# Simple Truth Serums for Massively Crowdsourced Evaluation Tasks

Vijay Kamble, Nihar Shah, David Marn, Abhay Parekh, Kannan Ramachandran, University of California, Berkeley

We consider the problem of eliciting truthful responses from agents in the absence of any known answers. This class of problems, pioneered by the peer-prediction method and the Bayesian truth serum, is now quite well studied in the literature. In this paper we propose new mechanisms that, unlike most works on this topic, require no extraneous elicitation from the agents, and furthermore allow the agents' beliefs to be (almost) arbitrary. Moreover, these mechanisms have the structure of output agreement mechanisms, which are simple, intuitive and have been quite popular in practice. These mechanisms operate under scenarios where the number of questions is large, and are suitable for most tasks in crowdsourcing and peer-grading.

## 1. INTRODUCTION

Systems that leverage the wisdom of the crowd are ubiquitous today. Recommendation systems such as Yelp and others, where people provide ratings and reviews for various entities, are used by millions of people across the globe [Luca 2011]. Commercial crowdsourcing platforms such as Amazon Mechanical Turk, where workers perform microtasks in exchange for payments over the Internet, are employed for a variety of purposes such as collecting labelled data to train machine learning algorithms [Raykar et al. 2010]. In massive open online courses (MOOCs), students' exams or homeworks are often evaluated by means of "peer-grading", where students grade each others' work [Piech et al. 2013].

A majority of applications in this domain involve a large number of evaluation tasks or questions that need to be answered, and every "agent" performs a subset of these tasks. For instance, a typical such tasks on a crowdsourcing platform such as Amazon Mechanical Turk comprise of labeling a large set of images for a for machine learning application. The problem of peer grading in massive open online courses (MOOCs) involves a similar procedure where a large number of answers provided by all the students have to be graded. We call these tasks "*Massively crowdsourced evaluation tasks*" or MCETs.

In this paper, we consider the strategic aspect of MCETs, with a goal of designing mechanisms that incentivize the agents to report their answers truthfully. The literature on this topic can largely be divided into two lines of research, depending on assumptions on the existence of so-called "*gold standard*" questions [Le et al. 2010; Chen et al. 2011]. These are a small subset of questions to which the principal either knows the answers apriori or can verify them accurately. Incentive design with such gold standard questions is easily facilitated by the use of *proper scoring rules* [Lambert and Shoham 2009; Shah et al. 2015].

The second line of research, that the present paper contributes to, makes no assumption about the existence of such gold standard questions and is more realistic in many applications of interest. The work on the peer-prediction method [Miller et al. 2005] pioneered mechanisms for incentivizing truthful reports in the absence of known answers under certain assumptions. Several works have subsequently built on these ideas to construct mechanisms with fewer assumptions, most notably the work on *Bayesian truth serum* [Prelec 2004; Witkowski and Parkes 2012b; Radanovic and Faltings 2013]. Most of these works rely on two critical drawbacks that seem to impede the widespread use of these mechanisms in practice.

First, along the lines of [Prelec 2004], these mechanisms elicit data that is extraneous to the question at hand. This extraneous information usually takes the form of predicting other agents' responses. From an agent's point of view, providing this additional information necessitates a two-fold or even higher increase in the effort. This increase may in turn also proportionally increase the principal's costs in crowdsourcing such tasks. Further, truthfulness holds under the assumption that these predictions are computed using Bayesian updates of subjective priors, which is arguably too much to ask from a typical agents, e.g., the workers on Mechanical Turk.

A second drawback is that most works assume "homogeneity" of the agents, which intuitively means that all agents are statistically similar in the way they answer any question, e.g. they do not have any relative biases or difference in abilities. To be more precise, consider a single question with options $o_1, \ldots, o_m$. Homogeneity assumes that conditional on any option $o_j$ being correct, any two agents have the same likelihood of giving different answers, and in particular of being correct or wrong. As we argue later, such an assumption is reasonable in the case of surveys, where an agent's answer to a question can be seen as an independent sample of the distribution of the answers in the population. But is certainly inappropriate in subjective evaluation tasks like rating movies or grading answers, in which systematic biases may exist because of differences in preferences, effort or abilities.

In this paper, we present incentive compatible mechanisms for MCETs involving both homogeneous and heterogeneous populations. In the homogeneous case, our mechanism elicits truthful answers from the agents without asking any additional questions. In the heterogeneous case, for the setting of binary-choice questions, our mechanism achieves this while making very mild assumptions on the structure of agents' beliefs. In order to achieve these objectives our mechanism leverages the typical characteristic of MCETs of the existence of a large number of similar evaluation tasks.

Our mechanisms share the structure of *output agreement mechanisms* (see [Von Ahn and Dabbish 2008] and [Von Ahn and Dabbish 2004]) that are simple, intuitive, and have indeed been quite popular in practice, but that suffer from a critical drawback of not being truthful. In an output agreement mechanism, two agents answer the same question, and they are both rewarded if their answers match. From the perspective of an agent, in the absence of any extraneous information, this almost incentivizes truthful reporting, since in most cases it is more likely that the other agent also has the same answer. But this is not the case when the agent believes that the answer he has is relatively unpopular and that a typical agent will have a different opinion. It is then tempting to report the answer that is more likely to be popular.

