Informed Truthfulness in Multi-Task Peer Prediction

VICTOR SHNAYDER, edX; Paulson School of Engineering, Harvard University
ARPIT AGARWAL, Indian Institute of Science, Bangalore
RAFAEL FRONGILLO, University of Colorado, Boulder
DAVID C. PARKES, Paulson School of Engineering, Harvard University

The problem of peer prediction is to elicit information from agents in settings without any objective ground truth against which to score reports. Peer prediction mechanisms seek to exploit correlations between signals to align incentives with truthful reports. A long-standing concern has been the possibility of uninformative equilibria. For binary signals, a multi-task mechanism [Dasgupta and Ghosh 2013] achieves *strong truthfulness*, so that the truthful equilibrium strictly maximizes payoff. We characterize conditions on the signal distribution for which this mechanism remains strongly-truthful with non-binary signals, also providing a greatly simplified proof. We introduce the *Correlated Agreement* (CA) mechanism, which handles multiple signals and provides *informed truthfulness*: no strategy profile provides more payoff in equilibrium than truthful reporting, and the truthful equilibrium is strictly better than any uninformed strategy (where an agent avoids the effort of obtaining a signal). The CA mechanism is maximally strongly truthful, in that no mechanism in a broad class of mechanisms is strongly truthful on a larger family of signal distributions. We also give a detail-free version of the mechanism that removes any knowledge requirements on the part of the designer, using reports on many tasks to learn statistics while retaining ϵ-informed truthfulness.

1. INTRODUCTION

We study the problem of information elicitation without verification ("peer prediction"). This challenging problem arises across a diverse range of multi-agent systems, in which participants are asked to respond to an information task, and where there is no external input available against which to score reports. Examples include completing surveys about the features of new products, providing feedback on the quality of food or the ambience in a restaurant, sharing emotions when watching video content, and peer assessment of assignments in Massive Open Online Courses (MOOCs).

The challenge is to provide incentives for participants to choose to invest effort in forming an opinion (a "signal") about a task, and to make truthful reports about their signals. In the absence of inputs other than the reports of participants, peer-prediction mechanisms make payments to one agent based on the reports of others, and seek to align incentives by leveraging correlation between reports (i.e., peers are rewarded for making reports that are, in some sense, predictive of the reports of others).

Some domains have binary signals, for example "was a restaurant noisy or not?", and "is an image violent or not?". We are also interested in domains with non-binary signals, for example:

This research is supported in part by a grant from Google, the SEAS TomKat fund, and NSF grant CCF-1301976. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone. Thanks to participants in seminars at IOMS NYU Stern, the Simons Instutite, the GSBE-ETBC seminar at Maastricht University, and reviewers for useful feedback. Author addresses: shnayder@eecs.harvard.edu, arpit.agarwal@csa.iisc.ernet.in, raf@colorado.edu, parkes@eecs.harvard.edu

EC'16, July 24–28, 2016, Maastricht, The Netherlands. ACM 978-1-4503-3936-0/16/07 ...$15.00.

http://dx.doi.org/10.1145/2940716.2940790

— *Image labeling.* Signals could correspond to answers to questions such as "Is the animal in the picture a dog, a cat or a beaver", or "Is the emotion expressed joyful, happy, sad or angry." These signals are categorical, potentially with some structure: 'joyful' is closer to 'happy' than 'sad', for example.
— *Counting objects.* There could be many possible signals, representing answers to questions such as ("are there 0, 1-5, 6-10, 11-100, or >100 people in the picture"?). The signals are ordered.
— *Peer assessment in MOOCs*. Multiple students evaluate their peers' submissions to an open-response question using a grading rubric. For example, an essay may be evaluated for clarity, reasoning, and relevance, with the grade for reasoning ranging from 1 ("wild flights of fancy throughout"), through 3 ("each argument is well motivated and logically defended.")

We do not mean to take an absolute position that external "ground truth" inputs are never available in these applications. We do however believe it important to understand the extent to which such systems can operate using only participant reports.

The design of peer-prediction mechanisms assumes the ability to make payments to agents, and that an agent's utility is linear-increasing with payment and does not depend on signal reports other than through payment. Peer prediction precludes, for example, that an agent may prefer to misreport the quality of a restaurant because she is interested in driving more business to the restaurant.[1] The challenge of peer prediction is timely. For example, Google launched *Google Local Guides* in November 2015. This provides participants with points for contributing star ratings and descriptions about locations. The current design rewards quantity but not quality and it will be interesting to see whether this attracts useful reports. After 200 contributions, participants receive a 1 TB upgrade of Drive storage (currently valued at $9.99/month.)

We are interested in *minimal* peer-prediction mechanisms, which require only signal reports from participants.[2] A basic desirable property is that truthful reporting of signals is a strict, correlated equilibrium of the game induced by the peer-prediction mechanism.[3] For many years, an Achilles heel of peer prediction has been the existence of additional equilibria that payoff-dominate truthful behavior and reveal no useful information [Dasgupta and Ghosh 2013; Jurca and Faltings 2009; Radanovic and Faltings 2015a]. An uninformative equilibrium is one in which reports do not depend on the signals received by agents. Indeed, the equilibria of peer-prediction mechanisms must always include an uninformative, mixed Nash equilibrium [Waggoner and Chen 2014]. Moreover, with binary signals, a single task, and two agents, Jurca and Faltings [2005] show that an incentive-compatible, minimal peer-prediction mechanism will always have an uninformative equilibrium with a higher payoff than truthful re-

[1]The payments need not be monetary; one could for example issue points to agents, these points conveying some value (e.g., redeemable for awards, or conveying status). On a MOOC platform, the payments could correspond to scores assigned as part of a student's overall grade in the class. What is needed is a linear relationship between payment (of whatever form) and utility, and expected-utility maximizers.

[2]While more complicated designs have been proposed (e.g. [Prelec 2004; Radanovic and Faltings 2015b; Witkowski and Parkes 2012]), in which participants are also asked to report their beliefs about the signals that others will report, we believe that peer-prediction mechanisms that require only signal reports are more likely to be adopted in practice. It is cumbersome to design user interfaces for reporting beliefs, and people are notoriously bad at reasoning about probabilities.

[3]It has been more common to refer to the equilibrium concept in peer-prediction as a Bayes-Nash equilibrium. But as pointed out by Jens Witkowski, there is no agent-specific, private information about payoffs (utility is linear in payment). In a correlated equilibrium, agents get signals and a strategy is a mapping from signals to actions. An action is a best response for a given signal if, conditioned on the signal, it maximizes an agent's expected utility. This equilibrium concept fits peer prediction: each agent receives a signal from the environment, signals are correlated, and strategies map signals into reported signals.

porting. Because of this, a valid concern has been that peer prediction could have the unintended effect that agents who would otherwise be truthful now adopt strategic misreporting behavior in order to maximize their payments.

In this light, a result due to Dasgupta and Ghosh [2013] is of interest: if agents are each asked to respond to multiple, independent tasks (with some overlap between assigned tasks), then in the case of binary signals there is a mechanism that addresses the problem of multiple equilibria. The binary-signal, multi-task mechanism is *strongly truthful*, meaning that truthful reporting yields a higher expected payment than any other strategy (and is tied in payoff only with strategies that report permutations of signals, which in the binary case means $1 \to 2, 2 \to 1$).

