(x,y) - P_{ij}(x,y)\right| + \frac{1}{2}\sum_{(x,y)\in\mathcal{X}^2}\left|\tilde{P}_i(x)\tilde{P}_j(y) - P_i(x)P_j(y)\right| \tag{13}$$

Also, notice that the second term in Eq. (13) can be bounded as

$$\begin{aligned} & \sum_{(x,y)\in\mathcal{X}^2}\left|\tilde{P}_i(x)\tilde{P}_j(y) - P_i(x)P_j(y)\right| \\ = & \sum_{(x,y)\in\mathcal{X}^2}\left|\left(\tilde{P}_i(x) - P_i(x)\right)\tilde{P}_j(y) + \left(\tilde{P}_j(y) - P_j(y)\right)P_i(x)\right| \\ \leq & \sum_{x\in\mathcal{X}}\left|\tilde{P}_i(x) - P_i(x)\right| + \sum_{y\in\mathcal{X}}\left|\tilde{P}_j(y) - P_j(y)\right| \end{aligned} \tag{14}$$

Therefore, in order to bound $\left|\widetilde{\text{twMI}}_{i,j} - \text{twMI}_{i,j}\right|$, we only have to bound the error of the estimated marginal distribution and the estimated joint distribution. Here, we use a standard result that any distribution with finite domain Ω is learnable within total variation distance d and with $1-\delta$ probability in $O\left(\frac{|\Omega|+\log(1/\delta)}{d^2}\right)$ samples. [9]. Therefore, we can learn the joint distribution with an error bounded by ϵ and probability $1-\delta$ with $O\left(\frac{|\mathcal{X}|^2+\log(1/\delta)}{\epsilon^2}\right)$ samples. Similarly, we can learn the marginal distribution with an error bounded by $\epsilon/2$ and probability $1-\delta$ with $O\left(\frac{4|\mathcal{X}|+4\log(1/\delta)}{\epsilon^2}\right)$ samples. Since we consider $|\mathcal{X}|$ as a constant, when $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ the following two upper bounds hold with probability $1-\delta$.

$$\sum_{(x,y)\in\mathcal{X}^2}\left|\tilde{P}_{ij}(x,y) - P_{ij}(x,y)\right| \leq \epsilon, \text{ and } \sum_{x\in\mathcal{X}}\left|\tilde{P}_i(x) - P_i(x)\right| \leq \frac{\epsilon}{2}.$$

Combining these two bounds with Eq. (13) and Eq. (14), we know that $\left|\widetilde{\text{twMI}}_{i,j} - \text{twMI}_{i,j}\right| \leq \epsilon$ (with probability $1-\delta$). Furthermore, we know that $\left|\widetilde{\text{twMI}}_{i,j} - \text{twMI}_{i,j}\right|$ is upper bounded by 1 always holds since both $\widetilde{\text{twMI}}_{i,j}$ and $\text{twMI}_{i,j}$ belong to $[0,1]$. Therefore,

Ineq 12 can be further written as

$$\left|\mathbb{E}_{j\in[n]\setminus\{i\}}\left[\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right]\right]-\mathbb{E}_{j\in\mathcal{R}\setminus\{i\}}\left[\mathbb{E}\left[\text{twMI}_{i,j}\right]\right]\right|$$
$$=\frac{2n}{n-1}\alpha(1-(1-\delta)\epsilon-\delta)+(1-\delta)\epsilon+\delta=2(1+\frac{1}{n-1})\alpha+\epsilon+\delta.$$

Then, what's left is to rewrite the error $\epsilon+\delta$ in terms of m. Furthermore, we want both ϵ and δ are asymptotically equal to zero as m is large. We know that $m=O(-\log\delta/\epsilon^2)$. An intuitive way is to set $\delta=O(1/m)$ and $\epsilon=O(\sqrt{\log m/m})$. Thus,

$$\left|\mathbb{E}_{j\in[n]\setminus\{i\}}\left[\mathbb{E}\left[\widetilde{\text{twMI}}_{i,j}\right]\right]-\mathbb{E}_{j\in\mathcal{R}\setminus\{i\}}\left[\mathbb{E}\left[\text{twMI}_{i,j}\right]\right]\right|$$
$$\leq 2(1+\frac{1}{n-1})\alpha+O\left(\sqrt{\frac{\log m}{m}}\right). \tag{15}$$

Therefore, combining Eq. (15) with Corollary 5.1, Mechanism 1 is $\left(4(1+\frac{1}{n-1})\alpha+O_n\left(\sqrt{\log m/m}\right)\right)$-informed truthful.