Our mechanism overcomes this drawback by giving proportionately higher rewards for answers that are relatively unlikely and lower rewards for answers that are more likely. These rewards are designed in such a way that before looking at the question, an agent is roughly indifferent between the different answers: a randomly chosen other agent is more likely to have the popular answer for the question but there is a low reward for matching, while he is less likely to have a relatively unpopular answer but there is a high reward for matching. But as soon as the agent sees the question and forms an answer, the conditional popularities from his perspective change in such a way that it becomes more profitable to report his opinion truthfully. In order to design such rewards, only prior statistics of the popularities are needed, which can be estimated from the answers to a large number of evaluation tasks. These are the essential key ideas that underly the design of our mechanisms.

The remainder of the paper is organized as follows. Section 2 presents a formal description of the model considered in the paper. Given the model, Section 3 puts our work in perspective of the existing literature. Section 4 and Section 5 contain the main results of the paper. Section 4 presents a mechanism to incentivize truthful reports without asking for additional information, assuming that the population is homogeneous. Section 5 then extends the results to a setting that does not make the homogeneity assumption. The paper concludes with a discussion in Section 6.

## 2. MODEL

Consider a population denoted by the set $\mathcal{M}$, with $M$ agents labelled $j = 1, \cdots, M$. Consider an *evaluation task* in which an agent $j$ in $\mathcal{M}$ interacts with an object and forms an evaluation taking values in a finite set $\mathcal{S} = (s_1, \cdots, s_K)$. Examples of object and evaluation pairs are: Movies/businesses → ratings (think of Yelp), images → labels (in crowdsourced labeling tasks), Answers → grades (in peer-grading). An agent's evaluation for an object is influenced by the unknown *attributes* of the object and the manner in which these attributes affect her evaluations, or in abstract terms, her *tastes* and preferences. Note that the attributes of an object capture everything about the object that could affect its evaluation and as such these attributes may or may not be observable/measurable. For example, in the case where a mathematical solution is being evaluated in a peer-grading platform, its attributes could be: elegance and conciseness: high, handwriting: poor, presentation: poor etc. Denote the hidden attributes of an object by the quantity $X$, which we will simply call the *type* of the object and assume that this type takes values in a finite universe $\mathcal{H} = \{h_1, \cdots, h_L\}$.

Denote agent $j$'s evaluation for the object by $Y_j \in \mathcal{S}$. The manner in which an object's attributes influence her evaluation is modeled by a conditional probability distribution over $Y_j$ given different values of $X$, i.e., $P(Y_j = s|X = h)$ for each $s \in \mathcal{S}$ and $h \in \mathcal{H}$. For notational convenience, we will denote this distribution by $p_j(s|h)$ and we will refer to it as the "filter" of person $j$.

Now consider a set-up where there are $N$ statistically similar objects, labeled $i = 1, \cdots, N$, that being evaluated. The type of object $i$ is denoted by $X^i$ and each $X^i$ is assumed to be drawn independently from a common probability distribution $P_X$ over $\mathcal{H}$. Let $\mathcal{M}^i \subseteq \mathcal{M}$ denote the set of persons that evaluate object $i$ and let $\mathcal{W}_j$ be the set of objects that a person $j$ evaluates. We assume that $|\mathcal{M}^i| \geq 2$ and $|\mathcal{W}_j| \leq Q$ for all $i$ and $j$, i.e., each object is evaluated by at least two people and each person evaluates a maximum of $Q$ objects. We assume that the sets $\mathcal{W}_j$ and $\mathcal{M}^i$ are fixed apriori. If an agent $j$ evaluates object $i$, let $Y_j^i$ denote her evaluation for that object. We assume that since the objects are similar, the filters of the agents are the same for evaluating the different objects, i.e., $P(Y_j^i = s|X^i = h) = P(Y_j^{i'} = s|X^{i'} = h) = p_j(s|h)$.

We also assume that the evaluations $Y_j^i$ by different $j$ are conditionally independent given $X^i$. Finally, we assume that the sets of random variables $\{X^i, \{Y_j^i : j \in \mathcal{M}^i\}\}$ for the different objects $i$ are mutually independent. In particular this implies that $Y_j^i$ and $Y_j^{i'}$ are independent for any person $j$ that has evaluated objects $i$ and $i'$. Note that the random variables $\{Y_j^i : j \in \mathcal{M}^i\}$ need not be independent unless conditioned on $X^i$.

Our goal is to design a payment mechanism that truthfully elicits evaluations from the population. With MCETs in mind, we are specifically interested in the case where $Q$ is small, $|\mathcal{M}^i|$'s are arbitrary, and $N$ is large. The mechanism designer is not assumed to have any knowledge of $P_X$ or the filters of the different people in the population. Further, we assume that every member of the population knows the structure of the underlying generating model, in particular the existence of a single $P_X$ that generates the type for each object, the conditional independence assumptions on the evaluations given the type for every object, and the independence of the evaluations across different objects. But the agents may not know, or may have different subjective beliefs about the values of $P_X$, the filters of others, and even their own filter. We present an example of this setting.

*Example* 2.1. **Peer-grading in MOOCs:** Peer-grading, where students evaluate their peers and these evaluations are processed to assign grades to every student, has been proposed as a scalable solution to the problem of grading in MOOCs. An important component of any such scheme is the design of incentives so that students are truthful when they grade others. For example, say that the answer of any student to a fixed question has some true grade $A$, $B$ or $C$, which can be taken to be the type of the answer. Suppose that apriori there is a distribution over the grade of any answer that is common to all answers (to a fixed question). Each answer is then graded by a few students (and in turn each student grades a few answers), who, depending on some given rubric and their abilities, form an opinion as to what grade should be assigned to the answer. Similarly there are thousands of such answers that are graded by other students. It is natural to assume that conditional on the true grade of an answer, the evaluations of different students who grade that answer are independent. Also it is natural to assume that the grades given by the students to different answers are independent. One then wants to design a mechanism that incentivizes the students to report their true opinions about the answers that they have graded.