We introduce a new, slightly weaker incentive property of *informed truthfulness*: no strategy profile provides more expected payment than truthful reporting, and the truthful equilibrium is strictly better than any uninformed strategy (where agent reports are signal-independent, and avoid the effort of obtaining a signal). Informed truthfulness is responsive to what we consider to be the two main concerns of practical peer prediction design:

(a) Agents should have strict incentives to exert effort toward acquiring an informative signal, and
(b) Agents should have no incentive to misreport this information.

Relative to strong truthfulness, the relaxation to informed truthfulness is that there may be other informed strategies that match the expected payment of truthful reporting. Even so, informed truthfulness retains the property of strong truthfulness that there can be no other behavior strictly better than truthful reporting.

The binary-signal, multi-task mechanism of Dasgupta and Ghosh is constructed from the simple building block of a *score matrix*, with a score of '1' for agreement and '0' otherwise. Some tasks are designated without knowledge of participants as bonus tasks. The payment on a bonus task is 1 in the case of agreement with another agent. There is also a penalty of -1 if the agent's report on another (non-bonus) task agrees with the report of another agent on a third (non-bonus) task. In this way, the mechanism rewards agents when their reports on a shared (bonus) task agree more than would be expected based on their overall report frequencies. Dasgupta and Ghosh remark that extending beyond two signals "is one of the most immediate and challenging directions for further work."

Our main results are as follows:

— We study the *multi-signal extension of the Dasgupta-Ghosh mechanism* (MSDG), and show that MSDG is strongly truthful for domains that are *categorical*, where receiving one signal reduces an agent's belief that other agents will receive any other signal. We also show that (i) this categorical condition is tight for MSDG for agent-symmetric signal distributions, and (ii) the peer grade distributions on a large MOOC platform do not satisfy the categorical property.
— We generalize MSDG, obtaining the *Correlated Agreement (CA) mechanism*. This provides informed truthfulness in general domains, including domains in which the MSDG mechanism is neither informed- nor strongly-truthful. The CA mechanism requires the designer to know the correlation structure of signals, but not the full signal distribution. We further characterize domains where the CA mechanism is strongly truthful, and show that no mechanism with similar structure and information requirements can do better.
— For settings with a large number of tasks, we present a *detail-free* CA *mechanism*, in which the designer estimates the statistics of the correlation structure from agent reports. This mechanism is informed truthful in the limit where the number of tasks

is large (handling the concern that reports affect estimation and thus scores), and we provide a convergence rate analysis for ϵ-informed truthfulness with high probability.

We believe that these are the first results on strong or informed truthfulness in domains with non-binary signals without requiring a large population for their incentive properties (compare with [Kamble et al. 2015; Radanovic and Faltings 2015a; Radanovic et al. 2016]). The robust incentives of the multi-task MSDG and CA mechanisms hold for as few as two agents and three tasks, whereas these previous papers crucially rely on being able to learn statistics of the distribution from multiple reports. Even if given the true underlying signal distribution, the mechanisms in these earlier papers would still need to use a large population, with the payment rule based on statistics estimated from reports, as this is critical for incentive alignment in these papers. Our analysis framework also provides a dramatic simplification of the techniques used by Dasgupta and Ghosh [2013].

In a recent working paper, Kong and Schoenebeck [2016] show that a number of peer prediction mechanisms that provide variations on strong-truthfulness can be derived within a single information-theoretic framework, with scores determined based on the information they provide relative to reports in the population (leveraging a measure of mutual information between the joint distribution on signal reports and the product of marginal distributions on signal reports). Earlier mechanisms correspond to particular information measures. Their results use different technical tools, and also include a different, multi-signal generalization of Dasgupta and Ghosh [2013] that is independent of our results, outside of the family of mechanisms that we consider in Section 5.2, and provides strong truthfulness in the limit of a large number of tasks.[4]

1.1. Related Work

The theory of peer prediction has developed rapidly in recent years. We focus on minimal peer-prediction mechanisms. Beginning with the seminal work of Miller et al. [2005], a sequence of results relax knowledge requirements on the part of the designer [Jurca and Faltings 2011; Witkowski and Parkes 2012], or generalize, e.g. to handle continuous signal domains [Radanovic and Faltings 2014]. Simple output-agreement, where a positive payment is received if and only if two agents make the same report (as used in the *ESP game* [von Ahn and Dabbish 2004]), has also received some theoretical attention [Jain and Parkes 2013; Waggoner and Chen 2014].

Early peer prediction mechanisms had uninformative equilibria that gave better payoff than honesty. Jurca and Faltings [2009] show how to remove uninformative, pure-strategy Nash equilibria through a clever three-peer design. Kong et al. [2016] show how to design strong truthful, minimal, single-task mechanisms with a known model when there are reports from a large number of agents.

In addition to Dasgupta and Ghosh [2013] and Kong and Schoenebeck [2016], several recent papers have tackled the problem of uninformative equilibria. Radanovic and Faltings [2015a] establish strong truthfulness amongst symmetric strategies in a large-market limit where both the number of tasks and the number of agents assigned to each task grow without bound. Radanovic et al. [2016] provide complementary theoretical results, giving a mechanism in which truthfulness is the equilibrium with highest payoff, based on a population that is large enough to estimate statistical properties of the report distribution. They require a self-predicting condition that limits the correlation between differing signals. Each agent need only be assigned a single

[4]While they do not state or show that the mechanism does not need a large number of tasks in any special case, the techniques employed can also be used to design a mechanism that is a linear transform of our CA mechanism, and thus informed truthful with a known signal correlation structure and a finite number of tasks (personal communication).

task. Kamble et al. [2015] describe a mechanism where truthfulness has higher payoff than uninformed strategies, providing an asymptotic analysis as the number of tasks grows without bound. The use of learning is crucial in these papers. In particular, they must use statistics estimated from reports to design the payment rule in order to align incentives. This is a key distinction from our work.[5] Witkowski and Parkes [2013] first introduced the combination of learning and peer prediction, coupling the estimation of the signal prior together with the shadowing mechanism.

Although there is disagreement in the experimental literature about whether equilibrium selection is a problem in practice, there is compelling evidence that it matters [Gao et al. 2014]; see Faltings et al. [2014] for a study where uninformed equilibria did not appear to be a problem.[6] Shnayder et al. [2016b] use replicator dynamics as a model of agent learning to argue that equilibrium selection is indeed important, and that truthfulness is significantly more stable under mechanisms that ensure it has higher payoff than other strategies. Orthogonal to concerns about equilibrium selection, Gao et al. [2016] point out a modeling limitation—when agents can coordinate on some other, unintended source of signal, then this strategy may be better than truthful reporting. They suggest randomly checking a fraction of reports against ground truth as an alternative way to encourage effort. We discuss this in Section 5.5.

Turning to online peer assessment for MOOCs, research has primarily focused on evaluating students' skill at assessment and compensating for grader bias [Piech et al. 2013], as well as helping students self-adjust for bias and provide better feedback [Kulkarni et al. 2013]. Other studies, such as the *Mechanical TA* [Wright et al. 2015], focus on reducing TA workload in high-stakes peer grading. A recent paper [Wu et al. 2015] outlines an approach to peer assessment that relies on students flagging overly harsh feedback for instructor review. We are not aware of any systematic studies of peer prediction in the context of MOOCs, though Radanovic et al. [2016] present experimental results from an on-campus experiment.

