



Information Elicitation and Aggregation: Theory, Behavior, and Application

Citation

Wang, Juntao. 2022. Information Elicitation and Aggregation: Theory, Behavior, and Application. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37372158>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences




DISSERTATION ACCEPTANCE CERTIFICATE


The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

“Information Elicitation and Aggregation: Theory, Behavior, and Application”

presented by: Juntao Wang

Signature 
Typed name: Professor Y. Chen

Signature 
Typed name: Professor D. Parkes

Signature 
Typed name: Professor Y. Liu

April 26, 2022

Information Elicitation and Aggregation: Theory, Behavior, and Application

A dissertation presented

by

Juntao Wang

to

John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

April 2022

© 2022 Juntao Wang
All rights reserved.

Dissertation Advisor:
Professor Yiling Chen

Author:
Juntao Wang

Information Elicitation and Aggregation: Theory, Behavior, and Application

Abstract

Recent decades have witnessed a broadening application of the wisdom of crowds, ranging from forecasting geopolitical events to estimating business variables to predicting the replicability of social science studies. In these applications, information is elicited from a crowd of human participants and then aggregated to generate final judgments or informed decisions. Two central problems in this process are the information elicitation problem, i.e., how we can elicit authentic and high-quality information from selfish information holders, and the information aggregation problem, i.e., how we can aggregate the noisy or even biased information that we collect into more accurate judgments or decisions.

The challenges in solving the above two problems vary with the concrete application scenarios. At a high level, these scenarios can be divided into two categories, the with-verification setting and the without-verification setting. In the with-verification setting, the principal can access (historical) ground truth information to verify the information quality of each participant and design elicitation and aggregation mechanisms accordingly. In contrast, the principal cannot access such ground truth information in the without-verification setting. In this thesis, I explore and make progress on information elicitation and aggregation problems under several specific scenarios in both settings.

In the with-verification setting, I study the betting scenario and propose the randomized wagering mechanisms for prediction elicitation. These mechanisms overcome an impossibility result for deterministic mechanisms that four desirable properties cannot be satisfied simultaneously. In the without-verification setting, I study both the probabilistic prediction elicitation and aggregation problems. I first extend the strictly proper scoring rules, the most prevalent information elicitation solution in the with-verification setting, into the without-verification setting and derive the surrogate scoring rules. These rules not only provide strong incentives for participants to report true beliefs but also characterize their prediction accuracy. I

further develop a forecast aggregation framework using the elicitation without-verification schemes such as the surrogate scoring rules to improve the aggregation accuracy consistently. This improvement is examined and verified on a diverse set of real-world human forecasting datasets.

Human behavior, involving how people understand elicitation questions, generate beliefs, and react to elicitation schemes, is overlooked in the literature on information elicitation and aggregation. This thesis also explores human behavior and its influence within this domain. In particular, I develop auction mechanisms to elicit truthful reporting of private signals when human players have bounded rationality in interdependent valuation auctions. I also conduct real-world elicitation and aggregation experiments that use laypeople to predict the replicability of social science studies. Several interesting experimental findings of laypeople's behaviors are analyzed and discussed.

Contents

Abstract	iii
Acknowledgments	ix
List of Tables	xi
List of Figures	xiii
List of Algorithms	xv
Citation to Previously Published Work	xvi
Introduction	1
1 Randomized Wagering Mechanisms	9
1.1 Introduction	9
1.2 Related works	11
1.3 Preliminaries	12
1.3.1 Strictly proper scoring rules and weighted score wagering mechanisms .	13
1.4 Randomized wagering mechanisms	14
1.4.1 Desirable properties	15
1.5 Lottery wagering mechanisms	18
1.6 Surrogate wagering mechanisms	19
1.6.1 Generic surrogate wagering mechanisms	19
1.6.2 SWM with error rate selection (SWME) and random partition SWME (RP-SWME)	21
1.6.3 Properties of SWME and RP-SWME	23
1.6.4 Wager with noisy ground truth	26
1.7 Extensions of SWM	27
1.7.1 Surrogate NAWM	27
1.7.2 Multi-outcome events	28
1.8 Evaluation	29
1.8.1 Simulation Setup	29
1.8.2 Comparison of efficiency of wagering mechanisms	30
1.8.3 Comparison of randomness properties of RP-SWME and LWS	32
1.9 Conclusion	34

2	Surrogate Scoring Rules	35
2.1	Introduction	35
2.2	Related work	37
2.3	Preliminaries	40
2.4	Model and mechanism design problem	41
2.4.1	Model of Information Structure	41
2.4.2	Mechanism design problem	44
2.5	Elicitation with a noisy estimate of ground truth	46
2.5.1	Surrogate scoring rules (SSR)	47
2.6	Elicitation without verification	50
2.6.1	Reference report and its property	51
2.6.2	Asymptotic setting	52
2.6.3	Finite sample analysis	57
2.7	Generalizations to multi-outcome tasks	60
2.7.1	Generalization of SSR	61
2.7.2	Generalization of SSR mechanisms	62
2.8	Empirical studies	63
2.8.1	Setting	63
2.8.2	Main results	67
2.9	Discussion	70
3	Forecast Aggregation via Peer Prediction	73
3.1	Introduction	73
3.2	Related Work	75
3.3	Setting	76
3.4	Aggregation Using PAS	77
3.5	Peer Prediction Methods for PAS	79
3.5.1	Mechanisms recovering the strictly proper scoring rules (SPSR)	80
3.5.2	Mechanisms rewarding the correlation	81
3.5.3	Peer prediction rewards and accuracy of agents	82
3.6	Empirical Studies	83
3.6.1	Experiment setup	83
3.6.2	Smaller but smarter crowd	86
3.6.3	Forecast aggregation performance on binary events	87
3.6.4	Forecast aggregation performance on multi-outcome events	91
3.7	Discussion and Future Directions	92

4	Cursed yet Satisfied Agents	95
4.1	Introduction	95
4.2	Related work	99
4.3	Model	101
4.3.1	The Winner’s Curse—A Behavioral Model	103
4.3.2	Incentive Properties for Cursed Agents and Other Desirable Properties	105
4.4	Preliminaries	109
4.4.1	C-EPIC-IR Mechanisms and Virtual Valuations	109
4.4.2	Deterministic C-EPIC-IR Mechanisms	110
4.5	Implications of Ex-post IR	111
4.6	Revenue Maximization	114
4.7	Welfare Maximization	117
4.7.1	Welfare Optimal Mechanism is not Budget Balanced	117
4.7.2	Masked Mechanisms	118
4.7.3	Ex-post Budget-Balance Implies No Positive Transfers	119
4.7.4	Optimal Mechanism	122
4.8	Conclusion and Future Directions	125
5	Can Laypeople Predict the Replicability of Social Science Studies without Expert Intervention: an Exploratory Study	127
5.1	Introduction	127
5.2	Methods	129
5.2.1	Materials	129
5.2.2	Procedure and incentive	131
5.2.3	Participants	132
5.3	Results	132
5.3.1	Participants’ Engagement	132
5.3.2	Participants’ perceptions about studies	134
5.3.3	Forecasting replicability	134
5.3.4	Forecasting using machine learning	136
5.4	Discussion	137
6	Conclusion	140
	References	143
	Appendix A Appendix to Chapter 2	166
A.1	Missing Figures	166
A.2	Missing Proofs	167
A.2.1	Proof of Lemma 2.2	167

A.2.2	Proof of Lemma 2.12	167
Appendix B	Appendix to Chapter 3	172
B.1	Forecast aggregation performance on small datasets	172
B.2	Missing tables	174
B.3	More details about the datasets	176
B.4	Missing Proofs	177
B.4.1	Proof of Theorem 3.2	177
B.5	Variational inference for crowdsourcing	179
Appendix C	Appendix to Chapter 4	182
C.1	Cursed Equilibrium in the Wallet-Game	182
C.2	Missing proofs	182
C.2.1	Proof of Proposition 4.3	182
C.2.2	Proof of Lemma 4.5	183
C.2.3	Proof of Proposition 4.21	184
C.2.4	Proof of Corollary 4.24	184
C.2.5	Proof of Lemma 4.26	185
C.2.6	Derivation of Equation 4.4	185

Acknowledgments

I would not have been able to write this thesis without the constant support and guidance of my amazing advisor, Yiing Chen. It is extraordinary how much Yiling has influenced my research interests while still allowing me to explore new research areas. I cannot thank you more for the patience, tolerance, and encouragement you offered during the toughest period of my PhD journey, and I will be forever grateful for your support.

I would also give my special thanks to Yang Liu and Alon Eden, who have been amazing friends and collaborators, broadened my research vision, and offered me concrete help and guidance in both research projects and career development. I enjoyed our collaboration immensely and hope to collaborate more in the future. In particular, I would like to thank Yang for hosting my 2019 summer visit to UCSC. I must thank David C. Parkes for timely help whenever needed and for sending information about wonderful seminars and research opportunities that I could not know on my own.

I was fortunately able to work with many excellent researchers when pursuing my PhD, Anna Dreber Almenberg, Thomas Pfeiffer, Michael Gordon, Michael Bishop, Charles Twardy, Debmalya Mandal, Goran Radanovic, Haifeng Xu, and Xiang Yan. I got tremendous inspiration and encouragement during our discussions. I am delighted and super grateful for having a warmful community of graduate students at the EconCS group, particularly, Chara, Debmalya, Hongyao, Shuran, Sophie, and Zhe. I really miss the life when we hung out together in person. Especially, I would like to thank Shuran for initiating our research collaboration and discussing our PhD progress together and thank Zhe for sharing his thesis latex template inherited from Debmalya with me to make my life much easier. I must also thank my friends, Yuantian Deng and Zudi Lin at SEAS Harvard, who made my academic life at Harvard much easier and more joyful.

I am immensely thankful to my undergraduate supervisor Xiaotie Deng, who brought me into the field of Game Theory and Mechanism Design and helped me develop my research career. I would also like to thank Shanghai Jiao Tong University and Harvard University for offering me an unparalleled environment to thrive and meet fantastic guys.

Finally, I would like to give my great thanks to my family. I cannot imagine I could finish a PhD without my wife Wenwen Li's support and help. I cannot be more fortunate to meet you at Shanghai Jiao Tong University and marry you at Harvard during my PhD journey. I also would like to thank my wife's parents, Fajun Li and Jun Ai, for their selfless support and belief in me. And everything will be impossible if I cannot meet you, my parents, Hongzhuan Wang and Meirong Yan.

List of Tables

1.1	A summary of properties of wagering mechanisms	17
2.1	Statistics about binary-outcome datasets from GJP, HFC and MIT datasets	64
3.1	The main abbreviations and the corresponding full names used in this chapter .	78
3.2	Statistics about the binary event datasets from GJP, HFC and MIT datasets	83
3.3	Statistics about the multiple-outcome event datasets from GJP and HFC datasets	84
3.4	The mean Brier scores (range [0, 2], the lower the better) of different aggregators on binary events of 14 datasets. The best mean Brier score among benchmarks on each dataset is marked by bold font. The mean Brier scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in green ; those outperforming the second best of benchmarks are highlighted in yellow ; the worst mean Brier scores over all aggregators on each dataset are highlighted in red	88
3.5	The two-sided paired <i>t</i> -test for the mean Brier scores and the mean log scores of each pair of a PAS-aided aggregator and a benchmark on binary events of 14 datasets. The first integer in each cell represents the number of datasets where the PAS-aided aggregator achieves significantly smaller mean score (with $p\text{-value} < 0.05$), while the second integer in each cell indicates the number of datasets where the benchmark achieves significantly smaller mean score. The cells where the # of outperforms exceeds the # of underperforms by at least 4 are highlighted in green	89
3.6	The mean Brier score and the mean log score of different aggregators on multi-outcome events of 6 datasets. The best mean score among benchmarks on each dataset is marked by bold font. The mean scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in green ; those outperforming the second best of benchmarks are highlighted in yellow ; the worst mean scores over all aggregators on each dataset are highlighted in red .	92

3.7	The two-sided paired <i>t</i> -test for mean Brier score and mean log score of each pair of a PAS-aided aggregator and a benchmark on multi-outcome events of 6 datasets. The first integer in each cell represents the number of datasets where the PAS-aided aggregator achieves the significantly smaller mean score (with $p\text{-value}<0.05$), while the second integer in each cell indicates the number of datasets where the benchmark achieves the significantly smaller mean score.	93
5.1	Statistics of the responses with different accessibility ratings. Spearman’s rank correlation coefficient ρ is calculated between participants’ replication predictions and actual replication outcomes.	135
5.2	Mean accuracy scores and AUC-ROCs on the training set and validation set when the mean replication prediction (Q8), mean accessibility (Q3) and mean surprisingness (Q4) and all of them are used as features respectively. Brackets show the 95% confidence intervals of the corresponding values.	137
B.1	The mean Brier scores (range [0, 2], the lower the better) of different aggregators on randomly sampled sub-datasets of 4 GJP datasets and 7 MIT datasets. The best mean Brier score among benchmarks on each dataset is marked by bold font. The mean Brier scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in green ; those outperforming the second best of benchmarks are highlighted in yellow ; the worst mean Brier scores over all aggregators on each dataset are highlighted in red	173
B.2	The mean and the standard deviation of the mean Brier scores and the mean log scores of the 10 PAS-aided aggregators and the benchmarks over 14 datasets. The bold font means that the data is significantly better than the counterparts of all benchmarks with $p\text{-value}<0.05$	174
B.3	The mean Brier scores of three statistical-inference-based aggregators on MIT datasets reported by McCoy and Prelec [MP17]. The Brier score has been re-scaled to the range [0,2] to align with ours. The bold font marks the five cases where theirs outperform the worst of our five mean-based PAS aggregators.	174
B.4	The mean log scores (the lower the better) of different aggregators on binary events of 14 datasets. The best mean score among benchmarks on each dataset is marked by bold font. The mean scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in green ; those outperforming the second best of benchmarks are highlighted in yellow ; the worst mean scores over all aggregators on each dataset are highlighted in red .	175

List of Figures

1.1	Average individual risk of each of five wagering mechanisms as a function of N under different prediction and wager models	31
1.2	Average individual risk of each of four mechanisms under events with multiple outcomes	32
1.3	Average money exchange rate of each of five wagering mechanisms as a function of N under different prediction and wager models	33
1.4	Average money exchange rate of each of four mechanisms under events with multiple outcomes	34
1.5	Std. variance of net-payoff as a function of prediction accuracy: RP-SWME v.s. LWS	34
1.6	Probability of winning money as a function of prediction accuracy: RP-SWME v.s. LWS	34
2.1	Regression of individuals' true accuracy and SSR score over 14 datasets under three different SPSR.	68
2.2	The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSR.	68
2.3	The number of datasets in each level of correlation (measured by Spearman's correlation coefficient) between individuals' peer prediction scores and different SPSR.	69
2.4	The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSR on sampled datasets (the correlation is averaged over 100 runs of random sampling).	69
2.5	The portion of top $t\%$ forecasters w.r.t. 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ forecasters selected by different methods (averaged over 14 datasets).	71
2.6	The portion of bottom 50% forecasters w.r.t. 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ users selected by different methods (averaged over 14 datasets).	71

3.1	The averages of the true mean Brier score of top forecasters selected by the five PAS and by the true Brier score.	86
3.2	The portions of overlapped agents, who are simultaneously selected by all of the five PAS and the true score.	86
3.3	The Brier score of the five mean-based PAS-aided aggregators with a varying number of selected top agents on dataset G2.	86
3.4	The mean and the standard deviation of the aggregation accuracy of the 10 PAS-aided aggregators (DMI/CA/PTS/SSR/PSR-aided \times Mean/Logit-based aggregators) and the benchmarks over 14 datasets.	90
A.1	The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSR when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.	166
A.2	The number of datasets in each level of correlation (measured by Spearman's correlation coefficient) between individuals' peer prediction scores and different SPSR when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.	166

List of Algorithms

1	Lottery Wagering Mechanisms	18
2	Surrogate Wagering Mechanisms	20
3	Error Rate Selection Algorithm	22
4	Random Partition SWME (RP-SWME)	23
5	SSR mechanisms (Sketch)	50
6	e_z^+, e_z^- solver	53
7	SSR mechanisms	56
8	PAS-aided aggregators	79

Citation to Previously Published Work

1. The work presented in Chapter 1 is based on the following publication:

Yiling Chen, Yang Liu, and Juntao Wang. "Randomized wagering mechanisms." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 1845-1852. 2019.

2. The work presented in Chapter 2 is based on the following publication:

Yang Liu, Juntao Wang, and Yiling Chen. "Surrogate scoring rules." In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 853-871. 2020.

3. The work presented in Chapter 3 is based on the following paper:

Juntao Wang, Yang Liu, and Yiling Chen. "Forecast aggregation via peer prediction." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, pp. 131-142. 2021.

4. The work presented in Chapter 4 is based on the following publication with full version posted on ArXiv:

Yiling Chen, Alon Eden, and Juntao Wang. "Cursed yet Satisfied Agents". In *13th Innovations in Theoretical Computer Science Conference (ITCS)*, Vol. 215, pp. 44:1–44:1. 2022.

Introduction

About a century ago, statistician Galton [Gal07] documented a very interesting phenomenon: in a game of guessing the weight of a displayed ox, the mean guess of a crowd of 800 participants matched exactly the 1197 pounds of the actual weight of the ox. This superior power of the collective intelligence in making judgments is now referred to as the wisdom of crowds [Sur05]. Nowadays, with the development of Internet technology, we can access the opinions of a large number of crowds much more easily. As a result, the wisdom of crowds has been applied to a wide array of domains to predict variables of interest and assist decision-making, such as forecasting geopolitical events [Tet+14; Fri+18], estimating business variables [CW86; LBD16], and even predicting the replicability of social science studies [Alt+19; HSW19]. However, there are two main challenges in successfully leveraging the wisdom of crowds, referred to as the information elicitation problem and the information aggregation problem:

1. **Information elicitation problem:** How can we incentivize participants to provide authentic and high-quality information?
2. **Information aggregation problem:** How can we aggregate the information provided by participants to make better judgments/decisions?

To illustrate, a typical scenario of information elicitation and aggregation is where a principal has a random variable of interest, e.g., whether it will rain tomorrow or whether the S&P 500 index will go up next month. The principal has access to a pool of participants who may hold relevant information about the variable of interest. In order to obtain this information, the principal can invite participants to provide their opinions, such as a probabilistic prediction, and in exchange, she offers the participants financial rewards as incentives. The informa-

tion elicitation problem is to design such reward mechanisms. The main objective of these mechanisms is to incentivize truthful reporting, i.e., each participant receives a strictly higher expected reward when reporting their true belief than manipulating their report. The principal then forms a synthetic final judgment about that random variable of interest based on the solicited information and takes informed actions based on the final judgment, e.g., whether to bring an umbrella or buy a stock. The information aggregation problem is to design algorithms to generate this synthetic final judgment, with an objective of maximizing the quality/accuracy of the judgment.

The broadening application of information elicitation and aggregation brings emerging challenges in achieving the above two objectives of elicitation and aggregation, especially with varied needs and constraints of different scenarios. Generally, these scenarios can be categorized into two folds: the *with-verification* setting and the *without-verification* setting, depending on whether the principal has access to certain ground truth information about the random variable of interest. Together with the distinction between elicitation and aggregation, we can partition the information elicitation and aggregation research into four sub-domains: $\{\text{Elicitation, Aggregation}\} \times \{\text{With verification, Without verification}\}$. Each domain differs in concrete objectives, challenges, and constraints and requires different technologies to tackle these problems. Orthogonal to these four sub-domains, human behavior is an overlooked sub-domain in information elicitation and aggregation. It considers problems such as how humans form their opinions, how they react to the elicitation methods, and how their behaviors influence the elicitation and the aggregation process. These factors are critical to the success of an information elicitation and aggregation system. In this thesis, I use a combination of theoretical and empirical approaches to address specific problems with each of the above sub-domains, except the domain of information aggregation with verification. I introduce the problems and research status of each domain, followed by the contributions of this thesis in the following.

Information elicitation with verification. Information elicitation with verification considers the scenarios where the principal can reward participants after knowing the ground truth of

the random variable of interest, such as in the geopolitical forecast tournaments. In this setting, the truthfulness incentive can be achieved using the strictly proper scoring rules [Win69; Sav71; JNW06; GR07b], a family of functions that take the reported information and the ground truth as inputs, and output a score whose expectation is strictly maximized when the report matches the true distribution of the random variable of interest. Meanwhile, In this setting, the truthfulness incentive also aligns with the accuracy of the provided information. The instances of strictly proper scoring rules, such as Brier score and log scoring rules, are also widely adopted as accuracy metrics of forecasts.

Recent works in this sub-domain focus on achieving specific properties in addition to truthfulness. For example, one potential issue of strictly proper scoring rules is that they require the principal to pay for the information while the payment grows linearly with the number of participants. Prediction markets [WZ04; Arr+08; CP10] and wagering mechanisms [Lam+08; Lam+15; Che+14; FPW17] are proposed to mitigate the budget problem. In prediction markets, participants trade securities to represent their probabilistic predictions about the random variable of interest, and the total payment is bounded by the difference between the final market price and the initial market price. Wagering mechanisms consider the betting markets, where participants are willing to wager on their reported information, and the principal can redistribute the wagers based on the ground truth to induce incentive instead of paying out of her own money. Another growing challenge is that in some scenarios, the principal has to take actions based on the information provided by the participants, and this action influences the ground truth information the principal can observe. Such an interaction loop creates more space for profitable manipulations of information reporting. Decision markets [OS10; Che+11] are proposed to address this issue.

My contributions. In the information elicitation with verification setting, I focus on the problem of designing wagering mechanisms. Wagering mechanisms consider betting scenarios where the participants have an amount of money they are willing to lose, referred to as the wager, to support their reported information and exchange the chance to win the wagers of other participants. There are generally four desirable properties in wagering mechanisms: individual rationality, which guarantees participants have a positive expected reward of participation from

their own perspective; truthfulness, which makes truthful reporting the most profitable strategy for each participant; budget balance, which guarantees that the principal needs not to pay using her own budget; and Pareto optimality, which means that the participants can make full use of their wagers, i.e., there is a chance that they can lose all their wagers or win all wagers from others. It has been shown that these four desirable properties cannot be satisfied simultaneously for deterministic wagering mechanisms. To overcome this impossibility result, I introduce randomness into wagering mechanisms and achieve these four properties simultaneously. The resulting mechanisms can also deal with the case where only a noisy signal of the ground truth is available and give the principal the control of the variance of participants' rewards.

Information aggregation with verification. Information aggregation with verification considers the problem of generating accurate predictions about the target random variables when the principal has access to some history prediction data of the same pool of participants and the corresponding ground truth. Traditional aggregation methods rely on identifying the correlation, noise, and bias in participants' predictions [WC92; Gun92; Cla+16] or identifying experts from the participants [CW86; Asp10; BC15] and use this information to aid aggregation. With the rise of machine learning technologies nowadays, this aggregation problem can be explored as a typical supervised learning problem with relatively limited training data. The abundant and ever-growing literature in supervised learning can be leveraged to attack this problem. This thesis will not cover this problem.

Information elicitation without verification. Information elicitation without verification considers the elicitation scenarios where the ground truth information is not available to the principal. In many scenarios, the ground truth is usually too expensive to obtain, e.g., whether a social science study can be replicated or not, or even impossible, e.g., do you like the restaurant. Without the ground truth, the truthfulness property is much harder to achieve, and there is no prevalent solution like the strictly proper scoring rules in the with-verification setting. Peer prediction mechanisms have been proposed as a means to achieve truthfulness property in this setting. Instead of using the ground truth to verify the quality of information,

these mechanisms assume certain knowledge of the correlation pattern between different participants' information and use their reported information to verify with each other's. This approach induces collusion among participants, resulting in different truthfulness notions based on the set of collusion strategies considered. Research in this area has been evolving in three dimensions: the knowledge requirement, the truthfulness notion, and the elicited information representation. As for the knowledge requirement, the pioneering work [MRZ05b] requires knowing the exact joint distribution between peer participants' observed signals, while the following works circumvent this strong requirement by either learning the knowledge from multiple homogeneous elicitation tasks [e.g. WP13; Shn+16; RFJ16; KS19; Kon20] or additionally asking participants' opinions about their peer participants' opinion distribution on the original elicitation task [e.g. Pre04a; WP12b; RF13b]. The truthfulness notions achieved have also evolved from making truthful reporting one of the Bayesian Nash equilibrium (BNE) of the reporting game [MRZ05b] to the BNE with the highest expected payment [e.g. KLS16; DG13; Shn+16; RFJ16] to the dominant uniform strategies [KS19; Kon20; GF19b]. The information representations elicited have also expanded from categorical signals to probabilistic predictions [e.g. WP12b; RF13b; KS18] to continuous statistic estimations [Kon+20; SY20a].

My contributions. Compared to the existing peer prediction mechanisms, the strictly proper scoring rules in the with-verification setting has the advantage of being able to measure the accuracy of the provided probabilistic predictions. In this thesis, I explore the problem of extending the strictly proper scoring rules into the without-verification setting. In particular, I propose the surrogate scoring rules, which recover the scores of strictly proper scoring rules in expectation when there are multiple homogeneous probabilistic forecasting tasks. Consequently, my mechanisms achieve the strongest truthfulness notion in the without-verification setting– the dominant uniform strategy truthfulness– to elicit probabilistic predictions. These mechanisms also inherit the capability of reflecting the prediction accuracy of the strictly proper scoring rules, which can be further used to aid forecast aggregation.

Information aggregation without verification. Information aggregation without verification considers the aggregation scenarios where the principal has no historical forecasting data

of the current participation population. The biggest challenge of information aggregation without verification is that the principal has little knowledge about the underlying information structure of the participants' provided information, e.g., how noisy the information is, what the bias looks like, and how the information sources of different participants overlap with each other [Sat+21]. Consequently, the principal cannot effectively utilize every piece of information they collect. This challenge also exists in the with-verification setting but is more salient in the without-verification setting, as some structure information can be inferred from the historical data. Research in this field has two main approaches. One is to identify systematic bias and noise in human predictions or the aggregation functions [Arm01; JW08; MLS12]. For example, a logit-mean aggregator has been proposed to extremize the mean prediction because human predictions are usually too conservative [Bar+14; Sat+14a]. The other is to identify the information structure by asking additional questions related to peer participants' opinions [e.g. Pre04b; PSM17; PS19] or inferring from participants' predictions on multiple homogeneous forecasting questions [e.g. LPI12; OVB14; LD14; MP17]. The first approach is robust as it uses generic human prediction patterns, but the accuracy improvement is limited. The second approach obtains high accuracy in specific settings but lacks robustness as it requires the data to follow the probabilistic model used in the inference.

My contributions. In this sub-domain, I propose a new aggregation approach that uses peer prediction mechanisms to robustly improve the aggregation accuracy compared to the existing approaches. Inspired by the surrogate scoring rules, which theoretically reflect the participants' prediction accuracy in the without-verification setting under certain assumptions, I identify the empirical correlation between the rewards of several peer prediction mechanisms, including the surrogate scoring rules, and the corresponding prediction accuracy of the participants. I further use these peer prediction mechanisms' rewards to select the underlying sophisticated forecasters from the participant pool and then apply generic aggregators over these selected forecasters. The new aggregation approach demonstrates consistent accuracy improvement over existing aggregators on 14 real-world forecasting datasets I tested. The success of this approach also sheds light on how we can utilize the multi-task information to aid aggregation. The existing approach to multi-task aggregation focuses on establishing probabilistic models on

the prediction data and directly inferring the ground truth. Such an approach lacks robustness because the inferred outcome may be dramatically wrong if the data deviates from the assumed probabilistic model. In contrast, my proposed approach uses multi-task information to select the top forecasters. Even if the selection might be inaccurate, the following generic aggregation layer will still offer acceptable performance.

Human behavior. Human behavior is an overlooked aspect of information elicitation and aggregation. When deploying elicitation and aggregation algorithms, the information provided by the participants is inevitably influenced by their behavioral patterns, e.g., whether they can correctly understand the questions and how they will react to the incentive mechanism. While most truthfulness designs in information elicitation implicitly assume that participants are fully rational, emerging empirical evidence has shown that human participants' behaviors deviate from the fully rational model [MT00; Rab13; BDL19]. Even if truthful reporting strictly dominates any other reporting strategy in the expected reward under a mechanism, human participants may still play some other suboptimal strategies [HRS16; Ree18]. Gao et al. [Gao+14] also found that some carefully incentivized peer prediction mechanisms even led to worse participant performance in controlled experiments. There are also several other interesting problems to investigate relevant to human behavior. For example, can we elicit rich information other than the direct predictions from the participants to help aggregate, and how the rich information interacts with the direct predictions? It is also a problem whether certain groups of people can understand the elicitation materials and provide useful information for specific applications. These behavior-related questions have not received broad attention yet.

My contributions. In this thesis, I conduct preliminary investigations of human behavior in information elicitation and aggregation. First, I investigate human behavior and its influence in a unique information elicitation scenario, the interdependent valuation auctions. In interdependent value auctions, the value of the item is jointly determined by the private signals held by each bidder. The principal needs to allocate the item to the bidders based on its valuation to achieve certain objectives such as social welfare maximization or revenue maximization. It is widely observed that instead of playing as fully rational agents, human bidders in interdepen-

dent valuation auctions tend to underestimate the contingency between other bidders' bids and their private signals, bidding sub-optimally and non-truthfully. To address this problem, I propose special interdependent valuation auctions making truthful reporting an equilibrium strategy for these bidders with the above behavioral bias. Furthermore, I show how much social welfare and revenue we must sacrifice to maintain a good incentive for human bidders.

Second, I explore a specific information elicitation and aggregation application, predicting the replicability of social science studies. Hoogeveen, Sarafoglou, and Wagenmakers [HSW19] found that laypeople can provide above-chance prediction accuracy in this application with the intervention of experts. However, expert intervention is the bottleneck to the scalability of the system. I investigate whether we can achieve similar prediction performance with laypeople without expert intervention but with rich information elicited. In particular, I run experiments with laypeople recruited online and elicit their surprisingness, understandability, and direct replication prediction about published social science studies. I obtain several interesting findings in these experiments related to how laypeople's provided rich information interact with their direct predictions and the ground truth.

Organization. The rest of the thesis is organized as follows. Chapter 1 introduces the randomized wagering mechanisms in the domain of information elicitation with verification. Chapter 2 talks about the surrogate scoring rules in the domain of elicitation without verification. Chapter 3 discusses the forecast aggregation methods using peer prediction in the aggregation without verification setting. Chapter 4 proposes the interdependent value auctions customized to human bidders with behavioral bias, while Chapter 5 reports the experimental findings in predicting the replicability of social science studies with laypeople. Finally, I conclude my works in Chapter 6.

Chapter 1

Randomized Wagering Mechanisms

1.1 Introduction

Wagering mechanisms [Lam+08; Lam+15; Che+14; FPW17; FP18] are one-shot betting mechanisms that allow a principal to elicit participating agents' beliefs about an event of interest without paying out of pocket or incurring a risk. Compared with prediction-market type of dynamic elicitation mechanisms, one-shot wagering is possibly preferred due to its simplicity. It is particularly designed for agents with immutable beliefs who "agree to disagree" and who do not update their beliefs. In a wagering mechanism, each agent submits a prediction for the event and specifies a wager, which is the maximum amount of money that the agent is willing to lose. Then after the event outcome is revealed, the total wagered money will be redistributed among the participants. Researchers have developed wagering mechanisms with various theoretical properties. In particular, Lambert et al. [Lam+08; Lam+15] proposed a class of weighted score wagering mechanisms (WSWM) that satisfy a set of desirable properties, including budget balance, individual rationality, incentive compatibility, sybilproofness, among others.¹ Chen et al. [Che+14] later proposed a no-arbitrage wagering mechanism (NAWM) that removes opportunities for participating agents to risklessly profit.

However, in both WSWM and NAWM, it has been observed that a participant only loses a very small fraction of his total wager even in the worst case. This seems to be undesirable

¹Definitions of some properties can be found in Section 1.4.

in practice as it is against the spirit of betting and a wager effectively loses its meaning as a budget. Freeman, Pennock, and Wortman Vaughan [FPW17] first formalized this observation by indicating that these mechanisms are not Pareto optimal, where Pareto optimality requires that there is no profitable side bet among participants before the allocation of a wagering mechanism is realized. They also proved an impossibility result: Pareto optimality cannot be satisfied together with individual rationality, weak budget balance and weak incentive compatibility. A double clinching auction (DCA) wagering mechanism [FPW17] was hence proposed to improve Pareto efficiency. The parimutuel consensus mechanism (PCM) has been shown to satisfy Pareto optimality [FP18], but violates incentive compatibility.

This chapter is another quest of wagering mechanisms with better theoretical properties. We expand the design space of wagering mechanisms to allow randomization on agent payoffs and ask whether we can achieve all aforementioned desirable properties, including Pareto optimality. We give a positive answer to this question: Our randomized wagering mechanisms are the first ones to achieve Pareto optimality along with other properties.