B.2 Proof of Theorem 5.3

To show that Mechanism 2 is asymptotically informed truthful, we will show the error of the inform truthfulness, denoted as ϵ, is asymptotically decreasing in n and m. With the batch learning algorithm introduced in Theorem 3.2, we will prove $\epsilon=O(\alpha/\sqrt{K})$, where $K=\lfloor m/(n-1)\rfloor$. To do so:

First, for a given rational agent i, the joint distributions between $\sigma_i(X_{i,s})$ and $\sigma_j(X_{j,s})$ are the same for all $j\in\mathcal{R}\setminus\{i\}$ and $s\in[m]$, and we denote this as simply $Q_{i,*}$. We use $\tilde{Q}_{i,*}$ to denote an estimation of $Q_{i,*}$ learned by $\mathcal{L}_{\text{batch}}$ with adversarial reports. Moreover, we use $R_{i,*}$ to denote the product of marginal distributions computed from $Q_{i,*}$, i.e. $R_{i,*}(x,y)=\sum_{z\in\mathcal{X}}Q_{i,*}(x,z)\cdot\sum_{w\in\mathcal{X}}Q_{i,*}(w,y)$. Similarly, $\tilde{R}_{i,*}$ is an estimation of $R_{i,*}$ computed from $\tilde{Q}_{i,*}$.

Following the idea of Corollary 5.1, we want to upper bound the difference between the expected payment (Eq. (8)) and the underlying twMI, i.e. the left hand of Eq. (5). This difference can be rewritten as follows.

$$\left|\mathbb{E}_{X,\boldsymbol{\sigma}_{\mathcal{R}},\boldsymbol{\sigma}_{\mathcal{A}}}\left[\mathcal{L}_i(\hat{X})\right]-\mathbb{E}_{j\in\mathcal{R}\setminus\{i\}}\left[\mathbb{E}_{X,\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_j}\left[\text{twMI}(\boldsymbol{\sigma}_i(X_i),\boldsymbol{\sigma}_j(X_j))\right]\right]\right|$$
$$=\frac{1}{2}\left|\mathbb{E}_{X,\boldsymbol{\sigma}_{\mathcal{R}},\boldsymbol{\sigma}_{\mathcal{A}}}\left[\left]|\tilde{Q}_{i,*}(x,y)-\tilde{R}_{i,*}(x,y)|-\sum_{x,y\in\mathcal{X}}|Q_{i,*}(x,y)-R_{i,*}(x,y)|\right]\right|\right.$$

Thus, in order to prove the theorem, it is sufficient to prove the following equation.

$$\left|\sum_{x,y\in\mathcal{X}}|\tilde{Q}_{i,*}(x,y)-\tilde{R}_{i,*}(x,y)|-\sum_{x,y\in\mathcal{X}}|Q_{i,*}(x,y)-R_{i,*}(x,y)|\right|=O\left(\frac{\alpha}{\sqrt{K}}\right).$$

The upper bound of the estimation error of $\tilde{Q}_{i,*}(x,y)$ can be obtained directly from the results of $\mathcal{L}_{\text{batch}}$ [22]. Let $\epsilon=\frac{\alpha}{\sqrt{K}}$. From the results of $\mathcal{L}_{\text{batch}}$ (Theorem 3.2), we know that if every agent is symmetric and $n=O((|\mathcal{X}|+K+\log(1/\delta))/\alpha^2)$, where $\mathcal{X}$ is the signal space, K is the size of each batch (in our case $K=\lfloor m/(n-1)\rfloor$), and $\delta\in(0,1)$, then with probability at least $1-\delta$ the error of learning the joint distribution is $O(\epsilon)$, i.e. $\left|Q_{i,*}(x,y)-\tilde{Q}_{i,*}(x,y)\right|=O(\epsilon)$ for $\forall x,y\in\Omega$. Next, we want to derive the upper bound of the error of $\tilde{R}_{i,*}(x,y)$.