2. MODEL

We consider two agents, 1 and 2, which are perhaps members of a larger population. Let $k \in M = \{1, \ldots, m\}$ index a task from a universe of $m \geq 3$ tasks to which one or both of these agents are assigned, with both agents assigned to at least one task. Each agent receives a signal when investing effort on an assigned task. The effort model that we adopt is binary: either an agent invests no effort and does not receive an informed signal, or an agent invests effort and incurs a cost and receives a signal.

Let S_1, S_2 denote random variables for the signals to agents 1 and 2 on some task. The signals have a finite domain, with $i, j \in \{1, \ldots, n\}$ indexing a realized signal to agents 1 and 2, respectively.

Each task is *ex ante* identical, meaning that pairs of signals are i.i.d. for each task. Let $P(S_1{=}i, S_2{=}j)$ denote the joint probability distribution on signals, with marginal probabilities $P(S_1{=}i)$ and $P(S_2{=}j)$ on the signals of agents 1 and 2, respectively. We assume exchangeability, so that the identity of agents does not matter in defining the signal distribution. The signal distribution is common knowledge to agents.[7]

[5] Cai et al. [2015] work in a different model, showing how to achieve optimal statistical estimation from data provided by self-interested participants. These authors do not consider misreports and their mechanism is not informed- (or strongly-) truthful and is vulnerable to collusion. Their model is interesting, though, in that it adopts a richer, non-binary effort model.

[6] One difference is that this later study was in a many-signal domain, making it harder for agents to coordinate on an uninformative strategy.

[7] We assume common knowledge and symmetric signal models for simplicity of exposition. Our mechanisms do not require full information about the signal distribution, only the correlation structure of signals, and can tolerate some user heterogeneity, as described further in Section 5.4.

We assume that the signal distribution satisfies *stochastic relevance*, so that for all $s' \neq s''$, there exists at least one signal s such that

$$P(S_1{=}s|S_2{=}s') \neq P(S_1{=}s|S_2{=}s''), \tag{1}$$

and symmetrically, for agent 1's signal affecting the posterior on agent 2's. If two signals are not stochastically relevant, they can be combined into one signal.

Our constructions and analysis will make heavy use of the following matrix, which encodes the correlation structure of signals.

Definition 2.1 (*Delta matrix*). The *Delta matrix* Δ is an $n \times n$ matrix, with entry (i, j) defined as

$$\Delta_{ij} = P(S_1{=}i, S_2{=}j) - P(S_1{=}i)P(S_2{=}j). \tag{2}$$

The Delta matrix describes the correlation (positive or negative) between different realized signal values. For example, if $\Delta_{1,2} = P(S_1{=}1, S_2{=}2) - P(S_1{=}1)P(S_2{=}2) = P(S_1{=}1)(P(S_2{=}2|S_1{=}1) - P(S_2{=}2)) > 0$, then $P(S_2{=}2|S_1{=}1) > P(S_2{=}2)$, so signal 2 is positively correlated with signal 1 (and by exchangeability, similarly for the effect of 1 on 2). If a particular signal value increases the probability that the other agent will receive the same signal then $P(S_1{=}i, S_2{=}i) > P(S_1{=}i)P(S_2{=}i)$, and if this holds for all signals the Delta matrix has a positive diagonal. Because the entries in a row i of joint distribution $P(S_1{=}i, S_2{=}j)$ and a row of product distribution $P(S_1{=}i)P(S_2{=}j)$ both sum to $P(S_1{=}i)$, each row in the Δ matrix sums to 0 as the difference of the two. The same holds for columns.

The CA mechanism will depend on the sign structure of the Δ matrix, without knowledge of the specific values. We will use a sign operator $\mathrm{Sign}(x)$, with value 1 if $x > 0$, 0 otherwise.[8]

Example 2.2. If the signal distribution is

$$P(S_1, S_2) = \begin{bmatrix} 0.4 & 0.15 \\ 0.15 & .3 \end{bmatrix}$$

with marginal distribution $P(S) = [0.55; 0.45]$, we have

$$\Delta = \begin{bmatrix} 0.4 & 0.15 \\ 0.15 & .3 \end{bmatrix} - \begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix} \cdot [0.55 \;\; 0.45] \approx \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 0.1 \end{bmatrix}, \text{ and } \mathrm{Sign}(\Delta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

An agent's *strategy* defines, for every signal it may receive and each task it is assigned, the signal it will report. We allow for mixed strategies, so that an agent's strategy defines a distribution over signals. Let R_1 and R_2 denote random variables for the *reports* by agents 1 and 2, respectively, on some task. Let matrices F and G denote the mixed strategies of agents 1 and 2, respectively, with $F_{ir} = P(R_1{=}r|S_1{=}i)$ and $G_{jr} = P(R_2{=}r|S_2{=}j)$ to denote the probability of making report r given signal i is observed (signal j for agent 2). Let $r_1^k \in \{1, \ldots, n\}$ and $r_2^k \in \{1, \ldots, n\}$ refer to the realized report by agent 1 and 2, respectively, on task k (if assigned).

Definition 2.3 (*Permutation strategy*). A *permutation strategy* is a deterministic strategy in which an agent adopts a bijection between signals and reports, that is, F (or G for agent 2) is a permutation matrix.

Definition 2.4 (*Informed and uninformed strategies*). An *informed strategy* has $F_{ir} \neq F_{jr}$ for some $i \neq j$, some $r \in \{1, \ldots, n\}$ (and similarly for G for agent 2). An *uninformed strategy* has the same report distribution for all signals.

[8] Note that this differs from the standard sign operator, which has value -1 for negative inputs.

Permutation strategies are merely relabelings of the signals; in particular, truthfulness (denoted $\mathbb{I}$ below) is a permutation strategy. Note also that by definition, deterministic uniformed strategies are those that give the same report for all signals.

Each agent is assigned to two or more tasks, and the agents overlap on at least one task. Let $M_b \subseteq M$ denote a non-empty set of "bonus tasks", a subset of the tasks to which both agents are assigned. Let $M_1 \subseteq M \setminus M_b$ and $M_2 \subseteq M \setminus M_b$, with $M_1 \cap M_2 = \emptyset$ denote non-empty sets of tasks to which agents 1 and 2 are assigned, respectively. These will form the "penalty tasks." For example, if both agents are assigned to each of three tasks, A, B and C, then we could choose $M_b = \{A\}$, $M_1 = \{B\}$ and $M_2 = \{C\}$.

We assume that tasks are *a priori* identical, so that there is nothing to distinguish two tasks other than their signals. In particular, agents have no information about which tasks are shared, or which are designated bonus or penalty. This can be achieved by choosing M_b, M_1 and M_2 randomly after task assignment. This can also be motivated in largely anonymous settings, such as peer assessment and crowdsourcing.

A *multi-task peer-prediction mechanism* defines a total payment to each agent based on the reports made across all tasks. The mechanisms that we study assign a total payment to an agent based on the sum of payments for each bonus task, but where the payment for a bonus task is adjusted downwards by the consideration of its report on a penalty task and that of another agent on a different penalty task.

For the mechanisms we consider in this paper, it is without loss of generality for each agent to adopt a uniform strategy across each assigned task. Changing a strategy from task to task is equivalent in terms of expected payment to adopting a linear combination over these strategies, given that tasks are presented in a random order, and given that tasks are equivalent, conditioned on signal. We prove this in extended version of the paper [Shnayder et al. 2016a].

Given this uniformity, we write $E(F, G)$ to denote the expected payment to an agent for any bonus task. The expectation is taken with respect to both the signal distribution and any randomization in agent strategies. Let $\mathbb{I}$ denote the truthful reporting strategy, which corresponds to the identity matrix.