Let $q_j^i$ denote the person $j$'s reported evaluation for object $i$. A payment (or scoring) mechanism is a set of functions $\{\tau_j : j \in \mathcal{M}\}$, one for each person in the population, that map the reports $\{q_j^i : i = 1, \cdots, N,\ j \in \mathcal{M}^i\}$ to real valued payments (or scores).

*Definition* 2.2. We say that a given payment mechanism $\{\tau_j : j \in \mathcal{M}\}$ is *detail-free Bayes-Nash incentive compatible* if for each $j \in \mathcal{M}$,

$$E[\tau_j(\{y_j^i : i \in \mathcal{W}_j\}, \{Y_{j'}^i : i = 1, \cdots, N,\ j' \in \mathcal{M}^i,\ j' \neq j\}) \mid Y_j^i = y_j^i,\ i \in \mathcal{W}_j,\ P_X,\ \{p_{j'}(s|h)\},\ j' \in \mathcal{M}]$$
$$\geq E[\tau_j(\{q_j^i : i \in \mathcal{W}_j\}, \{Y_{j'}^i : i = 1, \cdots, N,\ j' \in \mathcal{M}^i,\ j' \neq j\}) \mid Y_j^i = y_j^i,\ i \in \mathcal{W}_j,\ P_X,\ \{p_{j'}(s|h)\},\ j' \in \mathcal{M}],$$

for each $y_j^i \in \mathcal{S}$ for $i \in \mathcal{W}_j$, and for every specification of $P_X$ and $\{p_{j'}(s|h)\}$ for $j' \in \mathcal{M}$. We call it *strictly* detail-free Bayes-Nash incentive compatible if the above inequality is strict whenever $\{y_j^i : i \in \mathcal{W}_j\} \neq \{q_j^i : i \in \mathcal{W}_j\}$.

In other words, for our model, a mechanism is detail-free Bayes-Nash incentive compatible if truthful reporting is a Bayes-Nash equilibrium in the incomplete information game induced by the mechanism, for every possible specification of the underlying generating model for the evaluations that satisfies the independence assumptions discussed before: in particular, the specification of $P_X$ and the filters $p_j(s|h)$.

We will consider two types of generating models inspired by two types of applications encountered. This difference arises from the considerations for the variation in the tastes, preferences and biases of the population, resulting in differences in the manner in which they evaluate an object. We will in particular consider two cases:

— **Homogeneous population:** Consider a typical survey, e.g., suppose the government would like to find out what is the chance that a visit to the DMV office in a particular location faces a waiting time of more than 2 hours. This is a number $X$ that can be thought of as an attribute of the DMV and for simplicity, assume that it takes values in a finite set, say $[0, 0.1, 0.2, \cdots, 1]$. The evaluation of any agent $j$, $Y_j$ is just a value $\{0, 1\}$, with 1 denoting that she faced a wait time of greater than 2 hours. In this case it is natural to assume that $P(Y_j = 1|X = h) = h$, i.e., each person's evaluation is an independent sample of the hidden value $X$. This means that $p_j(s|h)$ does not depend on $j$, and is the same value $p(s|h)$ for everyone. In such a case, we say that the population is homogeneous, i.e., different agents form their evaluations in a statistically identical fashion, conditional on the type of the object. We will consider this case in Section 4.

— **Heterogeneous population:** In most subjective evaluations, the manner in which agents form evaluations differ considerably due to differences in preferences, abilities etc. So it is natural to assume that the filters vary across the population. We will consider this case in Section 5. We will see that in general it is impossible to design truthful mechanisms for this case unless some additional structural assumptions are made on the generating model. We will propose a structural assumption for the case $|\mathcal{S}| = 2$, i.e., in the case where the evaluations are binary, and design a truthful mechanism under this assumption.

## 3. RELATED WORK

The theory of elicitation of private evaluations or predictions of events has a rich history. In the standard setting, an agent possesses some private information in the form of an evaluation of some object or some informed prediction about an event, and one would like to elicit this private information. There are two categories of these problems. In the first category, the ground truth, e.g. true quality or nature of the object or the knowledge of the realization of the event that one wants to predict, is available or will be available at a later stage. In this case, the standard technique is to score an agent's reports against the ground truth, and proper scoring rules (see [Gneiting and Raftery 2007; Savage 1971; Lambert and Shoham 2009]) provide an elegant framework to do so. In the second category of problems, the ground truth is not known. In this case there is little to be done except to score these reports against the reports of other agents who have provided similar predictions about the same event. The situation is then inherently strategic, in which one hopes to sustain truthful reporting as an equilibrium of a game: assuming all the other agents provide their predictions truthfully, these predictions form an informative ensemble, and with a carefully designed rule that scores reports against this ensemble, one incentivizes any agent to also be truthful. The present work falls in this category.

