Informed Truthfulness for Multi-Task Peer Prediction (short paper)

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

We study the problem of information elicitation without verification ("peer prediction") [Miller et al. 2005]. This problem arises across a diverse range of 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 (or any such external input is costly). 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. Peer-prediction mechanisms make payments to an 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:

— *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."
— *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"?).
— *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.

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.)

An extended version of this paper was published at EC'16. The full version with proofs is Shnayder et al. [2016].
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 Institute, 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

We are interested in *minimal* peer-prediction mechanisms, which require only signal reports from participants.[1] A basic desirable property of peer prediction mechanisms is that investing effort and making truthful reports of signals is an equilibrium (e.g., a correlated equilibrium) of the game induced by the mechanism. For many years, the 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.[2] Because of this, a concern in regard to peer prediction is that these mechanisms could make things worse— participants who would otherwise be truthful may now misreport in order to maximize 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. Their binary-signal, multi-task mechanism is *strongly truthful*, meaning that truthful reporting yields a higher expected payment than any other strategy profile (and is tied in payoff only with permutation strategies, i.e. $1 \to 2, 2 \to 1$ for binary signals).

In this paper, we introduce the new incentive property of *informed truthfulness*: no strategy profile, even one involving coordination between agents, 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). Although slightly weakened from strong-truthfulness, informed truthfulness is responsive to 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 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:

---

[1] 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 minimal schemes will be more likely to adopted in practice, with it cumbersome for people to report beliefs.

[2] 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 a minimal peer-prediction mechanism will always have an uninformative equilibrium with a higher payoff than truthful reporting.

— The *Correlated Agreement (CA) mechanism*, which is informed-truthful and "detail free" in the sense that the design requires knowledge of only the correlation structure of signals, but not the full signal distribution. That is, it requires knowledge of which pairs of signals are positively correlated and which negatively correlated.
— We characterize domains where CA is strongly truthful, and show that CA is maximally strong truthful amongst mechanisms in a larger family and under the same knowledge requirements and establish a general impossibility result.
— We show that the Dasgupta-Ghosh mechanism is strongly truthful in multi-signal domains that are *categorical*, where receiving one signal reduces an agent's belief that another agent will receive any other signal. We also show that peer assessment domains do not satisfy this property.
— For settings with a large number of tasks, we extend the CA mechanism to simultaneously estimate the signal correlation structure from reports while scoring agents. This mechanism introduces no new strategic concerns (and even though reports are now used to design the score matrix), 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 incentive properties hold for as few as two agents and three tasks, whereas these previous papers rely on receiving an asymptotically large number of reports. Our analysis framework also provides a dramatic simplification of the techniques used by Dasgupta and Ghosh [2013].

In a contemporaneous 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, and provides strong truthfulness in the limit of a large number of tasks.[3]

Our work provides the foundation for the analysis of a large family of multi-task peer prediction mechanisms, and paves the way for extensions such as supporting non-binary models of effort, tolerating agent heterogeneity, and considering non-random task assignment. As emphasized by Gao et al. [2016], the theoretical model of peer prediction assumes that only the intended signal can be acquired. If there is some other, low-cost, high-entropy and high-agreement signal available (e.g., the shade of the top-left pixel in an image), then agents can coordinate on this unintended signal and achieve higher payoff. We expect peer-prediction methods to be useful in domains where (i) the intended signal is informative, i.e. high entropy and high-agreement, (ii) the effort for the intended signal is relatively low, and (iii) the population is diverse so that coordination on some other signal is challenging. A theoretical direction is to explore whether latent "taste models" [Dawid and Skene 1979] of agents can be used to promote higher agreement between agents and thus higher payoff from the intended behavior.

[3]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).

REFERENCES

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

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 20–28.

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).

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.

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

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51 (2005), 1359–1373.

Drazen Prelec. 2004. A Bayesian Truth Serum For Subjective Data. *Science* 306, 5695 (2004), 462.

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).

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. 2016. Informed Truthfulness in Multi-Task Peer Prediction. (2016). https://arxiv.org/abs/1603.03151

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*.