Claim B.1. $|\tilde{R}_{i,*}(x,y)-R_{i,*}(x,y)|=O(\epsilon)$.

Since $\tilde{R}_{i,*}(x,y)=\sum_{z\in\mathcal{X}}\tilde{Q}_{i,*}(x,z)\cdot\sum_{w\in\mathcal{X}}\tilde{Q}_{i,*}(w,y)$. The claim is true because every $\tilde{Q}_{i,*}$ is at most $O(\epsilon)$ away from $Q_{i,*}$, and $|\mathcal{X}|$ is considered as constant.

Thus, we have a upper bound of the error of the product of marginal distributions which is also $O(\epsilon)$. Now, we can write the average $\tilde{\text{MI}}$ as

$$\sum_{(x,y)\in\mathcal{X}^2}|\tilde{Q}_{i,*}(x,y)-\tilde{R}_{i,*}(x,y)|=\sum_{(x,y)\in\mathcal{X}^2}\left|Q_{i,*}(x,y)-R_{i,*}(x,y)\right|\pm O(\epsilon).$$

Note that the above derivation holds with probability $1-\delta$. However, by the same argument in section B.1, we know that with (the other) probability of δ, the payment is bounded at 1. Therefore, by increasing n, we can make δ arbitrarily small and the difference between the expected payment and the expected underlying twMI is bounded by $O(\frac{\alpha}{\sqrt{K}})$. Then combining this with Corollary 5.1, we complete the proof.

B.3 Optimizing Parameters for Symmetric Priors

We focus now on minimizing the parameters n and m. First, we can randomly select a small number of agents, pay those selected agents according to our mechanism, and pay the rest of agent zero. Second, we use the batch learning algorithm as a black box for the design of our mechanism, such that the truthfulness guarantee of our mechanism can be improved with any improvement in the batch learning algorithm. Finally, we integrate these two parts and rewrite Theorem 5.3 with a proof.

For the first part, note that the only requirement on the size of selected agents is the fraction of adversary in the group is approximately equal to the original fraction α with high probability. Therefore, the size of the selected group is depending on the error and independent of the total number of agents n. Consequently, the number of tasks m required is only depending on the error and independent of n.

Now, before we rewrite a new version of Theorem 5.3, we define the black-box version of the batch learning algorithm.

Definition B.2. Black-box Batch learning algorithm (α_b,h_b,ψ_b): If the fraction of adversarial agents is upper bounded, i.e. $\alpha<\alpha_b$, $\delta_b\in(0,1)$, $\epsilon_b>0$ and the size of each batch $k_b=\Omega(h_b(\alpha,\epsilon_b))$. There exists an $n_b=O(\psi_b(\alpha,\delta,k_b))$ where n_b batches of samples of which a $(1-\alpha)$ fraction of batches consists of k_b i.i.d. draws from a distribution P with support Ω ($|\Omega|$ is considered to be a constant), there is an algorithm $\mathcal{L}_{\text{b}}$ that returns a distribution $\tilde{P}$ such that

$$\|P-\tilde{P}\|_{\text{TV}}=O(\epsilon_b)$$

with probability at least $1-\delta_b$.

Any batch learning algorithm can be written in this form with three parameters, i.e. α_b, the upper bound of the adversarial fraction, h_b, a function of α and ϵ_b that determines the requirement of the size of batch to achieve error ϵ, and ψ_b, a function that determines the lower bound of the number of batches. Now, we are ready to write Theorem 5.3 in a black-box form.