In this category, majority of early literature has focused on the case where a single object is being evaluated. In a pioneering work, the peer-prediction method by [Miller et al. 2005] assumed that the population is homogeneous and the mechanism designer knows the agents' beliefs about the underlying generating model of evaluations. In this case they demonstrated the use of proper-scoring rules to design a truthful mechanism that utilizes the knowledge of these subjective beliefs. These mechanisms are minimal in the sense that they only require agents to report their evaluations. In another influential work, [Prelec 2004] considered a homogeneous population and designed an *oblivious* mechanism, famously termed Bayesian truth serum (BTS), that does not require the knowledge of the underlying generating model, but requires that the number of agents is large and that they have a common prior, i.e., they have the same beliefs about the underlying generating model and this fact is common knowledge. This mechanism is not minimal: apart from reporting their evaluations, agents are also required to report their beliefs about the reports of others. [Witkowski and Parkes 2012b] and [Radanovic and Faltings 2013] later used proper-scoring rules to design similar mechanisms for the case where the population size is finite. These mechanisms are again not minimal, and in fact it is known (see [Jurca and Faltings 2011], [Radanovic and Faltings 2013]) that no minimal mechanism that does not use the knowledge of the prior beliefs can incentivize truthful reporting of evaluations.

In the light of this impossibility, it is clear that any oblivious mechanism has to elicit some form of beliefs from the agents, and hence the agents' subjective beliefs about the beliefs of others become important in the equilibrium analysis. This is not the case with minimal mechanisms: one only needs to take into account an agent's first-order subjective beliefs about the underlying generating model. [Witkowski and Parkes 2012a] suggest a way around this difficulty for non-minimal mechanisms when the evaluations are binary.

It is the case in many applications in crowdsourcing, that one is interested in acquiring evaluations from a population for several similar objects. It is thus natural to explore the possibility of exploiting this statistical similarity to design better (e.g. minimal) mechanisms for jointly scoring these evaluation tasks. This is the context of the present work. Three major works in this area that have considered this case are [Witkowski and Parkes 2013], [Dasgupta and Ghosh 2013] and more recently, [Radanovic and Faltings 2015]. Both [Witkowski and Parkes 2013] and [Dasgupta and Ghosh 2013] only considered the case where the evaluations are binary. The former considered a homogeneous population while the latter considered a heterogeneous population, while both making specific assumptions on the generating model. [Radanovic and Faltings 2015] on the other hand have considered both homogeneous and heterogeneous populations.

For a homogeneous population with multiple objects, [Witkowski and Parkes 2013] try to utilize the statistical independence of the objects to estimate the prior distribution of evaluations and use that to compute payments using a proper scoring rule. In spirit, we are similar to this approach (and also [Prelec 2004]) in the sense that we use the law of large numbers to estimate some prior statistics and we get *asymptotic* incentive-compatibility, but we do not restrict ourselves to the binary setting. [Radanovic and Faltings 2015] recently have also designed a mechanism that is truthful in the general non-binary setting while requiring only a finite number of objects, again using proper scoring rules. In their mechanism, for computing the reward to an agent for evaluating a given object, a sample of evaluations of other agents for other objects of a fixed size needs to be collected, and an agent's reward can be non-zero only if this sample is sufficiently rich, i.e., it has an adequate representation of all the possible evaluations. Although our mechanism needs number of objects to be large, it has a much simpler structure.

For the case of heterogeneous population, [Radanovic and Faltings 2015] show that typically one cannot guarantee truthfulness with minimal elicitation. We prove a similar result for our setting. Nevertheless, [Dasgupta and Ghosh 2013] have designed truthful minimal mechanism for the case of binary evaluations for a very specific generating model: it is assumed that $\mathcal{H} = \mathcal{S}$ and for each agent the probability of correctly guessing the true type of the object is at least 0.5 and it does not depend on the type. Although we also consider the binary evaluations, we allow $\mathcal{H}$ to be arbitrary and our regularity condition is considerably weaker. Again we come up with a much simpler mechanism, albeit for the case where the number of evaluation tasks is large.

Another key aspect that differentiates the present work from others is that we do not use proper scoring rules in the design process. In fact, as would be clear from this discussion, proper scoring rules have been a ubiquitous and essential ingredient in most elicitation mechanisms studied so far. Indeed in most cases, the ingenuity in their design lies in cleverly proposing statistics that an agent's evaluation can be scored against using a proper scoring rule. Our mechanisms fall into a new framework that is a complete departure from this approach.

## 4. HOMOGENEOUS POPULATION

In this section, we will first consider the case where the population of agents is homogeneous. Consider the following mechanism: In the following theorem, we show that if $N$ is large, then truthful reporting of $Y^i_j$ is a Bayes-Nash equilibrium.

THEOREM 4.1. *Suppose that $N \to \infty$ and $Q$ is finite. Then the mechanism is detail-free Bayes-Nash incentive compatible. Suppose in addition that for any two evaluations $s_k$ and $s_l$, there is no $C \in \mathbb{R}$ such that $p(s_k|h) = Cp(s_l|h)$ for each $h \in \mathcal{H}$, then the mechanism is strictly detail-free Bayes-Nash incentive compatible.*

PROOF. Suppose that everyone but a person $j$ is truthful. Now since $N$ is large, and $|\mathcal{W}_j| \leq Q$, which is assumed to be small, by the mutual independence of $\{Y^i_j : j \in \mathcal{M}^i\}$