We first show that a simple randomized lottery-type implementation of existing wagering mechanisms (Lottery Wagering Mechanisms (LWM)) satisfy all desirable properties. In LWM, instead of receiving re-allocated money from a deterministic wagering mechanism, each agent receives a number of lottery tickets proportional to his payoff in the deterministic wagering mechanism. Then, the agent with the winning lottery wins the total wager (collected from all participants).

We then design another family of randomized wagering mechanisms, the Surrogate Wagering Mechanisms (SWM), by bringing insights from learning with noisy data [Nat+13; Sco15] to wagering mechanism design. A SWM first generates a “surrogate outcome” for each agent according to the true event outcome. An agent’s reported prediction is then evaluated using his surrogate but biased outcome together with a bias removal procedure such that in expectation the agent receives a score as if his prediction is evaluated against the true event outcome. Despite being randomized, SWM preserve the incentive properties of a deterministic wagering mechanism. We show that certain SWM satisfy all desirable properties of a wagering mechanism. Notably, SWM are robust to situations where only a noisy copy of the event outcome is

available - this property is due to the fact that we borrow the machinery from the literature of learning with noisy data in designing SWM. We believe that this is another unique contribution to the literature of wagering mechanisms.

1.2 Related works

The ability to elicit *information*, in particular predictions and forecasts about future events, is crucial for many application settings and has been studied extensively in the literature. Proper scoring rules [Bri50b; JNW06; MW76; Win69; GR07a] have been designed for this purpose, where each agent is rewarded by how well their reported forecasts predicted the true realized outcome (after the outcome is resolved). Later, the competitive scoring rules [KG04] and the parimutuel Kelly probability scoring rules [Joh07] adapt proper scoring rules to group competitive betting. Both mechanisms are budget balanced so that the principal doesn't need to pay any participant. These spur the further development of the previously discussed wagering mechanisms and the examination of their theoretical properties [Lam+08; Lam+15; Che+14; FPW17; FP18].

Our method used in lottery wagering mechanisms to transfer an arbitrary deterministic wagering mechanism into a randomized one, while maintaining the properties, is inspired by the method proposed in Witkowski et al. [Wit+18]. They study the incentive compatible forecasting competitions and propose to transfer the scores of multiple predictions under the strictly proper scoring rules into the odds of winning to maintain the incentive property. Lambert et al. [Lam+08] proposed a randomization method based on WSWM via randomly selecting strictly proper scoring rules and proper scoring rules with extreme values to increase the stake. However, this method does not generalize to other deterministic wagering mechanisms. Cummings, Pennock, and Wortman Vaughan [CPW16] proposed to apply differential privacy technology to randomize the payoff of wagering mechanisms in order to preserve the privacy of each agent's belief. However, their method does not maintain budget balance (in ex-post).

The idea of using randomization in wagering mechanism design is not entirely new, but

not thoroughly studied. Both Lambert et al. [Lam+08] and Cummings, Pennock, and Wortman Vaughan [CPW16] proposed certain types of randomized wagering mechanisms, but neither of the mechanisms satisfies Pareto optimality. The randomized wagering mechanisms first appeared in Lambert et al. [Lam+08]. There, the randomization is restricted to randomly selecting different scoring rules used in WSWM. It introduced this randomization in order to alleviate the the problem that in WSWM, agents only lose a small fraction of their wagers regardless of the event outcome. However, even with this randomization, an agent won't lose all his wager in the worst when the number of agents is finite. Cummings, Pennock, and Wortman Vaughan [CPW16] applied differential privacy technology to randomize the payoff of wagering mechanisms. Its goal is to preserve the privacy of agents' beliefs.

Our specific ideas of adding randomness as in the lottery-like wagering mechanisms are inspired by recent works on forecasting competition [Wit+18]. Our ideas of surrogate wagering mechanisms are inspired by surrogate scoring rules [LC18], and the literature on learning with noisy labels [Byl94; Nat+13; Sco15].

1.3 Preliminaries

In this section, we explain the scenario where a wagering mechanism applies and formally introduce the deterministic wagering mechanisms. Consider a scenario where a principal is interested in eliciting subjective beliefs from a set of agents $\mathcal{N} = \{1, 2, \dots, N\}$ about a random variable (event) X , which takes a value (outcome) in set $\mathcal{X} = \{0, 1, \dots, M - 1\}$, $M \geq 2$. The belief of each agent i is private, denoted as a vector of occurrence probabilities of each outcome $\mathbf{p}_i = (p_i^j)_{j \in \mathcal{X}} \in \Delta^{M-1}$. Following the previous work on wagering mechanism, we continue to adopt an immutable belief model for agents. Unlike in a Bayesian model, agents with immutable beliefs do not update their beliefs. The immutable belief model and the Bayesian model are two extremes of agent modeling for information elicitation, with the reality lies in between and arguably closer to the immutable belief side as people do "agree to disagree." Moreover, Lambert et al. [Lam+15] showed that while WSWM was designed for agents with immutable beliefs, it continued to perform well for Bayesian agents who have some innate

utility for trading.

The principal uses a *wagering mechanism* to elicit private beliefs of agents. In a wagering mechanism, each agent reports a probability vector $\hat{\mathbf{p}}_i \in \Delta^{M-1}$, capturing his belief, and wagers an amount of money $w_i \in \mathbb{R}_+$. Similar to Lambert et al. [Lam+08], we assume that wagers are exogenously determined for each agent and are not a strategic consideration. We use $\hat{\mathbf{p}}$ and \mathbf{w} to denote the reports and the wagers of all agents respectively, and use $\hat{\mathbf{p}}_{-i}$ and \mathbf{w}_{-i} to denote the reports and wagers of all agents other than agent i . In addition, we use W_S to denote $\sum_{i \in S} w_i$ for any set of agents $S \subseteq \mathcal{N}$. After an event outcome $x \in \mathcal{X}$ is realized, the wagering mechanism redistributes all the wagers collected from agents according to $\hat{\mathbf{p}}, \mathbf{w}, x$. The net-payoff of agent i is defined as the payoff or the money that agent i receives from the redistribution minus his wager. A wagering mechanism defines a net-payoff function $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$ for each agent i with wager constraint $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x) \geq -w_i$ and constraint $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x) = 0$ whenever $w_i = 0$. The two constraints ensure that no agent can lose more than his wager and no agent with zero wager can gain.

1.3.1 Strictly proper scoring rules and weighted score wagering mechanisms

Strictly proper scoring rules [GR07a] are scoring functions proposed and developed to truthfully elicit beliefs from risk-neutral agents. They are building blocks of many incentive compatible wagering mechanisms, such as WSWM and NAWM. A strictly proper scoring rule rewards a prediction $\hat{\mathbf{p}}_i$ by a score $s_x(\hat{\mathbf{p}}_i)$, according to the realization x of the random variable X . The scoring function $s_x(\cdot)$ is designed such that the expected payoff of truthful reporting is strictly larger than that of any other report, i.e., $\mathbb{E}_{X \sim \mathbf{p}_i}[s_X(\mathbf{p}_i)] > \mathbb{E}_{X \sim \mathbf{p}_i}[s_X(\hat{\mathbf{p}}_i)]$, $\forall \hat{\mathbf{p}}_i \neq \mathbf{p}_i$.

There is a rich family of strictly proper scoring functions, including Brier scores (for binary outcome event, $s_x(\hat{p}_i) = 1 - (\hat{p}_i - x)^2$, where \hat{p}_i is agent i 's report of $\mathbb{P}(X = 1)$), logarithmic and spherical scoring functions. Strictly proper scoring rules are closed under positive affine transformations.

WSWM [Lam+08] rewards an agent according to his wager and the accuracy of his prediction relative to that of other agents' predictions. The net-payoff of agent i in WSWM, is formally

defined as

$$\Pi_i^{\text{WS}}(\hat{\mathbf{p}}; \mathbf{w}; x) = \frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \left(s_x(\hat{\mathbf{p}}_i) - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N} \setminus \{i\}}} s_x(\hat{\mathbf{p}}_j) \right), \quad (1.1)$$

where $s_x(\cdot)$ is any strictly proper scoring rule bounded within $[0, 1]$. WSWM strictly encourages truthful reporting of predictions, because the net-payoff of agent i is a strictly proper scoring rule of his prediction. Meanwhile, $\sum_{i \in \mathcal{N}} \Pi_i^{\text{WS}}$ is always zero by the form of the net-payoff formula, no matter what $s_x(\cdot)$ is. This means that the budget balance property of Eqn. (1.1) doesn't depend on the scoring rules. Our proposed surrogate wagering mechanisms use the same general form of the net-payoff function (but a different scoring rule) to guarantee the ex-post budget balance.

1.4 Randomized wagering mechanisms

We introduce randomized wagering mechanisms as extensions of deterministic wagering mechanisms. Similar to deterministic wagering mechanisms, the net-payoff of an agent in randomized wagering mechanisms depends on all agents' predictions $\hat{\mathbf{p}}$ and wagers \mathbf{w} , as well as the realized outcome x . But different from deterministic wagering mechanisms, the net-payoffs are now random variables. For notational simplicity, we now use $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$ to represent the random variable of agent i 's net-payoff in a randomized wagering mechanism. We use $\pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$ to represent the realization of $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$. We use Π_i and π_i as abbreviations when $\hat{\mathbf{p}}; \mathbf{w}; x$ are clear in the context. We denote the maximum/minimum possible value of a random variable X by \bar{X}/\underline{X} . We denote the joint distribution of $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x), i \in \mathcal{N}$ by $\mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)$ and the marginal distribution of $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$ by $\mathcal{D}_i(\hat{\mathbf{p}}; \mathbf{w}; x)$.

Definition 1.1. *Given a set \mathcal{N} of agents, reports $\hat{\mathbf{p}}$ and wagers \mathbf{w} of agents and the event outcome x , a randomized wagering mechanism defines a joint distribution $\mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)$, and pays agent i by a net-payoff $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$, where $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x), i \in \mathcal{N}$ are jointly drawn from $\mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)$. Moreover, $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x) \geq -w_i$ and $\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x) = 0$ whenever $w_i = 0$.*

A deterministic wagering mechanism is a special case of randomized wagering mechanisms when $\mathcal{D}_i(\hat{\mathbf{p}}; \mathbf{w}; x)$ is a point distribution for all agent $i \in \mathcal{N}$.

1.4.1 Desirable properties

In the literature, several desirable properties of wagering mechanisms have been proposed in the deterministic context. Lambert et al. [Lam+08] introduced (a) individual rationality, (b) incentive compatibility, (c) budget balance, (d) sybilproofness, (e) anonymity, and (f) neutrality. Chen et al. [Che+14] introduced (g) no arbitrage. Freeman, Pennock, and Wortman Vaughan [FPW17] introduced (h) Pareto optimality. We extend these properties to the randomized context. These new properties reduce to the properties defined in the literature for the special case of deterministic wagering mechanisms.

(a) **Individual rationality** requires that each agent has nothing to lose in expectation by participating.

Definition 1.2. A randomized wagering mechanism is *individually rational (IR)* if $\forall i, \mathbf{p}_i, \mathbf{w}$, and $\hat{\mathbf{p}}_{-i}$, there exists $\hat{\mathbf{p}}_i$ such that

$$\mathbb{E}_{X \sim \mathbf{p}_i, \Pi_i \sim \mathcal{D}_i(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)} [\Pi_i(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)] \geq 0.$$

(b) **Incentive compatibility** requires that an agent's expected net-payoff is maximized when he reports honestly, regardless of other agents' reports and wagers.

Definition 1.3. A randomized wagering mechanism is *weakly incentive compatible (WIC)* if $\forall i, \mathbf{p}_i, \hat{\mathbf{p}}_i \neq \mathbf{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}$:

$$\begin{aligned} & \mathbb{E}_{X \sim \mathbf{p}_i, \Pi_i \sim \mathcal{D}_i(\mathbf{p}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)} [\Pi_i(\mathbf{p}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)] \\ & \geq \mathbb{E}_{X \sim \mathbf{p}_i, \Pi_i \sim \mathcal{D}_i(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)} [\Pi_i(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{-i}; \mathbf{w}; X)]. \end{aligned}$$

A randomized wagering mechanism is *strictly incentive compatible (SIC)* if the inequality is strict.

(c) **Ex-post budget balance** ensures that the principal does not need to subsidize the betting.

Definition 1.4. A randomized wagering mechanism is *weakly ex-post budget-balanced (WEBB)* if $\forall \hat{\mathbf{p}}, \mathbf{w}, x : \sum_{i \in \mathcal{N}} \pi_i(\hat{\mathbf{p}}, \mathbf{w}, x) \leq 0$ for any realization $(\pi_i)_{i \in \mathcal{N}}$ drawn from the joint distribution $\mathcal{D}(\hat{\mathbf{p}}, \mathbf{w}, x)$. A randomized wagering mechanism is *ex-post budget-balanced (EBB)* if the equality always holds.

(d) **Sybilproofness** requires that no agent can increase its expected net-payoff by creating fake identities and splitting his wager, regardless of other agents' reports and wagers.

Definition 1.5. Suppose agent i , instead of participating under one account with reported prediction $\hat{\mathbf{p}}_i$ and wager w_i , participates under $k > 1$ sybil accounts, with predictions and wagers $\{\hat{\mathbf{p}}_{i_l}, w_{i_l}\}_{l=1, \dots, k}$ such that $\hat{\mathbf{p}}_{i_l} = \hat{\mathbf{p}}_i, w_{i_l} \geq 0, \forall l = 1, \dots, k$ and $\sum_{l=1}^k w_{i_l} = w_i$. A randomized wagering mechanism is **sybilproof** if $\forall i, \hat{\mathbf{p}}, \mathbf{w},$ and x , and for all sybil reports $\hat{\mathbf{p}}_{i_1}, \dots, \hat{\mathbf{p}}_{i_k}$ and wagers w_{i_1}, \dots, w_{i_k} , we have

$$\begin{aligned} & \mathbb{E}_{\Pi \sim \mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)}[\Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)] \\ & \geq \mathbb{E}_{\Pi' \sim \mathcal{D}(\hat{\mathbf{p}}'; \mathbf{w}'; x)}\left[\sum_{l=1}^k \Pi_{i_l}(\hat{\mathbf{p}}'; \mathbf{w}'; x)\right]. \end{aligned}$$

where $\hat{\mathbf{p}}, \mathbf{w}$ and Π are the reports, wagers and net-payoffs when agent i participates under one account and $\hat{\mathbf{p}}', \mathbf{w}'$ and Π' are the reports, wagers and net-payoffs when agent i participates using k sybils.

(e) **Anonymity** requires that agents' identities do not affect their net-payoffs. Let $\sigma_{\mathcal{N}}$ be a permutation of the set of agents \mathcal{N} , and denote $\hat{\mathbf{p}}_{\sigma_{\mathcal{N}}}, \mathbf{w}_{\sigma_{\mathcal{N}}}$ the reports and wagers of agents after applying the permutation respectively. Denote $\mathcal{D}_{\sigma_{\mathcal{N}}}$ the joint distribution of net-payoffs of agents in \mathcal{N} after applying the permutation on agents.

Definition 1.6. A randomized wagering mechanism is **anonymous** if $\forall \sigma_{\mathcal{N}}, \hat{\mathbf{p}}, \mathbf{w}, x : \mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x) = \mathcal{D}_{\sigma_{\mathcal{N}}}(\hat{\mathbf{p}}_{\sigma_{\mathcal{N}}}; \mathbf{w}_{\sigma_{\mathcal{N}}}; x)$

(f) **Neutrality** requires that the net-payoffs do not depend on the labeling of the event outcomes. Let $\sigma_{\mathcal{M}}$ be a permutation of the set of outcomes \mathcal{M} . Denote by $\hat{\mathbf{p}}_i^{\sigma_{\mathcal{M}}}$ the reported prediction of agent i after we relabel the outcomes according to permutation $\sigma_{\mathcal{M}}$, and denote by $\sigma_{\mathcal{M}}(x)$ the new label of an outcome $x \in \mathcal{M}$.

Definition 1.7. A randomized wagering mechanism is **neutral** if $\forall \sigma_{\mathcal{M}}, \hat{\mathbf{p}}, \mathbf{w}, x :$

$$\mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x) = \mathcal{D}(\hat{\mathbf{p}}_1^{\sigma_{\mathcal{M}}}, \dots, \hat{\mathbf{p}}_N^{\sigma_{\mathcal{M}}}; \mathbf{w}; \sigma_{\mathcal{M}}(x)).$$

(g) **No arbitrage** requires that no agent can risklessly make a profit.

Definition 1.8. A randomized wagering mechanism has **no arbitrage** if $\forall i, \hat{\mathbf{p}}, \mathbf{w} (\mathbf{w} > \mathbf{0}), \exists x$ such that $\underline{\Pi}_i(\hat{\mathbf{p}}, \mathbf{w}, x) < 0$.

Mechanism	Budget Balance	Incentive Compatibility	Pareto Optimality	No Arbitrage
WSWM [Lam+08]	Strictly	Strictly	False	False
NAWM [Che+14]	Weakly	Strictly	False	True
DCA [FPW17]	Strictly	Weakly	False	True
PCM [FP18]	Strictly	False	True	True
Randomized WSWM [Lam+08]	Strictly	True	False	True
Private WSWM [CPW16]	False	True	False	True
LWS (ours)	Strictly	True	True	True
RP-SWME (ours)	Strictly	True	True	True

Table 1.1: A summary of properties of wagering mechanisms²

(h) **Pareto optimality** in economics refers to an efficient situation where no trade can be made to improve an agent’s payoff without harming any other agent’s payoff. In an IR wagering mechanism, agents with different beliefs can always form a profitable (in expectation) wagering game if they all have a positive budget. Freeman, Pennock, and Wortman Vaughan [FPW17] defined Pareto optimality of a wagering mechanism as a property that agents with different beliefs will each lose all of his wager under at least one of the event outcomes. This “worst-case” outcome might be different for different agents. Thus, before the event outcome is realized, no agent can commit to secure part of his wager from the mechanism and no additional profitable wagering game can be made. We define Pareto optimality for randomized wagering mechanisms in a similar spirit: no agents with different beliefs can commit to secure part of their wagers before the event outcome is realized.

Definition 1.9. A randomized wagering mechanism is **Pareto optimal (PO)** if $\forall \hat{\mathbf{p}}, \mathbf{w}, \forall i, j \in \mathcal{N}$ with $\hat{\mathbf{p}}_i \neq \hat{\mathbf{p}}_j, \exists l \in \{i, j\}$ and x , such that $\underline{\Pi}_l(\hat{\mathbf{p}}, \mathbf{w}, x) = -w_l$.

Properties of existing wagering mechanisms We summarize the properties of existing wagering mechanisms³ and ours in Table 1. No existing mechanism satisfies all properties (a)-(h). Moreover, Freeman, Pennock, and Wortman Vaughan [FPW17] showed an **impossibility**

²All of the mechanisms in this table satisfy individual rationality, anonymity, neutrality and sybilproofness.

³WSWM, NAWM, DCA, PCM, randomized WSWM [Lam+08], private WSWM [CPW16]

Mechanism 1 Lottery Wagering Mechanisms

- 1: Compute the payoff of each agent i under a DET: $\pi'_i \leftarrow w_i + \Pi_i(\hat{\mathbf{p}}; \mathbf{w}; x)$.
 - 2: Each agent has winning probability $\frac{\pi'_i}{\sum_{i \in \mathcal{N}} \pi'_i}$. Draw a lottery winner $i^* \in \mathcal{N}$.
 - 3: Winner i^* is assigned a net-payoff $\sum_{i \in \mathcal{N} \setminus \{i^*\}} w_i$ and any agent $j \neq i^*$ has a net-payoff $-w_j$.
-

result that for deterministic wagering mechanisms, it is impossible to achieve properties IR, WIC, WEBB, and PO simultaneously. For existing randomized wagering mechanisms, the randomized WSWM in Lambert et al. [Lam+08] only satisfies PO in the limit of large population of participants, and the private WSWM [CPW16] does not satisfy WEBB and PO.

1.5 Lottery wagering mechanisms

In this section we introduce a family of randomized wagering mechanisms, the *lottery wagering mechanisms* (LWM), which extends arbitrary deterministic wagering mechanisms into randomized wagering mechanisms. We will show that LWM easily preserve (the randomized version of) the properties of the underlying deterministic wagering mechanisms, while achieving Pareto optimality, overcoming the impossibility result.

In lottery wagering mechanisms, each agent receives a number of lottery tickets in proportion to the *payoff* he gets under a deterministic wagering mechanism, and a winner is drawn from all the lottery tickets to win the entire pool of wagers. The mechanisms are designed in a way such that the expected payoff of each agent is equal to his payoff in the underlying deterministic wagering mechanisms and each agent has a positive probability to lose all his wager. Hence, no profitable side bet exists and the mechanisms are Pareto optimal. We formally present the lottery wagering mechanism that extends an arbitrary deterministic wagering mechanism DET in Mechanism 1. To distinguish the payoff from the net-payoff, we denote the payoff of agent i by π'_i .

Lottery wagering mechanisms are powerful in obtaining desirable theoretical properties. We show in Theorem 1.1 that the lottery wagering mechanism that extends WSWM, namely Lottery Weighted Score wagering mechanism (LWS), satisfies all properties (a)-(h).

Theorem 1.1. *LWS satisfies all properties (a) - (h).*

We notice that although LWS satisfies all desirable properties, it can be unsatisfying because (1) agents have high variance in payoff and (2) except the winning agent, all other agents lose money. To alleviate these issues, we can mix LWS with WSWM by assigning each of them a probability to be executed. The resulting mechanism still satisfies all the properties (a)-(h). The probabilistic mixture allows us to adjust the variance of the payoffs as well as agents' winning probabilities in the resulting mechanism.

1.6 Surrogate wagering mechanisms

In this section, we propose the *surrogate wagering mechanisms* (SWM). We first introduce the generic SWM, then a variant of SWM that achieves the desirable theoretical properties and at the same time have moderate variance in payoffs and higher winning probabilities for accurate predictions. We then notice that randomization opens up the possibility of dealing with situations where only noisy ground truth is available. We discuss how to extend our results to this noisy setting.

1.6.1 Generic surrogate wagering mechanisms

A surrogate wagering mechanism consists of three main steps: (1) SWM generates a surrogate event outcome for each agent based on the true event outcome and a randomization device; (2) SWM evaluates each agent's prediction according to the surrogate event outcome using a designed scoring function such that the score is an unbiased estimate of the score derived by applying a strictly proper scoring rule to the ground truth outcome; (3) SWM applies WSWM to the scores based on the surrogate event outcome to determine the final net-payoff of each agent. Next, we explain these three steps in details. For clarity and simplicity of exposition, we consider only binary events, i.e., $\mathcal{X} = \{0, 1\}$, in this section. Extension to multi-outcome events will be introduced later.

Step 1. Surrogate event outcomes A SWM generates a surrogate event outcome \tilde{X}_i for each agent $i \in \mathcal{N}$. Denote $\tilde{\mathbf{X}} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N)$. \tilde{X}_i 's are drawn independently conditional on X , and are specified by SWM. The conditionally marginal distribution $\mathbb{P}(\tilde{X}_i|X), i \in \mathcal{N}$ can be

Mechanism 2 Surrogate Wagering Mechanisms

- 1: Collect the predictions $\hat{\mathbf{p}}$ and wagers \mathbf{w} .
 - 2: Select error rate $e_0^i, e_1^i \in [0, 1]$ and $e_0^i + e_1^i \neq 1, \forall i$.
 - 3: Generate surrogate outcome $\tilde{X}_i, \forall i$ such that $\mathbb{P}(\tilde{X}_i = 1|X = 0) = e_0^i, \mathbb{P}(\tilde{X}_i = 0|X = 1) = e_1^i$.
 - 4: Score each agent $i \in \mathcal{N}$ according to Eqn. (1.2).
 - 5: Pay each agent $i \in \mathcal{N}$ a net-payoff using Eqn. (1.3).
-

expressed by two parameters, the error rates of the surrogate outcome: $e_1^i = \mathbb{P}(\tilde{X}_i = 0|X = 1)$ and $e_0^i = \mathbb{P}(\tilde{X}_i = 1|X = 0)$. The conditionally marginal distribution $\mathbb{P}(\tilde{X}_i|X)$ can be any distribution satisfying $\forall i \in \mathcal{N} : e_1^i + e_0^i \neq 1$.⁴ We use \tilde{x} and \tilde{x}_i to denote the realization of $\tilde{\mathbf{X}}$ and \tilde{X}_i respectively.

Step 2. Computing unbiased scores Given a strictly proper scoring rule $s_x(\cdot)$ within $[0,1]$, SWM computes the score of agent i as $\varphi \circ s_{\tilde{x}_i}(\hat{p}_i)$, where

$$\varphi \circ s_{\tilde{x}_i}(\hat{p}_i) = \frac{(1 - e_{1-\tilde{x}_i}^i)s_{\tilde{x}_i}(\hat{p}_i) - e_{\tilde{x}_i}^i s_{1-\tilde{x}_i}(\hat{p}_i)}{1 - e_0^i - e_1^i}. \quad (1.2)$$

\tilde{x}_i is the realized surrogate event outcome for agent i . Lemma 1.2 shows that φ is an unbiased operator on the score $s_{\tilde{x}_i}(p_i)$ in the sense that $\mathbb{E}_{\tilde{X}_i|x}[\varphi \circ s_{\tilde{X}_i}(\hat{p}_i)] = s_x(\hat{p}_i)$.

Lemma 1.2 (Lemma 3.4 of Liu and Chen [LC18]). $\forall x \in \{0, 1\}, \forall \hat{p}_i, e_0^i, e_1^i \in [0, 1]$ and $e_0^i + e_1^i \neq 1$, we have $\mathbb{E}_{\tilde{X}_i|x}[\varphi \circ s_{\tilde{X}_i}(\hat{p}_i)] = s_x(\hat{p}_i)$.

Lemma 1.2 implies that if $s_x(\hat{p}_i)$ is a strictly proper scoring rule, then $\varphi \circ s_{\tilde{x}_i}(\hat{p}_i)$ is also a strictly proper scoring rule.

Step 3. Computing net-payoffs In the final step, SWM computes the net-payoff of agent i using WSWM and the unbiased score of agent i , i.e., replacing score $s_x(\hat{p}_i)$ in Eqn. (1.1) by score $\varphi \circ s_{\tilde{x}_i}(\hat{p}_i)$. Formally, we have

$$\Pi_i^{\text{SWM}}(\hat{\mathbf{p}}, \mathbf{w}, x) = \frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \left(\varphi \circ s_{\tilde{x}_i}(\hat{p}_i) - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N} \setminus \{i\}}} \varphi \circ s_{\tilde{x}_j}(\hat{p}_j) \right), \quad (1.3)$$

where x and $\tilde{x}_i, i \in \mathcal{N}$ are the event outcome and the surrogate event outcome for each agent i respectively.

⁴When $e_0 + e_1 = 1$, \tilde{X}_i turns out to be independent with X , and thus provides no information about X . We thus exclude $e_1^i + e_0^i = 1$.

We formally present SWM in Mechanism 2. According to Lemma 1.2 (applying to each score terms), we have $\forall i, x, \hat{\mathbf{p}}, \mathbf{w} : \mathbb{E}_{\Pi_i^{\text{SWM}} \sim \mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)}[\Pi_i^{\text{SWM}}(\hat{\mathbf{p}}; \mathbf{w}; x)] = \Pi_i^{\text{WS}}(\hat{\mathbf{p}}; \mathbf{w}; x)$. Because the deterministic WSWM satisfies properties ((a)-(f)) [Lam+08], SWM also satisfies these properties. A realization of the score $\varphi \circ s_{\hat{x}_i}(p_i)$ can be larger than 1, implying that agent i can lose (or win) more than what he can lose (or win) in the deterministic WSWM. However, we also notice that for some extreme values of error rates, the constraint $\underline{\Pi}_i(\hat{\mathbf{p}}; \mathbf{w}; x) \geq -w_i$ can be violated⁵, i.e., an agent may lose more than their wager, which makes SWM invalid. In the next section, we show that by selecting error rates in a subtle way, we can obtain all the properties (a)-(h) without violating the wager constraint $\underline{\Pi}_i(\hat{\mathbf{p}}; \mathbf{w}; x) \geq -w_i$.

1.6.2 SWM with error rate selection (SWME) and random partition SWME (RP-SWME)

We notice that according to Lemma 1.2, no matter which error rates e_0, e_1 are chosen, the unbiasedness property of SWM holds, i.e., $\mathbb{E}_{\Pi_i \sim \mathcal{D}(\hat{\mathbf{p}}; \mathbf{w}; x)}[\Pi_i^{\text{SWM}}(\hat{\mathbf{p}}; \mathbf{w}; x)] = \Pi_i^{\text{WSWM}}(\hat{\mathbf{p}}; \mathbf{w}; x)$. In other words, we can choose the error rates in an arbitrary way (even depending on $\hat{\mathbf{p}}, \mathbf{w}$) without changing the expected net-payoff⁶ of each agent under any realized event outcome. This gives us the flexibility to tune the maximum amount of money each agent can win or lose in the game, while preserving the properties ((a)-(f)) inherited from WSWM.

Given reports $\hat{\mathbf{p}}$ and wagers \mathbf{w} but not the event outcome x , the error rate pair that guarantees no wager violation under any outcome $x \in \mathcal{X}$ and any realization of the randomness induced by SWM may not be unique. We propose Algorithm 3 to select a pair of error rates e_0, e_1 after the reports and wagers are collected such that at least one agent loses all his wager in the worst case w.r.t. the outcome and the randomness of SWM. We name the mechanism as SWME when we use Algorithm 3 to select the error rates for SWM.

Lemma 1.3. *SWME has no wager violation and when there exists at least one report $\hat{p}_i \neq 0.5$, at least one of the agents loses all his wager in the worst case w.r.t. the event outcome and the randomness of SWME.*

Proof. In this proof, we use Brier Score as the scoring rule used by the mechanism, i.e., $s_x(\hat{p}_i) = 1 - (x - \hat{p}_i)^2$, and \hat{p}_i is agent i 's report of $\mathbb{P}(X = 1)$. The proof can be extended to other strictly proper scoring rule within $[0, 1]$.

⁵For example, in a wagering game, two agents both wager 1 and report 1 and 0, respectively. Let $s_x(\hat{p}_i) = 1 - (x - \hat{p}_i)^2, e_j^i = 0.4, i = 1, 2, j = 0, 1$. In the worst case of agent 1, the surrogate outcomes are realized as $\hat{x}_1 = 0, \hat{x}_2 = 1$. Then, $\pi_1 = -5 < -1$.

⁶The expectation is taken over the randomness of the mechanism conditioned on the event outcome.

Algorithm 3 Error Rate Selection Algorithm

- 1: Collect the predictions $\hat{\mathbf{p}}$ and wagers \mathbf{w} .
 - 2: $\forall i: s_i^w \leftarrow \min_{x \in \mathcal{X}} s_x(\hat{p}_i), s_i^b \leftarrow \max_{x \in \mathcal{X}} s_x(\hat{p}_i)$.
 - 3: For each agent $i \in \mathcal{N}$, compute $r_i: r_i \leftarrow \frac{1}{2} + \frac{(1 - \frac{w_i}{W_{\mathcal{N}}})(s_i^w - s_i^b) + \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N}}}(s_j^w - s_j^b)}{2(2 + s_i^w + s_i^b - \sum_{j \in \mathcal{N}} \frac{w_j}{W_{\mathcal{N}}}(s_j^w + s_j^b))}$
 - 4: If $\min_{j \in \mathcal{N}} \{r_j\} = 0.5$, set $e_1^i = e_0^i = 0, \forall i$, else set $e_1^i = e_0^i = \min_{j \in \mathcal{N}} \{r_j\}, \forall i$.
-

We first consider the corner case where all agents reports 0.5. It can be verified that in Algorithm 2, $\min_{i \in \mathcal{N}} r_i = 0.5$, and the algorithm sets $e_0^i = e_1^i = 0, \forall i$ and SWME is reduced to WSWM. Thus, no wager violation happens.

Next, we consider the scenario that $\exists i \in \mathcal{N}, \hat{p}_i \neq 0.5$. In this scenario, we first prove that, in Algorithm 2 $\forall i, r_i \in (0, 0.5)$.

We have $\forall i, s_i^w, s_i^b \in [0, 1], s_i^w \leq s_i^b$ (the equality only holds when $\hat{p}_i = 0.5$), $s_i^w + s_i^b \in [0.5, 1]$. Let

$$A = (1 - \frac{w_i}{W_{\mathcal{N}}})s_i^w - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N}}}s_j^b$$

and

$$B = (1 - \frac{w_i}{W_{\mathcal{N}}})(s_i^w + s_i^b) - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N}}}(s_j^w + s_j^b).$$

We have $r_i = \frac{1}{2} + \frac{2A-B}{2(2+B)}$, $A > -1, B \in (-1, 1)$ and $2A - B = \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N}}}(s_j^w - s_j^b) + (1 - \frac{w_i}{W_{\mathcal{N}}})(s_i^w - s_i^b) > 0$ (there exists at least one agent $i \in \mathcal{N}$ that $\hat{p}_i \neq 0.5$). Therefore, $\frac{2A-B}{2+B} \in (-1, 0)$. We have $r_i = \frac{1}{2} + \frac{2A-B}{2(2+B)} \in (0, 0.5)$.