Theorem B.3. *Given a batch learning algorithm $\mathcal{L}_b$ with parameters (α_b,h_b,ψ_b), in general RowdyCrowds$(\mathcal{P},\alpha)$ setting with*

symmetric prior and strategies, $\alpha < \alpha_b$, then, there exists an $n^* = O\left(\max\left\{\frac{\log 1/\epsilon^*}{(\alpha_b-\alpha)^2}, \psi_b(\alpha,\epsilon^*, h_b(\alpha,\epsilon^*))\right\}\right)$, such that if $n > n^*$ and $m > h_b(\alpha,\epsilon^*)n^*$, an $O(\epsilon^*)$-informed truthful mechanism exists.

The proof is mostly identical to Theorem 5.3. We omit the proof due to space constrain.

Theorem 5.3 is a special case of Theorem B.3 using $\mathcal{L}_{\text{batch}}$ (Theorem 3.2) as $\mathcal{L}_b$. We can recover Theorem 5.3 by setting $\alpha_b = \alpha_{\text{batch}} = 1/900$, $h_b(\alpha,\epsilon_b) = \alpha^2/\epsilon_b^2$, $\psi_b(\alpha,\delta,k_b) = (k_b + \log 1/\delta)/\alpha^2$. This leads to an $n^* = O\left(\max\left\{\frac{\log 1/\epsilon^*}{(\alpha^*-\alpha)^2}, \left(\left(\frac{1}{\epsilon^*}\right)^2 + \frac{\log 1/\epsilon^*}{\alpha^2}\right)\right\}\right)$, and the requirement on m becomes $\left(\frac{\alpha}{\epsilon^*}\right)^2 n^*$. This gives us the following corollary, which is another version of Theorem 5.3.

Corollary B.4. *In the general RowdyCrowds($\mathcal{P},\alpha$) setting,with symmetric prior and strategies, $\alpha < \alpha_{\text{batch}}$, then, there exists an $n^* = O\left(\max\left\{\frac{\log 1/\epsilon^*}{(\alpha_{\text{batch}}-\alpha)^2}, \left(\left(\frac{1}{\epsilon^*}\right)^2 + \frac{\log 1/\epsilon^*}{\alpha^2}\right)\right\}\right)$, such that if $n > n^*$ and $m > \left(\frac{\alpha}{\epsilon^*}\right)^2 n^*$, a $O(\epsilon^*)$-informed truthful mechanism exists.*

C PROOF OF THEOREM 6.2

With basic understanding of those three stages in Mechanism 3 in Sect. 6, the proof idea of Theorem 6.2 is straightforward. If we can estimate the parameters (w,Γ) accurately (Lemma C.2), and recover all latent labels $\{Y_s : s \in T_R\}$ correctly (Lemma C.6), we can pay an agent by the termwise mutual information between her reports and estimated latent labels, and asymptotically informed-truthfulness can be derived from Theorem 4.2.

Before we dive into the proof, we first need to control the fraction of adversary agents in all groups G_0, G_1 and G_2. Let $\alpha^{(g)}$ be the fraction of adversary in group G_g where $g = 0,1,2$. Since we partition the agents after they reports, in RowdyCrowds($\mathcal{P},\alpha$) we can use a Chernoff bound to show $\alpha^{(g)}$ are close to α for all g.

Lemma C.1. *Given RowdyCrowds($\mathcal{P},\alpha$) with n agents, for all $\epsilon_\alpha > 0$ and $\delta_\alpha > 0$, there exists $n_\alpha = \Omega\left(\frac{\log 1/\delta_\alpha}{\epsilon_\alpha^2}\right)$ such that if $n \geq n_\alpha$,*

$$\Pr[\forall g = 0,1,2, |\alpha^{(g)} - \alpha| > \epsilon_\alpha] \leq \delta_\alpha.$$

Now we prove Lemma C.2 and C.6. Due to the symmetry of Mechanism 3, without loss of generality, we only need to prove Theorem 6.2 in the perspective of agent i who is in G_2.