**Mechanism 1 (Homogeneous population):**

- The observations of all the people for the different objects are are solicited. Let these be denoted by $\{q^i_j\}$, where $q^i_j \in \mathcal{S}$.
- From each population $\mathcal{M}^i$, choose two persons $j_1$ and $j_2$ randomly, and for each possible evaluation $s_k \in \mathcal{S}$, compute the quantity

  $$f^i(s_k) = \mathbf{1}_{\{q^i_{j_1}=s_k\}}\mathbf{1}_{\{q^i_{j_2}=s_k\}}$$

  Then compute

  $$\bar{f}(s_k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f^i(s_k)}.$$
- For each evaluation $s_k$, fix a payment $r(s_k)$ that satisfies

  $$r(s_1)\bar{f}(s_1) = r(s_2)\bar{f}(s_2) = \cdots = r(s_K)\bar{f}(s_K).$$
- For each person $j$ in population $\mathcal{M}^i$, choose another person $j'$ from the same population. If their reports match, i.e. if $q^i_j = q^i_{j'} = s_k$, then the person $j$ gets a reward of $r(s_k)$. If the reports do not match, then $j$ gets 0 payment.

for the different objects $i$, we have by the strong law of large numbers (S.L.L.N.) and the continuous mapping theorem,

$$\bar{f}(s_k) \approx \sqrt{E(f^i(s_k))} = \sqrt{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2}.$$

The expected reward of person $j$ for evaluating object $i$, if she reports $q^i_j = s_l$ when her true evaluation is $s_k$ is

$$P(Y^i_{j'} = s_l|Y^i_j = s_k)r(s_l) = \frac{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)p(s_l|h)}{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)}r(s_l)$$

Thus $P(Y^i_{j'} = s_l|Y^i_j = s_k)r(s_l) \leq P(Y^i_{j'} = s_k|Y^i_j = s_k)r(s_k)$ if

$$\frac{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)p(s_l|h)}{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)}r(s_l) \leq \frac{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2}{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)}r(s_k)$$

i.e., if

$$\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)p(s_l|h) \leq \sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2\frac{r(s_k)}{r(s_l)},$$

or substituting, if

$$\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)p(s_l|h) \leq \sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2\frac{\bar{f}(s_l)}{\bar{f}(s_k)} = \sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2\frac{\sqrt{\sum_{h\in\mathcal{H}} P_X(h)p(s_l|h)^2}}{\sqrt{\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2}}.$$

$$= \sqrt{(\sum_{h\in\mathcal{H}} P_X(h)p(s_k|h)^2)(\sum_{h\in\mathcal{H}} P_X(h)p(s_l|h)^2)}.$$

which is always true by the Cauchy-Schwarz inequality. Thus person $j$ maximizes his expected payment by reporting each of his evaluations truthfully. Note that the equilibrium is not strict only when Cauchy-Schwarz inequality is not strict, which is the case if and only if for some $C \in \mathbb{R}$, $p(s_k|h) = Cp(s_l|h)$ for every $h \in \mathcal{H}$.

□

Note that if the condition for strictness is not satisfied, i.e., if for some $s_k$ and $s_l$, there is a $C$ such that $p(s_k|h) = Cp(s_l|h)$ for each $h \in \mathcal{H}$, then the evaluations $s_k$ and $s_l$ need not be distinguished since they contain the same information about $X$. In particular, $P(X = h|Y_j = s_k) = P(X = h|Y_j = s_l)$ for each $h \in \mathcal{H}$.

### 4.1. An alternative to the peer prediction method

In the case where the mechanism designer knows the underlying generating model, i.e. $P_X$ and $\{p(s|h)\}$, the mechanism can compute the rewards for each evaluation directly, without having to estimate statistics from evaluations for multiple objects. In order to do so, for each evaluation $s_k$, one defines

$$g(s_k) = \sqrt{\sum_{h \in \mathcal{H}} P_X(h) p(s_k|h)^2},$$

and defines payments $r(s_k)$ for the different evaluations such that:

$$r(s_1)g(s_1) = r(s_2)g(s_2) = \cdots = r(s_K)g(s_K).$$

In this case, our mechanism provides an alternative to the peer prediction method of [Miller et al. 2005], while using the simple structure of output agreement mechanisms and without using proper scoring rules.

## 5. HETEROGENEOUS POPULATION

We now consider the case of a heterogeneous population. In this case, we will be working under the following assumption:

**Assumption 1:** For each $j$, $\{p_j(s|h)\}$ is a random stochastic matrix of size $|\mathcal{H}| \times |\mathcal{S}|$ (recall that a stochastic matrix is one in which all the entries are non-negative and all the rows sum to 1), and these filters are independently but identically distributed for all $j$, with some distribution $\mathcal{Q}$ defined on a support $\mathcal{B}$, which is some subset of the set of all stochastic matrices of size $|\mathcal{H}| \times |\mathcal{S}|$.

We do not make any assumptions about the knowledge of this distribution: it may or may not be known to the population or the mechanism designer. With this additional assumption, we will modify the definition of a detail-free Bayes-Nash incentive compatible mechanism.