Next, we prove that if let r_i be a variable, and let $e_0^i = e_1^i = r_i$, the worst cast net-payoff π_i^w (w.r.t. the event outcome and the randomness of the mechanism) of agent i is a decreasing function of r_i .

In the worst case of agent i , $\varphi \circ s_{\hat{x}_i}(\hat{p}_i) = \frac{(1-r_i)s_i^w - r_i s_i^b}{1-2r_i}, \varphi \circ s_{\hat{x}_j}(\hat{p}_j) = \frac{(1-r_j)s_j^b - r_j s_j^w}{1-2r_j}$ and $\pi_i^w = w_i \frac{(A-Br_i)}{1-2r_i}$. We have $\frac{\partial \pi_i^w}{\partial r_i} = w_i \frac{2A-B}{(1-2r_i)^2} < 0$. Therefore, π_i^w is decreasing with r_i .

Finally, it is easy to verify that when $r_i = \frac{1}{2} + \frac{2A-B}{2(2+B)}$, $\pi_i^w = -w_i$.

Therefore, when we set for each agent $i \in \mathcal{N}, e_0^i = e_1^i = \min_{j \in \mathcal{N}} r_j$, no agent can lose more than his wager and agent $i^* = \operatorname{argmin}_{j \in \mathcal{N}} r_j$ loses all his wager in the worst case. \square

Note Lemma 1.3 does not imply PO for SWME - if there exist two agents who have different predictions and have wager left even in their own worst cases, they can form a profitable bet against each other. We propose a variant of SWME to fix this caveat as follows.

Random partition SWME (RP-SWME) Lemma 1.3 implies that when agents are partitioned into groups of two, there will not exist side bets. Meanwhile, a smaller number of agents imposes

Mechanism 4 Random Partition SWME (RP-SWME)

- 1: Partition agents into groups of two. If N is odd, leave one group with three agents.
 - 2: Run SWME for each group.
-

less restrictions in selecting the error rates, and thus each agent's wager can be fully leveraged in the randomization step. We would like to note that this is a very unique property of SWME: as both shown in Freeman, Pennock, and Wortman Vaughan [FPW17] and our experimental results, when the number of agents is small, existing wagering mechanisms (including DCA) all have low risk, i.e., have only a small portion of wager to lose in the worst case. This not only implies that SWME is particularly suitable for small group wagering but also points out a way to further improve the risk property of SWME, i.e. via randomly partitioning agents into smaller groups. We formally present the random partition SWME in Mechanism 4. We show in next section that RP-SWME achieves all properties (a)-(h).

1.6.3 Properties of SWME and RP-SWME

Theorem 1.4. *Both (SWME) and (RP-SWME) satisfy properties (a)-(g). (RP-SWME) satisfies (h).*

Proof. We prove the properties one by one.

(a) Individual rationality and (b) (strictly) incentive compatibility: First consider SWME. For an arbitrary profile of reports $\hat{\mathbf{p}}$ and wagers \mathbf{w} , Algorithm 3 outputs a profile \mathcal{E} of error rates of all agents. Denote by $\hat{\phi}_{\mathcal{E}}^i(\cdot)$ the corresponding surrogate function specified using the error rate profile \mathcal{E} for agent i . For each i and $j \in \mathcal{N}$:

$$\begin{aligned} \mathbb{E}_{X \sim p_i, \tilde{X}_j}[\hat{\phi}_{\mathcal{E}}^j \circ s_{\tilde{X}_j}(\hat{p}_j)] &= p_i \mathbb{E}_{\tilde{X}_j | X=1}[\hat{\phi}_{\mathcal{E}}^j \circ s_{\tilde{X}_j}(\hat{p}_j)] + (1 - p_i) \mathbb{E}_{\tilde{X}_j | X=0}[\hat{\phi}_{\mathcal{E}}^j \circ s_{\tilde{X}_j}(\hat{p}_j)] \\ &= p_i \cdot s_{X=1}(\hat{p}_j) + (1 - p_i) \cdot s_{X=0}(\hat{p}_j) = \mathbb{E}_{X \sim p_i}[s_X(\hat{p}_j)], \end{aligned}$$

using Lemma 1.2. Then, using the linearity of expectation, we have (here $\tilde{\mathbf{X}}$ encodes the randomness in Π_i^{SWME})

$$\begin{aligned} \mathbb{E}_{X \sim p_i, \tilde{\mathbf{X}}}[\Pi_i^{\text{SWME}}(\hat{\mathbf{p}}, \mathbf{w}, X)] &= \frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \left(\mathbb{E}_{X \sim p_i, \tilde{X}_i}[\hat{\phi}_{\mathcal{E}}^i \circ s_{\tilde{X}_i}(\hat{p}_i)] - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N} \setminus \{i\}}} \mathbb{E}_{X \sim p_i, \tilde{X}_j}[\hat{\phi}_{\mathcal{E}}^j \circ s_{\tilde{X}_j}(\hat{p}_j)] \right) \\ &= \mathbb{E}_{X \sim p_i} \left[\frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \left(s_X(\hat{p}_i) - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N} \setminus \{i\}}} s_X(\hat{p}_j) \right) \right] \\ &= \mathbb{E}_{X \sim p_i}[\Pi_i^{\text{WS}}(\hat{\mathbf{p}}, \mathbf{w}, X)]. \end{aligned}$$

Note the above holds for any possible reports ($\forall \mathcal{E}$). Thus the incentive properties, i.e., individual

rationality and strictly incentive compatibility of WSWM will preserve. The proof for RP-SWME is similar, with the only difference in that each agent's net-payoff is further averaged over the random partitions (but IR and SIC under each possible partition).

(c) Ex-post budget balance: This can be shown via writing down the sum of net-payoffs defined in Eqn. (1.3). Our note below Eqn. (1.1) also states that the budget balance property doesn't depend on the specific forms of the scoring functions therein. We formally present the deduction as follows:

$$\begin{aligned}
\sum_i \Pi_i^{\text{SWME}}(\hat{p}_i, w_i, \cdot) &= \sum_i \frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \left(\varphi \circ s_{\hat{x}_i}(\hat{p}_i) - \sum_{j \in \mathcal{N} \setminus \{i\}} \frac{w_j}{W_{\mathcal{N} \setminus \{i\}}} W_{\mathcal{N}} \cdot \varphi \circ s_{\hat{x}_j}(\hat{p}_j) \right) \\
&= \sum_i \left(\frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \varphi \circ s_{\hat{x}_i}(\hat{p}_i) - \sum_{j \neq i} \frac{w_j W_{\mathcal{N} \setminus \{j\}}}{W_{\mathcal{N}}} \cdot \frac{w_i}{W_{\mathcal{N} \setminus \{j\}}} W_{\mathcal{N}} \cdot \varphi \circ s_{\hat{x}_i}(\hat{p}_i) \right) \\
&= \sum_i \left(\frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \varphi \circ s_{\hat{x}_i}(\hat{p}_i) - \frac{w_i W_{\mathcal{N} \setminus \{i\}}}{W_{\mathcal{N}}} \varphi \circ s_{\hat{x}_i}(\hat{p}_i) \right) \\
&= 0.
\end{aligned}$$

The above also shows that for each group from the random partition of (RP-SWME), ex-post budget balance is satisfied. Thus, we also proved ex-post budget balance for (RP-SWME).

(d) Sybilproofness: In RP-SWME, any pair of agents with different beliefs have a positive probability to be partitioned into a sub-group. Applying Lemma 1.3, at least one of them loses all his wager in the worst case. Thus, by Definition 1.9, RP-SWME is PO.

Lemma 1.5. *If a (randomized) wagering mechanism \mathcal{W} is (weakly) budget-balanced, (weakly) incentive compatible, Sybilproof, then the mechanism \mathcal{W}^* that first uniformly randomly pairs agents in groups of two and then runs mechanism \mathcal{W} for each group is still Sybilproof.*

Proof. We prove the claim for the case that an agent is only allowed to create two identities. The claim holds in general, as we can always merge two identities into one without decreasing the payoff, following the result of the case of two.

Fixing an arbitrary belief \mathbf{p}_i of agent i , we denote the $E_i^{\mathcal{W}}(\hat{\mathbf{p}}, \mathbf{w}) := \mathbb{E}_{X \sim \mathbf{p}_i, \mathcal{D}^{\mathcal{W}}(\hat{\mathbf{p}}, \mathbf{w}, X)}[\Pi_i(\hat{\mathbf{p}}, \mathbf{w}, X = x)]$, where $\mathcal{D}^{\mathcal{W}}(\cdot)$ is the distribution specified by mechanism \mathcal{W} . Suppose an agent i divides its wager w_i into two wagers w_{i1}, w_{i2} , and reports two predictions $\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_{i2}$ correspondingly. We have

$\forall \hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_{i2}, w_{i1}, w_{i2}, \hat{\mathbf{p}}_{-i}, \mathbf{w}_{-i}, x,$

$$\begin{aligned}
& E_i^{\mathcal{W}^*}(\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_{i2}, \hat{\mathbf{p}}_{-i}, w_{i1}, w_{i2}, \mathbf{w}_{-i}) \\
&= \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_j, w_{i1}, w_j) + \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\hat{\mathbf{p}}_{i2}, \hat{\mathbf{p}}_j, w_{i2}, w_j) + \frac{1}{N} \left(E_{i1}^{\mathcal{W}}(\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_{i2}, w_{i1}, w_{i2}) + E_{i2}^{\mathcal{W}}(\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_{i2}, w_{i1}, w_{i2}) \right) \\
&\leq \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\hat{\mathbf{p}}_{i1}, \hat{\mathbf{p}}_j, w_{i1}, w_j) + \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\hat{\mathbf{p}}_{i2}, \hat{\mathbf{p}}_j, w_{i2}, w_j) \quad (\mathcal{W} \text{ is (weakly) budget balance}) \\
&\leq \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\mathbf{p}_i, \hat{\mathbf{p}}_j, w_{i1}, w_j) + \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\mathbf{p}_i, \hat{\mathbf{p}}_j, w_{i2}, w_j) \quad (\mathcal{W} \text{ is (weakly) incentive compatible}) \\
&\leq \sum_{j \neq i} \frac{1}{N} E_i^{\mathcal{W}}(\mathbf{p}_i, \hat{\mathbf{p}}_j, w_i, w_j) \quad (\mathcal{W} \text{ is sybilproof}) \\
&\leq \sum_{j \neq i} \frac{1}{N-1} E_i^{\mathcal{W}}(\mathbf{p}_i, \hat{\mathbf{p}}_j, w_i, w_j) = E_i^{\mathcal{W}^*}(\mathbf{p}_i, \hat{\mathbf{p}}_{-i}, w_i, \mathbf{w}_{-i})
\end{aligned}$$

Therefore, \mathcal{W}^* is sybilproof. \square

(e) Anonymity: For SWME, this proof can follow from the fact that the randomness (error rate selection) in SWME and tRP-SWME depends only on the reports and wagers of agents and do not depend on the identities of agents and the fact that the expected net-payoffs of agents are the same with those of WSWM (Corollary 1), which is anonymous [Lam+08]. RP-SWME only adds a random partition of agents in SWME and the partition does not depend on the identities of agents. Thus, RP-SWME is also anonymous.

(f) Neutrality: For SWME, this proof can follow from the fact that the randomness (error rate selection) in SWME and tRP-SWME depends only on the reports and wagers of agents and do not depend on the labeling of the outcomes and the fact that the expected net-payoffs of agents are the same with those of WSWM (Corollary 1), which is neutral [Lam+08]. RP-SWME only adds a random partition of agents in SWME and the partition does not depend on the labeling of the outcomes. Thus, RP-SWME is also neutral.

(g) Non-arbitrage opportunity: Now we prove that SWME does not allow arbitrage opportunity. The idea is simple and straight-forward: fix the set of prediction \mathbf{p}_{-i} and wagers \mathbf{w} . First notice the fact that under each possible realization $\tilde{x}_i, \tilde{x}_{-i}$ can be any possible realizations. Since $s_{\tilde{x}_i=1}(p_i)$ and $s_{\tilde{x}_i=0}(p_i)$ have opposite monotonicity, we know there does not exist an interval for risklessly predictions.

The above non-arbitrage opportunity is *ex-post*, but the arbitrage opportunity persists when agents evaluate the conditional expectation of his score with respect to the random flipping step (which is the

same as WSWM), which remains a concern when each agent participates in multiple event forecasts. This concern will be resolved when we apply the idea of surrogate wagering to the non-arbitrage wagering mechanism (NAWM). For details please refer to Section 1.7.1 .

For RP-SWME, it runs SWME on each pair of agents after the random partition. Therefore, agents also have no arbitrage opportunity.

(h) Pareto optimality: In RP-SWME, any pair of agents with different beliefs have a positive probability to be partitioned into a sub-group. Applying Lemma 1.3, at least one of them loses all his wager in the worst case. Thus, by Definition 1.9, RP-SWME is PO. \square

1.6.4 Wager with noisy ground truth

The above method also points out a way to implement a wagering mechanism with a noisy ground truth, as SWM is able to remove the noise in outcomes in expectation. The ability to wager with noisy ground truth provides informative information to agents who participated in a wagering mechanism immediately only when a noisy copy of outcome is available. We present the key idea below, while not re-defining all properties w.r.t. \hat{X} instead of X - the changes are rather straight-forward.

Suppose we know a noisy estimate \hat{X} on X , and denote the error rate of \hat{X} as \hat{e}_1, \hat{e}_0 (which we know, and agents trust us in knowing these two numbers), we will be able to reproduce our surrogate wager mechanism by plugging $\hat{X}, \hat{e}_1, \hat{e}_0$ into Eqn. (1.2), if we ignore the PO property for now. We similarly will have the wager violation issue pointed out earlier - we however do not have the control of the error rates directly. An easy fix is via the following affine transformation of the wagering scores: suppose under the worst case, the random flipping will incur $-scale \cdot w_i$ wager score (net-payoff) with $scale > 1$. We can then rescale every agent's wager score by $1/scale$. Note the above affine transformation does not affect the incentive and other properties of the original surrogate wagering mechanism, as $\mathbb{E}[\varphi \circ \Pi_i^{WS}(\cdot)] = \frac{1}{scale} \cdot \mathbb{E}[\Pi_i^{WS}(\cdot)]$.⁷ To achieve PO, we can further random partition agents into groups of two and flip on \hat{X} according to certain error rates \hat{e}_0^i, \hat{e}_1^i for each agent i . Let \tilde{X}_i be the flipped outcome.

⁷We didn't apply the scaling in SWME when there exists other options, as the scaling will effectively decrease the expected payment of each agent.

We can establish the error rates of \tilde{X}_i w.r.t. the ground truth X and \hat{e}_0^i, \hat{e}_1^i by following equations:

$$\begin{aligned} \mathbb{P}(\tilde{X}_i = 1|X = 0) &= \sum_{x \in \{0,1\}} \mathbb{P}(\tilde{X}_i = 1, \hat{X} = x|X = 0) \\ &= \sum_{x \in \{0,1\}} \mathbb{P}(\tilde{X}_i = 1|\hat{X} = x, X = 0) \cdot \mathbb{P}(\hat{X} = x|X = 0) \\ &= \hat{e}_0^i \cdot (1 - \hat{e}_0) + (1 - \hat{e}_1^i) \cdot \hat{e}_0, \end{aligned}$$

and similarly $\mathbb{P}(\tilde{X}_i = 0|X = 1) = \hat{e}_1^i \cdot (1 - \hat{e}_1) + (1 - \hat{e}_0^i) \cdot \hat{e}_1$. It's easy to see that when $\hat{e}_1 + \hat{e}_0 \neq 1$, we can tune the error rates of \tilde{X} via tuning \hat{e}_1^i, \hat{e}_0^i . This step corresponds to the error selection step in SWME, i.e., Algorithm 3.

1.7 Extensions of SWM

We discuss a couple of useful extensions of SWM: i). one is instead of building on WSWM, we show the idea of surrogate idea can also build upon another deterministic wagering mechanism NAWM. (ii). We extend our results to a multi-outcome setting.

1.7.1 Surrogate NAWM

We note that the bias removal procedure adopted in SWM does not rely the specific underlying wagering mechanism heavily. We demonstrate the idea with a non-arbitrage wagering mechanism (NAWM [Che+14])⁸.

Notice that since $\Pi_i^{\text{NA}}(\cdot)$ is not linear in the surrogate scores of each agent, the budget balance argument is not as easy as in the WSWM case. Nonetheless we notice the following fact proved in Chen et al. [Che+14]:

$$\Pi_i^{\text{NA}}(\hat{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, X = x) = \Pi_i^{\text{WS}}(\hat{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, X = x) - \Pi_i^{\text{WS}}(\bar{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, X = x)$$

where \bar{p}_i denotes the average prediction from $j \neq i$. Then we can safely apply the surrogate idea to the first WSWM scoring term:

$$\varphi \circ \Pi_i^{\text{NA}}(\hat{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, \tilde{X} = \tilde{x}) = \varphi \circ \Pi_i^{\text{WS}}(\hat{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, \tilde{X} = \tilde{x}) - \Pi_i^{\text{WS}}(\bar{p}_i, \hat{\mathbf{p}}_{-i}, \mathbf{w}, X = x)$$

⁸Though the randomization device already grants us the non-arbitrage property, we pick this mechanism for i. its simplicity for presentation, as NAWM also extends from WSWM. ii. we will show in experiments later that we empirically observe higher risk when applying this surrogate based randomized NAWM.

This mechanism will enjoy the higher risk property introduced by surrogate wagering, as well as the non-arbitrage (in conditional expectation) brought in by NAWM.

1.7.2 Multi-outcome events

For simplicity, our previous discussions focused largely on the binary outcome scenario. As promised, we now show that our results extend to the non-binary events. Recall that there are M outcomes, denoting as $[0, 1, 2, \dots, M - 1]$. Denote the following confusion matrix

$$C = \begin{bmatrix} c_{0,0} & c_{0,1} & \dots & c_{0,M-1} \\ c_{1,0} & c_{1,1} & \dots & c_{1,M-1} \\ \dots & \dots & \dots & \dots \\ c_{M-1,0} & c_{M-1,1} & \dots & c_{M-1,M-1} \end{bmatrix}$$

and each entries $c_{j,k}$ indicates the flipping probability for generating a surrogate outcome: $c_{j,k} = \Pr[\tilde{X}_i = k | X = j]$.

The core challenge of this extension is to find an unbiased operator φ . Writing out the conditions for unbiasedness (s.t. $\mathbb{E}_{\tilde{X}_i|x}[\varphi \circ s_{\tilde{X}_i=\tilde{x}_i}(\hat{\mathbf{p}})] = s_x(\hat{\mathbf{p}})$), we need to solve the following set of functions to obtain $\varphi(\cdot)$ (short-handing $\varphi \circ s_x(\hat{\mathbf{p}})$ as $\varphi_x(\hat{\mathbf{p}})$):

$$\begin{aligned} s_0(\hat{\mathbf{p}}) &= c_{0,0} \cdot \varphi_0(\hat{\mathbf{p}}) + c_{0,1} \cdot \varphi_1(\hat{\mathbf{p}}) + \dots + c_{0,M-1} \cdot \varphi_{M-1}(\hat{\mathbf{p}}) \\ s_1(\hat{\mathbf{p}}) &= c_{1,0} \cdot \varphi_0(\hat{\mathbf{p}}) + c_{1,1} \cdot \varphi_1(\hat{\mathbf{p}}) + \dots + c_{1,M-1} \cdot \varphi_{M-1}(\hat{\mathbf{p}}) \\ &\dots \\ s_{M-1}(\hat{\mathbf{p}}) &= c_{M-1,0} \cdot \varphi_0(\hat{\mathbf{p}}) + c_{M-1,1} \cdot \varphi_1(\hat{\mathbf{p}}) + \dots + c_{M-1,M-1} \cdot \varphi_{M-1}(\hat{\mathbf{p}}) \end{aligned}$$

Denote by $\mathbf{s}(\hat{\mathbf{p}}) = [s_0(\hat{\mathbf{p}}); s_1(\hat{\mathbf{p}}); \dots; s_{M-1}(\hat{\mathbf{p}})]$, and $\prime(\hat{\mathbf{p}}) = [\varphi_0(\hat{\mathbf{p}}); \varphi_1(\hat{\mathbf{p}}); \dots; \varphi_{M-1}(\hat{\mathbf{p}})]$. Then the above equation becomes equivalent with the following system of equation: $\mathbf{s}(\hat{\mathbf{p}}) = C \cdot \prime(\hat{\mathbf{p}})$. Choose a C with full rank. For instance when $M > 2$ we can set $\forall j, c_{j,j} = \frac{1}{2}$, $c_{j,k} = \frac{1}{2(M-1)}$, $k \neq j$ - not hard to verify that such a C is indeed full rank. Then we are ready to solve for $\prime(\hat{\mathbf{p}})$ as follows:

$$\prime(\hat{\mathbf{p}}) = C^{-1} \cdot \mathbf{s}(\hat{\mathbf{p}}). \quad (1.4)$$

With defining above unbiased surrogate operator, all other discussions generalize fairly straightforwardly - such a φ will give us the same equation as established in the lemma below for the non-binary event outcome setting:

Lemma 1.6. *Define $\varphi(\cdot)$ as in Eqn. (1.4), and flip \tilde{X}_i using C, x . Then $\mathbb{E}_{\tilde{X}_i|x}[\varphi \circ s_{\tilde{X}_i=\tilde{x}_i}(\hat{\mathbf{p}})] = s_x(\hat{\mathbf{p}})$.*

We include a detailed example of φ for three-outcome events below.

Example of φ for three-outcome events

Example 1.1. *An example with $M = 3$. Suppose we flip the outcome using the uniform-error confusion matrix:*

$$C = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \Rightarrow C^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

Therefore we obtain a closed-form of φ :

$$\begin{aligned} \varphi_0(\hat{\mathbf{p}}) &= 3\mathbf{s}_0(\hat{\mathbf{p}}) - \mathbf{s}_1(\hat{\mathbf{p}}) - \mathbf{s}_2(\hat{\mathbf{p}}) \\ \varphi_1(\hat{\mathbf{p}}) &= -\mathbf{s}_0(\hat{\mathbf{p}}) + 3\mathbf{s}_1(\hat{\mathbf{p}}) - \mathbf{s}_2(\hat{\mathbf{p}}) \\ \varphi_2(\hat{\mathbf{p}}) &= -\mathbf{s}_0(\hat{\mathbf{p}}) - \mathbf{s}_1(\hat{\mathbf{p}}) + 3\mathbf{s}_2(\hat{\mathbf{p}}) \end{aligned}$$

1.8 Evaluation

In this section, we evaluate LWS and RP-SWME with extensive simulations. We first compare the efficiency of LWS and RP-SWME with that of other existing deterministic (weakly) incentive compatible mechanisms WSWM, NAWM and DCA. The results show that the two randomized wagering mechanisms outperform the three deterministic wagering mechanism. Then, we compare the variance of payoff and the probability of winning money within the two randomized wagering mechanisms. The results show that RP-SWME is better than LWS in these two matrices.

1.8.1 Simulation Setup

We simulate both the binary events and the multi-outcome events. For binary events, we generated six sets of agents' predictions and wagers according to the combinations of three different prediction models and two different wager models. With a little abuse of notation, we denote that an event happens with probability q and that agent i believes that the event to predict will happen with probability p_i and will not happen with probability $1 - p_i$. We use three models to generate predictions $p_i, i \in \mathcal{N}$:

1. Uniform model: For each event, p_i is independently drawn from a uniform distribution over $[0, 1]$.
2. Logit-Normal model: This model assumes that p_i , when being mapped to the real line by a logit function as $\log\left(\frac{p_i}{1-p_i}\right)$, is independently drawn from a Normal distribution $\mathcal{N}(\log(\frac{q}{1-q})^{1/\alpha}, \sigma^2)$,

i.e., $p_i \sim \text{Logit-Normal} \left(\log\left(\frac{q}{1-q}\right)^{1/\alpha}, \sigma^2 \right)$. q, α, σ^2 are model parameters. This model is proposed and used to estimate the happening probability of the event in Satopää et al. [Sat+14a], where q is regarded as an estimator of the happening probability and α models the under-confident effect on human forecasters. Based on a real prediction dataset over 1300 forecasters and 69 geopolitical events collected in Satopää et al. [Sat+14a], this model outperforms most existing models to estimate the happening probability of events, which leads us to believe this model a good alternate to generate prediction data. In our simulations, we adopted $\alpha = 2$, which best fits the aforementioned real prediction dataset, $\sigma^2 = 1$, and q is drawn uniformly from $[0, 1]$ for each event.

3. Synthetic model: this synthetic model is introduced from a set of simulation studies in Ranjan and Gneiting [RG10], Allard, Comunian, and Renard [ACR12], and Satopää et al. [Sat+14a]. The model assumes that the happening probability of an event to be predicted by N is given by $q = \Phi\left(\sum_{i=1}^N u_i\right)$, where Φ is the cumulative distribution function of a standard normal distribution and u_i is independently drawn from $\mathcal{N}(0, 1)$. Each agent knows the true probability generating model and u_i but not $u_j, \forall j \neq i$. Accordingly, each agent's calibrated belief of the happening probability of the event is given by $p_i = \Phi\left(\frac{u_i}{\sqrt{2N-1}}\right)$.

We use two models to generate the wagers of agents:

1. Uniform model: All agents' wagers are equal to 1.
2. Pareto model: This model assumes that the wager w_i of agent i follows the Pareto distribution, which is often adopted to model the distribution of wealth in a population. In the simulations of Freeman, Pennock, and Wortman Vaughan [FPW17], the authors selected the shape parameter and scale parameter of the Pareto distribution as 1.16 and 1 correspondingly, which is the distribution depicted as "20% of the population has 80% of the wealth". We adopted the same parameters for comparison purpose.

For events with multiple outcomes, we simulated three sets of data with the number of possible outcomes 3, 6, 9 correspondingly. In each set, we drew the predictions from uniform distribution over the whole probability space and drew the wagers according to the Uniform model.

1.8.2 Comparison of efficiency of wagering mechanisms

We show that LWS and RP-SWME are more efficient than existing deterministic (weakly) incentive compatible mechanisms WSWM, NAWM and DCA. We evaluate the efficiency by two metrics: *Average*

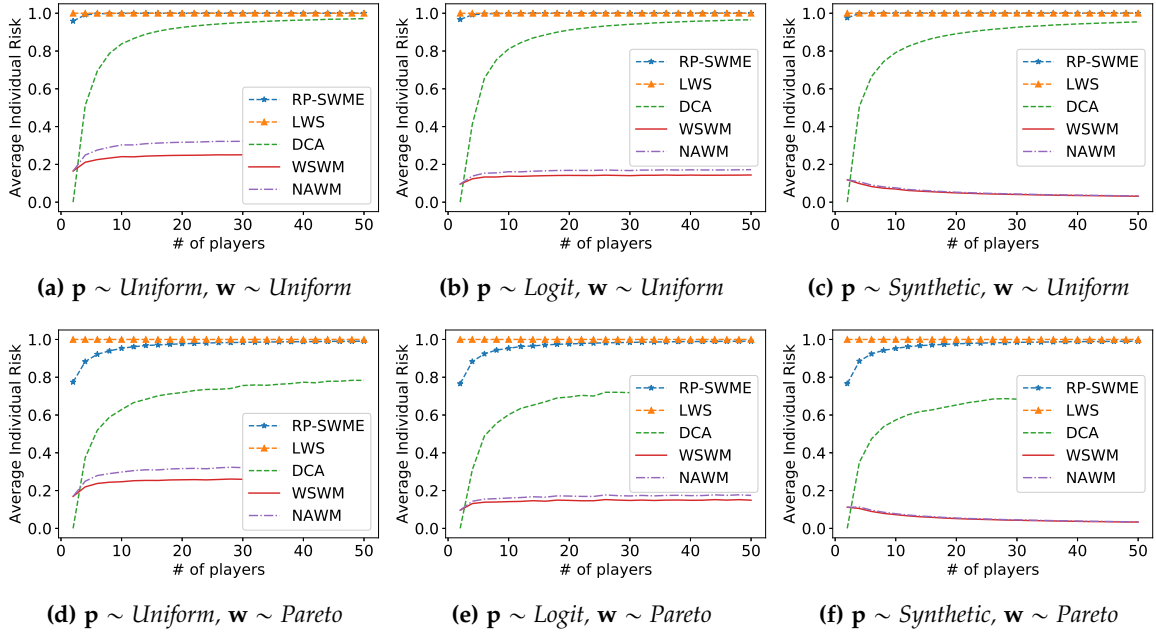


Figure 1.1: Average individual risk of each of five wagering mechanisms as a function of N under different prediction and wager models

individual risk and Average money exchange rate.

Individual risk is the percent of wager that an individual agent can lose in the worst case w.r.t. the event outcome and the randomness of the mechanisms. The average individual risk is an indicator of Pareto optimality, because the average individual risk equal to 1 (i.e., no one can commit to secure a positive wager before the wagering game) is a sufficient condition of Pareto optimality. Money exchange rate is the total amount of money exchanged in the game after the outcome of a wagering mechanism is realized, divided by the total amount of wagers. Average money exchange rate measures the efficiency of an average wagering game.

In our simulations, we vary the number of agents for 2 to 50 with a step of 2. For each number of agents, we randomly generate 1000 events and the agents' predictions and wagers for each of the six combinations of prediction models and wager models, and take the average of individual risk and money exchange rate over the 1000 events. When calculating the money exchange, we use the expectation of the money exchange over all possible outcomes according to the happening probability of each outcome. This happening probability is either specified in the model generating the predictions, or otherwise, drawn from a uniform distribution over the corresponding probability space.

In the simulations, both RP-SWME and LWS achieve the highest average individual risk (approximately 1) under all conditions (# of outcomes, # of agents, prediction models, and wager models) we

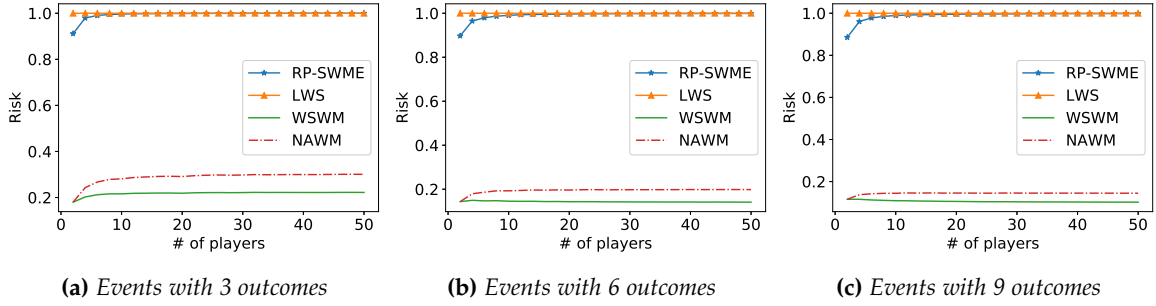


Figure 1.2: Average individual risk of each of four mechanisms under events with multiple outcomes

simulated (Figure 1.1, 1.2). In contrast, the best of the deterministic mechanisms DCA, only achieves an approximate 1 average individual risk when the wagers of agents are uniform and the number of participants is more than 30 (Figure 1.1a-1.1c). Its average individual risk drops to 0.6 when the wagers of agents follows the Pareto distribution (Figure 1.1d-1.1f). This result shows that the two randomized mechanisms effectively remove the opportunity for side bet and take use of all the wagers before the outcome is realized.

LWS doubles the money exchange rate of the second best alternative, hitting a more than 80% money exchange rate under all conditions we simulated (Figure 1.3, 1.4). On the other hand, RP-SWME also defeats the other two incentive compatible deterministic wagering mechanisms in expected money exchange under all conditions we simulated (Figure 1.3, 1.4). Meanwhile, it also outperforms DCA when the number of agents is small (Figure 1.3).

In particular, when the prediction follows the synthetic model, where the predictions are much closer to each other as the number of participants increases, the money exchange rate of the two incentive compatible deterministic wagering mechanisms, WSWM and NAWM converge to zero. However, the two randomized wagering mechanisms still keep a large money exchange rate (Figure 1.3c, 1.3f).

1.8.3 Comparison of randomness properties of RP-SWME and LWS

In this section, we compare the *standard variance* of payoffs and the *probability of not losing money* of RP-SWME and LWS. We evaluated these two metrics w.r.t. to the prediction accuracy, which is measured based on the distance of a prediction to the outcome, i.e., $\text{Accuracy} = 1 - |x - p_i|^9$.