C.1 Estimate the parameters

In the perspective of agent i in G_2, for the first stage, she only cares about the value of $\tilde{\Gamma}^{(0)}, \tilde{w}^{(0)}$ in G_0 illustrated in Figure 2.

Lemma C.2. *For all positive ϵ_{para} and δ_{para}, let $n_{\text{para}} = poly(\log 1/\delta_{\text{para}}, 1/\alpha_{\text{batch}})$ where constant α_{batch} is defined in Theorem 3.2. In group G_0, if the fraction of adversary $\alpha^{(0)} \leq \alpha_{\text{batch}}/3$, the number of agents in group 0, $\lfloor n/3 \rfloor \geq n_{\text{para}}$ and the number of tasks $|T_L| = \Omega(\epsilon_{\text{para}}^{-2} n^3)$, there exists a permutation $\pi^{(0)}$ on $\mathcal{Y}$ such that with probability at least $1 - \delta_{\text{para}}$*

$$\max_y |\tilde{w}_y^{(0)} - w_{\pi^{(0)}(y)}| \leq \epsilon_{\text{para}} \text{ and } \max_y \|\tilde{\Gamma}_y^{(0)} - \Gamma_{\pi^{(0)}(y)}\|_2 \leq \epsilon_{\text{para}},$$

where $\tilde{w}_y^{(0)}$ and $\tilde{\Gamma}_y^{(0)}$ are the estimated parameters learned from G_0.

The above lemma shows if the number of tasks in T_L is large enough, we can have an accurate estimation of w and Γ. The proof of Lemma C.2 consists of two parts: First, Lemma C.3 and C.4 ensure the estimated moments $\tilde{M}_2^{(0)}$ and $\tilde{M}_3^{(0)}$ are accurate. Then, in Lemma C.5, we show the spectral method can approximate the parameter (w,Γ) when the estimated moments are accurate. We omit the proof due to the space constrain.

Lemma C.3 (Estimate M_2). *Let the fraction of adversarial in group G_0 to be $\alpha^{(0)} < \alpha_{\text{batch}}/2$. For all $\delta > 0$, there exist $n_L \geq \lfloor n/3 \rfloor$ and $m_L \geq \lfloor m/2 \rfloor$ depending on $\log 1/\delta$ and $\alpha_{\text{batch}}/2$ such that the output $\tilde{M}_2^{(0)}$ is close to M_2. Formally,*

$$\sum_{x_1,x_2 \in \mathcal{X}} |\tilde{M}_2^{(0)}(x_1,x_2) - M_2(x_1,x_2)| = O\left(\alpha^{(0)}\sqrt{n_L^2/m_L}\right)$$

with probability at least $1-\delta$.

The above lemma follows directly from Theorem 3.2. A major difference is the upper bound of the fraction of adversary is changing from α_{batch} to $\alpha_{\text{batch}}/2$. This is because if the fraction of adversarial agents in group G_0 is at most $\alpha^{(0)}$, the fraction of the corrupted batches in that group is at most $2\alpha^{(0)}$.

By similar an augment we have the following lemma.

Lemma C.4 (Estimate M_3). *Let the fraction of adversarial in group G_0 $\alpha^{(0)} < \alpha_{\text{batch}}/3$ be a constant. For $\forall\delta > 0$, there exist $n_L \geq \lfloor n/3 \rfloor$ and $m_L \geq \lfloor m/2 \rfloor$ depending on $|\mathcal{X}|$, $\log 1/\delta$ and $\alpha_{\text{batch}}/3$. There is an algorithm that outputs $\tilde{M}_3$ close to M_3. Formally,*

$$\sum_{x_1,x_2,x_3 \in \mathcal{X}} |\tilde{M}_3^{(0)}(x_1,x_2,x_3) - M_3(x_1,x_2,x_3)| = O\left(\alpha^{(0)} \cdot \sqrt{n_L^3/m_L}\right)$$

with probability at least $1-\delta$.

The following lemma shows if the estimated M_2 and M_3 are accurate, we can recover (w,Γ) accurately.