*Definition* 5.1. Fix a subset $\mathcal{B}$ of the set of all stochastic matrices of size $|\mathcal{H}| \times |\mathcal{S}|$. We say that a given payment mechanism $\{\tau_j : j \in \mathcal{M}\}$ is *detail-free Bayes-Nash incentive compatible* if for each $j$,

$$E[\tau_j(\{y_j^i : i \in \mathcal{W}_j\}, \{Y_{j'}^i : i = 1, \cdots, N,\ j' \in \mathcal{M}^i,\ j' \neq j\}) \mid Y_j^i = y_j^i,\ i \in \mathcal{W}_j,\ \{p_j(s|h)\},\ P_X,\ \mathcal{Q}]$$
$$\geq E[\tau_j(\{q_j^i : i \in \mathcal{W}_j\}, \{Y_{j'}^i : i = 1, \cdots, N,\ j' \in \mathcal{M}^i,\ j' \neq j\}) \mid Y_j^i = y_j^i,\ i \in \mathcal{W}_j,\ \{p_j(s|h)\},\ P_X,\ \mathcal{Q}],$$

for each $y_j^i \in \mathcal{S}$ for $i \in \mathcal{W}_j$, and for every specification of $P_X$, of the distribution $\mathcal{Q}$ defined on the support $\mathcal{B}$, and $\{p_j(s|h)\} \in \mathcal{B}$. We call it *strictly* detail-free Bayes-Nash incentive compatible if the above inequality is strict whenever $\{y_j^i : i \in \mathcal{W}_j\} \neq \{q_j^i : i \in \mathcal{W}_j\}$.

We will first show that in this case, unless additional structural assumptions are made, it is impossible to design a strictly detail-free Bayes-Nash incentive compatible mechanism.

PROPOSITION 5.2. *Suppose that $\mathcal{B}$ is the entire space of stochastic matrices of size $|\mathcal{H}| \times |\mathcal{S}|$. Then there is no strictly detail-free Bayes-Nash incentive compatible mechanism for the resulting generating model.*

PROOF. (sketch) Assume a movie is being evaluated. Suppose that its type is in the set $\mathcal{H} = \{Action, Drama\}$ and that the set of evaluations is $\mathcal{S} = \{Good, Bad\}$. Consider an agent *Bart* who is an 'action lover', and another agent *Lisa* who is a 'drama lover', with respective filters shown in the Figure 1 below. Now observe that the filters are such that for any fixed distribution $\mathcal{Q}$ that contains them in its support, the conditional distribution over the evaluations of all the other agents in the population from the point of view of Bart, given that he evaluates the movie to be good, is the same as the conditional distribution over the evaluations of all the other agents in the population from the point of view of Lisa, given that she evaluates the movie to be bad. So any strictly detail-free Bayes-Nash incentive compatible mechanism that strictly incentivizes Bart to report 'Good' when his evaluation is 'Good' also simultaneously incentivizes Lisa to report 'Good' when her evaluation for the movie is 'Bad'. Hence there can be no mechanism that is strictly Bayes-Nash incentive compatible. □

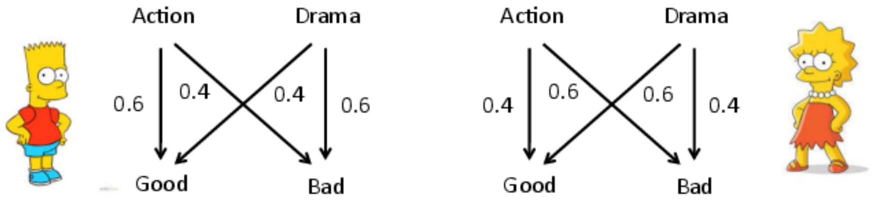

Fig. 1. Two filters.

### 5.1. Regular filters:

The previous result intuitively implies that if the preferences of the agents in the population varies too much, then designing a knowledge independent strictly truthful mechanism is impossible. One thus needs to impose some uniformity or *regularity* on the population filters for there to be any hope of designing such mechanisms. With this in mind, we look at the specific case of binary evaluations, i.e., $|\mathcal{S}| = 2$ and we impose a notion regularity on the filters that we now define. Consider the following generating model:

(1) $|\mathcal{S}| = 2$, i.e. evaluations take only 2 values. Let $\mathcal{S} = (s_1, s_2)$. We will call this the binary setting.

(2) For each $j$, $\{p_j(s|h)\}$ is a random stochastic matrix of size $|\mathcal{H}| \times 2$ drawn independently from some distribution $\mathcal{Q}$ defined on the support $\mathcal{B}$, where $\mathcal{B}$ has the property that each matrix in it satisfies the following *regularity* condition:
Suppose that there is a fixed ordering of types in the set $\mathcal{H}$, and w.l.o.g. we assume that it is

$$h_1 \succ h_2 \succ \cdots \succ h_L,$$

such that for any filter $\{p_j(s|h)\} \in \mathcal{B}$,

$$p_j(s_1|h) \geq p_j(s_1|h') \quad \text{if} \quad h \succ h' \tag{1}$$

for each $h \neq h' \in \mathcal{H}$. Note that this implies that $p_j(s_2|h) \leq p_j(s_2|h')$ if $h \succ h'$. Intuitively, this means that the types that are higher in the ordering are "closer" to $s_1$ and those lower in the ordering are in some sense "farther away" from $s_1$, and vice versa for $s_2$, so that an agent making a particular evaluation is more likely if the object type is "closer" to that evaluation. Note that for any particular filter $p_j(.|.)$ one can always define an ordering of the types such that this condition will be satisfied for that filter. But the regularity assumption says that there is one such fixed ordering of types for *all* filters in the support of the distribution. Although this distribution is not needed to be known to the population, the fact that the filters in its support satisfy this condition with respect to some fixed ordering of the types is commonly known. If this condition is satisfied, then we will simply say that the filters are *regular*. We say that the filters are *strictly regular* if the above inequality is strict for each $h$, $h' \in \mathcal{H}$. Denote

$$E(p_j(s|h)) = p(s|h).$$

We will call this the *ensemble* filter.

