# Surrogate Scoring Rule and a Dominant Truth Serum for Information Elicitation

YANG LIU, Harvard University
YILING CHEN, Harvard University

We present a method for information elicitation, where each agent truthfully reporting his information is a *dominant* strategy, even when there is no ground-truth verification. It is well known that truthful elicitation in dominant strategy can be achieved using proper scoring rules if we have access to the ground-truth (i.e. the outcome of the event) and truthful elicitation at a Bayesian Nash equilibrium can be achieved with various peer prediction mechanisms if ground-truth is not available. In this work, we first observe that if we have access to a random variable that is a noisy version of the ground-truth with known bias, we can design *surrogate scoring rules* to achieve truthful elicitation in dominant strategy. These surrogate scoring rules are inspired by the use of surrogate loss functions in machine learning and they remove bias from the noisy random variable such that in expectation a prediction is as if evaluated against the ground-truth. Built upon surrogate scoring rules, we develop a *dominant truth serum* (DTS) to achieve truthful elicitation in dominant strategy when agents' reports are the only source of information that we have access to. In DTS, reports from other agents are treated as noisy observations of the ground-truth and we develop a method to learn and then correct the biases in these reports without having access to the ground-truth. In expectation predictions in DTS are as if evaluated against the ground-truth.

## 1 INTRODUCTION

We are interested in eliciting private information from human agents with possibly subjective bias. For instance, we may be interested in collecting information about the following questions:

- (Q1) Will people land on Mars before 2030? (yes/no)
- (Q2) Will artificial intelligence dominate over human in all intellectual games by 2025? (yes/no)
- (Q3) Will the Nasdaq index go up tomorrow? (yes/no)
- (Q4) Will LeBron James leave Cavaliers next year? (yes/no)

The information we are interested in could be binary signals indicating whether agents believe the answer should be yes or no. Or it could be their predictions about how likely a yes event will happen. After the outcome reveals itself (at year 2030, 2025, tomorrow, next year respectively), the task left is to design a scoring function to score each agent's report using this realized outcome (ground-truth for the answer). Collectively there is a literature named strictly proper scoring rule [Brier, 1950, Gneiting and Raftery, 2007, Matheson and Winkler, 1976, Murphy and Winkler, 1970] designed exactly for this type of questions. To be more concrete, say we are interested in eliciting information about a binary event $y \in \{0,1\}$ from a set of human agents. Each of agents, say agent $i$, holds a noisy observation of $y$, in the form of $s_i \in \{0,1\}, i \in [N]$ (signal), or $p_i \in [0,1]$ the probability of event $\{y = 1\}$ happening under his belief (prediction). We unify the above cases and denote agent $i$'s information as $I_i$, and the information space as $I$. Above information structure is common knowledge among agents. In such a setting, a common way for incentivizing truthful report is to design a *strictly proper scoring function* $S : I \times \{0,1\} \to \mathbb{R}_+$ such that $\mathbb{E}[S(I_i, y)] > \mathbb{E}[S(\tilde{I}_i, y)]$, $\forall \tilde{I}_i \neq I_i$. [1]

It's worth to note the above elicitation game is only between us and each individual agent, and we call it a dominant strategy for each agent if truth-telling is indeed of their best interests. Strictly

[1] Classical strictly proper scoring rules were designed for eliciting predictions (probabilities), instead of a signal. But the ideas naturally extend to signal elicitation cases as shown in [Miller et al., 2005].

proper scoring rules have shown to be able to elicit good forecasts [Atanasov et al., 2015]. The question becomes much trickier when there is no access to the ground-truth $y$ (e.g., counterfactual questions), or the ground-truth can only arrive after extensive delays (e.g., Q1&2 we listed above), or it is too costly to obtain it (e.g., grading homework for a class with a large enrollment). The question we would like to answer in this paper is:

*Can we still achieve strict properness and dominant strategy elicitation even when the ground-truth is missing?*

We provide a positive answer to the setting where instead of the ground-truth signal, a noisy copy of it can be observed, and we know the noise level in this signal. Achieving this, we are able to extend the strictly proper scoring rules to the ones for the agnostic setting. We name such scoring rules as *surrogate scoring rules* (SSR). We show using surrogate scoring rules, we are able to achieve truthful elicitation in dominant strategy, without accessing the ground-truth. These surrogate scoring rules are inspired by the use of surrogate loss functions in machine learning and they remove bias from the noisy random variable such that in expectation a prediction is as if evaluated against the ground-truth [Angluin and Laird, 1988, Bylander, 1994, Natarajan et al., 2013, Scott, 2015, Scott et al., 2013]. This result immediately gives us an equilibrium implementation of truthful elicitation, if the error rates of other agents' reports are known. The idea looked similar to a recent work that proposes proper scoring rule [Witkowski et al., 2017] at the first sight. But we would like to note an unbiased proxy signal for the ground-truth is required to be known in above work, such that the existing strictly proper scoring rules can be applied directly – this is arguably hard to achieve in practice and makes the design task more challenging. Our design focuses on designing a surrogate scoring function that remove bias in any such noisy ground-truth, and thus can cope with a much more general setting.