Definition 2.5 (*Strictly Proper*). A multi-task peer-prediction mechanism is *proper* if and only if truthful strategies form a correlated equilibrium, so that $E(\mathbb{I}, \mathbb{I}) \geq E(F, \mathbb{I})$, for all strategies $F \neq \mathbb{I}$, and similarly when reversing the roles of agents 1 and 2. For *strict properness*, the inequality must be strict.

This insists that the expected payment on a bonus task is (strictly) higher when reporting truthfully than when using any other strategy, given that the other agent is truthful.

Definition 2.6 (*Strongly-truthful*). A multi-task peer-prediction mechanism is *strongly-truthful* if and only if for all strategies F, G we have $E(\mathbb{I}, \mathbb{I}) \geq E(F, G)$, and equality may only occur when F and G are both the same permutation strategy.

In words, strong-truthfulness requires that both agents being truthful has strictly greater expected payment than any other strategy profile, unless both agents play the same permutation strategy, in which case equality is allowed.[9] From the definition, it follows that any strongly-truthful mechanism is strictly proper.

Definition 2.7 (*Informed-truthful*). A multi-task peer-prediction mechanism is *informed-truthful* if and only if for all strategies F, G, $E(\mathbb{I}, \mathbb{I}) \geq E(F, G)$, and equality may only occur when both F and G are informed strategies.

[9]Permutation strategies seem unlikely to be a practical concern, since permutation strategies require coordination and provide no benefit over being truthful.

In words, informed-truthfulness requires that the truthful strategy profile has strictly higher expected payment than any profile in which one or both agents play an uninformed strategy, and weakly greater expected payment than all other strategy profiles. It follows that any informed-truthful mechanism is proper.

Although weaker than strong-truthfulness, informed truthfulness is responsive to the primary, practical concern in peer-prediction applications: avoiding equilibria where agents achieve the same (or greater) payment as a truthful informed agent but without putting in the effort of forming a careful opinion about the task. For example, it would be undesirable for agents to be able to do just as well or better by reporting the same signal all the time. Once agents exert effort and observe a signal, it is reasonable to expect them to make truthful reports as long as this is an equilibrium and there is no other equilibrium with higher expected payment. Informed-truthful peer-prediction mechanisms provide this guarantee.[10]

3. MULTI-TASK PEER-PREDICTION MECHANISMS

We define a class of multi-task peer-prediction mechanisms that is parametrized by a *score matrix*, $S : \{1, \ldots, n\} \times \{1, \ldots, n\} \to \mathbb{R}$, that maps a pair of reports into a score, the same score for both agents. This class of mechanisms extends the binary-signal multi-task mechanism due to Dasgupta and Ghosh [2013] in a natural way.

Definition 3.1 (*Multi-task mechanisms*). These mechanisms are parametrized by score matrix S.

(1) Assign each agent to two or more tasks, with at least one task in common, and at least three tasks total.
(2) Let r_1^k denote the report received from agent 1 on task k (and similarly for agent 2). Designate one or more tasks assigned to both agents as bonus tasks (set M_b). Partition the remaining tasks into penalty tasks M_1 and M_2, where $|M_1| > 0$ and $|M_2| > 0$ and M_1 tasks have a report from agent 1 and M_2 a report from agent 2.
(3) For each bonus task $k \in M_b$, pick a random $\ell \in M_1$ and $\ell' \in M_2$. The payment to both agent 1 and agent 2 for task k is $S(r_1^k, r_2^k) - S(r_1^\ell, r_2^{\ell'})$.
(4) The total payment to an agent is the sum total payment across all bonus tasks.[11]

As discussed above, it is important that agents do not know which tasks will become bonus tasks and which become penalty tasks. The expected payment on a bonus task for strategies F, G is

$$E(F,G) = \sum_{i=1}^{n}\sum_{j=1}^{n} P(S_1{=}i, S_2{=}j) \sum_{r_1=1}^{n}\sum_{r_2=1}^{n} P(R_1{=}r_1|S_1{=}i)P(R_2{=}r_2|S_2{=}j)S(r_1,r_2)$$
$$- \sum_{i=1}^{n}\sum_{j=1}^{n} P(S_1{=}i)P(S_2{=}j) \sum_{r_1=1}^{n}\sum_{r_2=1}^{n} P(R_1{=}r_1|S_1{=}i)P(R_2{=}r_2|S_2{=}j)S(r_1,r_2)$$

[10]For simplicity of presentation, we do not model the cost of effort explicitly, but it is a straightforward extension to handle the cost of effort as suggested in previous work [Dasgupta and Ghosh 2013]. In our proposed mechanisms, an agent that does not exert effort receives an expected payment of zero, while the expected payment for agents that exert effort and play the truthful equilibrium is strictly positive. With knowledge of the maximum possible cost of effort, scaling the payments appropriately incentivizes effort.

[11]A variation with the same expected payoff and the same incentive analysis is to compute the expectation of the scores on all pairs of penalty tasks, rather than sampling. We adopt the simpler design for ease of exposition. This alternate design would reduce score variance if there are many non-bonus tasks, and may be preferable in practice.

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}S(r_1,r_2)F_{ir_1}G_{jr_2}. \tag{3}$$

The expected payment can also be written succinctly as $E(F,G) = \mathrm{tr}(F^\top \Delta G S^\top)$. In words, the expected payment on a bonus task is the sum, over all pairs of possible signals, of the product of the correlation (negative or positive) for the signal pair and the (expected) score given the signal pair and agent strategies.

For intuition, note that for the identity score matrix which pays \$1 in the case of matching reports and \$0 otherwise, agents are incentivized to give matching reports for signal pairs with positive correlation and non-matching reports for signals with negative correlation. Now consider a general score matrix S, and suppose that all agents always report 1. They always get $S(1,1)$ and the expected value $E(F,G)$ is a multiple of the sum of entries in the Δ matrix, which is exactly zero. Because individual rows and columns of Δ also sum to zero, this also holds whenever a single agent uses an uninformed strategy. In comparison, truthful behavior provides payment $E(\mathbb{I},\mathbb{I}) = \sum_{ij}\Delta_{ij}S(i,j)$, and will be positive if the score matrix is bigger where signals are positively correlated than where they are not.

While agent strategies in our model can be randomized, the linearity of the expected payments allows us to restrict our attention to deterministic strategies.

LEMMA 3.2. *For any world model and any score matrix S, there exists a deterministic, optimal joint strategy for a multi-task mechanism.*

The proof is in the extended version of the paper, and relies on solutions to convex maximization problems being extremal. Lemma 3.2 has several consequences:

— It is without loss of generality to focus on deterministic strategies when establishing strongly truthful or informed truthful properties of a mechanism.
— There is a deterministic, perhaps asymmetric equilibrium, because the optimal solution that maximizes $E(F,G)$ is also an equilibrium.
— It is without loss of generality to consider deterministic deviations when checking whether or not truthful play is an equilibrium.

We will henceforth assume deterministic strategies. By a slight abuse of notation, let $F_i \in \{1,\ldots,n\}$ and $G_j \in \{1,\ldots,n\}$ denote the reported signals by agent 1 for signal i and agent 2 for signal j, respectively. The expected score then simplifies to

$$E(F,G) = \sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}S(F_i,G_j). \tag{4}$$

We can think of deterministic strategies as mapping signal pairs to reported signal pairs. Strategy profile (F,G) picks out a report pair (and thus score) for each signal pair i,j with its corresponding Δ_{ij}. That is, strategies F and G map signals to reports, and the score matrix S maps reports to scores, so together they map signals to scores, and we then dot those scores with Δ.