In the evaluation, we run 10000 wagering instances under these two mechanisms and recorded the

⁹We use it as measurement of accuracy for two reasons: i. it is linear in prediction p_i , ii. it has an inject to Brier Score

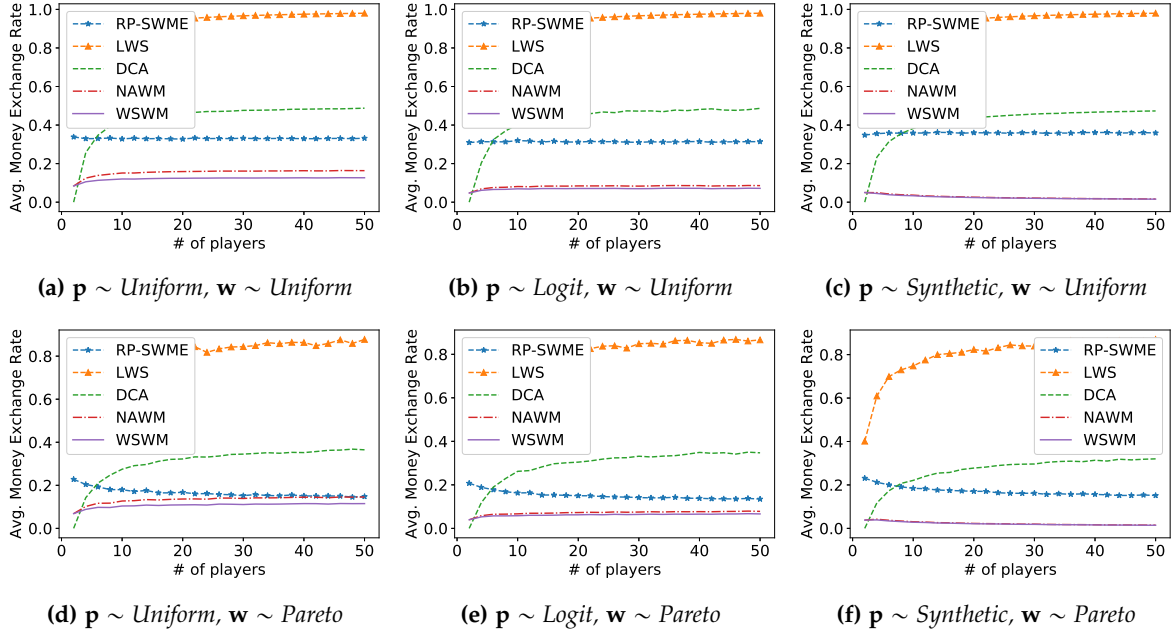


Figure 1.3: Average money exchange rate of each of five wagering mechanisms as a function of N under different prediction and wager models

prediction accuracy of each agent in each instance and the corresponding net-payoff. Then, we group these agents into 10 groups that correspond to 10 consecutive accuracy intervals. In each group, we calculate the standard variance and the percent of agents winning money. For fair comparison, we normalize the net-payoff of each agent by its own wager.

We simulate binary events. We generate two set of simulated data. In both sets, we varied the number of agents from 2 to 50 with a set of 2, and under each number, we generated 10000 instances. In each instance, the agents' predictions are drawn from the Uniform model, while the wagers are drawn from the Uniform model in one set and drawn from the Pareto model in the other set.

Our results show that under all conditions we simulate, RP-SWME has a much smaller variance in agents' net-payoff and the variance is steady across agents with different prediction accuracy. In contrast, the LWS has a much larger variance in net-payoff, which increases with the prediction accuracy (Figure 1.5). On the other hand, RP-SWME has a much larger probability of not losing money and this probability increases with the prediction accuracy, while LWS has a much smaller such probability (Figure 1.6). In brief, while both RP-SWME and LWS can effectively improve the efficiency of wagering, RP-SWME provides much less uncertainty than LWS does and thus, may be regarded as a more attractive alternative for deterministic wagering mechanisms.

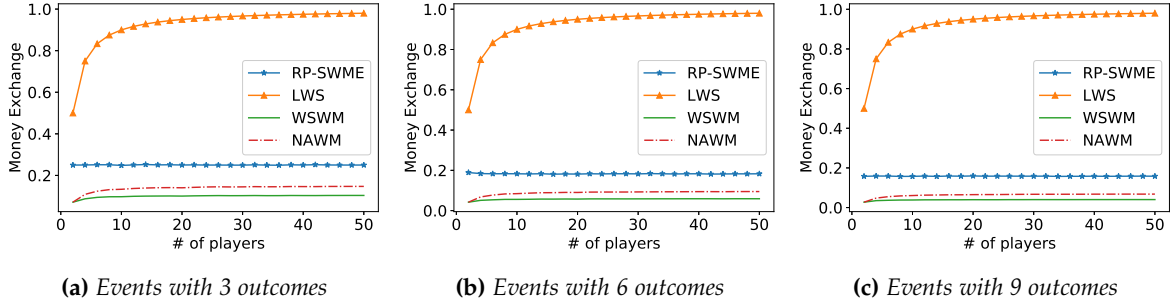


Figure 1.4: Average money exchange rate of each of four mechanisms under events with multiple outcomes

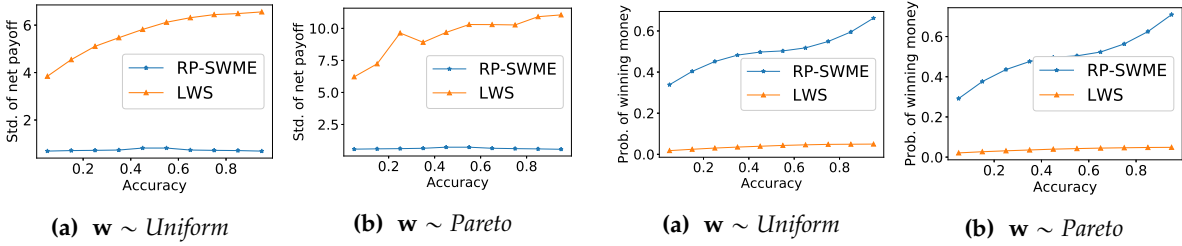


Figure 1.5: Std. variance of net-payoff as a function of prediction accuracy: RP-SWME v.s. LWS

Figure 1.6: Probability of winning money as a function of prediction accuracy: RP-SWME v.s. LWS

1.9 Conclusion

We extend the design of wagering mechanism to its randomized space. We propose two of them: Lottery Wagering Mechanisms (LWM) and Surrogate Wagering Mechanisms (SWM). We demonstrate the power of randomness by theoretically proving that they both satisfy a set of desirable properties, including Pareto efficiency which is missing in exiting wagering literature. We also carried out extensive experiments to support our theoretical findings. SWM is also robust to noisy outcomes. In particular, as shown by simulations, surrogate wagering mechanisms have reasonably small standard variance in agents' payoff and low probability for agents to lose all their wagers.

Chapter 2

Surrogate Scoring Rules

2.1 Introduction

Accurate assessment of random variables of interest (e.g. how likely the S&P 500 index will go up next week) plays a crucial role in a wide array of applications, including computational finance [DP97], geopolitical forecasting [Tet+14; Fri+18], weather and climate forecasting [GR05], and the prediction of the replicability of social science studies [Alt+19; HSW19]. Since such assessments are often elicited from people, how to incentivize people to provide accurate assessments has been a topic of great scientific interests.

For settings where the principal will have access to the ground truth (e.g. after a week, knowing whether the S&P 500 index actually went up), strictly proper scoring rules (SPSR) [Bri50a; Win69; Sav71; JNW06; GR07b] have been developed to elicit probabilistic assessments and evaluate them against the ground truth. SPSR have two desirable properties. First, they incentivize truthful information reporting: the SPSR score of an agent's reported prediction is strictly maximized in the agent's expectation if she truthfully reveals her prediction. Second, the SPSR score of a prediction measures the quality of the prediction in the sense that the closer the prediction is to the underlying, unknown true distribution of the random event, the higher the expected score.

However, in many applications, the ground truth is not available in time or at all. For example, geopolitical events usually take months to resolve [Tet+14], and whether a study will be successfully replicated is not known if a replication test of it is not attempted. In this chapter, we extend the literature of SPSR to the information elicitation *without* verification (IEWV) settings, where the principal has no access to the ground truth and still wants to elicit private probabilistic beliefs. We ask the following

research question:

Can we extend SPSR to scoring mechanisms that can achieve truthful elicitation of probabilistic information and quantify the quality of the elicited information for IEWV?

Witkowski et al. [Wit+17] explored this question in a single-task setting (i.e., having a single random variable of interest to predict). When an unbiased proxy to the true probability distribution of the ground truth is available, they generalized SPSR to *proper proxy scoring rules*, which score a prediction against the unbiased proxy while maintaining the two properties of SPSR. However, when the principal only has access to agents' reports, it remains an open problem how such an unbiased proxy can be constructed without affecting the incentive properties.

In this chapter, we study the research question in a multi-task setting, where a principal wants to predict multiple random variables that are similar a priori. We provide a positive answer to the question. In our solution, the principal only needs to know the order of the prior probability of each possible outcome (e.g., for binary random variables, the more likely outcome) and does not need to have an unbiased proxy for each task. Specifically, we develop a family of scoring mechanisms that utilize the similarity of tasks and the conditional independence of agents' beliefs to construct a biased proxy of the ground truth, and then, score a prediction against this proxy by removing the bias w.r.t. the underlying SPSR that one wants to recover. Our proxy is explicitly constructed only from agents' reported predictions. As a result, we achieve the dominant uniform strategy truthfulness [GF19b] in eliciting probabilistic predictions, where truthful reporting is the strict best strategy when each agent adopts the same strategy across all tasks. Furthermore, the scores of our mechanisms recover the scores of SPSR in expectation. To the best of our knowledge, our work provides the first meta solution that enables applications of any SPSR to the IEWV setting without relying on access to unbiased proxies of the ground truth. We name our solution *Surrogate Scoring Rules (SSR)*.

As a building block, we first introduce SSR for a stylized setting where the principal has access to a noisy estimate of the ground truth, as well as the estimate's error rates, to evaluate the elicited information. We show that SSR preserve the same information quantification and truthful elicitation properties just as SPSR, despite the lack of access to the ground truth. These surrogate scoring rules are inspired by the use of surrogate loss functions in machine learning [AL88; Byl94; Sco+13; Nat+13; Sco15]. They remove the bias from the noisy estimate of the ground truth such that in expectation a report is as if evaluated against the ground truth.

Building on the above bias correction step, when the principal only has access to agents' reports and the order of the prior probabilities of each outcome, we develop the *SSR mechanisms* for the multi-task

setting to achieve information quantification and the dominant uniform strategy truthfulness when the principal has sufficiently many tasks and agents. Our mechanisms rely on an estimation procedure to accurately estimate the average bias in the peer agents' reports. With the estimation, a random peer agent's report can serve as a noisy estimate of the ground truth, and SSR can then be applied to achieve the two desired properties. We evaluate the empirical performance of the SSR mechanisms using 14 real-world human forecast datasets. The results show that SSR effectively recover SPSR scores but using only agents' reports.

We summarize our contributions as follows:

- We extend SPSR to a family of scoring mechanisms, the SSR mechanisms, that operate in the IEWV setting. The SSR mechanisms only require access to peer reports and the order of the prior probabilities of the ground truth being each outcome, and they can truthfully elicit probabilistic beliefs. An SSR mechanism can build upon any SPSR and quantifies in expectation the value of the elicited information just as the corresponding SPSR does as if it had access to the ground truth. Therefore, our work complements the proper scoring rule literature and expands the application of SPSR in challenging elicitation settings where the ground truth is unavailable.
- For the IEWV setting, most existing mechanisms focus on incentivizing truthful reporting of categorical signals via rewarding the correlation between two agents' reports. Our SSR mechanisms complement this literature from two perspectives. First, SSR mechanisms induce dominant uniform strategy truthfulness in eliciting probabilistic predictions instead of categorical signals. Second, instead of scoring a prediction by assessing the correlation between two agents' reports, SSR mechanisms score predictions according to their prediction accuracy against the unknown ground truth. This property encourages agents to search for more accurate forecasts.
- We evaluate the empirical performance of SSR mechanisms on 14 real-world human prediction datasets. The results show that SSR mechanisms can better reflect the true accuracy of agents in terms of SPSR scores than other existing mechanisms designed for IEWV.

2.2 Related work

The most relevant literature to this work is on *strictly proper scoring rules* (SPSR) and *peer prediction*. SPSR are designed to elicit subjective beliefs about random variables when the principal can evaluate agents' predictions after the random variables are realized. Brier [Bri50a] proposed the widely used Brier score to quantify the quality of forecasts. Subsequent work studied other SPSR and developed

several characterizations of SPSR [Win69; Sav71; JNW06; GR07b].

Peer prediction refers to a collection of mechanisms developed for incentivizing truthful reporting in IEWV. Our SSR mechanisms are additions to this collection. The core idea of peer prediction is to leverage peer reports as references to score an agent’s report. The pioneer work [MRZ05b] considered a single-task elicitation setting where each agent observes a private signal associated with a single task of interest, and a principal who knows the joint distribution of these signals wants to elicit the exact realizations of the signals. It proposed the first mechanism where truthful reporting is a Bayesian Nash Equilibrium (BNE). Following this work, Jurca and Faltings [JF07; JF09] proposed mechanisms where truthful reporting is a BNE with a strictly higher payment than any other pure-strategy equilibrium. Kong, Ligett, and Schoenebeck [KLS16] proposed a mechanism, in which truthful reporting is the BNE with the highest payoff for agents among all equilibria on a binary-outcome task. Frongillo and Witkowski [FW16] characterized all mechanisms that admit a truthful reporting equilibrium in this setting. Another research thread for single-task elicitation asks agents to answer additional questions in addition to providing their signal. The Bayesian Truth Serum [Pre04b] additionally asks the agents to report their beliefs about other agents’ reports and then uses this additional information to score the answer of each agent. The advantage of this approach is that the principal needs not to know the joint distribution of agents’ signals and that the additional information can be used to identify the correct answer to the question [PSM17]. However, this approach introduces extra work for the agents. For interested readers, this line of research has been further developed by other studies [RF13a; WP12b; Ril14; SY20b].

To relax the requirement on the principal’s knowledge of the signal distribution, many recent peer prediction studies have focused on a multi-task setting, where there exists a set of i.i.d. tasks, allowing the principal to leverage the statistical patterns in agents’ reports to incentivize truthful reporting. Our work falls into this category. The multi-task setting was simultaneously developed by Dasgupta and Ghosh [DG13] and Witkowski and Parkes [WP13]. The latter was the first to explicitly estimate relevant aspects of agents’ belief models from agents’ reports (which this work also uses), while the former achieves provably stronger equilibrium properties. In the mechanism of Dasgupta and Ghosh [DG13], the truthful reporting equilibrium has the highest expected payoff for agents among all equilibria when eliciting binary signals. Radanovic, Faltings, and Jurca [RFJ16] and Shnayder et al. [Shn+16] extended the mechanism of Dasgupta and Ghosh [DG13] to elicit categorical signals while maintaining the same incentive property. More recent studies have achieved the dominant uniform strategy truthfulness in the multi-task setting. Parallel to our work, Kong and Schoenebeck [KS19] developed a framework to design mechanisms to elicit general signals as long as certain notions of mutual information can be

estimated from agents’ reports. Their mechanisms, which includes the mechanism of Shnayder et al. [Shn+16] as a special case, are dominant uniform strategy truthful when there is an infinite number of tasks. Kong [Kon20] further achieved this truthfulness property with a finite number of tasks for eliciting categorical signals. Kong et al. [Kon+20] and Schoenebeck and Yu [SY20a] proposed dominant uniform strategy truthful mechanisms to elicit continuous signals with normal distributions and with general full-support marginal distributions, respectively. When there is a noisy estimate of the ground truth with a known confusion matrix, Goel and Faltings [GF19b] proposed a mechanism that also achieves the dominant uniform strategy truthfulness; the reward of an agent in the mechanism is an affine transformation of the the agent’s correctness rate over all classes. In comparison, our dominant uniform strategy truthfulness mechanisms focus on eliciting posterior beliefs of the ground truth and the rewards in our mechanisms recover in expectation the accuracy of agents in terms of the SPSR. Instead of assuming availability of an estimate of the confusion matrix, we construct an estimate from the agents’ reports, assuming the principal knows the order of the prior probabilities of each possible outcome of the ground truth.

There are a few studies also focusing on eliciting probabilistic predictions like this work. Among these studies, Witkowski and Parkes [WP12a] and Radanovic and Faltings [RF14] consider single-task elicitation and ask agents to report additional information as required by the Bayesian Truth Serum [Pre04b]. The two mechanisms proposed make truthful reporting an ex-post equilibrium and a BNE, respectively. Kong and Schoenebeck [KS18] provided a mechanism to elicit probabilistic predictions for the multi-task setting. Although truthful reporting is an equilibrium strategy under their mechanism, the mechanism is not dominant uniform strategy truthful. When the principal has access to an unbiased proxy of the ground truth, the proxy scoring rules developed by Witkowski et al. [Wit+17] can be used to elicit probabilistic predictions for the single-task setting as what SPSR offer with access to the ground truth. In this case, proxy scoring rules score a prediction against the unbiased proxy using a SPSR, and the expected score is equal to the expected score given by the SPSR using the ground truth up to a positive affine transformation [FK19]. In comparison, our mechanisms also offer a meta approach to recover the score for any SPSR. Our mechanisms do not require access to an unbiased proxy but a set of i.i.d. tasks.

Finally, our work borrows ideas from the machine learning literature on learning with noisy data [e.g. Nat+13; FV14; Sco15; RW15]. At a high level, our goal aligns with the goal in learning from noisy labels – both aim to evaluate a prediction when the ground truth is missing, but instead a noisy signal of the ground truth is available. Our work addresses the additional challenge that the error rate of the noisy signal remains unknown a priori.

2.3 Preliminaries

Before we introduce our model of information elicitation without verification, we first briefly introduce strictly proper scoring rules (SPSR), which are designed for the well-studied information elicitation with verification settings. We highlight two nice properties of SPSR: i. SPSR quantify the value of information and ii. SPSR is incentive compatible for elicitation. Our goal is to develop scoring rules that match these properties for the more challenging without verification settings. Our solutions build upon the understanding of SPSR.

SPSR are designed for eliciting subjective distributions of random variables when the principal can reward agents after the realization of the random variables. SPSR apply to eliciting predictions for any random variables, but we introduce them for binary random variables in this section because most of this chapter focuses on the binary case. Let $Y \in \{0, 1\}$ represent a binary event. An agent has a subjective belief $p \in [0, 1]$ for the likelihood of $Y = 1$. When the agent reports a probabilistic prediction $q \in [0, 1]$ of Y being 1, the principal rewards the agent using a scoring function $S(q, y)$ that depends on both the agent's report q and the realized outcome of Y . Strict properness of $S(\cdot, \cdot)$ is defined as follows.

Definition 2.1. *A function $S : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ that maps a reported belief q and the ground truth Y into a score is a strictly proper scoring rule if it satisfies $\mathbb{E}[S(p, Y)] > \mathbb{E}[S(q, Y)]$, for all $p, q \in [0, 1]$ and $p \neq q$. Both expectations are taken with respect to $Y \sim \text{Bernoulli}(p)$.*

There is a rich family of strictly proper scoring rules, including the Brier score ($S(q, Y) = 1 - (q - Y)^2$), the log scoring rule ($S(q, Y) = \log(q)$ if $Y = 1$ and $S(q, Y) = \log(1 - q)$ if $Y = 0$) and the spherical scoring rules [GR07b].

Incentive compatibility of SPSR The definition of SPSR immediately gives incentive compatibility. If an agent's belief is p , reporting p truthfully uniquely maximizes her expected score.

SPSR quantify the value of information Another nice property of SPSR is that they quantify the value/accuracy of reported predictions. To give a rigorous argument, we use an indicator vector \mathbf{y} of length 2 to represent the realization of Y , with 1 at the Y -th position and 0 otherwise. That is, $\mathbf{y} = (0, 1)$ if $Y = 1$ and $\mathbf{y} = (1, 0)$ if $Y = 0$. We use a probability vector $\mathbf{q} = (1 - q, q)$ to represent probability q . By the representation theorem [McC56; Sav71; GR07b], any strictly proper scoring rule can be characterized using a corresponding strictly convex function G as follows: $S(\mathbf{q}, \mathbf{y}) = G(\mathbf{y}) - D_G(\mathbf{y}, \mathbf{q})$, where D_G is the Bregman divergence function of G . Now consider the unknown true distribution of Y , denoted by

$\mathbf{p}^* = (1 - p^*, p^*)$. The expected score for an agent predicting \mathbf{q} is

$$\mathbb{E}[S(\mathbf{q}, \mathbf{y})] = \mathbb{E}[G(\mathbf{y})] - \mathbb{E}[D_G(\mathbf{y}, \mathbf{q})],$$

where all three expectations are taken over $Y \sim \text{Bernoulli}(p^*)$. This means that the maximum score an agent can receive in expectation is $\mathbb{E}_{Y \sim \text{Bernoulli}(p^*)}[G(\mathbf{y})]$, which happens when the agent's report $\mathbf{q} = \mathbf{p}^*$. Moreover, a prediction \mathbf{q} with a smaller divergence $\mathbb{E}_{y \sim \mathbf{p}^*}[D_G(\mathbf{y}, \mathbf{q})]$ receives a higher score in expectation. Intuitively, $\mathbb{E}_{Y \sim \text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})]$ characterizes how "far away" \mathbf{q} is from the true distribution of Y under divergence function D_G . This implies that a strictly proper scoring rule S qualifies the accuracy of a prediction \mathbf{q} based on the corresponding divergence function. When S is taken as the Brier scoring rule, the corresponding Bregman divergence is the quadratic function, and $\mathbb{E}_{Y \sim \text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})] = \|\mathbf{p}^* - \mathbf{q}\|^2$, implying that a prediction \mathbf{q} closer to \mathbf{p}^* according to ℓ_2 norm receives a higher score in expectation. When S is taken as the log scoring rule, the corresponding Bregman divergence is the KL-divergence, D_{KL} , which is also called the relative entropy, and $\mathbb{E}_{Y \sim \text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})] = D_{KL}(\mathbf{p}^* \parallel \mathbf{q}) + H(\mathbf{p}^*)$, where H is the entropy function. A prediction with a smaller KL-divergence from \mathbf{p}^* receives a higher score in expectation. This property of SPSR allows the principal to take an expert's average score over a set of prediction tasks as a proxy of his average accuracy and rank experts accordingly.

2.4 Model and mechanism design problem

We consider a multi-task setting for the information elicitation without verification (IEWV) problem. Under this setting, we aim to develop scoring mechanisms that are incentive compatible and are able to quantify the value of elicited information, recovering the two desirable properties that SPSR achieve in the presence of the ground truth. In this section, we formally introduce the information structure of our setting and the exact mechanism design problem we consider.

2.4.1 Model of Information Structure

A principal has a set of tasks $[m] = \{0, \dots, m - 1\}$. Each task asks for a prediction for an independent random variable of interest, denoted by Y_k , $k \in [m]$. For now, we assume that these random variables to predict are binary variables, i.e., $Y_k \in \{0, 1\}$, $\forall k \in [m]$. We will generalize our results to (non-binary) categorical random variables in Section 2.7. There is a set of informed agents $[n] = \{0, \dots, n - 1\}$. Each agent $i \in [n]$ privately observes a random signal $O_{i,k}$ generated by Y_k for each task $k \in [m]$, and thus holds a posterior belief about Y_k , represented by $P_{i,k} := \Pr[Y_k = 1 | O_{i,k}]$. The posterior $P_{i,k}$ is a random

variable as the signal $O_{i,k}$ is a random variable. Furthermore, we make following main assumptions on the information structure among the signals and ground truth.

Assumption 2.1. *Tasks are independent and similar a priori, that is, the joint distribution of $(O_{1,k}, \dots, O_{n,k}, Y_k)$ is i.i.d. for all tasks $k \in [m]$.*

This assumption is natural when the set of tasks are of similar nature, for example, tasks to predict the replicability of multiple studies published in the same journal and the same year. In this example, readers may a priori hold the same journal-wide belief about the features and the replicability of each study. After reading the journal, each agent receives a private signal for each individual study, which allows her to provide a more informed prediction for that study. This assumption is common for multi-task IEWV.¹

Based on Assumption 2.1, each Y_k has the same prior, denoted by $p := \Pr[Y_k = 1]$. Also, for a fixed agent i , the distribution of signal $O_{i,k}$ conditioned on Y_k on each task $k \in [m]$ is the same. We use \mathcal{D}_i^+ and \mathcal{D}_i^- to denote this conditional distribution for agent i for conditions $Y_k = 1$ and $Y_k = 0$, respectively. We assume that $\mathcal{D}_i^+ \neq \mathcal{D}_i^-$, otherwise, observation $O_{i,k}$ is independent and uninformative to Y_k . Each agent forms her posterior belief $P_{i,k}$ using the prior p and the conditional distributions \mathcal{D}_i^+ and \mathcal{D}_i^- . We require no knowledge of \mathcal{D}_i^+ and \mathcal{D}_i^- for the principal and the agents other than agent i . Furthermore, we assume that agents' signals are independent conditioned on the ground truth.

Assumption 2.2. *For each task, agents' signals are mutually independent conditioned on the ground truth, i.e., $\forall k \in [m], \Pr [O_{1,k}, \dots, O_{n,k} | Y_k] = \prod_{i \in [n]} \Pr [O_{i,k} | Y_k]$.*

This assumption excludes the scenarios where agents have some form of "side information" to coordinate their reports. With "side information", it is impossible to have any mechanism that can truthfully elicit agents' predictions without access to ground truth. This issue has been noted by Kong and Schoenebeck [KS18] and Kong [Kon20] for tasks with ground truth and the same assumption has been adopted. Finally, we make a technical assumption about the prior p and the principal's knowledge.

Assumption 2.3. *The prior $p \neq 0.5$ and the principal knows whether $p > 0.5$ or not.*

We do not assume that the principal knows the exact prior p of tasks but assume that she knows whether $p > 0.5$ or $p < 0.5$. This one binary-bit of information helps the principal distinguish between

¹Kong and Schoenebeck [KS18] and Kong [Kon20] consider information elicitation for objective questions (i.e., questions where an objective ground truth exists). They make the same assumption as Assumption 2.1. Other studies (e.g., [DG13; Shn+16; RFJ16; KS19; Kon20]) consider information elicitation for subjective questions (i.e., questions with no objective ground truth, e.g., how do you rate the movie?). These studies also assume that the joint distributions of agents' signals are the same across all tasks.

the set of truthful predictions and the set of inverted predictions (i.e. everyone reporting $1 - p_{i,k}$ instead of $p_{i,k}$), which otherwise is impossible to distinguish. In practice, this information is usually easy to obtain. In the example of predicting the replicability of studies, this assumption only requires that the principal knows whether the majority of the studies can be replicated or not. The assumption $p \neq 0.5$ is a technical condition we need in order to distinguish the truthful reporting scenario from the inverted reporting scenario.

We also assume that the posterior $P_{i,k}$ for any agent i on any task k is different under different realizations of private signal $O_{i,k}$. This assumption is without loss of generality, because different realizations of $O_{i,k}$ which lead to the same posterior $P_{i,k}$ for agent i on task k also lead to the same posterior about any other agent's signal $O_{j,k}$ for agent i due to Assumption 2.2. Therefore, we can merge multiple realizations of $O_{i,k}$ that lead to the same posterior $P_{i,k}$ into one realization without influencing agent i 's belief about other agents' signals and the ground truth. Consequently, it is without loss of generality to assume that there exists a one-to-one correspondence between the realization of an agent's signal $O_{i,k}$ and her posterior $P_{i,k}$. According to this one-to-one correspondence and Assumptions 2.1 and 2.2, the following two conditions hold for $P_{i,k}$ for $i \in [n], k \in [m]$.

Proposition 2.1. *Under Assumptions 2.1 and 2.2, the following two conditions hold for agents' beliefs $P_{i,k}, i \in [n], k \in [m]$.*

1. $P_{1,k}, \dots, P_{n,k}$ and Y_k are independent of their own counterparts across tasks $k \in [m]$ but have the same joint distribution, i.e., $(P_{1,k}, \dots, P_{n,k}, Y_k)$ are i.i.d. across tasks $k \in [m]$.
2. For each task $k \in [m]$, $P_{1,k}, \dots, P_{n,k}$ are independent conditioned on Y_k , i.e., $\Pr[P_{1,k}, \dots, P_{n,k} | Y_k] = \prod_{i \in [n]} \Pr[P_{i,k} | Y_k], \forall k \in [m]$.

The first condition in Proposition 2.1 implies that an agent has the same expertise level across different tasks, as the joint distribution of her posterior belief and the ground truth is the same across tasks. The second condition implies that given the ground truth, each agent's probabilistic prediction is independent. The two conditions in Proposition 2.1 in fact characterize a broader space of information structure than the space captured by Assumptions 2.1 and 2.2. The former space includes the information structure where each task has a different prior but the distribution of the posterior beliefs of each agent are still the same across tasks. Our theoretical results hold for the model with this more broader information structure space characterized by the two conditions in Proposition 2.1 and with Assumption 2.3, where p refers to the mean prior over all tasks.

2.4.2 Mechanism design problem

We consider the multi-task peer prediction mechanisms where the principal assigns each task k to a subset $[n_k] \subseteq [n]$ of agents, collects a single probabilistic prediction $q_{i,k} \in [0, 1]$ from each agent i assigned with task k , and pays each agent based on all predictions collected from all agents. We use $[m_i] \subseteq m$ to denote the set of tasks assigned to agent i . We use $q_{i,k} = \emptyset$ to denote that agent i has not been assigned to task k . Such a multi-task peer prediction mechanism can be formally expressed as a function $R : \{\emptyset \cup [0, 1]\}^{n \times m} \rightarrow \mathbb{R}^n$, which maps a prediction profile on all tasks and all agents to a vector of total payments of all agents. We restrict our attention to anonymous mechanisms that give each prediction from an agent an independent payment like SPSR. Thus, a mechanism that we consider can be fully expressed by a score function $R : \{\emptyset \cup [0, 1]\} \times \{\emptyset \cup [0, 1]\}^{n-1 \times m} \rightarrow \mathbb{R}$, which maps a single prediction $q_{i,k}$ of agent i on task k and a profile of predictions from all other agents into a single reward score for that prediction, and agent i 's total reward is the sum of the scores she obtains across the tasks she is assigned with.

Agents have no obligation to report their true beliefs. Instead, given a mechanism, an agent can report strategically to maximize her expected payment. As there exists a one-to-one correspondence between an agent's signal and her posterior on a single task in our model, we can define an agent's reporting strategy on a single task without loss of generality as a function that maps her posterior to a distribution where her reported prediction is drawn from.

Definition 2.2. Let $\Delta_{[0,1]}$ be the space of all probability distributions over $[0, 1]$. The strategy of an agent i on task k is a mapping $\sigma_i : [0, 1] \rightarrow \Delta_{[0,1]}$, which maps her posterior belief $P_{i,k}$ into a distribution $\sigma_i(P_{i,k})$ over $[0, 1]$, from which the agent draws the reported prediction $Q_{i,k}$.

We use the upper case $Q_{i,k}$ to denote the reported prediction when we want to emphasize that the reported prediction is a random variable determined by an agent's posterior belief and her reporting strategy jointly, otherwise, $q_{i,k}$ is used. We further assume that each agent adopts the same mixed strategy across all assigned tasks.

Assumption 2.4. (Uniform Strategy) For any agent $i \in [n]$, she adopts the same strategy $\sigma_i(\cdot)$ over all assigned tasks $k \in [m_i]$.

This assumption is reasonable as we assume that tasks are a priori similar to each other. We use $\sigma_i(\cdot)$ to denote the reporting strategy adopted by agent i on all tasks she answers and use σ_{-i} to denote the strategy profile used by all agents except agent i . Furthermore, we use $\mathbb{E}[R(q_{i,k}; \sigma_{-i})]$ to denote the expected score that agent i receives for reporting $q_{i,k}$ when other agents use strategy profile σ_{-i} , where

the expectation is taken over the randomness in ground truth, other agents' signals and strategies and in the mechanism itself. We use $\mathbb{E}[R(\sigma_i; \sigma_{-i})]$ to denote the expected reward of agent i when her report is also a random variable generated by her belief $P_{i,k}$ and reporting strategy σ_i .

In this chapter, our goal is to design a mechanism $R(\cdot)$ in the IEWV setting with similar properties that SPSR have for the information elicitation with verification setting: *quantification of the value of information and incentive compatibility*.

Quantifying value of information The score of each prediction should reflect the true accuracy of the prediction, similar to what SPSR achieve. That is, for all i, k and $q_{i,k}$ and for any true distribution of the ground truth Y_k , $\mathbb{E}[R(q_{i,k}; \sigma_{-i})] = f(E_{Y_k}[S(q_{i,k}, Y_k)])$ holds for a SPSR $S(\cdot)$ and a strictly increasing function f , where the two expectations are taken over the true distributions of the random variables in the two expressions at each side of the equality. This design goal pursues that the score that an agent receives for a prediction in IEWV recovers what the agent would receive with a SPSR (with access to the ground truth) in expectation.

Incentive Compatibility. A mechanism satisfies incentive compatibility to some extent if truthful reporting is a strategy that maximizes an agent's expected utility under certain conditions. We pursue the dominant uniform strategy truthfulness [GF19b], where truthful reporting is a dominant strategy if we restrict the strategy space with the uniform strategy assumption (Assumption 2.4).