It is instructive to think about what regularity boils down to for the case where $|\mathcal{H}| = |\mathcal{S}| = 2$. In this case, w.l.o.g., we can assume that $\mathcal{H} = \mathcal{S} = \{h_1, h_2\}$. Then the regularity condition will be satisfied if either:

—$p_j(h_1|h_1) \geq p_j(h_1|h_2)$ (and thus $p_j(h_2|h_2) \geq p_j(h_2|h_1)$) or if

—$p_j(h_1|h_2) \geq p_j(h_1|h_1)$ (and thus $p_j(h_2|h_1) \geq p_j(h_2|h_2)$).

In the context of peer grading, the first condition intuitively makes sense. It basically says that a person judging that the answer deserves a grade A is more likely if the true grade is A, than if the true grade is something else. Note that this is different from saying that

$$p_j(h_1|h_1) \geq p_j(h_2|h_1),$$

which says for example that if the true grade is A, then it is more likely that a person thinks it is A than it is something else. In fact, in this case, this condition of $p_j(h_1|h_1) \geq p_j(h_2|h_1)$ and $p_j(h_2|h_2) \geq p_j(h_1|h_2)$ imply that $p_j(h_1|h_1) \geq \frac{1}{2}$, $p_j(h_2|h_2) \geq \frac{1}{2}$, $p_j(h_2|h_1) \leq \frac{1}{2}$ and $p_j(h_1|h_2) \leq \frac{1}{2}$. It follows that $p_j(h_1|h_1) \geq p_j(h_1|h_2)$, thereby proving that the condition (1) is strictly weaker (in the case $|\mathcal{H}| = |\mathcal{S}| = 2$).

Consider the following mechanism: We then have the following result:

THEOREM 5.1. *Suppose that $N \to \infty$ and $Q$ is finite. Also suppose that the filters are strictly regular. Then the mechanism is strictly detail-free Bayes-Nash incentive compatible.*

PROOF. Suppose a person $j$ in population $i$ makes an observation $Y^i_j = s_1$. Since $N$ is large, and $|\mathcal{W}_j \leq Q|$, which is small, by the mutual independence of $\{Y^i_j : j \in \mathcal{M}^i\}$ for the different objects $i$, we have by the S.L.L.N.,

$$\bar{f}(s_k) \approx E(f^i(s_k)) = \sum_{h_l} P_X(h_l) p(s_k|h_l)$$

for $k = 1, 2$. Next we have

$$P(X^i = h_l | Y^i_j = s_1) = \frac{P_X(h_l) p_j(s_1 \mid h_l)}{\sum_{h_l} P_X(h_l) p_j(s_1|h_l)} = P_X(h_l) t(s_1, h_l),$$

**Mechanism 2 (Heterogeneous population with binary evaluations and regular filters):**

- The observations of all the people for the different objects are are solicited. Let these be denoted by $\{q_j^i\}$.
- From each population $\mathcal{M}^i$, a person $j$ is picked at random. For each of the two evaluations $s_1$ and $s_2$, define
$$f_i(s_k) = \mathbf{1}_{\{q_j^i = s_k\}}.$$
Then compute
$$\bar{f}(s_k) = \frac{1}{N}\sum_{i=1}^{N} f_i(s_k).$$
- For the evaluations $s_1$ and $s_2$, fix payments $r(s_1)$ and $r(s_2)$ that satisfy
$$r(s_1)\bar{f}(s_1) = r(s_2)\bar{f}(s_2).$$
- For each person $j$ in population $\mathcal{M}^i$, choose another person $j'$ in the same population. Then if $q_j^i = q_{j'}^i = s_k$, then person $j$ receives a payment of $r(s_k)$. If their reports do not match, then person $j$ gets 0 payment.

where we define

$$t(s_1, h_l) \triangleq \frac{p_j(s_1 \mid h_l)}{\sum_{h_l} P_X(h_l) p_j(s_1|h_l)}.$$

Because of strict regularity of the filters, we have that $t(s_1, h_1) > t(s_1, h_2) > \cdots > t(s_1, h_L)$ and further $t(s_1, h_1) > 1$ and $t(s_1, h_L) < 1$.

Now from the point of view of person $i$, the distribution of the report of a randomly chosen person $j'$ is

$$P(Y_{j'}^i = s_1 | Y_j^i = s_1) = \sum_{h_l} P(X^i = h_l | Y_j^i = s_1) p(s_1|h_l) = \sum_{h_l} P_X(h_l) t(s_1, h_l) p(s_1|h_l),$$

and $P(Y_{j'}^i = s_2 | Y_j^i = s_1) = 1 - P(Y_{j'}^i = s_1 | Y_j^i = s_1)$. This holds because $Y_j^i$ for different persons $j$ are conditionally independent given the type $X^i$. Next we have

$$P(Y_{j'}^i = s_1 | Y_j^i = s_1) - \bar{f}(s_1) \ = \sum_{h_l} p(s_1|h_l)\Big(P_X(h_l) t(s_1, h_l) - P_X(h_l)\Big).$$

We want to show that this quantity is positive, i.e. if an agent has an evaluation $s_1$ for an object, then the posterior probability that another agent also has the same evaluation for that object increases relative to the prior. Now since $t(s_1, h_1) > t(s_1, h_2) > \cdots > t(s_1, h_L)$ and $t(s_1, h_1) > 1$ while $t(s_1, h_L) < 1$, there must be some $l^*$ such that $t(s_1, h_l) \geq 1$ for all $l \leq l^*$ while $t(s_1, h_l) < 1$ for all $l > l^*$. Then we have that

$$\sum_{h_l; l \leq l^*} P_X(h_l)(t(s_1, h_l) - 1) > 0. \tag{2}$$

Further, since $\sum_{h_l} P_X(h_l)(t(s_1, h_l) - 1) = 0$, we have that

$$\sum_{h_l; l \leq l^*} P_X(h_l)(t(s_1, h_l) - 1) = - \sum_{h_l; l > l^*} P_X(h_l)(t(s_1, h_l) - 1). \qquad (3)$$