4. THE DASGUPTA-GHOSH MECHANISM

We first study the natural extension of the Dasgupta and Ghosh [2013] mechanism from binary to multi-signals. This multi-task mechanism uses as the score matrix S the identity matrix ('1' for agreement, '0' for disagreement.)

Definition 4.1 (*The Multi-Signal Dasgupta-Ghosh mechanism (*MSDG*)*). This is a multi-task mechanism with score matrix $S(i,j) = 1$ if $i = j$, 0 otherwise.

Example 4.2. Suppose agent 1 is assigned to tasks $\{A, B\}$ and agent 2 to tasks $\{B, C, D\}$, so that $M_b = \{B\}, M_1 = \{A\}$ and $M_2 = \{C, D\}$. Now, if the reports on B are both 1, and the reports on A, C, and D were $0, 0$, and 1, respectively, the expected payment to each agent for bonus task B is $1 - (1 \cdot 0.5 + 0 \cdot 0.5) = 0.5$. In contrast, if both agents use an uninformed coordinating strategy and always report 1, the expected score for both is $1 - (1 \cdot 0.5 + 1 \cdot 0.5) = 0$.

The expected payment in the MSDG mechanism on a bonus task is

$$E(F, G) = \sum_{i,j} \Delta_{ij} \mathbb{1}_{[F_i = G_j]}, \tag{5}$$

where $\mathbb{1}_{x=y}$ is 1 if $x = y$, 0 otherwise. An equivalent expression is $\mathrm{tr}(F^\top \Delta G)$.

Definition 4.3 (*Categorical model*). A world model is *categorical* if, when an agent sees a signal, all other signals become less likely than their prior probability; i.e., $P(S_2 = j | S_1 = i) < P(S_2 = j)$, for all i, for all $j \neq i$ (and analogously for agent 2). This implies positive correlation for identical signals: $P(S_2 = i | S_1 = i) > P(S_2 = i)$.

Two equivalent definitions of categorical are that the Delta matrix has positive diagonal and negative off-diagonal elements, or that $\mathrm{Sign}(\Delta) = \mathbb{I}$.

THEOREM 4.4. *If the world is categorical, then the* MSDG *mechanism is strongly truthful and strictly proper. Conversely, if the Delta matrix Δ is symmetric and the world is not categorical, then the* MSDG *mechanism is not strongly truthful.*

PROOF. First, we show that truthfulness maximizes expected payment. We have $E(F, G) = \sum_{i,j} \Delta_{ij} \mathbb{1}_{[F_i = G_j]}$. The truthful strategy corresponds to the identity matrix $\mathbb{I}$, and results in a payment equal to the trace of Δ: $E(\mathbb{I}, \mathbb{I}) = \mathrm{tr}(\Delta) = \sum_i \Delta_{ii}$. By the categorical assumption, Δ has positive diagonal and negative off-diagonal elements, so this is the sum of all the positive elements of Δ. Because $\mathbb{1}_{[F_i = G_j]} \leq 1$, this is the maximum possible payment for any pair of strategies.

To show strong truthfulness, first consider an asymmetric joint strategy, with $F \neq G$. Then there exists i s.t. $F_i \neq G_i$, reducing the expected payment by at least $\Delta_{ii} > 0$. Now consider symmetric, non-permutation strategies $F = G$. Then there exist $i \neq j$ with $F_i = F_j$. The expected payment will then include $\Delta_{ij} < 0$. This shows that truthfulness and symmetric permutation strategies are the only optimal strategy profiles. Strict properness follows from strong truthfulness.

For the tightness of the categorical assumption, first consider a symmetric Δ with positive off-diagonal elements Δ_{ij} and Δ_{ji}. Then agents can benefit by both "merging" signals i and j. Let $\bar{F}$ be the strategy that is truthful on all signals other than j, and reports i when the signal is j. Then $E(\bar{F}, \bar{F}) = \Delta_{ij} + \Delta_{ji} + \mathrm{tr}(\Delta) > E(\mathbb{I}, \mathbb{I}) = \mathrm{tr}(\Delta)$, so MSDG is not strongly truthful. Now consider a Δ where one of the on-diagonal entries is negative, say $\Delta_{ii} < 0$. Then, because all rows and columns of Δ must add to 0, there must be a j such that $\Delta_{ij} > 0$, and this reduces to the previous case where "merging" i and j is useful. □

For binary signals ('1' and '2'), any positively correlated model, such that $\Delta_{1,1} > 0$ and $\Delta_{2,2} > 0$, is categorical, and thus we obtain a substantially simpler proof of the main result in Dasgupta and Ghosh [2013].

4.1. Discussion: Applicability of the MSDG mechanism

Which world models are categorical? One example is a noisy observation model, where each agent observes the "true" signal t with probability q greater than $1/n$, and otherwise makes a mistake uniformly at random, receiving any signal $s \neq t$ with probability

$(1-q)/(n-1)$. Such model makes sense for classification tasks in which the classes are fairly distinct. For example, we would expect a categorical model for a question such as "Does the animal in this photo swim, fly, or walk?"

On the other hand, a classification problem such as the ImageNet challenge [Russakovsky et al. 2015], with 1000 nuanced and often similar image labels, is unlikely to be categorical. For example, if "Ape" and "Monkey" are possible labels, one agent seeing "Ape" is likely to increase the probability that another says "Monkey", when compared to the prior for "Monkey" in a generic set of photos. The categorical property is also unlikely to hold when signals have a natural order, which we dub *ordinal* worlds.

Example 4.5. If two evaluators grade essays on a scale from one to five, when one decides that an essay should get a particular grade, e.g. one, this may increase the likelihood that their peer decides on that or an adjacent grade, e.g. one or two. In this case, the sign of the delta matrix would be

$$\mathrm{Sign}(\Delta) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{6}$$

Under the MSDG mechanism, evaluators increase their expected payoff by agreeing to always report one whenever they thought the score was either one or two, and doing a similar "merge" for other pairs of reports. We will return to this example below.

The categorical condition is a stronger requirement than previously proposed assumptions in the literature (see extended version of the paper). To see if it is a reasonable assumption in practice, we look at the correlation structure in a dataset from a large MOOC provider, focusing on 104 questions with over 100 submissions each, for a total of 325,523 assessments from 17 courses. Each assessment consists of a numerical score, which we examine, and an optional comment, which we do not study here. As an example, one assessment task for a writing assignment asks how well the student presented their ideas, with options "Not much of a style at all", "Communicative style", and "Strong, flowing writing style", and a paragraph of detailed explanation for each. These correspond to 0, 1, and 2 points on this rubric element.[12]

We estimate Δ matrices on each of the 104 questions from the assessments. We can think about each question as corresponding to a different signal distribution, and assessing a particular student's response to the question as an information task that is performed by several peers. The questions in our data set had five or fewer rubric options (signals), with three being most common (Figure 1L).

This analysis confirms that the categorical condition only holds for about one third of our three-signal models and for none of the larger models (Figure 1L). We also computed the average Δ matrix for each model size, as visualized in Figure 1R. The bands of positive correlation around the diagonal are typical of what we refer to as an ordinal rather than categorical domain.