Formally, in IEWV, a dominant uniform strategy truthful mechanism is a mechanism where truthful reporting on each task maximizes an agent's expected reward no matter what uniform strategies the other agents play and strictly maximize the agent's expected reward if other agents' reports are also informative.² Let σ_i^* be the truthful reporting strategy for agent i , i.e., σ_i^* is the function that maps a belief p_i to a distribution where all probability mass is put on p_i . Let $\bar{Q}_{-i,k} := \frac{1}{n-1} \sum_{j \neq i} Q_{j,k}$ be the mean of all agents' reported predictions on task k except agent i 's. Note that $\bar{Q}_{-i,k}$ is a random variable, because of the randomness in reporting strategy σ_j and the randomness in signal $O_{j,k}$ for all $j \neq i$. We

²In a standard dominant truthful mechanism, truthful reporting strictly maximizes the agent's expected reward no matter what strategies other agents play. In IEWV, however, if all peer agents report predictions independently w.r.t. the ground truth on each task, then there will be no information available for the mechanism to incentivize truthful reporting. Therefore, it is inevitable to allow a dominant truthful mechanism in IEWV to pay truthful reporting strictly more only when the peer reports are informative about the ground truth. For example, in studies [KS19; Kon20], the dominant uniform strategy truthful mechanism is defined to be a mechanism that pays truthful reporting strictly more only when for each agent, there exists at least one peer agent reporting truthfully. We will see later that in our definition, we do not require that there is at least one peer agent reporting truthfully. We allow all peer agents to play non-truthfully, but require the mean of peer agents' reports to be informative with respect to the ground truth.

say that $\bar{Q}_{-i,k}$ is informative about the ground truth if $\mathbb{E}[\bar{Q}_{-i,k}|y_k = 1] \neq \mathbb{E}[\bar{Q}_{-i,k}|y_k = 0]$. We formally define the dominant uniform strategy truthful mechanisms as follows.

Definition 2.3. (*Dominant uniform strategy truthfulness*). A mechanism $R(\cdot)$ is dominant uniform strategy truthful if $\forall i \in [n], \forall k \in [m_i], \forall \{\mathcal{D}_j^+, \mathcal{D}_j^-\}_{j \in [n]}$ and for any realization $o_{i,k}$ of signal $O_{i,k}$: $\mathbb{E}[R(\sigma_i^*; \sigma_{-i})|O_{i,k} = o_{i,k}] \geq \mathbb{E}[R(\sigma_i; \sigma_{-i})|O_{i,k} = o_{i,k}]$ for any uniform strategy $\sigma_i \neq \sigma_i^*$ and any uniform strategy profile of other agents σ_{-i} , and the inequality holds strictly for any uniform strategy profile σ_{-i} under which $\bar{Q}_{-i,k}$ is informative about Y_k .

In Definition 2.3, we characterize the condition that peers' reports are informative by that the expectation of the mean of peers' reports on a task differs when conditioned on different realizations of the ground truth.

2.5 Elicitation with a noisy estimate of ground truth

Before we develop mechanisms with the two desirable properties we pursue, in this section we first obtain these two properties under a very stylized setting: *elicitation with a noisy estimate of ground truth*. In this stylized setting, we introduce surrogate scoring rules as an effective solution. These scoring rules will be the building blocks of our mechanisms designed for the general setting.

This stylized setting has only one event Y and one agent i , who observes a signal O_i generated from distribution $\mathcal{D}_i(Y)$ and forms a posterior $P_i = \Pr[Y = 1|O_i]$. The principal in this setting has access to a noisy estimate $Z \in \{0, 1\}$ of the ground truth Y , although she has no access to the exact realization of Y . The noisy estimate Z is characterized by two *error rates*, e_z^+ and e_z^- , defined as $e_z^+ := \Pr[Z = 0|Y = 1]$, $e_z^- := \Pr[Z = 1|Y = 0]$, which are the probabilities that Z mismatches Y under the two realizations of Y . The principal knows the realization Z and the exact error rates e_z^+, e_z^- . The principal cannot expect to do much if Z is independent of Y . Therefore, we assume that Z and Y are stochastically relevant, an assumption commonly adopted on the relation between a signal and the ground truth in the information elicitation literature [MRZ05b].

Definition 2.4. A random variable Z is stochastically relevant to a random variable Y if the distribution of Y conditioned on Z differs for different realizations of Z .

The following lemma shows that the stochastic relevance condition directly translates to a constraint on the error rates, that is, $e_z^+ + e_z^- \neq 1$. This lemma can be proved immediately by writing out the distribution of Y conditioned on Z in terms of the two error rates e_z^+, e_z^- and the prior of Z .

Lemma 2.2. *The noisy estimate Z is stochastically relevant to the ground truth Y if and only if $e_z^+ + e_z^- \neq 1$.*

The goal of the principal in this setting is to design a scoring rule to elicit the posterior P_i truthfully based on this noisy estimate Z and the error rates e_z^+, e_z^- . We define the design space of the scoring rules with the noisy estimate as follows.

Definition 2.5. *Given a noisy estimate Z of ground truth Y with error rates $(e_z^+, e_z^-) \in [0, 1]^2$, a scoring rule against the noisy estimate of the ground truth is a function $R : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ that maps a prediction $q_i \in [0, 1]$ and a realized noisy estimate $z \in \{0, 1\}$ to a score. The function R can depend on the two error rates (e_z^+, e_z^-) .*

Adopting the terminology from the scoring rule literature, we refer to strict properness of a scoring rule against a noisy estimate of ground truth as the property that the rule assigns a strictly better expected score to a truthful prediction of the ground truth than to a non-truthful prediction.

Definition 2.6. *A scoring rule $R(q_i, Z)$ against a noisy estimate Z of ground truth is strictly proper for eliciting an agent's posterior belief generated by signal O_i if it holds for all realizations o_i of O_i and the posterior $p_i = \Pr[Y = 1 | O_i = o_i]$ that*

$$\mathbb{E}_{Z|O_i=o_i}[R(p_i, Z)] > \mathbb{E}_{Z|O_i=o_i}[R(q_i, Z)], \forall q_i \in [0, 1] (q_i \neq p_i).$$

2.5.1 Surrogate scoring rules (SSR)

In this section, we present our solution, the *surrogate scoring rules* (SSR), for this stylized setting. SSR are a family of scoring rules that evaluate a prediction against a noisy estimate of ground truth. For any distribution of the ground truth and any stochastically relevant noisy estimate of the ground truth, the expected score that SSR give to the prediction, with expectation taken over the randomness of the noisy estimate, is equal to (up to a monotonic increasing transformation) the expected score that a SPSR gives to the same prediction, with expectation taken over the randomness of the ground truth. We will see that SSR are strictly proper under mild conditions.

Definition 2.7 (Surrogate Scoring Rules). *$R : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ is a surrogate scoring rule if for some strictly proper scoring rule $S : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ and a strictly increasing function $f : \mathbb{R} \rightarrow \mathbb{R}$, it holds that $\forall p_i, q_i, e_z^+, e_z^- \in [0, 1]$ and $e_z^+ + e_z^- \neq 1$, $\mathbb{E}_Z[R(q_i, Z)] = f(\mathbb{E}_Y[S(q_i, Y)])$, where Y is the ground truth drawn from Bernoulli(p_i) and Z is a noisy estimate of Y with error rates e_z^+, e_z^- .*

Definition 2.7 defines the SSR $R(\cdot)$ as scoring rules that help us remove the bias in Z and return us the same score given by a SPSR in expectation. The idea of SSR is borrowed from the machine learning

literature on learning with noisy data [Byl94; Nat+13; Sco15; Men+15; RW15]. SSR can be viewed as a particular class of the proxy scoring rules proposed by Witkowski et al. [Wit+17]. Witkowski et al. [Wit+17] achieve properness of proxy scoring rules by plugging in an *unbiased* proxy of the ground truth to a SPSR. With SSR, we directly work with biased proxy and design scoring functions to de-bias the noise in the proxy. We have the following strict properness result for SSR straightforwardly:

Theorem 2.3. *Given the prior p of the ground truth Y and a private signal O_i , SSR $R(q_i, Z)$ against a noisy estimate Z is strictly proper for eliciting the posterior $P_i = \Pr[Y = 1|O_i]$ if Z and O_i are independent conditioned on Y , and Z is stochastically relevant to Y .*

We provide an implementation of SSR, which we call SSR_α :

$$R(q_i, Z = 1) = \frac{(1 - e_z^-) \cdot S(q_i, 1) - e_z^+ \cdot S(q_i, 0)}{1 - e_z^+ - e_z^-}, \quad (2.1)$$

$$R(q_i, Z = 0) = \frac{(1 - e_z^+) \cdot S(q_i, 0) - e_z^- \cdot S(q_i, 1)}{1 - e_z^+ - e_z^-}, \quad (2.2)$$

where S can be any strictly proper scoring rule. This SSR implementation is inspired by Natarajan et al. [Nat+13]. As can be seen from Eqs. 2.1 and 2.2, the knowledge of the error rates e_z^+, e_z^- is crucial for defining SSR_α . Moreover, SSR_α has the property that the expected score $\mathbb{E}_{Z|Y}[R(q_i, Z)]$ conditioned on the realization of the ground truth Y is exactly the same as the score $S(q_i, Y)$ given by the SPSR. More formally, we have the following lemma.

Lemma 2.4 (Lemma 1, [Nat+13]). *For SSR_α , ground truth Y and noisy estimate Z , $\forall q_i, e_z^+, e_z^- \in [0, 1]$ and $e_z^+ + e_z^- \neq 1, \forall y \in \{0, 1\} : \mathbb{E}_{Z|Y=y}[R(q_i, Z)] = S(q_i, Y)$.*

Proof. Lemma 1 in [Nat+13] proves the statement for $e_z^+ + e_z^- < 1$. For completeness, we provide the proof for $e_z^+ + e_z^- \neq 1$ here. Let $q_i \in [0, 1]$ be an arbitrary prediction. When $y = 1$, we have

$$\begin{aligned} \mathbb{E}_{Z|Y=1}[R(q_i, Z)] &= (1 - e_z^+)R(q_i, 1) + e_z^+ R(q_i, 0) \\ &= (1 - e_z^+) \frac{(1 - e_z^-)S(q_i, 1) - e_z^+ S(q_i, 0)}{1 - e_z^+ - e_z^-} + e_z^+ \frac{(1 - e_z^+)S(q_i, 0) - e_z^- S(q_i, 1)}{1 - e_z^+ - e_z^-} \\ &= \frac{((1 - e_z^+)(1 - e_z^-) - e_z^+ e_z^-) S(q_i, 1)}{1 - e_z^+ - e_z^-} \\ &= S(q_i, 1) \end{aligned}$$

When $y = 0$, we have

$$\begin{aligned}
\mathbb{E}_{Z|Y=0}[R(q_i, Z)] &= e_z^- R(q_i, 1) + (1 - e_z^-) R(q_i, 0) \\
&= e_z^- \frac{(1 - e_z^-) S(q_i, 1) - e_z^+ S(q_i, 0)}{1 - e_z^+ - e_z^-} + (1 - e_z^-) \frac{(1 - e_z^+) S(q_i, 0) - e_z^- S(q_i, 1)}{1 - e_z^+ - e_z^-} \\
&= S(q_i, 0)
\end{aligned}$$

□

Intuitively, the linear transformation in SSR_α ensures that, in expectation, the prediction q_i is scored as if it was scored against the ground truth Y under the underlying SPSR. We would like to note that other surrogate loss functions designed for learning with noisy labels can also be leveraged to design SSR. With the conditional unbiasedness property of SSR_α , we can formally claim that SSR_α is a surrogate scoring rule, as stated in Theorem 2.5 below.

Theorem 2.5. *SSR_α is a surrogate scoring rule and $\forall p_i, q_i, e_z^+, e_z^- \in [0, 1] (e_z^+ + e_z^- \neq 1)$, $\mathbb{E}_Z[R(q_i, Z)] = \mathbb{E}_Y[S(q_i, Y)]$, where Y is the ground truth drawn from Bernoulli(p_i) and Z is the noisy estimate of ground truth Y with error rate e_z^+, e_z^- .*

Proof. As shown by Lemma 2.4, for SSR_α , we have $\forall p_i, q_i, e_z^+, e_z^- (e_z^+ + e_z^- \neq 1)$ and $\forall y \in \{0, 1\}$, $\mathbb{E}_{Z|Y=y}[R(q_i, z)] = S(q_i, Y = y)$, we have immediately

$$\mathbb{E}_Z[R(q_i, z)] = \mathbb{E}_Y \left[\mathbb{E}_{Z|Y}[R(q_i, Z)] \right] = \mathbb{E}_Y[S(q_i, Y)].$$

□

With Theorem 2.5 we know that SSR_α quantifies the quality of information of a prediction just as the underlying strictly proper scoring rule S does. Furthermore, SSR_α has the following variance:

Theorem 2.6. *Let $p_z := \Pr[Z = 1]$. For a fixed prediction $q_i \in [0, 1]$, SSR_α suffers the following variance:*

$$\mathbb{E}_Z[R(q_i, Z) - \mathbb{E}_z[R(q_i, Z)]]^2 = \frac{p_z \cdot (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} \cdot (S(q_i, 1) - S(q_i, 0))^2. \quad (2.3)$$

Mechanism 5 SSR mechanisms (Sketch)

- 1: For each task k , we uniformly randomly pick at least 3 agents, assign task k to them and collect their predictions.
 - 2: For each agent i and each task k the agent answers, we construct a reference report $Z_{i,k}$ using the agent's peer agents' reports, and estimate the error rates $e_{z_{i,k}}^+$ and $e_{z_{i,k}}^-$ for $Z_{i,k}$.
 - 3: Pay each agent i for her prediction $q_{i,k}$ on task k by SSR $R(q_{i,k}, Z_{i,k})$ if $e_{z_{i,k}}^+ + e_{z_{i,k}}^- \neq 1$, and pay 0, otherwise.
-

Proof.

$$\begin{aligned} & \mathbb{E}_Z [R(q_i, Z) - \mathbb{E}_Z [R(q_i, Z)]]^2 \\ &= p_z \left(R(q_i, 1) - (p_z R(q_i, 1) + (1 - p_z) R(q_i, 0)) \right)^2 \\ & \quad + (1 - p_z) \left(R(q_i, 0) - (p_z R(q_i, 1) + (1 - p_z) R(q_i, 0)) \right)^2 \\ &= p_z (1 - p_z)^2 (R(q_i, 1) - R(q_i, 0))^2 + (1 - p_z) p_z^2 (R(q_i, 0) - R(q_i, 1))^2 \\ &= p_z (1 - p_z) (R(q_i, 0) - R(q_i, 1))^2 \\ &= \frac{p_z (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} \left((1 - e_z^-) S(q_i, 1) - e_z^+ S(q_i, 0) - ((1 - e_z^+) S(q_i, 0) - e_z^- S(q_i, 1)) \right)^2 \\ &= \frac{p_z (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} (S(q_i, 1) - S(q_i, 0))^2 \end{aligned}$$

□

2.6 Elicitation without verification

The results in the previous section are built upon the fact that there exists a noisy estimate of ground truth with known error rates. In this section, we apply the idea of SSR to the IEWV setting. A reasonable way to do so is to use agents' reports as the source of the noisy estimate. Although the principal does not know the exact bias in agents' reports, we find a way to construct such a noisy proxy of ground truth and estimate its error rates. We refer to this noisy proxy as the *reference report*. Applying SSR with this reference report, we can finally get a family of mechanisms which are dominant uniform strategy truthful and which also quantify the value of information in agents' reports as what SPSR do. Within this family, we can choose different underlying SPSR for SSR to get different mechanisms. We call this family of mechanisms *SSR mechanisms*. We present a sketch of our SSR mechanisms in Mechanism 5.

The challenge of designing such mechanisms is to construct the reference report $Z_{i,k}$ in Mechanism 5

and successfully estimate its error rates $e_{z_{i,k}}^+, e_{z_{i,k}}^-$. In the following sections, we show how to construct this reference report and estimate its error rates.

2.6.1 Reference report and its property

Recall that we use $Q_{j,k}$ to denote the reported prediction of agent j on task k , which is generated by agent j 's posterior belief $P_{j,k}$ and reporting strategy σ_j . Let $S_{j,k} \in \{0, 1\}$ be a binary signal independently drawn from $\text{Bernoulli}(Q_{j,k})$. We refer to $S_{j,k}$ as the *prediction signal* of agent j on task k . We construct the reference report $Z_{i,k}$ for agent i as follows: *We uniformly randomly pick an agent j from agent i 's peer agent set $[n] \setminus \{i\}$, collect agent j 's prediction $Q_{j,k}$, and draw a prediction signal $S_{j,k} \sim \text{Bernoulli}(Q_{j,k})$. We use this $S_{j,k}$ as the reference report $Z_{i,k}$ for agent i on task k .*

Conditioned on all peer agents' reports $Q_{j,k}, j \in [n] \setminus \{i\}$, the distribution of $Z_{i,k}$ is $\text{Bernoulli}(\bar{Q}_{-i,k})$, because we pick a prediction signal from all peer agents uniformly randomly. Recall that in our model, $Q_{i,k} \sim \sigma_i(P_{i,k}), i \in [n], k \in [m]$. Due to Proposition 2.1 and Assumption 2.4, $\bar{Q}_{-i,k}$ is i.i.d. across tasks $k \in [m]$ and is independent to agent i 's posterior $P_{i,k}$ conditioned on the ground truth Y_k for any task k . Therefore, $Z_{i,k}, k \in [m]$ that we construct have the following two properties.

Lemma 2.7. $\forall i \in [n], k \in [m], Z_{i,k}$ is independent to agent i 's posterior $P_{i,k}$ conditioned on Y_k .

This property ensures that $Z_{i,k}$ can be used as the conditionally independent noisy estimate of the ground truth in Theorem 2.3 and thus, SSR against $Z_{i,k}$ is strictly proper for eliciting the posterior belief $P_{i,k}$.

Lemma 2.8. *For any strategy profile agents play, the reference reports of a single agent i for any $i \in [n]$ are i.i.d. across tasks and have the same error rates w.r.t. their corresponding ground truth Y_k , i.e., $\forall \sigma_1, \dots, \sigma_n, \forall i \in [n], \exists e_i^+, e_i^- \in [0, 1], \forall k \in [m] : \Pr[Z_{i,k} = 0 | Y_k = 1] = e_i^+, \Pr[Z_{i,k} = 1 | Y_k = 0] = e_i^-$.*

This lemma shows that the error rates of the reference reports of an agent i are the same across all tasks. This property allows the estimation of the error rates using the multi-task prediction data. In the following sections, we introduce the estimation of the error rates and complete our mechanisms. We prove Lemma 2.7 and 2.8 below.

Proof. Proposition 2.1 and Assumption 2.4 directly imply that 1) for each task, $Q_{1,k}, \dots, Q_{n,k}$ are mutually independent conditioned on the ground truth Y_k , and 2) $(Q_{1,k}, \dots, Q_{n,k}, Y_k)$ are i.i.d across tasks $k \in [M]$. As $Z_{i,k}$ is independently drawn from $\text{Bernoulli}(\bar{Q}_{-i,k})$, we immediately have that 1') for each task $k \in [m]$, $Z_{i,k}$ is independent to $O_{i,k}$ and thus to $P_{i,k} := \Pr[Y_k = 1 | O_{i,k}]$, and 2') $(Z_{i,k}, Y_k)$ have the same joint

distribution for $k \in [m]$. As a result of 2'), $Z_{i,k}, k \in [m]$ have the same error rates w.r.t. the corresponding Y_k . □

2.6.2 Asymptotic setting

To better deliver our idea for error rates estimation, we start with an asymptotic setting with infinite amounts of tasks and agents, i.e., $m, n \rightarrow \infty$. We will later provide a finite sample justification for our mechanism. Based on Lemma 2.8, the reference reports of an agent on different tasks have the same distribution and error rates. Therefore, we focus on estimating the error rates of the reference report of agent i on a generic task k , while we use Z to denote this reference report, omitting the subscripts i and k , and use e_z^+, e_z^- to denote its error rates.

Our estimation algorithm resembles the “method of moments.” We establish three equations on the first- to the third-order statistics, of which the parameters can be expressed by the unknown error rates e_z^+, e_z^- . We show that the three equations, with knowing the true parameters (which is true in the asymptotic setting), together uniquely determine e_z^+, e_z^- . Thus, we can solve the three equations to obtain e_z^+, e_z^- . In the next section, we argue that in the finite sample setting, with imperfect estimates of the parameters of the three questions, the solution from these three perturbed equations still approximate the true values of e_z^+, e_z^- with guaranteed accuracy.

To construct these three equations, we make the following preparation. Let $s_{j,k}$ be the realization of the prediction signal $S_{j,k}$ of agent j on task k , and let $\mathcal{S}_{-i} := \{s_{j,k}\}_{j \neq i, k \in [M]}$ be the realization profile of all prediction signals from all peer agents of agent i . On a generic task k , we draw three random variables Z_1, Z_2, Z_3 . Z_1 represents the realization of a prediction signal uniformly randomly drawn from the set of all prediction signals $\{s_{j,k}\}_{j \neq i}$ on task k except agent i 's. Z_2 represents the realization of another uniformly randomly picked prediction signal from set $\{s_{j,k}\}_{j \neq i}$ but excluding Z_1 . Similarly, Z_3 represents the realization of another uniformly randomly picked prediction signal from set $\{s_{j,k}\}_{j \neq i}$ but excluding Z_1 and Z_2 . Because agents' reports are conditionally independent, Z_1, Z_2, Z_3 are also independent conditioned on the ground truth. Moreover, Z_1 and the reference report Z have the same error rates, as they are generated by the same random process. With infinite number of agents, Z_2 and Z_3 also have the same error rates as Z . Furthermore, (Z_1, Z_2, Z_3) is i.i.d. across different tasks, according to Proposition 2.1 and Assumption 2.4. Therefore, with infinite number of tasks (and thus infinite number of samples from the joint distribution Z_1, Z_2, Z_3), we can know the exact distribution parameters of any statistics about Z_1, Z_2 and Z_3 . We can then establish the following three equations.

Algorithm 6 e_z^+, e_z^- solver

Input: $\alpha_{-i}, \beta_{-i}, \gamma_{-i}, \mathbb{1}(p > 0.5)$ **Output:** e_z^+, e_z^-

1: Compute the following quantities:

$$a := \frac{\gamma_{-i} - \alpha_{-i}\beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2}, \quad b := \frac{\alpha_{-i}\gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2}.$$

2: Let

$$\underline{x} := \frac{a - \sqrt{a^2 - 4b}}{2}, \quad \bar{x} := \frac{a + \sqrt{a^2 - 4b}}{2}, \quad p' := \frac{\alpha_{-i} - \underline{x}}{\bar{x} - \underline{x}}$$

3: If $\mathbb{1}(p' > 0.5) = \mathbb{1}(p > 0.5)$, then $e_z^+ = 1 - \bar{x}$, $e_z^- = \underline{x}$, else $e_z^+ = 1 - \underline{x}$, $e_z^- = \bar{x}$.

1. First-order equation: The first equation is based on the distribution of Z . Let $\alpha_{-i} := \Pr[Z = 1]$. α_{-i} can be expressed as a function of e_z^+, e_z^- via spelling out the conditional expectation:

$$\alpha_{-i} = p \cdot \Pr[Z = 1|Y = 1] + (1 - p) \cdot \Pr[Z = 1|Y = 0] = p \cdot (1 - e_z^+) + (1 - p) \cdot e_z^-. \quad (2.4)$$

2. Matching between two prediction signals: The second equation is based on a second-order statistic called the matching probability. We consider the matching-on-1 probability of Z_1, Z_2 , i.e., the matching-on-1 probability of the prediction signals from two uniformly randomly picked peer agents of agent i). Let $\beta_{-i} := \Pr[Z_1 = 1, Z_2 = 1]$. It can be written as a function of e_z^+, e_z^- as follows:

$$\begin{aligned} \beta_{-i} &= p \cdot \Pr[Z_1 = 1, Z_2 = 1|Y = 1] + (1 - p) \cdot \Pr[Z_1 = 1, Z_2 = 1|Y = 0] \\ &= p \cdot \Pr[Z_1 = 1|Y = 1] \cdot \Pr[Z_2 = 1|Y = 1] + (1 - p) \cdot \Pr[Z_1 = 1|Y = 0] \Pr[Z_2 = 1|Y = 0] \\ &= p \cdot (1 - e_z^+)^2 + (1 - p) \cdot (e_z^-)^2. \end{aligned} \quad (2.5)$$

3. Matching among three prediction signals: The third equation is obtained by going one order higher. We check the matching-on-1 probability over three prediction signals Z_1, Z_2, Z_3 uniformly randomly drawn from three different peer agents on the same task. Let $\gamma_{-i} := \Pr[Z_1 = Z_2 = Z_3 = 1]$. Similar to Eq. 2.5, we have:

$$\gamma_{-i} = p \cdot (1 - e_z^+)^3 + (1 - p) \cdot (e_z^-)^3. \quad (2.6)$$

Notice that all three parameters $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ can be perfectly estimated using \mathcal{S}_{-i} with infinite number of tasks and agents, yet without accessing any of the ground truth. With the knowledge of these three parameters, we prove the following:

Theorem 2.9. p, e_z^-, e_z^+ are uniquely identified by Eqs. 2.4-2.6 under Assumption 2.3 ($p \neq 0.5$ and the principal knows whether $p > 0.5$ or not). The solution is in the closed form shown in Algorithm 6.

Proof. Let $x^- := e_z^-, x^+ := 1 - e_z^+$. Recall the three equations we have

$$\alpha_{-i} = (1 - p) \cdot x^- + p \cdot x^+ \quad (2.7)$$

$$\beta_{-i} = (1 - p) \cdot (x^-)^2 + p \cdot (x^+)^2 \quad (2.8)$$

$$\gamma_{-i} = (1 - p) \cdot (x^-)^3 + p \cdot (x^+)^3 \quad (2.9)$$

We can rewrite the three equations as:

$$\alpha_{-i} - x^+ = (1 - p)(x^- - x^+) \quad (2.10)$$

$$\beta_{-i} = (1 - p)(x^- - x^+)(x^- + x^+) + (x^+)^2 \quad (2.11)$$

$$\gamma_{-i} = (1 - p)(x^- - x^+) \left((x^-)^2 + x^- \cdot x^+ + (x^+)^2 \right) + (x^+)^3 \quad (2.12)$$

Plugging Eq. 2.10 into Eqs. 2.11 and 2.12 and re-organizing the two equations, we have respectively:

$$\beta_{-i} = \alpha_{-i}(x^- + x^+) - x^- \cdot x^+ \quad (2.13)$$

$$\gamma_{-i} = \alpha_{-i} \left((x^- + x^+)^2 - x^- \cdot x^+ \right) - x^- \cdot x^+ (x^- + x^+) \quad (2.14)$$

Let

$$x^- + x^+ = a, \quad x^- \cdot x^+ = b,$$

then we have $a = \frac{b + \beta_{-i}}{\alpha_{-i}}$ from Eq. 2.13. Note that a is well defined, as o.w. if $\alpha_{-i} = 0$, we have to have $x^- = x^+ = 0$ which leads to $e_z^- + e_z^+ = 1$, a contradiction.

Substituting $x^- + x^+$ and $x^- \cdot x^+$ with $\frac{b + \beta_{-i}}{\alpha_{-i}}$ and b correspondingly in Eq. 2.14, we have

$$\alpha_{-i} \cdot \left(\frac{(b + \beta_{-i})^2}{(\alpha_{-i})^2} - b \right) - b \cdot \frac{b + \beta_{-i}}{\alpha_{-i}} = \gamma_{-i} \quad (2.15)$$

$$\Rightarrow \frac{(b + \beta_{-i})^2}{\alpha_{-i}} - b \cdot \alpha_{-i} - \frac{b^2}{\alpha_{-i}} - \frac{b \cdot \beta_{-i}}{\alpha_{-i}} = \gamma_{-i} \quad (2.16)$$

$$\Rightarrow \left(\frac{\beta_{-i}}{\alpha_{-i}} - \alpha_{-i} \right) b = \gamma_{-i} - \frac{(\beta_{-i})^2}{\alpha_{-i}} \Rightarrow b = \frac{\alpha_{-i} \gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2} \quad (2.17)$$

Thus, $a = \frac{b + \beta_{-i}}{\alpha_{-i}} = \frac{\gamma_{-i} - \alpha_{-i} \beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2}, b = \frac{\alpha_{-i} \gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2}$. Then from $x^- + x^+ = a, x^- \cdot x^+ = b$, we have

$$x^+ = \frac{a \pm \sqrt{a^2 - 4b}}{2}, x^- = \frac{a \mp \sqrt{a^2 - 4b}}{2}, p = \frac{\alpha_{-i} - x^-}{x^+ - x^-}$$

Thus, we have two pairs of solutions for the error rates and the prior:

$$e_{z,(1)}^+ = 1 - \frac{a + \sqrt{a^2 - 4b}}{2}, e_{z,(1)}^- = \frac{a - \sqrt{a^2 - 4b}}{2}, p^{(1)} = \frac{\alpha_{-i} - e_{z,(1)}^-}{1 - e_{z,(1)}^+ - e_{z,(1)}^-}$$

$$e_{z,(2)}^- = 1 - e_{z,(1)}^+, e_{z,(2)}^+ = 1 - e_{z,(1)}^-, p^{(2)} = 1 - p^{(1)}$$

As in these two solutions, the values for the prior is symmetric w.r.t. 0.5. Thus, by Assumption 2.3, the principal can identify the unique correct solution from the two. □

We can continue to establish higher-order equations. However, we show that they do not provide additional information about the three unknown variables, $p, e_z^+,$ and e_z^- .

Theorem 2.10. *Any higher order (≥ 4) matching equations can be expressed by the first- to the third-order equations, Eqs. 2.4-2.6.*

Proof. We follow the shorthand notations as in the proof of Theorem 2.9. The n -th equation is

$$\Pr[Z_1 = \dots = Z_n = 1] = (1 - p)(x^-)^n + p(x^+)^n.$$

For $n \geq 4$, the right-hand of the equation can be expressed as

$$\begin{aligned} (1 - p)(x^-)^n + p(x^+)^n &= \left((1 - p)(x^-)^{n-1} + p(x^+)^{n-1} \right) (x^- + x^+) \\ &\quad - x^- \cdot x^+ \left((1 - p)(x^-)^{n-2} + p(x^+)^{n-2} \right) \\ &= \Pr[Z_1 = \dots = Z_{n-1}] (x^- + x^+) \\ &\quad - \Pr[Z_1 = \dots = Z_{n-2}] x^- \cdot x^+ \end{aligned}$$

As we know from the proof of Theorem 2.9, $x^- + x^+$ and $x^- \cdot x^+$ are uniquely determined by the first three equations, i.e., Eqs. 2.4-2.6 (no matter whether Assumption 2.3 is made or not). Therefore, by induction starting from $n = 4$, the n -th equation can be expressed by the first three equations. □

Now we have completed our SSR mechanisms. The full version of the mechanisms is presented in Mechanism 7. Intuitively speaking, Theorem 2.9 shows that without ground truth data, knowing how frequently agents' predictions reach consensus with each other will help us characterize the (average) subjective biases in their reports. Furthermore, it implies that SSR mechanisms are asymptotically (in m, n) preserving the information quantification property that strictly proper scoring rules have, i.e., $\mathbb{E}_Z[R(q_{i,k}, Z)] = \mathbb{E}_Y[S(q_{i,k}, Y)]$, and that SSR mechanisms induce truthful reporting as the unique best

Mechanism 7 SSR mechanisms

- 1: For each task k , uniformly randomly pick at least 3 agents, assign task k to them, collect their reported predictions $q_{i,k}$ and generate the prediction signal $S_{i,k}$ for each prediction.
 - 2: For each agent i and each task k the agent answers, uniformly randomly select one prediction signal $S_{j,k}$ from her peers' prediction signals on the same task and let the reference report $Z_{i,k} := S_{j,k}$.
 - 3: Establish Eqs. 2.4-2.6 and solve out the error rates $e_{z_i}^-, e_{z_i}^+$ for $Z_{i,k}$ for any k using Algorithm 2.3.
 - 4: Pay each agent i 's prediction $q_{i,k}$ on each task k she answers by applying SSR_α with $q_{i,k}$ and the noisy estimate $Z_{i,k}$ with error rates $e_{z_i}^+, e_{z_i}^-$ if $e_{z_i}^+ + e_{z_i}^- \neq 1$, and pay 0, otherwise.
-

uniform strategy for an agent, when Z is informative (i.e., $1 - e_z^+ - e_z^- \neq 0$), and as a best strategy otherwise. Formally, we have the following theorem.