Lemma C.5 (Lemma 4 [4]). *There exists a constant K depending on Γ and w, such that for all $\epsilon_1 \leq 1/2$, if $\|\tilde{M}_2^{(0)} - M_2\|_{op}$ and $\|\tilde{M}_3^{(0)} - M_3\|_{op}$[4] are both less than $K\epsilon_1$, there exists a permutation π on latent labels $\mathcal{Y}$ such that*

$$\max_y |\tilde{w}_y^{(0)} - w_{\pi(y)}| \leq \epsilon_1 \text{ and } \max_y \|\tilde{\Gamma}_y^{(0)} - \Gamma_{\pi(y)}\|_2 \leq \epsilon_1.$$

C.2 Infer the latent labels

For the second stage, we simply compute the most likely label given the parameters and the reports (of the appropriate group of agents).

For lemmas below, we consider agent i in group G_2 and a tasks j in T_R. The proof is by a union bound and Chernoff bound argument. We omit the proof here due to space constrain.

Lemma C.6. *Given a symmetric DS model with parameter (w,Γ), recall that $\gamma_{DS} = \frac{\min_{y\neq y'} D_{KL}(\Gamma_y;\Gamma_{y'})}{4\max_{y,x}|\log \Gamma_{y,x}|}$, and $\epsilon_{kl} = \frac{1}{8}\min_{x,y}\Gamma_{y,x}$ $\min\left\{4, \min_{y\neq y'} D_{\text{KL}}(\Gamma_y;\Gamma_{y'})\right\}$. If $\tilde{\Gamma}^{(0)}$ is accurate where $\max_y \|\tilde{\Gamma}_y^{(0)} -$*

[4]Here if M is a matrix, $\|M\|_{op}$ is the operator norm (largest singular value). When M is in $\mathbb{R}^{C\times C\times C}$, $\|M\|_{op} = \frac{1}{C}\sum_{i\in C}\|M_{i,\cdot,\cdot}\|_{op}$ which is the average operator norm over all C unfoldings.

$\Gamma_y\|_2 \le \epsilon_{kl}$, and the fraction of adversaries in G_1, $\alpha^{(1)}$, is less than γ_{DS}, then for all $s \in T_R$

$$\Pr[\tilde{Y}_s^{(1)} \neq Y_s] \le \exp(-\Theta(n_L)),$$

thus, the latent labels for all the tasks are correct with probability greater than $1 - m_L \exp(-\Theta(n_L))$.

In Lemma C.5, we can only show the estimation of parameter is accurate *up to some permutation*. Therefore, we can only recover the latent variable Y_s accurately up to some permutation. However, this is sufficient to our mechanism, because the mutual information is invariant under permutations.

C.3 Proof of Theorem 6.2

To prove the theorem, we need to show our payment in Mechanism 3 with the latent label recovery mapping f from reports $\mathcal{X}^{n'} \to \mathcal{Y}$ is a *robust mutual information estimation pair* for any $\epsilon_1^* > 0$ and $\epsilon_2^* > 0$ when the number of tasks m and the number of agents n are large enough. Formally, given a sample $X \in \mathcal{X}^{n'}$ with the latent label $Y \in \mathcal{Y}$ from $(\mathcal{X}, \mathcal{Y}, w, \Gamma)$, the latent label recovery mapping f maps X to Y.[5]

First we show Eq. (2): for any adversary strategy $\boldsymbol{\sigma}_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$, the expected payment of agent $i \in \mathcal{R}$ satisfies:

$$u_i^{P,\mathcal{L}}(\boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) = \mathop{\mathbb{E}}_{X, \boldsymbol{\tau}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}} \left[\mathcal{L}_i(\hat{X})\right] \ge \mathbb{E}_{f\sim F}[\mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i}))] - \epsilon_1^*.$$

Let $\mathcal{E}$ be the event that the estimation $\tilde{Y}_s$ is equal to the latent label Y_s for all $s \in T_R$ up to a permutation π on $\mathcal{Y}$,

$$\mathcal{E} := \{\exists \pi, \forall s \in T_R, \tilde{Y}_s^{(1)} = \pi(Y_s)\}.$$

When $\mathcal{E}$ happens, the latent labels are all correctly recovered up to some permutation π. Mechanism 3 approximately pays agent i with the termwise mutual information between her reports and latent labels, $\mathrm{twMI}_{\mathrm{TV}}(X_i; Y) - \epsilon_1^*/2$ when m is large enough because the termwise mutual information is invariant under permutation. On the other hand, if the event $\mathcal{E}$ fails, agent i loses at most 1, because the termwise mutual information is bounded by 1.