5. HANDLING THE GENERAL CASE

In this section, we present a mechanism that is informed-truthful for general domains. We then discuss when it is strongly-truthful, give a version of it requiring no domain knowledge, and discuss other considerations.

[12]While we only see student reports, we take as an assumption that these reasonably approximate the true world model. As MOOCs develop along with valuable credentials based on their peer-assessed work, we believe it will nevertheless become increasingly important to provide explicit credit mechanisms for peer assessment.

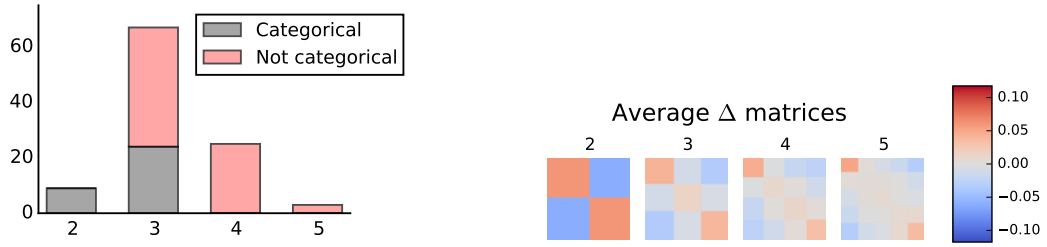

Fig. 1: Left: MOOC peer assessment is an ordinal domain, with most models with three or more signals not categorical. Right: Averaged Δ matrices, grouped by the number of signals in a domain. The positive diagonals show that users tend to agree on their assessments. For models of size 4 and 5, the ordinal nature of peer assessment is clear (e.g., an assessment of 2/5 is positively correlated with an assessment of 3/5).

5.1. The Correlated Agreement Mechanism

Based on the intuition given in Section 3, and the success of MSDG for categorical domains, it seems promising to base the construction of a mechanism on the correlation structure of the signals, and in particular, directly on Δ itself. This is precisely our approach. In fact, we will see that essentially the simplest possible mechanism following this prescription is informed-truthful for *all* domains.

Definition 5.1 (CA *mechanism*). The *Correlated Agreement (*CA*) mechanism* is a multi-task mechanism with score matrix $S = \mathrm{Sign}(\Delta)$.

THEOREM 5.2. *The* CA *mechanism is informed-truthful and proper for all worlds.*

PROOF. The truthful strategy F^*, G^* has higher payment than any other pair F, G:

$$E(F^*, G^*) = \sum_{i,j} \Delta_{i,j} S(i,j) = \sum_{i,j:\Delta_{ij}>0} \Delta_{i,j} \geq \sum_{i,j} \Delta_{i,j} S(F_i, G_j) = E(F, G),$$

where the inequality follows from the fact that $S(i,j) \in \{0,1\}$.

The truthful score is positive, while any uninformed strategy has score zero. Consider an uninformed strategy F, with $F_i = r$ for all i. Then, for any G,

$$E(F, G) = \sum_i \sum_j \Delta_{i,j} S(r, G_j) = \sum_j S(r, G_j) \sum_i \Delta_{i,j} = \sum_j S(r, G_j) \cdot 0 = 0,$$

where the next-to-last equality follows because rows and columns of Δ sum to zero. □

While informed-truthful, the CA mechanism is not always strictly proper. As discussed at the end of Section 2, we do not find this problematic; let us revisit this point. The peer prediction literature makes a distinction between proper and strictly proper, and insists on the latter. This comes from two motivations: (i) properness is trivial in standard models: one can simply pay the same amount all the time and this would be proper (since truthful reporting would be as good as anything else); and (ii) strict properness provides incentives to bother to acquire a useful signal or belief before making a report. Neither (i) nor (ii) is a critique of the CA mechanism; consider (i) paying a fixed amount does not give informed truthfulness, and (ii) the mechanism provides strict incentives to invest effort in acquiring a signal.

Example 5.3. Continuing with Example 4.5, we can see why CA is not manipulable. CA considers signals that are positively correlated on bonus tasks (and thus have a positive entry in Δ) to be matching, so there is no need to agents to misreport to ensure

Fig. 2: The blue and red nodes represent signals of agent 1 and 2, respectively. An edge between two signals represents that there is positive correlation between those signals. Left: A signal distribution for an image classification task with clustered signals. Right: A signal distribution for a MOOC peer assessment task or object counting task with ordinal signals and without clustered signals.

matching. In simple cases, e.g. if only the two signals 1 and 2 are positively correlated, they are "merged," and reports of one treated equivalently to the other. In cases such as Equation 6, the correlation structure is more complex, and the result is not simply merging.

5.2. Strong Truthfulness of the CA Mechanism

The CA mechanism is always informed truthful. In this section we characterize when it is also strongly truthful (and thus strictly proper), and show that it is maximal in this sense across a large class of mechanisms.

Definition 5.4 (*Clustered signals*). A signal distribution has *clustered signals* when there exist at least two identical rows or columns in $\mathrm{Sign}(\Delta)$.

Equivalently, two signals i and i' of an agent are clustered if i is positively correlated with the same set of matched agent's signals as i'.

Example 5.5. See Figure 2. The first example corresponds to an image classification task where there are categories such as "Monkey", "Ape", "Leopard", "Cheetah" etc. The signals "Monkey" and "Ape" are clustered: for each agent, seeing one is positively correlated with the other agent having one of the two, and negatively correlated with the other possible signals. The second example concerns models with ordinal signals, such as peer assessment or counting objects. In this example there are no clustered signals for either agent. For example, signal 1 is positively correlated with signals 1 and 2, while signal 2 with signals 1, 2, and 3.

LEMMA 5.6. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then a joint strategy where at least one agent uses a non-permutation strategy and matches the expected score of truthful reporting exists if and only if there are clustered signals.*

PROOF. Suppose clustered signals, so there exists $i \neq i'$ such that $\mathrm{Sign}(\Delta_{i,\cdot}) = \mathrm{Sign}(\Delta_{i',\cdot})$. Then if agent 2 is truthful, agent 1's expected score is the same for being truthful or for reporting i' whenever she receives either i or i'. Formally, consider the strategies $G = \mathbb{I}$ and F formed by replacing the i-th row in $\mathbb{I}$ by the i'-th row. Observe that $S(i,j) = S(F_i, G_j)$ as the i-th and i'-th row in S are identical. Hence, $E(F,G) = E(\mathbb{I},\mathbb{I})$. The same argument holds for clustered signals for agent 2.

If the world does not have clustered signals, any agent using a non-permutation strategy leads to lower expected score than being truthful. Suppose F is a non-permutation strategy, such that $E(F, G) = E(\mathbb{I}, \mathbb{I})$ for some G. Then there exist signals $i \neq i'$ such $F_i = F_{i'} = r$, for some r. No clustered signals implies that $\exists j$ such that $\text{Sign}(\Delta_{i,j}) \neq \text{Sign}(\Delta_{i',j})$. Let $G(j) = j'$, for some j'. Without loss of generality assume that $\Delta(i, j) > 0$, then we get $\Delta(i', j) < 0$ as $\Delta(i', j) \neq 0$. The score for signal pair $(S_1 = i, S_2 = j)$ is $S(r, j')$ and for $(S_1 = i', S_2 = j)$ is also $S(r, j')$. Either $S(r, j') = 1$ or $S(r, j') = 0$. In both cases the strategy profile F, G will lead to a strictly smaller expected score as compared to the score of truthful strategy, since $\Delta(i, j) > 0$ and $\Delta(i', j) < 0$. Similarly, we can show that if the second agent uses a non-permutation strategy, that also leads to strictly lower expected scores for both agents. □

We now give a condition under which there are asymmetric permutation strategy profiles that give the same expected score as truthful reporting.