Theorem 2.11. *Under Assumptions 2.1-2.4, SSR mechanisms are dominant uniform strategy truthful with infinite number of tasks and agents. Furthermore, for any agent i and task k , if the average prediction of all other agents are informative, i.e., $e_z^+ + e_z^- \neq 1$ for the noisy estimate of the ground truth $Z_{i,k}$ constructed for agent i , then the expected score of SSR mechanisms for agent i 's prediction on a task is equal to the expected score given by the corresponding strictly proper scoring rule S : $\forall q_{i,k} \in [0, 1], \mathbb{E}_{Z_{i,k}}[R(q_{i,k}, Z_{i,k})] = \mathbb{E}_{Y_k}[S(q_{i,k}, Y_k)]$.*

Proof. Recall that in Assumption 2.4, we assume that each agent adopts the same reporting strategy across tasks. As long as this assumption is satisfied, for an agent i , no matter what exact strategy the other agents play, we can always correctly estimate the error rates e_z^+ and e_z^- of the reference report Z constructed for agent i , according to Theorem 2.9. Furthermore, by Lemma 2.7, Z is independent to agent i 's belief conditioned on the ground truth. Therefore, according to Theorem 2.3, when $e_z^+ + e_z^- \neq 1$, i.e., the other agents' average prediction is informative about the ground truth Y , SSR give agent i 's prediction $q_{i,k}$ a reward unbiased to the expected reward given by the corresponding SPSR, i.e., $\forall q_{i,k}, \mathbb{E}_{Z_{i,k}}[R(q_{i,k}, Z_{i,k})] = \mathbb{E}_{Y_k}[S(q_{i,k}, Y_k)]$. Consequently, truthful reporting strictly maximizes the expected reward of agent i . When $e_z^+ + e_z^- = 1$, i.e., the other agents' average prediction is uninformative about Y_k for task k , SSR mechanisms always reward agent i zero, where truthful reporting also maximizes the expected reward of agent i . Thus, SSR mechanisms are dominant uniform strategy truthful. \square

Remark 2.1. *Theorems 2.9 and 2.11 rely on Proposition 2.1 and Assumptions 2.3 and 2.4. Proposition 2.1 and Assumption 2.4 guarantee that there exists a similar information pattern across the predictions of different tasks that we can learn to infer the ground truth. Therefore, they can be hardly relaxed in IEVW settings. For Assumption 2.3, we'd like to argue that at least one bit of information is needed in order to distinguish the case where agents are truthfully reporting from the case where agents are misreporting by reverting their observations.*

This is because for any distribution of the observed reports of agents resulted by a world with parameters (p, e_z^+, e_z^-) and with agents reporting truthfully, there always exists the following counterfactual world achieving the same distribution of the observed reports of agents: a world with parameters $(1 - p, 1 - e_z^-, 1 - e_z^+)$ and with agents misreporting predictions via relabelling $0 \rightarrow 1$ and $1 \rightarrow 0$. Thus, the mechanism designer cannot tell the two worlds apart from only the observed reports. Some studies [KS19; Kon20] relax Assumption 2.3 by allowing the truthful reporting strategy to weakly dominate this “relabeling equilibrium”.

We will show in the next section, SSR mechanisms are also dominant uniform strategy truthful with finite number of tasks and agents under mild conditions. Several remarks follow. (1) We would like to emphasize again that for an agent i , both Z and $R(\cdot)$ come from the prediction signals of her peer agents’ reports \mathcal{S}_{-i} : Z is directly picked from \mathcal{S}_{-i} ; $R(\cdot)$ depends on the error rates e_z^+ and e_z^- of Z , which are also learnt from \mathcal{S}_{-i} . (2) When making reporting decisions under SSR mechanisms, agents can choose to be oblivious of how much error presents in others’ reports, because truthful reporting is the dominant strategy, i.e., no matter what uniform reporting strategy other agents play, truthful reporting always maximizes the expected reward. This removes the practical concern of implementing truthful reporting as a particular Nash Equilibrium when there exists a non-truthful reporting equilibrium. (3) Another salient feature of SSR mechanisms is that they transfer the cognitive load of having prior knowledge from the agent side to the mechanism designer side. Yet we do not assume the designer has exact knowledge of the prior either (but the knowledge of whether the prior is greater than 0.5 or not); instead we will leverage the power of estimation from reported data to achieve our goals.

2.6.3 Finite sample analysis

With finite m, n , we use the same procedure as shown in Algorithm 6 to estimate the error rates e_z^+, e_z^- for each agent, except that we cannot have the exact value for $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ but only with finite-sample estimates for them. Specifically, for agent i , letting k_1, k_2, k_3 (which could be different on different tasks) be the three agents whose prediction signals are selected as Z_1, Z_2, Z_3 on each task $k \in [M]$,³ we estimate:

$$\widetilde{\alpha}_{-i} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = 1)}{m}, \quad \widetilde{\beta}_{-i} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = S_{k_2,k} = 1)}{m}, \quad \widetilde{\gamma}_{-i} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = S_{k_2,k} = S_{k_3,k} = 1)}{m}.$$

We then use these three values to replace $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$, respectively, in Algorithm 6 to solve Eqs. 2.4-2.6. We denote the resulted error rates as \widetilde{e}_z^+ and \widetilde{e}_z^- , and the corresponding SSR_α using these error rates as $\widetilde{R}(\cdot)$.

³In practice, we only need to assign task k to these three randomly selected agents.

There are two reasons that these finite-sample estimates \widetilde{e}_z^+ and \widetilde{e}_z^- are not equal to the exact true error rates e_z^+ and e_z^- for Z . First, in constructing Eqs. 2.4-2.6, the error rates of two randomly picked prediction signals Z_2, Z_3 will not have the exactly same error rates with Z , as these signals come from a slightly different agent pool. Second, $\widetilde{\alpha}_{-i}, \widetilde{\beta}_{-i}, \widetilde{\gamma}_{-i}$ are not exactly equal to $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ with finite samples. However, we will show that the errors induced by these two factors in estimating the error rates diminish with m and n . Consequently, the SSR computed using $\widetilde{e}_z^+, \widetilde{e}_z^-$ also have a small and diminishing error towards the SSR computed with the exact error rates e_z^+, e_z^- .

Lemma 2.12. $\widetilde{e}_z^+, \widetilde{e}_z^-$ given by Algorithm 6 using $\widetilde{\alpha}_{-i}, \widetilde{\beta}_{-i}, \widetilde{\gamma}_{-i}$ satisfy that for an arbitrary $\delta \in (0, 1)$, with probability at least $1 - \delta$, $|\widetilde{e}_z^+ - e_z^+| \leq \epsilon$, $|\widetilde{e}_z^- - e_z^-| \leq \epsilon$ for some $\epsilon = O\left(\frac{1}{n} + \sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$, which can be made arbitrarily small with increasing m and n .

Proof Sketch. We present the high-level idea of our proof here and defer the complete proof to the appendix. We consider the two aforementioned errors separately. Both of them can be transformed to a diminishing error attaching to the evaluation of α_{-i}, β_{-i} , and γ_{-i} . This diminishing noise in α_{-i}, β_{-i} , and γ_{-i} can then be transformed into a diminishing error in the final solution of e_z^+, e_z^- . \square

Next, we show that the deviations of the rewards of SSR mechanisms due to the imperfect estimation of the error rates in the finite sample case can also be bounded to be arbitrarily small. We first deal with a special case: even if $1 - e_z^+ - e_z^-$ is far from zero, the estimated $1 - \widetilde{e}_z^+ - \widetilde{e}_z^-$ in the denominator of SSR_α can be arbitrary close to zero by coincidence. In this case, agents can have unbounded scores, which may be far from the exact scores agents should obtain when the estimation is perfect. To address this special case, the principal can select a threshold κ greater but close to zero, and pay agents zero when $|1 - \widetilde{e}_z^+ - \widetilde{e}_z^-| < \kappa$ instead of just when $1 - \widetilde{e}_z^+ - \widetilde{e}_z^- = 0$. As a result, the final reward of each agent is always bounded. Next, we introduce a lemma we will use in our proof.

Lemma 2.13. $\forall l_1, l_2, t_1, t_2 \in [-1, 1], t_1, t_2 \neq 0, \left| \frac{l_1}{t_1} - \frac{l_2}{t_2} \right| \leq \frac{|l_1 - l_2| + |t_1 - t_2|}{|t_1 t_2|}$

Proof. $\left| \frac{l_1}{t_1} - \frac{l_2}{t_2} \right| = \left| \frac{l_1 t_2 - l_2 t_1}{t_1 t_2} \right| = \left| \frac{l_1 t_2 - l_2 t_2 + l_2 t_2 - l_2 t_1}{t_1 t_2} \right| \leq \frac{|t_2| |l_1 - l_2| + |l_2| |t_2 - t_1|}{|t_1 t_2|} \leq \frac{|l_1 - l_2| + |t_1 - t_2|}{|t_1 t_2|}$ \square

This lemma is an extension to Lemma 7 of [LL15], which considers the case where all variables are non-negative. Now we present our main theorem about the diminishing error in estimating the SSR scores.

Theorem 2.14. For a bounded SPSR $S(\cdot)$ with supremum $\max S$, for an arbitrary $\delta \in (0, 1)$, and some $\epsilon = O\left(\frac{1}{n} + \sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$ such that with probability at least $1 - \delta$, $|\widetilde{e}_z^+ - e_z^+| \leq \epsilon$, $|\widetilde{e}_z^- - e_z^-| \leq \epsilon$, let m and n be

sufficiently large such that $\epsilon \leq |1 - e_z^- - e_z^+|/4$, the SSR mechanism built upon $S(\cdot)$ satisfies, with probability at least $1 - \delta$, that

$$|\tilde{R}(q_{i,k}, Z) - R(q_{i,k}, Z)| \leq \frac{12 \max S}{\Delta^2} \cdot \epsilon, \quad \forall i \in [n], k \in [m], q_{i,k} \in [0, 1], Z \in \{0, 1\},$$

where $\Delta = |1 - e_z^- - e_z^+|$. Furthermore, taking over all the randomness in the score, we have

$$\left| \mathbb{E}[\tilde{R}(q_{i,k}, Z)] - \mathbb{E}[S(q_{i,k}, Z)] \right| = O\left(\frac{1}{N} + \sqrt{\frac{\ln m}{m}}\right), \quad \forall i, k.$$

Proof. This proof is straight-forward following the error rate bounding result (Lemma 2.12). We use $\text{sgn}(Z), Z \in \{0, 1\}$ as the superscript, where $\text{sgn}(0)$ refers to super script “-” and $\text{sgn}(1)$ refers to super script “+”.

Consider an arbitrary agent i and a task k , we have

$$\begin{aligned} |\tilde{R}(q_{i,k}, Z) - R(q_{i,k}, Z)| &= \left| \left(\frac{1 - e_z^{\widetilde{\text{sgn}(1-Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{1 - e_z^{\text{sgn}(1-Z)}}{1 - e_z^+ - e_z^-} \right) S(q_{i,k}, Z) \right. \\ &\quad \left. - \left(\frac{e_z^{\widetilde{\text{sgn}(Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{e_z^{\text{sgn}(Z)}}{1 - e_z^+ - e_z^-} \right) S(q_{i,k}, 1 - Z) \right| \\ &\leq \left| \frac{1 - e_z^{\widetilde{\text{sgn}(1-Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{1 - e_z^{\text{sgn}(1-Z)}}{1 - e_z^+ - e_z^-} \right| \max S \\ &\quad + \left| \frac{e_z^{\widetilde{\text{sgn}(Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{e_z^{\text{sgn}(Z)}}{1 - e_z^+ - e_z^-} \right| \max S \end{aligned}$$

Since $\epsilon \leq (1 - e_z^- - e_z^+)/4$, we know that

$$|1 - \widetilde{e}_z^+ - \widetilde{e}_z^-| \geq |1 - e_z^- - e_z^+|/2$$

Thus, with probability at least $1 - \delta$,

$$\begin{aligned} \left| \frac{1 - e_z^{\widetilde{\text{sgn}(1-Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{1 - e_z^{\text{sgn}(1-Z)}}{1 - e_z^+ - e_z^-} \right| &\leq \frac{\left| e_z^{\widetilde{\text{sgn}(1-Z)}} - e_z^{\text{sgn}(1-Z)} \right| + \left| \widetilde{e}_z^+ + \widetilde{e}_z^- - e_z^+ - e_z^- \right|}{|(1 - \widetilde{e}_z^+ - \widetilde{e}_z^-)(1 - e_z^+ - e_z^-)|} \\ &\leq \frac{3\epsilon}{|(1 - \widetilde{e}_z^+ - \widetilde{e}_z^-)(1 - e_z^+ - e_z^-)|} \leq \frac{6\epsilon}{\Delta^2} \end{aligned}$$

In above inequalities, the first “ \leq ” follows Lemma 2.13, the second follows Lemma 2.12, and the third follows $|1 - \widetilde{e}_z^+ - \widetilde{e}_z^-| \geq |1 - e_z^- - e_z^+|/2$. Similarly, we have

$$\left| \frac{e_z^{\widetilde{\text{sgn}(Z)}}}{1 - \widetilde{e}_z^+ - \widetilde{e}_z^-} - \frac{e_z^{\text{sgn}(Z)}}{1 - e_z^+ - e_z^-} \right| = \frac{6\epsilon}{\Delta^2}$$

Plugging back, we have proved the claim that with probability at least $1 - \delta$,

$$|\tilde{R}(q_{i,k}, Z) - R(q_{i,k}, Z)| \leq \frac{12\epsilon \cdot \max S}{\Delta^2}, \quad \forall q_{i,k} \in [0, 1], Z \in \{0, 1\}.$$

As $\mathbb{E}[S(q_{i,k}, Z)] = \mathbb{E}[R(q_{i,k}, Z)]$, letting $\delta = \frac{1}{m}$, we have the expected error $|\mathbb{E}[\tilde{R}(q_{i,k}, Z)] - \mathbb{E}[S(q_{i,k}, Z)]|$ bounded by $O\left(\left(1 - \frac{1}{m}\right)\left(\frac{1}{n} + \sqrt{\frac{\ln m}{m}}\right) + \frac{1}{m}\right) = O\left(\frac{1}{m} + \sqrt{\frac{\ln m}{m}}\right)$. □

Theorem 2.14 indicates that the errors of the expected scores given by SSR mechanisms w.r.t. the expected score given by the underlying SPSR can be made arbitrary small with sufficiently large m and n . As a result, for arbitrarily discretized report space of a prediction, SSR mechanisms are still dominant uniform strategy truthful with finite but sufficiently large m and n . To see this, we can make the error smaller than the minimum absolute difference of the SPSRs of any two allowed probability reports. In such way, there will be no beneficial deviation for agents to report non-truthfully. This result considers the reality that in real surveys, agents are often allowed to specify at most two decimal digits for probabilistic predictions.

Corollary 2.15. *For discretized report space of probabilistic predictions, SSR mechanisms which are built upon bounded SPSR are dominant uniform strategy truthful for finite but sufficiently large m and n .*

2.7 Generalizations to multi-outcome tasks

In this section, we discuss how to extend SSR and SSR mechanisms to the multi-outcome multi-task setting. A multi-outcome task asks agents to provide predictions about a multi-outcome random variable Y , which takes value from a finite support set $[c] = \{0, \dots, c-1\}$ with $c > 2$. A noisy estimate $Z \in [c]$ of the ground truth Y is characterized by a confusing matrix:

$$E_Z = \begin{bmatrix} e_{0,0} & e_{0,1} & \dots & e_{0,c-1} \\ e_{1,0} & e_{1,1} & \dots & e_{1,c-1} \\ \dots & \dots & \dots & \dots \\ e_{c-1,0} & e_{c-1,1} & \dots & e_{c-1,c-1} \end{bmatrix},$$

where $e_{u,v}$ represents the flipping probability of Z w.r.t. Y , i.e., $e_{u,v} = \Pr[Z = v | Y = u], \forall u, v \in [c]$.

2.7.1 Generalization of SSR

The surrogate scoring rules for a task with c outcomes are defined as follows. Let Δ^{c-1} be the $(c-1)$ -dimension probability simplex, i.e., $\Delta^{c-1} := \{(x_0, \dots, x_{c-1}) \mid \sum_{i=0}^{c-1} x_i = 1, x_0, \dots, x_{c-1} \geq 0\}$.

Definition 2.8 (Surrogate Scoring Rules). $R : \Delta^{c-1} \times [c] \rightarrow \mathbb{R}$ is a surrogate scoring rule for a c -outcome task if for some strictly proper scoring rule $S : \Delta^{c-1} \times [c] \rightarrow \mathbb{R}$ and a strictly increasing function $f : \mathbb{R} \rightarrow \mathbb{R}$, the following equation holds:

$$\forall \mathbf{p}, \mathbf{q} \in \Delta^{c-1}, \forall E_Z \in [0, 1]^{c \times c} (E_Z \text{ is invertible}) : \underline{\mathbb{E}_Z[R(\mathbf{q}, Z)]} = f(\mathbb{E}_Y[S(\mathbf{q}, Y)]),$$

where the ground truth Y is drawn from Categorical(\mathbf{p}) and Z is a noisy estimate of Y with confusing matrix E_Z .

We have the following theorem immediately.

Theorem 2.16. Given the prior \mathbf{p} of the ground truth Y and a private signal O_i , SSR $R(\mathbf{q}, z)$ with a noisy estimate Z of the ground truth is strictly proper for eliciting an agent's posterior $\mathbf{p}_i := \Pr[Y|O_i]$ if Z and O_i are independent conditioned on Y and E_Z is invertible.

Now we give an implementation of SSR, SSR_α , for a c -outcome task. Let $S(\mathbf{q}_i)$ be the vector of SPSR scores for a prediction $\mathbf{q}_i \in \Delta^{c-1}$ under each realization of Y , i.e., $S(\mathbf{q}_i) := (S(\mathbf{q}_i, Y = 0), \dots, S(\mathbf{q}_i, Y = c-1))$. Similarly, let $R(\mathbf{q}_i) := (R(\mathbf{q}_i, Z = 0), \dots, R(\mathbf{q}_i, Z = c-1))$. Our implementation SSR_α goes as follows:

$$R(\mathbf{q}_i) := (E_Z)^{-1} S(\mathbf{q}_i)$$

Clearly, for SSR_α we have $S(\mathbf{q}_i) = E_Z \cdot R(\mathbf{q}_i)$, which gives

$$\forall v \in [c], S(\mathbf{q}_i, Y = v) = \sum_{k=0}^{c-1} e_{v,k} R(\mathbf{q}_i, Z = k) = \mathbb{E}_{Z|Y=v}[R(\mathbf{q}_i, Z)].$$

Lemma 2.17. For SSR_α : $\forall v \in [c], \mathbb{E}_{Z|Y=v}[R(\mathbf{p}_i, Z)] = S(\mathbf{p}_i, Y = v)$

The following theorem follows immediately.

Theorem 2.18. SSR_α is a surrogate scoring rule for a multi-outcome task, and for any distribution $\mathbf{p} \in \Delta^{c-1}$ of the ground truth Y and for any invertible confusing matrix E_Z of a noisy estimate Z of the ground truth, we have $\forall \mathbf{q} \in \Delta^{c-1}, \mathbb{E}_Z[R(\mathbf{q}, Z)] = \mathbb{E}_Y[S(\mathbf{q}, Y)]$.

We include a detailed example of SSR_α for a three-outcome task below.

Example 2.1. Let $c = 3$ and let the confusing matrix of a noisy signal Z being

$$E_z = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \Rightarrow (E_z)^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

We obtain a closed-form of SSR_α :

$$R(\mathbf{q}, Z = 0) := 3S(\mathbf{q}, 0) - S(\mathbf{q}, 1) - S(\mathbf{q}, 2)$$

$$R(\mathbf{q}, Z = 1) := -S(\mathbf{q}, 0) + 3S(\mathbf{q}, 1) - S(\mathbf{q}, 2)$$

$$R(\mathbf{q}, Z = 2) := -S(\mathbf{q}, 0) - S(\mathbf{q}, 1) + 3S(\mathbf{q}, 2)$$

2.7.2 Generalization of SSR mechanisms

SSR mechanisms can also be extended to multi-outcome tasks and maintain the two properties we pursue: the dominant uniform strategy truthfulness and qualifying the value of information as what SPSR do.

We consider the same setting of information structures under Assumptions 2.1-2.3, except that $Y_k, k \in [m]$ in these assumptions are c -outcome categorical random variables, agents' beliefs are categorical distributions, and that in Assumption 3, the prior probabilities of Y_k being each outcome are different and the principal knows the order of these prior probabilities. As we have shown that SSR can be extended to multi-outcome events, to construct the corresponding SSR mechanism, we just need to construct the corresponding noisy estimate Z of the ground truth and estimate the confusion matrix E_z for multi-outcome tasks.

The noisy estimate Z for an agent i on task k can be constructed similarly as the counterpart in the binary case, i.e., we uniformly randomly pick an agent $j \neq i$ and draw $Z \sim \text{Categorical}(\mathbf{q}_{j,k})$, where $\mathbf{q}_{j,k}$ is the reported distribution of Y_k from agent j . Then, the confusion matrix can also be estimated using the method of moments. However, as there are $c^2 - 1$ unknown parameters in the confusion matrix E_z and the prior \mathbf{p} of Y_k , we have to establish $c^2 - 1$ equations. These equations could be solved numerically. These $c^2 - 1$ equations will have $c!$ real-value symmetric solutions, each corresponds to a permutation of the labeling of the c outcomes. To identify the unique solution that yields the true confusion matrix and the prior of Y , i.e., to identify the correct labeling of the outcomes, the principal has to know the order of the prior probabilities of Y_k being each outcome, as what we assume in Assumption 2.3 for multi-outcome tasks. Thus, with the multi-task variant of Assumptions 2.1-2.3, we can still construct a noisy estimate Z of the ground truth, estimate its confusion matrix, and apply SSR to obtain unbiased

estimates of agents' scores given by the underlying SPSR.

Despite the positive result in theory, there are some caveats of applying SSR mechanisms to multi-outcome tasks. First, Assumption 2.1 essentially assume that the confusion matrix of an agent is homogeneous across different tasks. However, as there is no clear correspondence between the labels of the outcomes of different tasks, the confusion matrix of the noisy estimate Z for an agent is less likely to be homogeneous across different tasks. Therefore, the real data can deviate far from Assumption 2.1. Second, as there are more parameters in the confusion matrix to estimate in the multi-task case than in the binary case, we need a much larger number of agents and tasks and denser predictions to maintain decent estimation accuracy. Third, to apply a SSR mechanism to multi-outcome tasks, these tasks have to have the same number of outcomes. However, in most crowd forecasting projects, the number of multi-outcome tasks with the same number of outcomes is much smaller than the number of binary questions and may not be sufficient to make accurate estimation of the confusion matrix. These caveats leave a massive space for future research.

2.8 Empirical studies

Using 14 real-world human forecasting datasets, we empirically examine the performance of SSR mechanisms in revealing agents' prediction accuracy in terms of SPSR. We focus on three aspects: the unbiasedness of SSR, the correlation of SSR scores to SPSR scores, and the accuracy of SSR in selecting true top forecasters in terms of SPSR. We also compare the performance of SSR mechanisms to several existing peer prediction mechanisms. The overall results show that our SSR mechanisms have an advantage in recovering SPSR.

2.8.1 Setting

Datasets

We conduct our experiments on 14 datasets from three human forecasting and crowdsourcing projects: the Good judgment Project (GJP), the Hybrid Forecasting Competition (HFC), and the human judgment datasets collected by MIT. These three projects differ in participant population, forecasting topics, and elicitation methods, offering a rich environment for empirical evaluation.

GJP datasets [Ata+16] The GJP data consists of four datasets for geopolitical forecasting questions. The four datasets, denoted by G1~G4, were collected from 2011 to 2014, respectively. They contain

Items	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
# of questions (original)	94	111	122	94	44	86	203	50	50	50	80	80	90	90
# of agents (original)	1972	1238	1565	7019	79	317	222	51	32	33	39	25	20	20
After applying the filter														
# of questions	94	111	122	94	44	86	203	50	50	50	80	80	90	90
# of agents	1409	948	1033	3086	79	316	222	51	32	33	39	25	20	20
Avg. # of answers per question	851	533	369	1301	71	295	220	51	32	33	39	18	20	20
Avg. # of answers per agent	57	62	44	40	39	80	201	50	50	50	80	60	90	90
Majority vote correct ratio (%)	0.90	0.92	0.95	0.96	0.93	0.93	0.86	0.58	0.76	0.74	0.61	0.68	0.62	0.72

Table 2.1: Statistics about binary-outcome datasets from GJP, HFC and MIT datasets

different sets of forecasting questions and forecasters.

HFC datasets [IAR19] We use the forecast data of team participants in the Hybrid Forecasting Competition. The data consists of three datasets, denoted by H1~H3, referring to the forecasting data collected in the preseason competition, the first competition, and the second competition, respectively. The the preseason competition lasted half a year, and the two formal competitions lasted around one year. The three datasets have different forecasting questions and partially overlapped participating teams.

MIT datasets [PSM17] The MIT data consists of seven datasets, denoted by M1a, M1b, M1c, M2, M3, M4a, M4b, respectively. Each dataset uses one of four sets of questions and has a different participant pool. The questions range from guessing the capital of each state and predicting the price interval of artworks to some trivia questions. The forecasters were students in class and colleagues in labs. In datasets M1a, M1b, M4a, M4b, forecasters report only binary votes on forecasting questions. In datasets M1c, M2, M3, forecasters give probabilistic predictions.

Both GJP and HFC projects allow participants to make daily forecasts. For testing peer prediction mechanisms in our setting, we only need to use a single prediction for each participant on a forecasting question. In our experiments, we use the final prediction of each participant made on each question and ignore the other predictions in these two projects. Also, we focus on the forecasting questions which have binary outcomes in these datasets. To have a relatively stable estimation over the accuracy of agents, we filter out participants who made predictions on less than 15 questions. The basic statistics of these datasets are presented in Table 2.1.

SPSR

We consider three SPSR, the Brier score, the log scoring rule, and the rank-sum scoring rule, because of their usage in practice and connections to machine learning concepts. The first two are the most widely adopted scoring rules. They are equivalent to two main loss functions, the squared error and the cross-entropy loss, respectively, used in the machine learning community. The rank-sum scoring rule can be written as an affine transformation of the area under the receiver operating characteristic curve (AUC-ROC),⁴ which is also a widely adopted accuracy metric in the machine learning community.

In our experiments, we adopt the conventional formula of the Brier score used in the GJP and HFC projects. The Brier score ranges from 0 to 2, with a smaller score corresponds to higher accuracy. This is different from using SPSR as a payment method, where the higher the better. We can transfer between these two usages by applying a negative scalar. We orient the log scoring rule and the rank-sum score rule in the same direction as the Brier score, with a minimum (best) score of 0. The exact formula for each scoring rule is as follows: Recall that $q_{i,k}$ and Y_k are agent i 's prediction and the ground truth for task k , respectively, and $[m_i]$ is the set of tasks answered by agent i .

- **Brier score:** $S^{\text{Brier}}(q_{i,k}, Y_k) = 2(q_{i,k} - Y_k)^2$. We use the mean Brier score, $\frac{1}{m_i} \cdot \sum_{k \in [m_i]} S^{\text{Brier}}(q_{i,k}, Y_k)$, to represent an agent's overall accuracy under the Brier score over the set of tasks she answered.
- **Log scoring rule:** $S^{\text{log}}(q_{i,k}, Y_k) = Y_k \log(q_{i,k}) + (1 - Y_k) \log(1 - q_{i,k})$. We use the mean log score, $\frac{1}{m_i} \sum_{k \in [m_i]} S^{\text{log}}(q_{i,k}, Y_k)$, to represent an agent's overall accuracy under the long scoring rule over the tasks she answered. As the log scoring rule is unbounded when the forecast predicts the opposite of the ground truth, we change all forecasts of 1 to 0.99 and forecasts of 0 to 0.01 to ensure that the score is always a real number.
- **Rank-sum scoring rule** is a multi-task scoring rule. For a single task k , it assigns a score

$$S^{\text{rank}}(q_{i,k}, y_k) = -y_k \cdot \psi \left(q_{i,k} | \{q_{i,k'}\}_{k' \in [m_i]} \right),$$

where $\psi \left(q_{i,k} | \{q_{i,k'}\}_{k' \in [m_i]} \right) := \sum_{k' \in [m_i]} \mathbb{1}(q_{i,k'} < q_{i,k}) - \sum_{k' \in [m_i]} \mathbb{1}(q_{i,k'} > q_{i,k})$ is the rank of prediction $q_{i,k}$ among all predictions from agent i . Then, agent i 's rank-sum score S_i^{rank} is defined as:

⁴The affine transformation coefficients are determined by the numbers of the tasks with ground truth 1 and with ground truth 0. (according to Eqs. 12 and 13, [Par+16]). Thus, when evaluating agents' prediction accuracy on the same set of answered questions, the rank-sum scoring rule is equal to the AUC-ROC for each agent up to the same affine transformation determined by the ground truth of the questions. However, the AUC-ROC itself is not an SPSR, as when considering the incentive, the affine transformation coefficients may differ in different agents' beliefs.

$S_i^{\text{rank}} = \sum_{k \in [m_i]} S^{\text{rank}}(q_{i,k}, Y_k)$.⁵ The range of the score increases with the number of answered tasks quadratically, thus we use the normalized score $1 + \frac{4}{m_i^2} S_i^{\text{rank}}$ with range $[0, 2]$.

Treatments

Though existing peer prediction methods are not designed for recovery of SPSR, we add comparisons to them for completeness of our study.⁶ In particular, we would like to understand whether in practice SSR mechanisms have the advantage of revealing the true scores given by SPSR while not accessing ground truth information.

In our experiments, we consider four popular existing peer prediction methods, serving as comparisons to SSR: proxy scoring rules (PSR) with extremized mean [Wit+17], peer truth serum (PTS) [RFJ16], correlated agreement (CA) [Shn+16], determinant mutual information (DMI) [Kon20].

PSR is to directly apply the SPSR w.r.t. an unbiased proxy of the ground truth. When the principal knows no unbiased proxy, Witkowski et al. [Wit+17] recommend using the extremized mean of the reported predictions to serve as the proxy. In our experiments, we adopt the same formula for the extremized mean as in their experiments [Wit+17], i.e., $\frac{\bar{q}_k^2}{\bar{q}_k^2 + (1 - \bar{q}_k)^2}$, where \bar{q}_k is the average reported prediction on task k . Using different SPSR as the underlying scoring rule, we can get different PSR and SSR.

PTS, CA, and DMI do not depend on SPSR and are designed to elicit categorical labels instead of probabilistic predictions. So we make the following adaption for them to take probabilistic predictions as inputs. Our adaption is based on the fact that in essence, these mechanisms all appreciate the joint distribution of agents' reported labels to compute the scores: For a task k , an agent who reports probability $P_{i,k}$ believes that the true label of the task has probability $P_{i,k}$ to be 1. Therefore, on this task, the joint probability of agent i 's believed true label and agent j 's believed true label both being 1 is $P_{i,k}P_{j,k}$, assuming their believed true labels are independent conditioned on their predictions. By this way, we can compute the joint distribution of the believed true labels of two peer agents on each task and their joint distribution over the whole dataset is the mean of their joint distributions on each task. Using this joint distribution over the whole dataset, we can compute the scores for PTS, CA, and DMI directly. This adaption method for PTS, CA, and DMI turns out to give better correlations between

⁵The AUC-ROC of agent i is $\frac{1}{2} \left(1 - \frac{1}{m_i^+ (m_i - m_i^+)} S_i^{\text{rank}} \right)$, where $m_i^+ := \sum_{k' \in [m_i]} \mathbb{1}(Y_{k'} = 1)$ (given by Eqs. 12 and 13, [Par+16]).

⁶We do not intend to claim our mechanism is better in any sense, as it would be an unfair comparison since the goals were different in each design of these mechanisms. For example, the mechanisms [Shn+16; Kon20] can characterize determinant mutual information between an agent's reports and the underlying ground truth.

the scores of these three mechanisms and the true SPSR scores than the alternative adaption method of using the mostly likely categorical labels indicated by the probabilistic predictions as inputs for these mechanisms (see how the correlations shown in Figs. A.1 and A.2 in the appendix (most likely labels as inputs) compare to the correlations shown in Figs. 2.2 and 2.3.).

2.8.2 Main results

Unbiasedness of SSR Our theorem shows that under certain assumptions, the reward of an SSR mechanism is unbiased to the reward of the SPSR that the SSR mechanism is built upon. However, it is unclear to what extent this unbiasedness holds in real datasets where these assumptions are unlikely to hold strictly. Therefore, we empirically examine the concrete relationship between SSR scores and the corresponding SPSR scores.

Fig. 2.1 plots the score pairs received by forecasters in each of the 14 datasets. Each score pair represents the SPSR score and the SSR score that an individual forecaster receives in a single dataset. As can be seen, under each of the three SPSR we test, the SSR scores demonstrate a salient linear relationship to the true SPSR scores. We further draw a linear regression curve between the SSR scores and the true scores for each of the three SPSR of interest (the blue curves in Fig. 2.1). To draw this linear regression curve, we first cluster the score pairs into different groups based on the value of the SPSR scores and compute the center point (the mean score pair) for each group, represented by the orange triangles in Fig. 2.1. Then, we regress on these center points.⁷ The three regression curves demonstrate a slope of 0.74, 0.73, and 0.84, respectively, all with an intercept near 0. This result indicates that though the SSR scores are not exactly unbiased in real data, they still follow an affine transformation of the true SPSR scores with decent approximate unbiasedness.