Therefore, to complete the proof of Eq. (2), we only need to prove

Claim C.7.

$$\Pr[\mathcal{E}] \ge 1 - \epsilon_1^*/2 \tag{16}$$

We can show this by taking n and m large enough and using union bound on Lemma C.1, C.2 and C.6. Due to space limit, we omit the details here.

Now we show Eq. (3) is satisfied: for all $\boldsymbol{\sigma}_{\mathcal{R}}$ and i

$$u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) = \mathop{\mathbb{E}}_{X, \boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}} \left[\mathcal{L}_i(\hat{X})\right] \le \mathbb{E}_{f\sim F}[\mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i}))] + \epsilon_2^*.$$

Recall that agent i is in group G_2. Regardless of the value of $\tilde{\Gamma}$, Eq. (11), is a function of reports in group G_1. Therefore, by data processing inequality, fix an arbitrary $\tilde{\Gamma}$ the mutual information between $\tilde{Y}^{(1)}$ in Eq. (11) and i's report is smaller than the mutual information between the reports in group G_1 and i's reports. Let $\tilde{f}_{\tilde{\Gamma}}$ denote such mapping, and we have

$$\mathrm{twMI}(X_i; \tilde{f}_{\tilde{\Gamma}}(X_{G_1})) \le \mathrm{twMI}(X_i; X_{G_1}) \le \mathrm{twMI}(X_i; X_{-i}).$$

Moreover, because $\{X_{i,s}\}_i$ are mutually independent conditional on the latent variable Y_s for all s, by data processing inequality

$$\mathrm{twMI}(X_i; X_{-i}) \le \mathrm{twMI}(X_i; Y) \le \mathrm{twMI}(X_i; \boldsymbol{f}(X_{-i})) + \epsilon_2^*/2$$

where the last inequality comes from rerunning the above argument so that $\boldsymbol{f}(X_{-i}) = Y_i$ with probability $1 - \delta$ where $\delta < \epsilon_2^*/2$. Finally, by Lemma 3.1, we complete the proof by taking m large enough such that for any $\tilde{\Gamma}$

$$u_i^{P,\mathcal{L}}(\boldsymbol{\sigma}_{\mathcal{R}}, \boldsymbol{\sigma}_{\mathcal{A}}) \le \mathrm{twMI}(X_i; \tilde{f}_{\tilde{\Gamma}}(X_{G_1})) + \epsilon_2^*.$$

C.4 Parameter Optimization for Symmetric DS Model

Similar to Appendix B.3, here we have three steps to optimize the parameters. First, by Lemma C.1, n should be lower bounded such that we can randomly select a small number of agents so long as the fraction of adversarial agents in it is basically unchanged. Second, if n and m satisfy the requirements of the black-box batch learning algorithm (Definition B.2), it can guarantee the parameters of the symmetric DS model will be recovered with small error. Third, we write the latent label recovery algorithm in Appendix C.2 as a black box, which leads to an additional requirement on n and m. Integrating these three parts together, we rewrite Theorem 6.2 in terms of the black-box parameters and n and m can be optimized.