Definition 5.7 (*Paired permutations*). A signal distribution has *paired permutations* if there exist distinct permutation matrices P, Q s.t. $P \cdot \text{Sign}(\Delta) = \text{Sign}(\Delta) \cdot Q$.

LEMMA 5.8. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then there exist asymmetric permutation strategy profiles with the same expected score under the* CA *mechanism as truthful reporting if and only if the signal distribution has paired permutations.*

Lemma 5.6 shows that when the world has clustered signals, the CA mechanism cannot differentiate between individual signals in a cluster, and is not strongly truthful. Similarly, Lemma 5.8 shows that under paired permutations this mechanism is not able to distinguish whether an agent is reporting the true signals or a particular permutation of the signals. In domains without clustered signals and paired permutations, all strategies (except symmetric permutations) lead to a strictly lesser score than truthful strategies, and hence, the CA mechanism is strongly truthful.

The CA mechanism is informed truthful, but not strongly truthful, for the image classification example in Figure 2 as there are clustered signals in the model. For the peer assessment example, it is strongly truthful because there are no clustered signals and a further analysis reveals that there are no paired permutations.

A natural question is whether we can do better by somehow 'separating' clustered signals from each other, and 'distinguishing' permuted signals from true signals, by giving different scores to different signal pairs, while retaining the property that the designer only needs to know $\text{Sign}(\Delta)$. Specifically, can we do better if we allow the score for each signal pair $(S_1 = i, S_2 = j)$ to depend on i, j in addition to $\text{Sign}(\Delta_{ij})$? We show that this extension does not add any additional power over the CA mechanism in terms of strong truthfulness.

THEOREM 5.9. *If $\Delta_{ij} \neq 0$, $\forall i, j$, then* CA *is maximally strong truthful amongst multi-task mechanisms that only use knowledge of the correlation structure of signals, i.e. mechanisms that decide $S(i, j)$ using $\text{Sign}(\Delta_{ij})$ and index (i, j).*

We use Lemmas 5.6 and 5.8 to argue that if a model has neither clustered signals nor paired permutations then CA is strongly truthful. To show maximality we prove that if a model has either clustered signals or paired permutations then there do not exist any strongly truthful multi-task mechanisms that only use knowledge of the correlation structure. The proof is included in the extended paper.

This result shows that if a multi-task mechanism only relies on the correlation structure and is strongly truthful in some world model then the CA mechanism will also be strongly truthful in that world model. Therefore, even if one uses $2 \cdot n^2$ parameters in the design of scoring matrices from $\text{Sign}(\Delta)$, one can only be strongly truthful in the worlds where CA mechanism is strongly truthful, which only uses 2 parameters.

A remaining question is whether strongly truthful mechanisms can be designed when the score matrix can depend on the exact value of the Δ matrix. We answer this question negatively; see the extended paper.

THEOREM 5.10. *There exist symmetric signal distributions such that no multi-task mechanism is strongly truthful.*

Figure 3 evaluates the sign structure of the Δ matrix for the 104 MOOC questions described earlier. The CA mechanism is strongly truthful up to paired permutations when signals are not clustered, and thus in roughly half of the worlds.

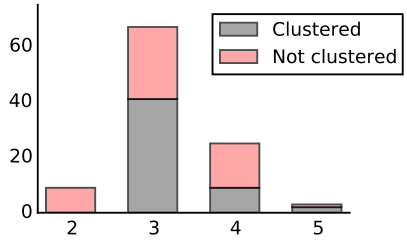

Fig. 3: Number of MOOC peer assessment models with clustered signals (CA is informed truthful) and without clustered signals (CA is strongly truthful up to paired permutations).

5.3. Detail-Free Implementation of the CA Mechanism

So far we have assumed that the CA mechanism has access to the sign structure of Δ. In practice, the signs may be unknown, or partially known (e.g. the designer may know or assume that the diagonal of Δ is positive, but be uncertain about other signs).

The CA mechanism can be made detail-free in a straightforward way by estimating correlation and thus the score matrix from reports; it remains informed truthful if the number of tasks is large (even allowing for the new concern that reports affect the estimation of the distribution and thus the choice of score matrix.)

Definition 5.11 (*The* CA *Detail-Free Mechanism (CA-DF)*). As usual, we state the mechanism for two agents for notational simplicity:

(1) Each agent completes m tasks, providing m pairs of reports.
(2) Randomly split the tasks into sets A and B of equal size.
(3) Let T^A, T^B be the empirical joint distributions of reports on the bonus tasks in A and B, with $T^A(i,j)$ the observed frequency of signals i,j. Also, let T^A_M, T^B_M be the empirical marginal distribution of reports computed on the penalty tasks in A and B, respectively, with $T^A_M(i)$ the observed frequency of signal i. Note that we only take one sample per task to ensure the independence of samples.
(4) Compute the empirical estimate of the Delta matrix, based on reports rather than signals: $\Gamma^A_{ij} = T^A(i,j) - T^A_M(i)T^A_M(j)$, and similarly for Γ^B.
(5) Define score matrices, *swapping task sets*: $S^A = \text{Sign}(\Gamma^B)$, $S^B = \text{Sign}(\Gamma^A)$. Note that S^A does not depend on the reports on tasks in A.
(6) Apply the CA mechanism separately to tasks in set A and set B, using score matrix S^A and S^B for tasks in set A and B, respectively.

LEMMA 5.12. *For all strategies F, G and all score matrices $S \in \{0,1\}^{n\times n}$, $E(S^*, \mathbb{I}, \mathbb{I}) \geq E(S, F, G)$ in the multi-task mechanism, where $E(S,F,G)$ is the expected score of the mechanism with a fixed score matrix S.*

PROOF. The expected score for arbitrary score matrix and strategies is:

$$E(S,F,G) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij} S(F_i, G_j)$$

The expected score for truthful reporting with S^* is

$$E(S^*, \mathbb{I}, \mathbb{I}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij} \operatorname{Sign}(\Delta)_{ij} = \sum_{i,j:\Delta_{ij}>0} \Delta_{ij} \geq \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij} S(F_i, G_j),$$

where the inequality follows because S is a 0/1 matrix. □

The lemma gives the main intuition for why CA-DF is informed truthful for large m: even if agents could set the score matrix completely independently of their strategies, the "truthful" score matrix S^* is the one that maximizes payoffs. To get a precise result, the following theorem shows that a score matrix "close" to S^* will be chosen with high enough probability. The proof is in the extended version of the paper.

THEOREM 5.13 (MECHANISM CA-DF IS (ϵ, δ)-INFORMED TRUTHFUL). *Let $\epsilon > 0$ and $\delta > 0$ be parameters. Then there exists a number of tasks $m = O(n^3 \log(1/\delta)/\epsilon^2)$ (for n signals), such that with probability at least $1 - \delta$, there is no strategy profile with expected score more than ϵ above truthful reporting, and any uninformed strategy has expected score strictly less than truthful. Formally, with probability at least $1 - \delta$, $E(F, G) \leq E(\mathbb{I}, \mathbb{I}) + \epsilon$, for all strategy pairs F, G; for any uninformed strategy F_0 (equivalently G_0), $E(F_0, G) < E(\mathbb{I}, \mathbb{I})$.*

5.4. Agent heterogeneity

The CA mechanism only uses the signs of the entries of Δ to compute scores, not the exact values. This means that the results can handle some variability across agent "sensing technology," as long as the sign structure of the Δ matrix is uniform across all pairwise matchings of peers. In the binary signal case, this reduces to agents having positive correlation between their signals, giving exactly the heterogeneity results in Dasgupta and Ghosh [2013]. Moreover, the agents themselves do not need to know the detailed signal model to know how to act; as long as they believe that the scoring mechanism is using the correct correlation structure, they can be confident in investing effort and simply report their signals truthfully.