Now we have:

$$\begin{aligned}
& P(Y^i_{j'} = s_1 | Y^i_j = s_1) - \bar{f}(s_1) \\
&= \sum_{h_l; l \leq l^*} p(s_1|h_l) P_X(h_l)(t(s_1, h_l) - 1) + \sum_{h_l; l > l^*} p(s_1|h_l) P_X(h_l)(t(s_1, h_l) - 1) \\
&\geq p(s_1|h_{l^*}) \sum_{h_l; l \leq l^*} P_X(h_l)(t(s_1, h_l) - 1) + p(s_1|h_{l^*+1}) \sum_{h_l; l > l^*} P_X(h_l)(t(s_1, h_l) - 1) \\
&= \Big( p(s_1|h_{l^*}) - p(s_1|h_{l^*+1}) \Big) \sum_{h_l; l \leq l^*} P_X(h_l)(t(s_1, h_l) - 1) > 0.
\end{aligned}$$

The first inequality follows from the fact that $P_X(h_l)t(s_1, h_l) - P_X(h_l) \geq 0 \, (< 0)$ for $l \leq l^* \, (> l^*)$ and that $p(s_1|h_1) > p(s_1|h_2) > \cdots > p(s_1|h_L)$ by the strict regularity of the ensemble filter. The second equality follows from (2) and (3). Thus, we also have that $P(Y^i_{j'} = s_2 | Y^i_j = s_1) - \bar{f}(s_2) < 0$. Hence the expected payoff if a person $j$ reports $s_1$, if he observes $s_1$ is

$$\begin{aligned}
\phi_{s_1|s_1} &= P(Y^i_{j'} = s_1 | Y^i_j = s_1) r(s_1) > \bar{f}(s_1) r(s_1) \\
&= \bar{f}(s_2) r(s_2) > P(Y^i_{j'} = s_2 | Y^i_j = s_1) r(s_2) = \phi_{s_2|s_1}.
\end{aligned}$$

Thus being truthful maximizes his expected payment. □

## 6. DISCUSSION

In this paper, we presented mechanisms for obtaining truthful reports with minimal elicitation Our mechanisms support the setting where agents are assumed to be homogeneous, and also support heterogeneous workers when questions are of binary-choice format. The mechanisms rely on the existence of many questions, a feature commonly encountered in the settings of crowdsourcing and peer-grading. Interestingly, our mechanisms are built under a novel framework that is a significant departure from the traditional setup of proper scoring rules.

Our broader objective is to construct mechanisms that incentivize truthful reports in the absence of any ‘gold standard’ questions, that are also very viable in practice. The results of this paper take a step in this direction by eliminating two of the critical assumptions that were found in most earlier works. While most of our theory in this paper is asymptotic in the number of questions, in the future, we intend to examine the effect of a restriction on the number of questions. We hope to obtain analytical approximation guarantees, and also perhaps a better understanding of these effects via empirical evaluations. A second question left open in this manuscript is that of $|\mathbb{S}| > 2$ in the heterogeneous setting. It would be interesting to find the right notion of “regularity” of the agents’ filters that would allow one to design a strictly detail-free Bayes-Nash incentive compatible mechanism in that case.

## REFERENCES

Chen, J. J., Menezes, N. J., Bradley, A. D., and North, T. 2011. Opportunities for crowdsourcing research on amazon mechanical turk. *Interfaces 5,* 3.

Dasgupta, A. and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 319–330.

Gneiting, T. and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

Jurca, R. and Faltings, B. 2011. Incentives for answering hypothetical questions. In *Workshop on Social Computing and User Generated Content, EC-11*. Number EPFL-CONF-197783.

Lambert, N. and Shoham, Y. 2009. Eliciting truthful answers to multiple-choice questions. In *ACM conference on Electronic commerce*. 109–118.

Le, J., Edmonds, A., Hester, V., and Biewald, L. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 21–26.

Luca, M. 2011. Reviews, reputation, and revenue: The case of yelp. com. *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper* 12-016.

Miller, N., Resnick, P., and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. 2013. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science 306,* 5695, 462–466.

Radanovic, G. and Faltings, B. 2013. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*. Number EPFL-CONF-197486. 833–839.

Radanovic, G. and Faltings, B. 2015. Incentives for Subjective Evaluations with Private Beliefs. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. 2010. Learning from crowds. *The Journal of Machine Learning Research 11*, 1297–1322.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66,* 336, 783–801.

Shah, N. B., Zhou, D., and Peres, Y. 2015. Approval voting and incentives in crowdsourcing. In *International Conference on Machine Learning (ICML)*.

Von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

Von Ahn, L. and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM 51,* 8, 58–67.

Witkowski, J. and Parkes, D. C. 2012a. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 964–981.

Witkowski, J. and Parkes, D. C. 2012b. A robust Bayesian truth serum for small populations. In *AAAI*.

Witkowski, J. and Parkes, D. C. 2013. Learning the prior in minimal peer prediction. In *3rd Workshop on Social Computing and User Generated Content at the ACM Conference on Electronic Commerce*.