We also notice that under all three SPSR, the SSR scores tend to underestimate the true scores by around 20%. As the SSR scores follow an affine transformation of the SPSR scores empirically, this underestimation can possibly be mitigated by applying a constant scaling factor (e.g., 1.25 as suggested by our regression) without influencing the incentive properties of the SSR mechanisms.

⁷The reason for clustering score pairs before regression is that the SPSR scores of forecasters are not distributed evenly within the range of the SPSR score, with most forecasters' SPSR scores falling in the low range of the SPSR score. Consequently, drawing the regression curve directly on all score pairs will mainly reflect the regression pattern in the low range of the SPSR score instead of the whole range. In fact, for each of the three SPSR tested, the corresponding SSR mechanism obtains a regression slope closer to 1 at the low range of the SPSR score.

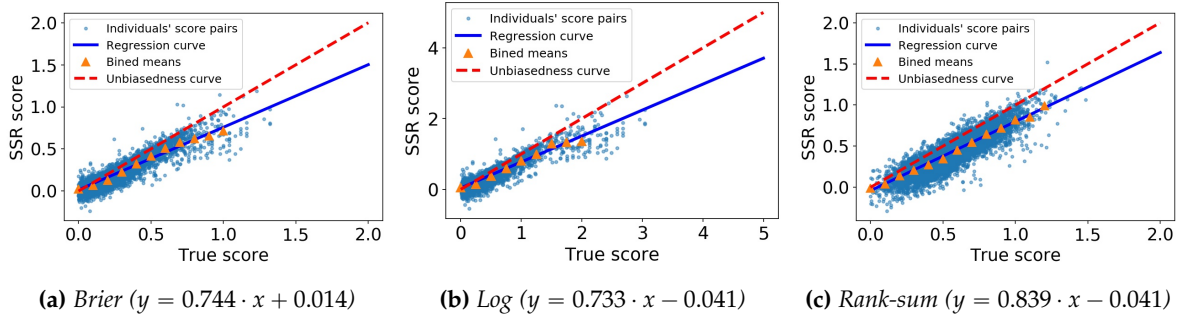


Figure 2.1: Regression of individuals' true accuracy and SSR score over 14 datasets under three different SPSR.

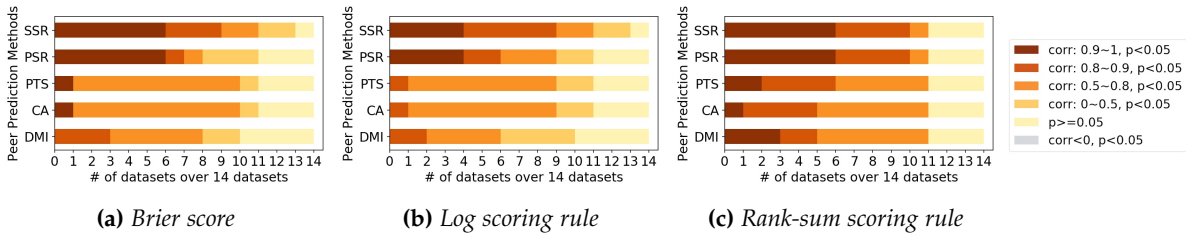


Figure 2.2: The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSR.

Correlation with SPSR We compare the correlations between agents' SPSR scores and the scores given by the five peer prediction mechanisms we test. We first measure the correlations on each dataset independently using Pearson's correlation coefficient (*corr*) and then classify them into different levels based on the value of the coefficient. Finally, we count the number of datasets at different correlation levels for each peer prediction mechanism and present the results in Fig. 2.2. As can be seen, all five peer prediction mechanisms achieve a strong correlation ($\text{corr} > 0.5$) to the SPSR on half of the 14 datasets, while the SSR mechanisms demonstrate an even stronger correlation pattern. In particular, the SSR mechanisms achieve a very strong correlation ($\text{corr} > 0.8$) on 9 out of the 14 datasets under all three SPSR, and achieve correlations in more datasets than other mechanisms for each of the following levels: $\text{corr} > 0.9$, $\text{corr} > 0.8$, and $\text{corr} > 0.5$. The advantage of SSR in the correlation to the SPSR is most salient under the Brier score and is more salient when compared to the PTS, CA, DMI mechanisms than compared to the PSR mechanisms. We observe similar results using Spearman's rank correlation test (Fig. 2.3), which implies that SSR mechanisms also rank the agents similarly to SPSR.

The performance of SSR mechanisms in reflecting the true SPSR scores depends on the accuracy of estimating the error rates of the constructed noisy estimate of ground truth in SSR mechanisms. This estimation accuracy depends on the number of prediction samples that SSR mechanisms have

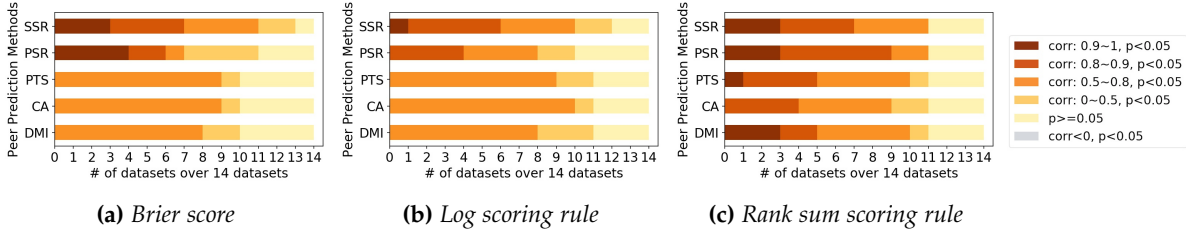


Figure 2.3: The number of datasets in each level of correlation (measured by Spearman’s correlation coefficient) between individuals’ peer prediction scores and different SPSR.

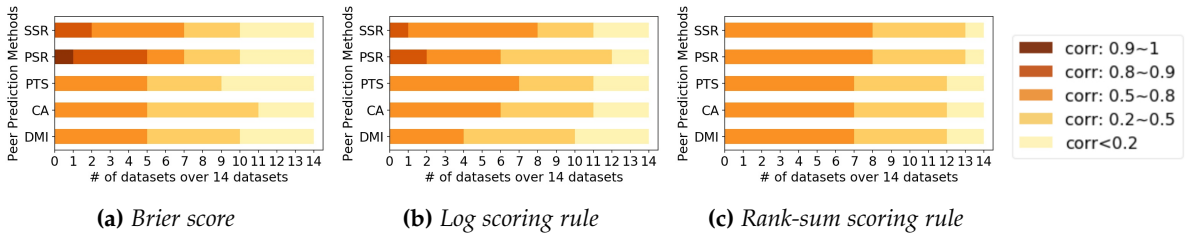


Figure 2.4: The number of datasets in each level of correlation (measured by Pearson’s correlation coefficient) between individuals’ peer prediction scores and different SPSR on sampled datasets (the correlation is averaged over 100 runs of random sampling).

access to. In our previous experiments, each task receives a considerable number of predictions (no less than 20 on average), which may give an edge to the SSR mechanisms. However, a principal with a limited budget can often collect only a small number of predictions for each task. Therefore, we are also interested in comparing the performance of SSR mechanisms to other peer prediction mechanisms when each task receives only a limited number of predictions. To simulate this scenario, for each original dataset, we sample a subset of users to create a new dataset such that each new dataset has an average of 4~5 predictions per task with a minimum of 3 predictions, which is the minimum number of predictions per task required by our SSR mechanisms.⁸ Fig. 2.4 shows the correlation results of each peer prediction mechanism based on the average Pearson’s correlation coefficient over 100 runs of random sampling. As can be seen, overall, the correlations between each peer prediction mechanism and the three SPSR in these sampled datasets decrease when compared to the corresponding correlations in the original datasets. SSR mechanisms still maintain a strong correlation ($\text{corr}>0.5$) over half of

⁸To ensure a minimum of 3 predictions per task, we removed a small number of tasks that receive less than 3 predictions by this sampling method. Over the 100 runs of random sampling, around 20 tasks are removed on average from each GJP dataset, and no more than 2 tasks are removed from each of the other datasets in each run. This sampling operation keeps a decent number of predictions for each agent, which allows a stable computation for the scores of SSR, PTS, CA, and DMI mechanisms.

the 14 datasets, while the other mechanisms do not. However, the performance difference of SSR and other mechanisms shrinks. The PSR mechanisms outperform SSR mechanisms at two correlation levels, $\text{corr} > 0.8$ and $\text{corr} > 0.9$, under the Brier score and the log scoring rule. In fact, the single-task PSR mechanisms demonstrate smaller correlation decreases, indicating that they are more robust to the number of predictions than the other four multi-task mechanisms.

Expert identification SPSR are sometimes used to identify top forecasters to assign prizes, e.g., in projects GJP and HFC. Moreover, accurate identification of true top forecasters without access to the ground truth can help improve the aggregation accuracy, when a principal wants to aggregate forecasters’ predictions in to a final prediction for each task [WLC19]. Therefore, we examine to what extent different peer prediction scores can identify top-performing experts in terms of the true SPSR, without access to the ground truth. We first rank the forecasters according to one of the three SPSR (when the rank-sum scoring rule is chosen, we use the AUC-ROC instead to evaluate agents’ true accuracy, because as an accuracy metric instead of an incentive device, AUC-ROC is much more popular than the rank-sum scoring rule). We focus on two metrics about expert identification: (i) the percentage of top $t\%$ forecasters identified by the SPSR in the top $t\%$ forecasters selected by a peer prediction method, and (ii) the percentage of below-average forecasters (the bottom 50% forecasters) under the SPSR in the top $t\%$ forecasters selected by a peer prediction method. The results are shown in Figs. 2.5 and 2.6. We find that for both the Brier score and the log score, there are more true top $t\%$ forecasters in the top $t\%$ forecasters selected by SSR than in the top $t\%$ forecasters selected by other peer prediction mechanisms, when $t\%$ ranges from 5% to 50%. Meanwhile, there are less true below-average forecasters in the top $t\%$ forecasters under SSR and PSR mechanisms than under the other peer prediction mechanisms. For AUC-ROC, while the SSR mechanism maintains a relatively smaller number of true below-average forecasters in its top 10% to 15% forecasters, all five peer prediction mechanisms perform similarly, which echos the correlation results under the rank-sum scoring rule, where the five peer prediction mechanisms all achieve strong correlation in most of the datasets (Fig. 2.2c).

2.9 Discussion

In this chapter, we propose the SSR mechanisms such that truthful reporting one’s posterior belief is a dominant strategy in the multi-task IEWV setting, when each agent uses a consistent reporting strategy across all tasks. Moreover, the reward of a prediction given by an SSR mechanism quantifies the value of information in expectation as if the prediction is assessed by the corresponding SPSR with access

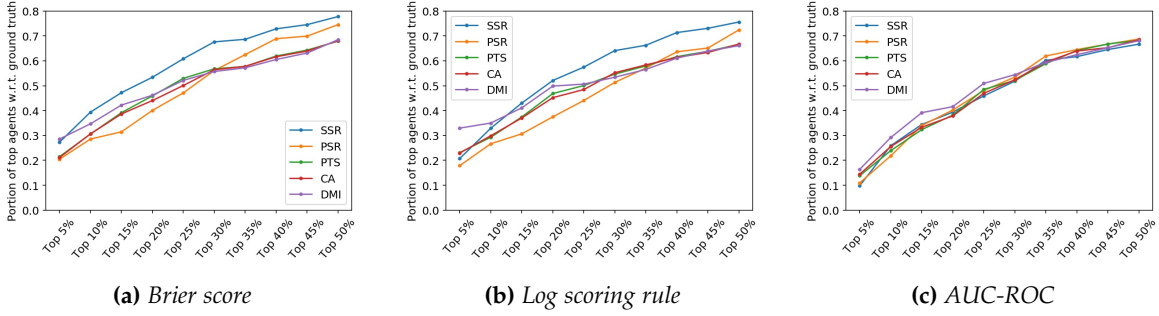


Figure 2.5: The portion of top $t\%$ forecasters w.r.t. 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ forecasters selected by different methods (averaged over 14 datasets).

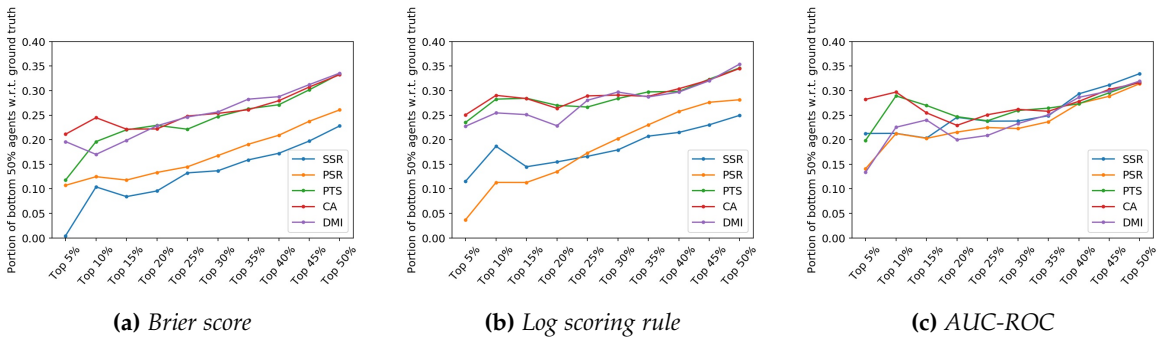


Figure 2.6: The portion of bottom 50% forecasters w.r.t. 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ users selected by different methods (averaged over 14 datasets).

to the ground truth. Because of these two properties, our mechanisms are particularly suitable for information elicitation scenarios where using SPSR to reward agents are favored but the ground truth is not available in time, such as forecasting long-term geopolitical events and predicting the replicability of social science studies.

There are also some limitations of applying our models and mechanisms. First, our Assumption 2.2 requires agents' signals on a task to be independent conditioned on the ground truth Y . This implies that our SSR mechanisms only apply to scenarios where there exists such an objective ground truth or where there is no objective ground truth but the agents' signals are correlated only through a single latent variable. An example of the latter is asking an agent how likely an essay is well-written or not. Although whether an essay is well-written or not may not have an objective answer, as long as the agents' signals are independent conditioned on a latent variable that captures the real quality of the essay, our mechanisms should incentivize truthful reporting as a dominant strategy when all agents adopt uniform strategies across tasks. In comparison, most existing multi-task peer prediction mechanisms [e.g. RFJ16;

[Shn+16](#); [Kon20](#)] that elicit categorical signals do not require agents' signals to be correlated only through a latent variable. Instead, they allow a broader correlation pattern (e.g., self-predicting [[RFJ16](#)]) or arbitrary correlations as long as signals are not completely independent [e.g. [Shn+16](#); [Kon20](#)].

Second, to estimate the error rates of the noisy estimate of the ground truth, our mechanisms require at least three reports for each task. In contrast, several multi-task mechanisms [e.g. [DG13](#); [RFJ16](#); [Shn+16](#); [Kon20](#)] only need one peer agent to achieve their incentive properties. Moreover, the variance of the rewards of SSR mechanisms depends on the number of tasks and reports that the mechanisms have access to. A relatively large number of tasks and reports is needed to obtain a low-variance reward for each agent. As can be seen from our empirical study, although SSR mechanisms still maintain better correlations to the true SPSR scores than the other mechanisms when there are only a few reports per task, SSR mechanisms have a more salient correlation decrease when compared to the case where each task receives a sufficient number of answers. SSR mechanisms are more sensitive to the size of the dataset. However, our analysis suggests that as long as agents adopt uniform strategies across tasks, it is possible to learn the statistical patterns of agents' reports without influencing the incentive. Therefore, a future direction to mitigate SSR mechanisms' sensitivity to the amount of data is to develop or adopt more sophisticated estimation algorithms that require fewer tasks and reports to achieve stable performance.

Chapter 3

Forecast Aggregation via Peer Prediction

3.1 Introduction

Forecasting is one of the main areas where collective intelligence is frequently garnered. In crowd forecasting, a pool of human participants are invited to make forecasts on a set of prediction questions of interest and the solicited forecasts are then aggregated to obtain final predictions. Crowd forecasting has been widely applied in solving challenging forecasting tasks such as forecasting geopolitical events [Ata+16], predicting the replicability of social science studies [Liu+20], diagnosing skin lesions [PSM17] and labeling training sets for machine classifiers [LPI12].

Aiming to more effectively leverage collective intelligence in forecasting, we focus on improving multi-task forecast aggregation in this chapter. We consider a minimal-information setting where each participant offers a single prediction to each forecasting question of a subset of total forecasting questions, and no other information such as participants' historical performance is available. By exploring only hidden information in participants' predictions over multiple questions, we develop a family of aggregation methods that robustly improves the accuracy of the final predictions across a variety of datasets.

The minimal-information setting requires the least effort to collect information and put almost no constraints on crowdsourcing workflow. Our methods can be used during the cold-start stage of long-term forecasting [Ata+16], where no event has been resolved yet to evaluate participants' performance. They can also serve as elegant benchmarks for developing more complex aggregators when additional information is available.

Our approach is to leverage peer forecasts to generate a proxy evaluation of each forecaster's

performance that potentially positively correlates with her true performance. We call such proxy evaluations peer assessment scores (PAS). We then develop PAS-aided aggregators that build upon simple aggregators, such as mean. Our PAS-aided aggregators set larger weights in the simple aggregators on predictions from forecasters who obtain higher PAS.

The question then boils down to how to generate credible PAS evaluations. We are blessed by recent advances in the *peer prediction* literature. Peer prediction mechanisms are a family of reward mechanisms designed to use only peer reports on forecasting questions to motivate crowd forecasters to provide truthful or high-quality forecasts in the absence of the ground truth [MRZ05a]. While they are primarily developed for the purpose of forecast elicitation, Liu, Wang, and Chen [LWC20] and Kong [Kon20] revealed theoretically that the rewards given by their mechanisms correlate positively with the prediction accuracy (defined using the ground truth) under certain conditions. Liu, Wang, and Chen [LWC20] also showed empirical evidence of this correlation for several other peer prediction mechanisms. These mechanisms are potentially tools to use to construct the PAS-aided aggregators.

In this chapter, we explore the use of five recently proposed peer prediction mechanisms [RFJ16; Shn+16; Wit+17; LWC20; Kon20] as PAS. After showing their theoretical properties in recovering the forecasters' true performance, we thoroughly examine the empirical performance of PAS-aided aggregators built upon them. We employ 14 real-world human forecast datasets and two widely-adopted accuracy metrics, the Brier score and the log score. We compare the performance of these PAS-aided aggregators with four representative existing aggregators that neither require knowing the ground truth of resolved historical forecasting questions: the mean aggregator [JW08; MLS12], the logit-mean aggregator, which is based on the idea of extremization of predictions [ACR12; Sat+14a; Bar+14], a statistical-inference-based aggregator [LPI12], and the minimal pivoting aggregator, which is based on "surprising popularity." [PSM17; PS19]

Our results reveal: 1) Though each of the above four existing aggregators has strong performance on specific datasets, none of them has consistent, robust performance across all datasets. 2) In contrast, our PAS-aided aggregators demonstrate a significant and consistent improvement in the aggregation accuracy compared to the four existing aggregators. 3) These PAS-aided aggregators adopt a very intuitive (*explainable*) and straightforward (*generically applicable*) strategy to incorporate PAS: select top forecasters according to their PAS and apply the mean or the logit-mean aggregator to the predictions of these selected forecasters. 4) Moreover, this improvement is observed when any one of the five peer prediction mechanisms is used as PAS, and there is no statistically significant difference found in the improvements when different PAS are used. 5) The above results demonstrate the possibility of discovering a smaller but smarter crowd in real-time forecast aggregation without accessing any ground

truth outcomes.

We want to emphasize that aggregation without access to historical ground truth information is an incredibly challenging problem. One cannot expect that there is a universal aggregator that has the best performance on all datasets. There isn't. Instead, we hope to devise aggregators that perform well and robustly on different datasets. The significance of our work is three-fold. First, it provides a framework to select forecasts to achieve more robust and accurate aggregation. Second, our method can be used as a booster to aggregators in almost all multi-task forecast aggregation scenarios since it has minimal information requirements. Third, our work reveals a new and meaningful application of peer prediction methods - as scoring mechanisms to identify top experts and to improve forecast aggregation.

3.2 Related Work

Our work considers the multi-task forecast aggregation setting, where there is a set of (independent) judgement questions to forecast and each participant forecasts on multiple questions. A large part of the forecast aggregation literature considers the single-task setting, where all participants predict about a simple forecasting question. The methods and aggregators designed for the single-task setting are also often used in the multi-task forecast setting directly. Single-task aggregators include the mean, median, their trimmed variants [Gal07; Cle89; JW08; MLS12], the aggregators that extremize the mean predictions [RG10; Bar+14; ACR12; Sat+14a], and the "surprising-popularity"-based aggregators [PSM17; PS19; PS20], which use the additionally collected participants' estimates about the other participants' forecasts to help aggregation. The aggregators proposed in our work also use single-task aggregators as building blocks. When there are multiple forecasting questions, the aggregation problem can also be viewed as learning a universal pattern between forecasters' predictions and the latent ground truth across forecasting questions. Therefore, statistical inference methods [LPI12; OVB14; LD14; MP17] are also customized and developed to aggregate forecasts in the multi-task setting. Our work includes both single-task aggregators and statistical-inference-based aggregators as benchmarks. We introduce more details about different aggregators in the benchmark selection part in Section 3.6.1.

Our proposed aggregators use the heterogeneity of participants' expertise to improve aggregation accuracy. There is a large literature, including [CW86; GMS14; Asp10; BC15; Sat+14b], which explores this idea but in the case where forecasters' historical performance is available, or where the forecasting is conducted in a dynamic manner where forecasting questions are resolved sequentially, and the resolution can be used to aggregate unresolved questions. In contrast, we consider the scenario where no ground truth information is available, i.e., aggregated predictions are requested before any forecasting

question is resolved. Wang et al. [Wan+11] consider the same scenario. However, they assume that there exists a known logical dependence between the outcomes of different forecasting questions.

Our idea of peer assessment scores, which aims to measure a forecaster’s prediction accuracy in the absence of ground truth information, is derived from multi-task peer prediction mechanisms [Pre04a; MRZ05a; WP12b; RFJ16; KLS16; Aga+17; Wit+17; GF19a; LWC20], a family of mechanisms used to determine forecasters’ rewards on multiple forecasting questions before any question resolves. For binary-vote judgement questions, Kurvers et al. [Kur+19] proposed a measure of similarity of forecasters’ votes, which is also empirically correlated with forecasters’ true accuracy. In this work, we investigate the use of five representative peer prediction methods to generate PAS.

3.3 Setting

We consider the scenario with a set \mathcal{N} of agents recruited to make forecasts on a set \mathcal{M} of events (forecasting questions).

Events. We consider binary events (sometimes called tasks).¹ Each event i is represented by a random variable $Y_i \in \{0, 1\}$, denoting the event outcome (ground truth). We assume that Y_i is drawn from a Bernoulli distribution $\text{Bern}(q_i)$ with an unknown $q_i \in [0, 1]$. To illustrate, consider an event i as “Will Democrats win the 2024’s election?” The outcome is either “Yes” ($Y_i = 1$) or “No” ($Y_i = 0$), and $q_i = 0.5$ means that the outcome is random (at the time of forecasting) and the Democrats has 50% chance to win.

Agents. Each agent (indexed by j) forecasts on a subset of events $\mathcal{M}_j \subseteq \mathcal{M}$. \mathcal{M}_j could either be assigned by the principal or be constructed by agent j herself. We use $\mathcal{N}_i \subseteq \mathcal{N}$ to denote the subset of agents who forecast on event i . We use $p_{i,j} \in [0, 1] \cup \{\emptyset\}$ to denote the probabilistic prediction made by agent j on event i for $Y_i = 1$, with $p_{i,j} = \emptyset$ denoting agent j provides no forecast on event i . Meanwhile, we let $\mathbf{p}_i = (p_{i,j})_{j \in \mathcal{N}_i}$ and $P = \{p_{i,j}\}_{i \in \mathcal{M}, j \in \mathcal{N}}$.

The forecast aggregation problem. The forecast aggregation problem is to design an aggregation function $F : ([0, 1] \cup \{\emptyset\})^{|\mathcal{M}| \times |\mathcal{N}|} \rightarrow [0, 1]^{|\mathcal{M}|}$, which maps the prediction profile P of all agents on all events to an aggregated prediction profile $\{\hat{q}_i\}_{i \in \mathcal{M}}$, where $\hat{q}_i \in [0, 1]$ is the aggregated prediction for event i . The design goal is to make the aggregated predictions as accurate as possible. The accuracy

¹Our methods and results can be extended to multi-outcome events in two ways. Please refer to Section 3.6.4.

of predictions is evaluated against the corresponding ground truth of the forecasted events, which are expected to be revealed some time after the aggregation.

Our aggregators will use two popular existing single-task aggregators as building blocks: the mean (Mean) and the logit-mean (Logit) [Sat+14a]. Mean has empirically proved robustness [JW08], while Logit extremizes the predictions of Mean and demonstrates significantly higher accuracy on some human forecast datasets [Sat+14a]. We introduce the weighted versions of the two aggregators that we will use as follows. For a single event i with a prediction profile \mathbf{p}_i and a weight vector $(w_j)_{j \in \mathcal{N}_i}$, we have

- $F_i^{\text{Mean}}(\mathbf{p}_i) = \sum_{j \in \mathcal{N}_i} w_j p_{i,j}$,
- $F_i^{\text{Logit}}(\mathbf{p}_i) = \text{sigmoid} \left(\frac{\alpha}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} w_j \text{logit}(p_{ij}) \right)$ and $\alpha = 2$ [Sat+14a].

The Logit aggregator first maps probabilistic predictions into the log-odds space using the logit function, the inverse function of the sigmoid function. It then takes the weighted average and applies a scaling factor to further extremize the prediction. Finally, it maps the prediction back into a probability using the sigmoid function. Empirically, Satopää et al. [Sat+14a] recommended a scaling factor of 2.

Prediction accuracy metrics The accuracy of forecasts is typically evaluated using the strictly proper scoring rules (SPSR) [GR07a]. Two widely-adopted rules are the Brier score and the log score. We use them to evaluate our aggregators' performance in our experiments. For a prediction \hat{q}_i and ground truth Y_i on an event i , we evaluate the two scores as follows:

- Brier score²: $S^{\text{Brier}}(\hat{q}_i, Y_i) = 2(\hat{q}_i - Y_i)^2$.
- Log score: $S^{\text{log}}(\hat{q}_i, Y_i) = -Y_i \log(\hat{q}_i) - (1 - Y_i) \log(1 - \hat{q}_i)$.

With above formulas, a lower scores refer to a higher accuracy. The Brier score ranges from 0 to 2. The log score ranges from 0.1 to 4.61.³ An uninformative prediction of 0.5 receives a Brier score of 0.5 and a log score of 0.69 regardless of the event outcome.

3.4 Aggregation Using PAS

We now formalize the notion of *peer assessment scores (PAS)*, and introduce our aggregation framework that uses PAS. We defer the introduction of concrete instantiations of PAS that lead to good aggregation performance into the next section. We list the abbreviations that we frequently use hereafter in Table 3.1.

²We adopt the same formula for the Brier score as in the Good Judgment Project [e.g., Ata+16]

³The log score is unbounded when the prediction is 0 or 1. We thus map predictions of 1 (0) to 0.99 (0.01).

Abbr.	Full name	Abbr.	Full name
DMI	Determinant mutual information mechanism	SPSR	Strictly proper scoring rules
CA	Correlated agreement mechanism	PAS	Peer assessment scores
PTS	Peer truth serum mechanism	BS	Brier score
SSR	Surrogate scoring rule mechanism	VI	Variational inference aggregator
PSR	Proxy scoring rule mechanism	MP	Minimal pivoting aggregator

Table 3.1: *The main abbreviations and the corresponding full names used in this chapter*

In short, and in different to the true accuracy that is evaluated against the ground truth, PAS assess a prediction against the other agents’ predictions. Thus, unlike the true accuracy, PAS can be computed for all crowdsourcing forecasting scenarios, with no additional information (e.g., the ground truth) required. Formally, a peer assessment score on an event set \mathcal{M} and an agent set \mathcal{N} is a scoring function $R : ([0, 1] \cup \{\emptyset\})^{|\mathcal{M}| \times |\mathcal{N}|} \rightarrow [0, 1]^{|\mathcal{N}|}$ that maps the prediction profile P of all agents on all events into a score s_j for each agent $j \in \mathcal{N}$. The score s_j should reflect the average prediction accuracy of agent j .

Bearing this notion of PAS in mind, we introduce our aggregation framework. The intuition of our framework is straightforward: In aggregation, if we rely more on predictions from agents with higher accuracy indicated by PAS, we shall hopefully derive more accurate aggregated predictions. In general, we can incorporate PAS into an aggregation process via three steps:

1. Compute a PAS score s_j for each agent $j \in \mathcal{N}$.
2. Choose a weight scheme that weight agents’ predictions based on the scores $s_j, j \in \mathcal{N}$.
3. Choose a base aggregator and apply the weight scheme to generate final predictions.

Each step features multiple design choices, which will influence the aggregation accuracy and can be customized case by case. In Step 1, there are multiple alternatives to compute PAS. Ideally, the computed PAS should reflect the true accuracy of agents. In Step 2, the weight scheme can be, for example, either ranking the agents by PAS and selecting a subset of top agents to aggregate (*ranking & selection*), or applying a softmax function to PAS to obtain weights. In Step 3, we can apply different base aggregators that can incorporate the weight scheme, such as weighted Mean or Logit.

We call the aggregators following the above framework the *PAS-aided aggregators*. We present the detailed PAS-aided aggregators that we will test in this chapter in Algorithm 8. In Step 1, we use five different peer prediction mechanisms (DMI, CA, PTS, SSR, and PSR) to compute PAS, which will be introduced in the next section. In Step 2, we choose the ranking & selection scheme rather than the softmax weight, as the former can be applied to any base aggregator and its hyper-parameter, the percent of top agents selected, has an straightforward physical interpretation. In our experiments, these

Algorithm 8 PAS-aided aggregators

- 1: Compute PAS (using one of DMI, CA, PTS, SSR, PSR) based on all predictions.
 - 2: Rank agents according to PAS.
 - 3: For each event i , select the predictions from top $\max(10\% \cdot |\mathcal{N}|, 10)$ agents who predict on that event, and run Mean or Logit aggregator on these predictions.
-

two weight schemes show similar performance with best-tuned hyper-parameters. In Step 3, we use Mean and Logit as the base aggregator.

3.5 Peer Prediction Methods for PAS

Peer prediction mechanisms are a family of emerging reward mechanisms designed to incentivize crowd workers to truthfully report their private signals (e.g., probabilistic predictions or votes on the outcome) in the absence of ground truth information. These mechanisms can be expressed by a function $R : ([0, 1] \cup \emptyset)^{|\mathcal{M} \times \mathcal{N}|} \rightarrow [0, 1]^{|\mathcal{M}|}$ that maps forecasters' prediction profile P to a reward R_j for each forecaster j . The function $R(\cdot)$ is carefully designed so that an agent's expected reward according to her belief about others' reports (formed by her private signal) will be maximized when she reports truthfully.

While most peer prediction scores do not necessarily reflect prediction accuracy, we selectively review five peer prediction mechanisms in this section and provide theoretical support for using them as PAS — scores of these five mechanisms each correlate with accuracy of agents according to some metric. The core intuition of these peer prediction mechanisms to achieve truthful elicitation is to quantify and reward the correlations among participants' predictions that are associated with the ground truth of the forecasting questions, instead of rewarding the simple similarity between participants' predictions. As a result, forecasters with predictions containing more information about the ground truth tend to receive a better score in expectation. This property makes them ideal candidates to serve as PAS.

Two *assumptions* are often required for these mechanisms to work:

- A1. Events are independent and a priori similar, i.e., the joint distribution of agents' private signals and the ground truth is the same across events.
- A2. For each event, agents' private signals are independent conditioned on the ground truth.

These two assumptions resemble the requirements for using statistical inference methods to infer the ground truth: there exists a consistent pattern between the ground truth and agents' predictions across tasks. The difference is that these two conditions do not restrict the pattern to follow some generative models specified by the inference methods. In the following paragraphs, we first introduce these five

peer prediction mechanisms and then show why their rewards may correlate with agents' true prediction accuracy. We divide the five mechanisms into two categories.

3.5.1 Mechanisms recovering the strictly proper scoring rules (SPSR)

When SPSR are reoriented such that a higher score corresponds to higher accuracy, they can serve reward schemes to incentivize truthful reporting [GR07a]. But they require the ground truth information to compute. Surrogate scoring rules (SSR) [LWC20] and proxy scoring rules (PSR) [Wit+17] are two peer prediction mechanisms that try to recover the SPSR from participants' reports, thus providing two methods to estimate the prediction accuracy of agents in the minimal information setting. Both mechanisms estimate a proxy of ground truth from participants' forecasts and assess their forecasts against this proxy. To introduce SSR and PSR, we use $S(\cdot)$ to denote an arbitrary SPSR.