We then further relax the assumption of knowing such a noisy ground-truth and its error rates, and move to the second setting, where any additional information can only be elicited from other peer agents. This question enjoys wider popularity and has been attempted continuously in the past decade. Particularly, peer prediction [Dasgupta and Ghosh, 2013, Jurca and Faltings, 2006, Miller et al., 2005, Prelec, 2004, Radanovic and Faltings, 2013, Witkowski and Parkes, 2012], a class of mechanisms that have been developed recently, are often adopted for incentivizing truthful or high-quality contributions from strategic sources when the quality of the contributions cannot be verified. In light of above discussions, we will also refer to this setting as the peer prediction setting.

Built upon surrogate scoring rules, we develop a multi-task mechanism, *dominant truth serum* (DTS), to achieve truthful elicitation in dominant strategy in the peer prediction setting. From a high level, in DTS, we exposit peer agents' information as "noisy ground-truth', instead of seeing them as "peer reports". We hope to first "learn" or reason about how much noise is contained in this noisy ground-truth ( as a result of other agents' reporting strategies, but remains unknown a-priori). Then we aim to apply our obtained results for surrogate scoring rule using this knowledge, such that we will enjoy the dominant strategy reasoning for this peer prediction setting.

The contributions of this work summarize as follows. (1) We extend the classical proper scoring rule setting for information elicitation to its *agnostic* setting, where only a noisy copy of the ground-truth is available. We name this type of scoring rules as *surrogate scoring rules*. This result complements the literature by providing a systematic way of quantifying the value of elicited information, even when ground-truth is missing. (2) We apply surrogate scoring rule to the peer prediction setting to obtain a dominant truth serum, with which it is always of human agents' best interest to truthfully report, regardless of other agents' actions. (3) Our results unify both signal elicitation with prediction elicitation, resolving one technical challenge in peer prediction for eliciting continuous signals. (4) Our work also establishes some interesting conceptual connection between

human bias and fact. We demonstrate that human bias is theoretically inferable, without access to ground-truth, and without assuming any parametric models of human biased data. This observation may be of independent interests for other studies. (5) It is worth to mention that our method requires *minimal* amount of information – we do not require agents to report anything, other than their private signal or prediction of the event.

Besides complementing the literature via extending the scoring rule results to a more robust and noisy setting, our results are of significant application merits. Peer prediction has seen very little empirical success, see e.g. [Gao et al., 2014], in spite of their thriving theoretical results. One conjecture of the failure in practice is that existing approaches often establish their truthfulness based on an equilibrium notion, that is it is agent's best interest to truthfully reveal their answer, if all other agents are truthfully reporting, as typically adopted in the game theory context. This equilibrium notion exposes such mechanisms very vulnerable in practice, for several good reasons. First hardly all other agents would be behaving rationally and be playing the equilibrium strategy, and hardly agents trust the others would do so. Secondly, due to the existence of multiple equilibria besides the "good equilibrium" where agents truthfully report, agents face the equilibrium selection issue [Banks and Sobel, 1987, Harsanyi et al., 1988]. Without coordination, this will lead to a chaotic selection of actions. These issues have become more concerned recently in presence of malicious manipulations when applying these elicitation mechanisms to situations where parties of agents have incentives to mislead others' opinion. Consider a political opinion elicitation problem. Due to manipulations of adversarial information (for either party's interests), agents' incentives for telling the truth will severely degrade, and surely such information sources are not "rational". Thus we see a great need in pushing the solution concept to a more robust regime.

In addition, often it is the case that peer prediction scoring functions measure the correlations between reported information from agents. These scores do not necessarily reflect the quality of agents' reports; while strictly proper scores often do, as agents' reports are calibrated against the ground-truth. This raises concern when we would like to better interpret the scores to agents in the peer prediction setting. Both SSR and DTS address the question of how to quantify the quality of elicited information when ground-truth is missing, as effectively each agent's reported elicitation is calibrated against the ground-truth in expectation.