Definition C.8. Black-box DS model latent label recovery algorithm (α_r, d_r, ψ_r):
Given a symmetric DS model with parameter (w, Γ), let α_r and d_r be the parameters given by the algorithm, which could be functions of w, Γ. Suppose for any task, there are n_r samples of which a $(1-\alpha)$ fraction are i.i.d. draws from the distribution $w_y\Gamma_y$ given the latent label of that task is y, while the remaining α fraction are adversarialy controlled. If $\alpha < \alpha_r$, given an estimation of Γ, $\tilde{\Gamma}$, such that $\max_y \|\tilde{\Gamma}_y - \Gamma_y\|_2 \le d_r$, $n_r = \Omega(\psi_r(\epsilon_r))$ and $m_r = o(1/\epsilon_r)$ then there is a recovery algorithm $\mathcal{L}_{\mathrm{r}}$ s.t. the latent label of any task $s \in [m_r]$ can be recovered with probability at least $1 - O(\epsilon_r)$.

As an example, we use the maximum likelihood estimation in the previous proof which serves as a special case of this black box algorithm. Here, α_r is the upper bound of the fraction of adversaries required by the algorithm. The input of the latent label recovery algorithm is an estimation of the distribution Γ. The upper bound of the tolerance of the error of the input distribution is denoted as d_r. Finally, ψ_r determines the lower bound of the number of samples. Now, we write Theorem 6.2 in the black-box form.

Theorem C.9. *Given a batch learning algorithm $\mathcal{L}_{\mathrm{b}}$ with parameters (α_b, h_b, ψ_b), and a DS model latent label recovery algorithm $\mathcal{L}_{\mathrm{r}}$ with parameters (α_r, d_r, ψ_r), let $\alpha^* = \min\{\alpha_b/3, \alpha_r\}$. For RowdyCrowds$(\mathcal{P}_{DS}, \alpha)$ setting on the symmetric DS model, if $\alpha \le \alpha^*$, there exists an $n^* = O\left(\max\left\{\frac{\log 1/\epsilon^*}{(\alpha^*-\alpha)^2}, (\psi_b(\alpha, \epsilon^*, h_b(\alpha, d_r)))^{\frac{1}{2}}, \psi_r(\epsilon^*)\right\}\right)$, such that if $n > n^*$, $m > h_b(\alpha, d_r)(n^*)^3$ and $m = o(1/\epsilon^*)$, then we have a $O(\epsilon^*)$-informed truthful mechanism.*

[5] Such mapping is not well-defined since we cannot always recover the latent label from a finite number of signals. However, as we show in the results, it is sufficient to approximately recover the latent label which is possible as the number of signal is large enough.

The proof is basically identical to the proof of Theorem B.3, and we omit it for space constrain.

Using $\mathcal{L}_{\text{batch}}$ (Theorem 3.2) as $\mathcal{L}_b$ and the maximum likelihood estimator as $\mathcal{L}_r$, we can recover Theorem 6.2. Let $\alpha_b = \alpha_{\text{batch}} = 1/900$, $h_b(\alpha, \epsilon_b) = \alpha^2/\epsilon_b^2$, $\psi_b(\alpha, \delta, k_b) = (k_b + \log 1/\delta)/\alpha^2$, and $\alpha_r = \gamma_{DS}$, $d_r = \epsilon_{kl}$, $\psi_r(\epsilon_r) = \log 1/\epsilon_r$, where γ_{DS} and ϵ_{kl} are defined in Lemma C.6. Plugging these in Theorem C.9 we obtain the following corollary, which is another version of Theorem 6.2.

Corollary C.10. *Let $\alpha^* = \min\{\alpha_{\text{batch}}/3, \gamma_{DS}\}$. Then, for any RowdyCrowds($\mathcal{P}_{DS}, \alpha$) setting on the symmetric DS model, if $\alpha \le \alpha^*$, there exists an $n^* = O\left(\max\left\{\frac{\log 1/\epsilon^*}{(\alpha^*-\alpha)^2}, \left(\frac{1}{d_r^2} + \frac{\log 1/\epsilon^*}{\alpha^2}\right)^{\frac{1}{2}}\right\}\right)$, such that if $n > n^*$, $m > \left(\frac{\alpha}{d_r}\right)^2 (n^*)^3$ and $m = o(1/\epsilon^*)$, then we have a $O(\epsilon^*)$-informed truthful mechanism.*