5.5. Unintended Signals

Finally, we discuss a seemingly pervasive problem in peer prediction: in practice, tasks may have many distinctive attributes on which agents may base their reports, in addition to the intended signal, and yet all models in the literature assume away the possibility that agents can choose to acquire such unintended signals. For example, in online peer assessment where students are asked to evaluate the quality of student assignments, students could instead base their assessments on the length of an essay or the average number of syllables per word. In an image categorization system, users could base their reports on the color of the top-left pixel, or the number of kittens present (!), rather than on the features they are asked to evaluate. Alternative assessments can benefit agents in two ways: they may require less effort, and they may result in higher expected scores via more favorable Delta matrices.[13]

We can characterize when this kind of manipulation cannot be beneficial to agents in the CA mechanism. The idea is that the amount of correlation coupled with variability across tasks should be large enough for the intended signal. Let η represent a particular *task evaluation strategy*, which may involve acquiring different signals from the task than intended. Let Δ^η be the corresponding Δ matrix that would be designed if this was the signal distribution. This is defined on a domain of signals that may be distinct from that in the designed mechanism. In comparison, let η^* define the

[13] This issue is related to the perennial problem of spurious correlations in classification and regression.

task evaluation strategy intended by the designer (i.e., acquiring signals consistent with the mechanism's message space), coupled with truthful reporting. The expected payment from this behavior is $\sum_{ij:\Delta_{ij}^{\eta^*}>0} \Delta_{ij}^{\eta^*}$.

The maximal expected score for an alternate task evaluation strategy η may require a strategy remapping signal pairs in the signal space associated with η to signal pairs in the intended mechanism (e.g., if the signal space under η is different than that provided by the mechanism's message space). The expected payment is bounded above by $\sum_{ij:\Delta_{ij}^{\eta}>0} \Delta_{ij}^{\eta}$. Therefore, if the expected score for the intended η^* is higher than the maximum possible score for other η, there will be no reason to deviate.

6. CONCLUSION

We study the design of peer prediction mechanisms that leverage signal reports on multiple tasks to ensure informed truthfulness, where truthful reporting is the joint strategy with highest payoff across all joint strategies, and strictly higher payoff than all uninformed strategies (i.e., those that do not depend on signals or require effort). We introduce the CA mechanism, which is informed-truthful in general multi-signal domains. The mechanism reduces to the Dasgupta and Ghosh [2013] mechanism in binary domains, is strongly truthful in categorical domains, and maximally strongly truthful among a broad class of multi-task mechanisms. We also present a detail-free version of the mechanism that works without knowledge of the signal distribution while retaining ϵ-informed truthfulness. Interesting directions for future work include: (i) adopting a non-binary model of effort, and (ii) combining learning with models of agent heterogeneity.

REFERENCES

Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. 2015. Optimum Statistical Estimation with Strategic Data Sources. In *Proceedings of The 28th Conference on Learning Theory*. 280–296.

Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *WWW13*. 1–17.

Boi Faltings, Pearl Pu, and Bao Duy Tran. 2014. Incentives to Counter Bias in Human Computation. In *HCOMP 2014*. 59–66.

Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan P Adams. 2014. Trick or Treat : Putting Peer Prediction to the Test. In *EC'14*.

Xi Alice Gao, R. James Wright, and Kevin Leyton-Brown. 2016. Incentivizing Evaluation via Limited Access to Ground Truth : Peer Prediction Makes Things Worse. Unpublished, U. British Columbia. (2016).

Shaili Jain and David C Parkes. 2013. A Game-Theoretic Analysis of the ESP Game. *ACM Transactions on Economics and Computation* 1, 1 (2013), 3:1–3:35.

Radu Jurca and Boi Faltings. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *WINE05*, Vol. 3828 LNCS. 268–277.

Radu Jurca and Boi Faltings. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research* 34, 1 (2009), 209–253.

Radu Jurca and Boi Faltings. 2011. Incentives for Answering Hypothetical Questions. In *Workshop on Social Computing and User Generated Content, EC-11*.

Vijay Kamble, Nihar Shah, David Marn, Abhay Parekh, and Kannan Ramachandran. 2015. Truth Serums for Massively Crowdsourced Evaluation Tasks. (2015). http://arxiv.org/abs/1507.07045

Yuqing Kong and Grant Schoenebeck. 2016. A Framework For Designing Information Elicitation Mechanism That Rewards Truth-telling. (2016). http://arxiv.org/abs/

1605.01021
Yuqing Kong, Grant Schoenebeck, and Katrina Ligett. 2016. Putting Peer Prediction Under the Micro(economic)scope and Making Truth-telling Focal. *CoRR* abs/1603.07319 (2016). http://arxiv.org/abs/1603.07319
Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and self assessment in massive online classes. *ACM TOCHI* 20, 6 (Dec 2013), 1–31.
Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51 (2005), 1359–1373.
Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. *EDM* (2013).
Drazen Prelec. 2004. A Bayesian Truth Serum For Subjective Data. *Science* 306, 5695 (2004), 462.
Goran Radanovic and Boi Faltings. 2014. Incentives for Truthful Information Elicitation of Continuous Signals. In *AAAI'14*. 770–776.
Goran Radanovic and Boi Faltings. 2015a. Incentive Schemes for Participatory Sensing. In *AAMAS 2015*.
Goran Radanovic and Boi Faltings. 2015b. Incentives for Subjective Evaluations with Private Beliefs. *AAAI'15* (2015), 1014–1020.
Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for Effort in Crowdsourcing using the Peer Truth Serum. *ACM TIST* January (2016).
Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (April 2015), 1–42.
Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. 2016a. Informed Truthfulness in Multi-Task Peer Prediction. (2016). https://arxiv.org/abs/1603.03151
Victor Shnayder, Rafael Frongillo, and David C. Parkes. 2016b. Measuring Performance Of Peer Prediction Mechanisms Using Replicator Dynamics. *IJCAI-16* (2016).
Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *CHI'04*. ACM, New York, NY, USA, 319–326.
Bo Waggoner and Yiling Chen. 2014. Output Agreement Mechanisms and Common Knowledge. In *HCOMP'14*.
Jens Witkowski and David C Parkes. 2012. A Robust Bayesian Truth Serum for Small Populations. In *AAAI'12*.
Jens Witkowski and David C Parkes. 2013. Learning the Prior in Minimal Peer Prediction. In *EC'13*.
James R Wright, Chris Thornton, and Kevin Leyton-Brown. 2015. Mechanical TA : Partially Automated High-Stakes Peer Grading. In *SIGSCE'15*.
William Wu, Christos Tzamos, Constantinos Daskalakis, Matthew Weinberg, and Nicolaas Kaashoek. 2015. Game Theory Based Peer Grading Mechanisms For MOOCs. In *Learning@Scale 2015*.