### 1.1 Related works

The most relevant literature to our paper is *strictly proper scoring rule* and *peer prediction*. Scoring rules were developed for eliciting truthful prediction (probability) [Savage, 1971]. The pioneer works include [Brier, 1950], where Brier scoring rule was proposed to verify the qualities of forecasts. Competitive scoring rule is proposed in [Kilgour and Gerchak, 2004] that each agent is scored according to their relative forecasting accuracy to the average population. Closed-form characterizations are given for strictly proper scoring rules in [Gneiting and Raftery, 2007]. Motivated by the complexity of reporting the probability for events with large outcome space, property elicitation, such as mean, variance etc of the forecast distributions, has been studied more recently [Frongillo and Kash, 2015, Lambert et al., 2008]. Though in this paper we will not cover the property elicitation setting, our results are ready to extend to cover this case. We leave this for future discussions.

The core idea of peer prediction is to score each agent based on another reference report elicited from the rest of agents, and to leverage on the stochastic correlation between different agents' information. This line of research started with the celebrated *Bayesian Truth Serum* work [Prelec, 2004], where a surprisingly popular answer methodology is shown to be able to incentivize agents to truthfully report even they believe they hold the minority answer (but more likely to be true in their own opinion). The seminal work [Miller et al., 2005] established that strictly proper scoring

rule [Gneiting and Raftery, 2007] can be adopted in the peer prediction setting for eliciting truthful reports (but the mechanism designer need to know details of agents' model); following which a sequence of followed up works have been done to relax the assumptions that have been imposed therein [Radanovic and Faltings, 2013, Witkowski and Parkes, 2012]. More recently, [Dasgupta and Ghosh, 2013, Witkowski et al., 2013] formally introduced and studied an effort sensitive model for binary signal data elicitation. Particularly [Dasgupta and Ghosh, 2013] proposed a multi-task peer prediction mechanism that can help remove undesirable equilibria that lead to low quality reports. These results are further strengthened and extended to a non-binary signal setting in [Shnayder et al., 2016].

It is worth to mention that our work borrows ideas from the machine learning literature on learning with noisy data [Natarajan et al., 2013, Scott, 2015]. From a high level's perspective, our goal in this paper aligns with the goal in learning with noisy data – both aim to evaluate a prediction when the ground-truth is missing, but instead a noisy signal of the ground-truth is available.

## REFERENCES


Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

Pavel D Atanasov, Phillip Rescober, Eric Stone, Samuel A Swift, Emile Servan-Schreiber, Philip E Tetlock, Lyle Ungar, and Barbara Mellers. 2015. Distilling the wisdom of crowds: Prediction markets versus prediction polls. (2015).

Jeffrey S Banks and Joel Sobel. 1987. Equilibrium selection in signaling games. *Econometrica: Journal of the Econometric Society* (1987), 647–661.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthey Weather Review* 78, 1 (1950), 1–3.

Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*. ACM, 340–347.

Anirban Dasgupta and Arpita Ghosh. 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.

Rafael Frongillo and Ian A Kash. 2015. Vector-valued property elicitation. In *Conference on Learning Theory*. 710–727.

Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan Prescott Adams. 2014. Trick or treat: putting peer prediction to the test. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 507–524.

Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

John C Harsanyi, Reinhard Selten, and others. 1988. A general theory of equilibrium selection in games. *MIT Press Books* 1 (1988).

R. Jurca and B. Faltings. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce (EC '06)*. ACM, 190–199.

D Marc Kilgour and Yigal Gerchak. 2004. Elicitation of probabilities using competitive scoring rules. *Decision Analysis* 1, 2 (2004), 108–113.

N.S. Lambert, D.M. Pennock, and Y. Shoham. 2008. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC '08)*. ACM, 129–138.

James E Matheson and Robert L Winkler. 1976. Scoring rules for continuous probability distributions. *Management science* 22, 10 (1976), 1087–1096.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359 –1373.

Allan H Murphy and Robert L Winkler. 1970. Scoring rules in probability assessment and evaluation. *Acta psychologica* 34 (1970), 273–286.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*. 1196–1204.

Dražen Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695 (2004), 462–466.

G. Radanovic and B. Faltings. 2013. A Robust Bayesian Truth Serum for Non-Binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI '13)*.

Leonard J Savage. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

Clayton Scott. 2015. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels.. In *AISTATS*.

Clayton Scott, Gilles Blanchard, Gregory Handy, Sara Pozzi, and Marek Flaska. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising.. In *COLT*. 489–511.

V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. 2016. Informed Truthfulness in Multi-Task Peer Prediction. *ACM EC* (March 2016). arXiv:cs.GT/1603.03151

Jens Witkowski, Pavel Atanasov, Lyle H Ungar, and Andreas Krause. 2017. Proper proxy scoring rules.

Jens Witkowski, Yoram Bachrach, Peter Key, and David C. Parkes. 2013. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*.

J. Witkowski and D. Parkes. 2012. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI '12)*.