Surrogate scoring rules (SSR) For a prediction $p_{i,j}$ from agent j , SSR randomly draws a binary signal Z from other agents' forecasts on the same task as the proxy to evaluate $p_{i,j}$, with $Z \sim \text{Bern}\left(\frac{\sum_{k \in \mathcal{N}_i \setminus \{j\}} p_{i,k}}{|\mathcal{N}_i| - 1}\right)$. The bias of Z to ground truth Y_i can be represented by two error rates $e_0 = \mathbb{P}(Z = 1 | Y_i = 0)$ and $e_1 = \mathbb{P}(Z = 0 | Y_i = 1)$. Assumptions A1 and A2 guarantee that the error rates of Z for agent j are the same across different tasks. Based on this property, Liu, Wang, and Chen [LWC20] provided an algorithm to accurately estimate e_0 and e_1 using participants' forecasts on multiple events. SSR then assess a prediction $p_{i,j}$ using a de-bias formula for $S(\cdot)$ to get an unbiased estimate for $S(\cdot)$ with Z . For prediction $p_{i,j}$, we have

$$R_{i,j}^{\text{SSR}}(p_{i,j}, Z) = \frac{(1 - e_{1-Z})S(p_{i,j}, z) - e_Z S(p_{i,j}, 1 - Z)}{(1 - e_0 - e_1)}.$$

Consequently, $\mathbb{E}_{Z|Y_i} [R_{i,j}^{\text{SSR}}(p_{i,j}, Z)] = S(p_{i,j}, Y_i)$.

Proxy scoring rules (PSR) In contrast to SSR, PSR directly apply SPSR $S(\cdot)$ to an agent's forecast against a proxy \hat{Y}_i of the ground truth to obtain the reward score, i.e., $R_{i,j}^{\text{PSR}}(p_{i,j}, \hat{Y}_i) = S(p_{i,j}, \hat{Y}_i)$. Witkowski et al. [Wit+17] showed that as long as the proxy \hat{Y}_i is unbiased to the ground truth, the proxy scoring rule gives an positive affine transformation of $S(\cdot)$, maintaining the incentive property. In practice, Witkowski et al. [Wit+17] recommended using an extremized mean prediction as the proxy when there is no explicit unbiased proxy of ground truth available.

3.5.2 Mechanisms rewarding the correlation

Determinant mutual information mechanism (DMI) [Kon20], correlated agreement (CA) [Shn+16], and peer truth serum (PTS) [RFJ16] are three mechanisms that reward agents by examining their forecasts' correlation to their peers'. Their core idea is to reward by a correlation metric that measures the agreement degree between agents' forecasts that are introduced through the ground truth, while excludes the agreement degree introduced by pure chance. In this way, an agent who independently manipulates her reports regardless the ground truth can only decrease her agreement with other agents. The computation of the expected reward under these three mechanisms for an agent j relies on the joint voting distribution between agent j and an uniformly randomly selected peer agent k . Given a prediction $p_{i,j}$, agent j 's vote on event i can be viewed as drawn from $\text{Bern}(p_{i,j})$. Thus, the joint voting probability of agent j voting u and agent k voting v for any $u, v \in \{0, 1\}$ can be computed empirically as

$$\hat{d}_{u,v}^{j,k} = \frac{1}{|\mathcal{M}_{j,k}|} \sum_{i \in \mathcal{M}_{j,k}} p_{i,j}^u (1 - p_{i,j})^{1-u} p_{i,k}^v (1 - p_{i,k})^{1-v},$$

where $\mathcal{M}_{j,k}$ is the subset of forecasting questions answered by both agents. We use $\hat{D}^{j,k} = \left(\hat{d}_{u,v}^{j,k} \right)_{u,v \in \{0,1\}}$ to denote the entire joint voting distribution of agent j and k . In the following paragraphs, we review how these three mechanisms reward agent j given the peer agent k .

Determinant mutual information mechanism (DMI) DMI measures the correlation using the determinant mutual information [Kon20]. Let $\mathcal{M}'_{j,k}, \mathcal{M}''_{j,k}$ be two disjoint subsets of $\mathcal{M}_{j,k}$, and let \hat{D}', \hat{D}'' be the joint voting distribution computed on these two subsets separately. DMI rewards agent j by an unbiased estimate to the squared determinant mutual information between agents j and k :

$$R_j^{\text{DMI}} = \eta \det(D') \cdot \det(D''), \quad (3.1)$$

where η is a normalization coefficient.

Correlated agreement (CA) CA rewards an agent j by

$$R_j^{\text{CA}} = \sum_{u \in \{0,1\}} \sum_{v \in \{0,1\}} |\hat{d}_{u,v}^{j,k} - \hat{d}_u^j \cdot \hat{d}_v^k|, \quad (3.2)$$

where $\hat{d}_u^j = \sum_{v \in \{0,1\}} \hat{d}_{u,v}^{j,k}$ is the marginal distribution of agent j reporting u estimated from the data. R_j^{CA} rewards the correlation by measuring the gap between the overall matching probability (represented by $\hat{d}_{u,v}^{j,k}$) and the matching probability caused by pure chance (represented by $\hat{d}_u^j \cdot \hat{d}_v^k$).

Peer Truth Serum (PTS) PTS rewards agent j by the matching probability of her votes to the peer agent k 's votes. PTS mitigates the effect of a match caused by pure chance via rewriting the matching probability under different vote realizations. Let $\bar{p}_{-j,u}$ be the average marginal probability of voting u of all agents except j . PTS rewards agent j by

$$R_j^{\text{PTS}} = \hat{d}_{0,0}^{j,k} / \bar{p}_{-j,0} + \hat{d}_{1,1}^{j,k} / \bar{p}_{-j,1}. \quad (3.3)$$

3.5.3 Peer prediction rewards and accuracy of agents

In this section, we formally show that the five peer prediction mechanisms reflect forecasters' true accuracy. First, SSR and PSR reflect the underlying accuracy of predictions due to the unbiasedness of their rewards w.r.t. the (affine transformation of) SPSR that they are built upon. As a direct corollary of their unbiasedness, we have the following.

- Proposition 3.1.**
1. Under Assumptions A1 and A2, SSR ranks the agents in the order of their mean SPSR that SSR is built upon asymptotically ($|\mathcal{M}|, |\mathcal{N}| \rightarrow \infty$).
 2. When there is an unbiased estimate of the ground truth and all agents are scored with the same unbiased estimate, PSR ranks the agents in the order of their mean SPSR that PSR is built upon asymptotically ($|\mathcal{M}| \rightarrow \infty$).

Second, the mechanisms, DMI, CA, PTS, reflect the accuracy of each agent because they essentially try to capture the *informativeness* of agents forecasts, i.e., the correlation between the agents' forecasts that is established through the ground truth instead of the pure chance. More specifically, we have the following proposition.

Proposition 3.2. Under Assumptions A1 and A2, and assuming agents report truthfully, the expected rewards of DMI, CA, PTS reflect a certain accuracy measure of agents. In particular,

1. DMI ranks the agents in the order of their reports' squared determinant mutual information [Kon20] w.r.t. the ground truth asymptotically ($|\mathcal{M}|, |\mathcal{N}| \rightarrow \infty$).
2. CA ranks the agents in the order of their reports' determinant mutual information w.r.t. the ground truth asymptotically ($|\mathcal{M}|, |\mathcal{N}| \rightarrow \infty$).
3. PTS ranks the agents in the inverse order of their signals' expected weighted 0-1 loss w.r.t. the ground truth outcome asymptotically ($|\mathcal{M}|, |\mathcal{N}| \rightarrow \infty$), when the binary answer drawn from the mean prediction of all agents has a true positive rate and a true negative rate both above 0.5.

Items	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
# of questions	94	111	122	94	72	80	86	50	50	50	80	80	90	90
# of agents	1409	948	1033	3086	484	551	87	51	32	33	39	25	20	20
Avg. # of ans. per ques.	851	534	369	1301	188	252	33	51	32	33	39	18	20	20
Avg. # of ans. per agent	56.74	62.46	43.55	39.63	28.03	36.5	32.8	49.88	49.96	50	79.97	60	90	89.5
Maj. vote correct ratio	0.90	0.92	0.95	0.96	0.88	0.86	0.92	0.58	0.76	0.74	0.61	0.68	0.62	0.72

Table 3.2: Statistics about the binary event datasets from GJP, HFC and MIT datasets

Item 1 in Proposition 3.2 follows straightforwardly from Theorem 6.4 in [Kon20]. We present the proofs for the items 2 and 3 in Appendix B.4. We note that mutual information does not directly imply accuracy in the binary case. For example, a random variable $Y'_i = 1 - Y_i$ contains all information w.r.t. the ground truth Y_i . But Y'_i is clearly not an accurate prediction of ground truth Y_i . However, when agents' forecasts $p_{i,j}$ are positively correlated to the ground truth Y_i , i.e., agents' predictions are better than random guess, then the mutual information does rank forecasts in the correct order, i.e., ranking the perfect prediction ($p_{i,j} = Y_i$) the highest and ranking random ones the lowest.

3.6 Empirical Studies

Our theoretical results suggest that the five peer prediction methods can effectively identify participants who predict more accurately than others under certain assumptions. In practice, however, it is often challenging or impossible to know to what extent these assumptions hold. Therefore, we conduct extensive experiments to study the performance of our PAS-aided aggregators. We use a diverse set of 14 real-world human forecast datasets and adopt two widely used accuracy metrics, the Brier score and the log score. We first introduce our experimental setup, then examine the effectiveness of PAS in selecting top performing forecasters, and finally present a comprehensive evaluation of our aggregators' performance. We first focus on binary events and then discussion our results on multi-outcome events in Section 3.6.4.

3.6.1 Experiment setup

Datasets

Our 14 test datasets consist of 4 datasets from the Good Judgement Projects (GJP) collected from 2011 to 2014 [GJP16], 3 datasets from the Hybrid Forecasting Competition (HFC) of varied populations [IAR19], and 7 MIT datasets [PSM17]. These datasets vary in several dimensions, including dataset size, sparsity, topics, collecting environment, and participants' performance. Together they offer a rich environment

Items	G1	G2	G3	G4	H1	H2	H3
# of questions	8	24	42	43	81	80	86
# of agents	1409	948	1033	3086	484	551	87
Avg. # of ans. per question	945.25	566.25	341.8	1104.58	136.30	202.99	26.03
Avg. # of ans. per agent	5.37	14.34	13.9	15.39	22.81	30.20	29.32
Maj. vote correct ratio	0.88	0.96	0.90	0.88	0.57	0.61	0.68

Table 3.3: Statistics about the multiple-outcome event datasets from GJP and HFC datasets

for evaluating the performance of aggregators.

The GJP and the HFC collected predictions about real-world issues involving geopolitics and economics via year-long online forecast contests. In these contests, forecasting questions were opened, closed, and resolved dynamically, and forecasters’ accuracy can be evaluated using previously resolved questions and used to aggregate predictions of remaining open questions. In contrast, the MIT datasets are static prediction datasets, where participants predict on a set of questions all at once. The topics include the capital of states, the price interval of arts, and the diagnosis of skin lesions. The MIT datasets also contain additionally solicited predictions that participants made about other participants’ predictions. This information enables one to apply the surprising-popularity-based aggregators.

We focus on the minimal-information aggregation setting. Therefore, we ignore the temporal information in the GJP and HFC datasets and only use each individual’s final forecast on each forecasting question.⁴ We also ignore the additional information solicited in MIT datasets when applying our aggregators, but use it for a surprising-popularity-based benchmark aggregator. We filter out participants with less than 15 predictions and questions with less than 10 answers from these datasets. This operation only removed a few forecasting questions in the HFC datasets with no sufficient predictions to make meaningful aggregation. We summarize the main statistics about the binary events of the 14 datasets after filtering in Table 3.2 and the multi-outcome events in Table 3.3. More details about datasets can be found in Appendix B.3.

Benchmark aggregators

In addition to the two base aggregators, Mean and Logit, which are widely-used in the minimal-information aggregation setting [Sat+14a; JW08], we also use two other types of aggregators as our benchmarks, the inference-based methods and the surprising-popularity-based methods.

- *Inference-based methods* contain a wide range of minimal-information multi-task aggregators. These methods establish parameterized models to characterize the latent features of forecasters such as

⁴We obtain similar qualitative results when the first forecasts or the average forecasts are used.

their biases towards the ground truth probability and the variances in their beliefs. Then, they infer these parameters as well as the ground truth using the forecasts across all events. In this type of aggregators, we use the *variational inference for crowdsourcing (VI)* method as a benchmark. It is a go-to approach to aggregate predictions in the machine learning community. We use the estimate ground truth probabilities given by VI as its predictions. Details of VI are included in Appendix B.5. Other sophisticated methods in this category include the cultural consensus model [OVB14], the cognitive hierarchy model [LD14], and the multi-task statistical surprising popularity method [MP17]⁵. We will also compare to the performance these aggregators reported by McCoy and Prelec [MP17] on the MIT datasets.

- *Surprising-popularity-based methods* are not minimal-information aggregators, but they represent a new trend of forecast aggregation [PSM17; PS19]. They require forecasters to additionally predict other forecasters’ predictions about the events of interest. Using this additional information, these methods can identify commonly shared information in participants’ forecasts and avoid counting them multiple times in the aggregation. The typical aggregator in this category refers to the surprisingly-popular algorithm [PSM17]. We use a more recent variant, called the *minimal pivot (MP)* method, as our benchmark. It has a better performance in generating probabilistic predictions. It has a simple form: the aggregated prediction equals two times the mean of the participants’ forecasts minus the mean of the participants’ predictions about other participants’ average prediction.

Median is another popular aggregator in the minimal information setting. In our test, its performance is always between the performance of Mean and Logit. Thus, we omit our results about median.

Implementation of PAS-aided aggregators

In our experiments, we evaluate 10 PAS-aided aggregators. Each PAS-aided aggregator uses one of the five peer prediction mechanisms (DMI, CA, PTS, SSR, PSR) to compute PAS and then incorporate the PAS into one of the two base aggregators (the Mean and Logit) using the rank&selection scheme. These PAS-aided aggregators have a single hyper-parameter—the number of top participants selected for each forecasting question. We set it to be the larger one of 10 and 10% percent of the total number of users. This hyper-parameter is shared among all PAS-aided aggregators on all datasets. Meanwhile, for SSR and PSR aggregators, we set the SPSR they are built upon as the metric SPSR. We use the output of the

⁵This aggregator combines both inference and surprising-popularity.

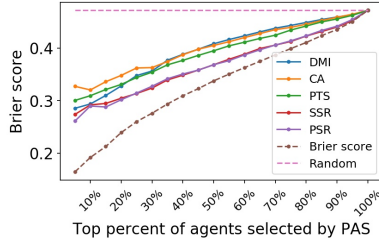


Figure 3.1: The averages of the true mean Brier score of top forecasters selected by the five PAS and by the true Brier score.

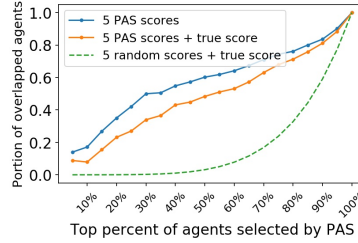


Figure 3.2: The portions of overlapped agents, who are simultaneously selected by all of the five PAS and the true score.

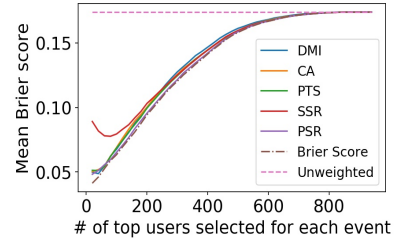


Figure 3.3: The Brier score of the five mean-based PAS-aided aggregators with a varying number of selected top agents on dataset G2.

VI aggregator as the proxy used in PSR.⁶ All these aggregators are described in Algorithm 8.

3.6.2 Smaller but smarter crowd

Before we dive into the comprehensive comparison between our PAS-aided aggregators and benchmarks, we first examine the effectiveness of PAS in identifying top forecasters and the influence of the number of top forecasters selected to the aggregation.

Fig. 3.1 shows the average prediction accuracy of the top forecasters selected by the five PAS (DMI, CA, PTS, SSR, PSR) over the 14 datasets. For all five PAS, the average of the true mean Brier scores of the selected top forecasters steadily increases (from around 0.3 to around 0.45) when we gradually enlarge the selection range from top 5% to all forecasters. This result indicates that all five PAS scores effectively rank the forecasters in the order of their true performance. We also notice that at each level of top forecasters selected, the mean accuracy of top forecasters selected by different PAS is very similar. We further examine the overlap of these top forecasters. The result (Fig. 3.2) suggests that the sets of top forecasters selected by different PAS scores have considerable overlap, and among these overlapped forecasters, the portion of the actual top forecasters is also remarkable. For example, as shown in Fig. 3.2, around 50% of forecasters are common among the top 30% forecasters under different PAS scores, and in these common forecasters, 60% forecasters are the actual top 30% forecasters (because at the level of top 30%, 30% forecasters are shared by all 5 PAS together with the true Brier score). This result further confirms that the five PAS can identify true top performers and that they have similar abilities in doing so.

Next, we examine how the number of top forecasters selected by PAS influences the aggregation accuracy. Overall, we observe that the accuracy of the PAS-aided aggregators peaks at a certain top

⁶We also tested using proxies (e.g, the mean of agents' predictions and the extremized mean [Wit+17]) in PSR, while using VI as the proxy gives us the best result.

percent (usually at top 5% to top 20%) and outperforms the accuracy of the base aggregator that they are built upon. We illustrate this observation with dataset G2 in Fig. 3.3, which also shows the accuracy of a Brier-score-(BS)-aided aggregator. The performance of this BS-aided aggregator shows the “in hindsight” performance we could achieve if the peer assessment is as accurate as if we knew the ground truth. In this particular dataset, the PAS-aided aggregators perfectly recover this “in hindsight” performance of the BS-aided aggregator (Fig. 3.3).

Overall, these results confirm prior findings which show that there often exists a smaller but smarter crowd whose mean prediction outperforms that of the entire crowd (e.g. “superforecasters” [Mel+15] and [GMS14]). Our contribution is to demonstrate that we can identify this set of smarter forecasters using only their prediction information.

3.6.3 Forecast aggregation performance on binary events

In this section, we present our main experimental results—the aggregation performance of our 10 PAS-aided aggregators against the benchmark aggregators on binary events of the 14 datasets. Our extensive evaluation highlights the following findings:

1. The performance of the four benchmark aggregators varies significantly across datasets, confirming the difficulty of forecast aggregation in the minimal-information setting.
2. The PAS-aided aggregators not only have higher overall accuracy than the benchmarks but also perform more stably and robustly across datasets.
3. While the performance of the 10 PAS-aided aggregators is not statistically different, the Mean-based PAS-aided aggregators tend to have higher accuracy and lower variance than the Logit-based PAS-aided aggregators.

Our main results are shown in Table 3.4 and Table 3.5. Table 3.4 shows the accuracy of the 10 PAS-aided aggregators and the benchmark aggregators on each dataset under the Brier score. As can be seen, 9 out of 10 PAS-aided aggregators outperform the best of the benchmarks on at least 5 datasets, and the remaining one outperforms the best benchmark on 4 datasets. Furthermore, each of the 5 PAS-aided Mean aggregators outperforms the second-best benchmark on at least 12 out of 14 datasets. Moreover, no PAS-aided aggregator underperforms the worst benchmark on any dataset, with only one exception of the PSR-aided Logit aggregator on dataset M1a. This is a significant improvement as we can see that though these benchmark aggregators are carefully designed for aggregating forecasts in the minimal information setting, none of them has stable performance across datasets.

Base aggr.	PAS	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
Mean	DMI	.125	.068	.071	.066	.219	.196	.110	.326	.126	.114	.434	.429	.535	.282
	CA	.127	.069	.073	.071	.200	.195	.126	.340	.126	.114	.454	.443	.536	.282
	PTS	.122	.069	.070	.066	.188	.192	.116	.359	.125	.114	.474	.443	.536	.282
	SSR	.137	.079	.072	.063	.164	.188	.122	.359	.116	.114	.474	.436	.522	.303
	PSR	.133	.065	.070	.059	.175	.187	.116	.459	.108	.107	.472	.451	.536	.278
Logit	DMI	.113	.053	.072	.037	.199	.194	.115	.517	.056	.058	.425	.545	.702	.325
	CA	.109	.053	.066	.036	.162	.191	.119	.547	.056	.058	.482	.569	.686	.325
	PTS	.109	.053	.071	.036	.172	.191	.120	.587	.066	.058	.508	.569	.686	.325
	SSR	.106	.053	.072	.039	.132	.187	.118	.587	.046	.058	.518	.556	.701	.422
	PSR	.106	.054	.071	.039	.182	.195	.117	.715	.037	.028	.535	.579	.686	.376
Mean (benchmark)		.206	.174	.114	.151	.212	.184	.143	.452	.347	.347	.480	.441	.473	.333
Logit (benchmark)		.116	.080	.066	.065	.136	.174	.122	.681	.433	.357	.500	.562	.663	.485
VI (benchmark)		.213	.072	.082	.085	.306	.325	.163	.595	.037	.000	.841	.610	.733	.345
MP (benchmark)		N/A	N/A	N/A	N/A	N/A	N/A	N/A	.425	.251	.232	.479	.471	.609	.491

Table 3.4: The mean Brier scores (range [0, 2], the lower the better) of different aggregators on binary events of 14 datasets. The best mean Brier score among benchmarks on each dataset is marked by bold font. The mean Brier scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in green; those outperforming the second best of benchmarks are highlighted in yellow; the worst mean Brier scores over all aggregators on each dataset are highlighted in red.

Table 3.5 provides the number of datasets on which one aggregator statistically outperforms the other for each pair of PAS-aided aggregators and benchmarks. Each of the 10 PAS-aided aggregators, especially the Mean-based PAS-aided aggregators, statistically outperforms each benchmark on at least 4 more datasets than it underperforms, with a maximum of 9 more datasets. Similar results are observed under the log scoring rule (Table B.4, Appendix B.2 and Table 3.5). Next, we give a more detailed review of the experimental results.

Performance of the benchmarks. The Logit aggregator performs better than the other benchmarks on the GJP and HFC datasets, but performs worse on the MIT datasets, while the Mean aggregator performs in the other directions. This is likely because that the questions in MIT datasets are more challenging than those in the GJP and HFC datasets (e.g., see the correctness ratio of majority vote shown in Table 2.1), and the Logit aggregator, which extremizes the mean prediction, further worsens the situation. VI predicts almost flawlessly on datasets M1b, M1c, but is outperformed by uninformative guess (predicting 0.5) on M2, M3, and M4a. This is likely because the accuracy of VI heavily depends on the extent to which the data follows the assumed generative model that VI uses to infer the ground truth. MP has a relatively stable performance on the MIT datasets, but on some of these datasets, it is outperformed by VI and Mean.

PAS-aided aggregators vs. Mean and Logit. As can be seen in Table 3.5, the PAS-aided aggregators outperform the Mean and the Logit aggregators with statistical significance on most datasets. Dataset H2

Base aggr.	PAS	Brier Score				Log Score			
		Mean	Logit	VI	MP	Mean	Logit	VI	MP
Mean	DMI	10, 1	7, 1	5, 2	5, 0	10, 1	7, 2	8, 2	6, 0
	CA	8, 1	6, 1	5, 2	4, 0	8, 1	6, 2	8, 2	5, 0
	PTS	9, 1	6, 1	5, 2	4, 0	9, 1	6, 2	9, 2	5, 0
	SSR	8, 1	6, 0	6, 2	5, 0	8, 1	6, 3	7, 2	4, 0
	PSR	8, 1	6, 1	5, 2	3, 0	8, 1	6, 2	9, 2	4, 0
Logit	DMI	6, 2	6, 1	2, 0	3, 1	6, 2	4, 1	6, 0	3, 1
	CA	6, 2	4, 0	3, 0	3, 1	7, 3	5, 0	5, 0	3, 2
	PTS	6, 2	4, 0	3, 0	3, 2	6, 3	3, 0	5, 0	3, 2
	SSR	7, 2	4, 0	3, 0	2, 2	7, 4	2, 0	5, 1	2, 3
	PSR	6, 3	4, 1	4, 0	3, 2	6, 4	4, 1	5, 1	3, 3

Table 3.5: The two-sided paired *t*-test for the mean Brier scores and the mean log scores of each pair of a PAS-aided aggregator and a benchmark on binary events of 14 datasets. The first integer in each cell represents the number of datasets where the PAS-aided aggregator achieves significantly smaller mean score (with p -value <0.05), while the second integer in each cell indicates the number of datasets where the benchmark achieves significantly smaller mean score. The cells where the # of outperforms exceeds the # of underperforms by at least 4 are highlighted in green.

is the only exception where Mean and Logit are not outperformed by any PAS-aided aggregator under the Brier score. However, a closer look shows that the accuracy difference of these two aggregators in H2 is minimal (within 0.02). This advantage of the PAS-aided aggregators over the Mean and the Logit aggregators is because of the use of cross-task information when computing the PAS, i.e., the top forecasters are truly identified by these PAS using agents’ forecasts on multiple tasks. These empirical results suggest that one can safely replace the Mean and Logit with the PAS-aided aggregators and expect an accuracy improvement in most cases (if a sufficient number⁷ of predictions are collected from each forecaster to compute the PAS).

PAS-aided aggregators vs. VI and other inference-based methods We notice that although VI ranks the worst in many datasets, the number of datasets on which VI statistically underperforms each PAS-aided aggregator is smaller than those numbers of the other benchmarks (Table 3.5). This is because VI tends to output extreme predictions (close to 0 or 1) and thus receives extreme accuracy scores (e.g., close to 0 or 2 under the Brier score), requiring more events to draw statistically significant conclusions. Also, as we have mentioned, the performance of VI varies significantly across different datasets (Table 3.4). If one is uncertain about whether the data follows the generative model assumed by VI, the PAS-aided aggregators (especially the SSR-/PSR-aided aggregators) are better choices. They

⁷We will discuss this number in the next section.

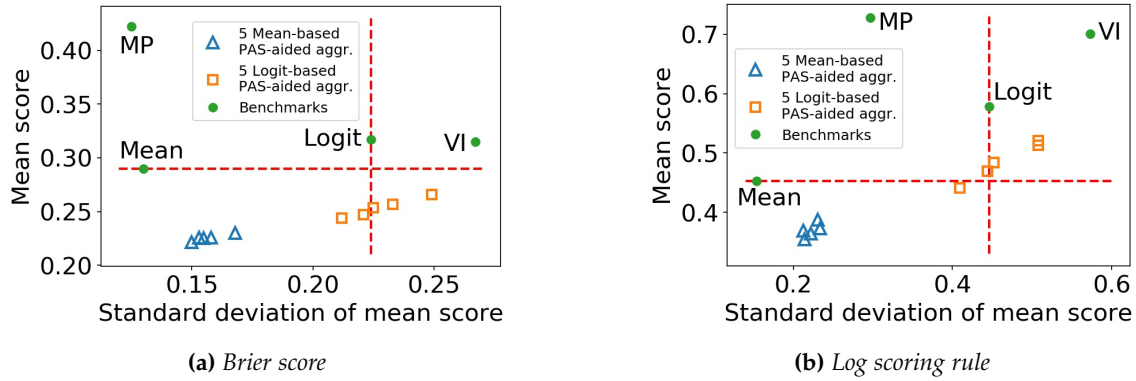


Figure 3.4: The mean and the standard deviation of the aggregation accuracy of the 10 PAS-aided aggregators ($DMI/CA/PTS/SSR/PSR$ -aided \times Mean/Logit-based aggregators) and the benchmarks over 14 datasets.

perform much closer to VI than the other benchmark aggregators on datasets where VI makes almost perfect predictions (datasets M1b, M1c), and perform more stably on datasets where VI makes extremely wrong predictions (datasets M2, M3, M4a).

McCoy and Prelec [MP17] reported the mean Brier score (with range [0,1]) of three other inference-based aggregators (the cultural consensus model, the cognitive hierarchy model and the multi-task statistical surprising popularity method) on MIT datasets (Table B.3, Appendix B.2). Based on their reports, only the multi-task statistical surprising popularity method outperforms our PAS-aided aggregators on one more datasets than what VI does. However, this method requires forecasters to provide additional predictions beyond the predictions of the events of interest just as other surprising-popularity-based aggregators.

PAS-aided aggregators vs. MP MP generally performs better than other benchmarks on the 7 MIT datasets, as it uses the additionally solicited information available these datasets. However, Table 3.5 still shows a salient advantage of PAS-aided Mean aggregators over MP. This result implies that when forecasters make predictions on multiple events, the cross-task information leveraged by the PAS scores may be more powerful in facilitating aggregation than the additionally solicited information used in MP.

Finally, we find no significant difference in the performance of PAS-aided aggregators that use different PAS. In particular, under the Brier score, no PAS-aided aggregator statistically outperforms another on more than three datasets if the same base aggregator is used. This is likely because different PAS have similar abilities in identifying the top forecasters as we have shown in Fig. 3.2.

Average performance across datasets

We present the mean and the standard deviation of the accuracy of our 10 PAS-aided aggregators and benchmarks over the 14 datasets in Fig. 3.4 (Concrete data can be found in Table B.2, Appendix B.2). As can be seen, all PAS-aided aggregators have better mean accuracy under the Brier score than all benchmarks. In particular, the five Mean-based PAS-aided aggregators outperform all benchmarks with statistical significance ($p < 0.05$) under both the Brier score and the log scoring rule.⁸ Moreover, the five Mean-based aggregators also show much smaller variances than the Logit and VI aggregators under both accuracy metrics, suggesting that the Mean-based PAS-aided aggregators are more stable than these two benchmarks. Within PAS-aided aggregators, the Mean-based ones appear to be more accurate and stable than the Logit-based ones, while the differences are not statistically significant. We conjecture that as the PAS already select out the forecasters with more accurate predictions, the extremization provided by the Logit base aggregator no longer benefits for any accuracy improvement, but only increases the aggregation variance.

These findings suggest that one can expect better accuracy and smaller performance variance when using PAS-aided aggregators instead of the benchmark aggregators. Moreover, the Mean-based PAS-aided aggregators, especially the Mean-based DMI-aided aggregator, are likely to produce the best aggregation outcomes. We also evaluated PAS-aided aggregators on smaller datasets that were sampled from the 14 original datasets. These datasets have 20 events and 30 or 50 participants. We observe similar improvements of the PAS-aided aggregators over the benchmarks. This result suggests that the PAS-aided aggregators may also mitigate the cold-start problem in long-term forecast aggregation settings, where only a small set of forecasts is available with no ground truth yet revealed. We present the details of this experiment in Appendix B.1.

3.6.4 Forecast aggregation performance on multi-outcome events

Our 10 PAS-aided aggregators can be extended to aggregate forecasts on multi-outcome events, because the 5 PAS scores and the two base aggregators can be extended to multi-outcome events [Sat+14a; RFJ16; Shn+16; Wit+17; LWC20; Kon20]. However, the performance of these multi-outcome-event extensions may not be as good as their binary counterparts for two reasons. First, in the multi-outcome event settings, there are more latent variables to be estimated in the PAS scores, while the number of the samples (the multi-outcome events to forecast and the predictions collected) are usually smaller than

⁸The only exceptions are the PSR-aided aggregator under the Brier score, and the SSR-/PSR-aided aggregators under the log score when compared to the MP aggregator, as the MP aggregator only applies to 7 MIT datasets.

Base aggr.	PAS	Brier Score						Log Score					
		G2	G3	G4	H1	H2	H3	G2	G3	G4	H1	H2	H3
Mean	DMI	.099	.136	.115	.522	.527	.402	.219	.287	.264	.975	.986	.779
	CA	.103	.165	.123	.516	.526	.400	.229	.343	.283	.956	.985	.770
	PTS	.099	.139	.114	.509	.528	.403	.218	.291	.260	.947	.988	.771
	SSR	.136	.145	.109	.516	.524	.419	.320	.296	.254	.956	.966	.785
	PSR	.097	.126	.101	.521	.530	.406	.208	.255	.227	.969	.980	.763
Logit	DMI	.067	.131	.067	.488	.506	.442	.129	.233	.138	.909	.960	.878
	CA	.069	.136	.067	.484	.509	.439	.131	.249	.141	.887	.967	.866
	PTS	.065	.129	.065	.478	.512	.444	.127	.233	.135	.879	.974	.879
	SSR	.083	.127	.067	.493	.507	.461	.188	.225	.149	.894	.939	.898
	PSR	.069	.125	.061	.496	.518	.448	.129	.220	.130	.913	.962	.865
Mean (benchmark)		.243	.232	.239	.534	.526	.445	.509	.484	.490	.992	.981	.839
Logit (benchmark)		.147	.149	.161	.500	.505	.462	.298	.295	.309	.921	.947	.893
VI (benchmark)		.083	.190	.186	.864	.780	.633	.202	.448	.438	1.996	1.803	1.417

Table 3.6: The mean Brier score and the mean log score of different aggregators on multi-outcome events of 6 datasets. The best mean score among benchmarks on each dataset is marked by bold font. The mean scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in **green**; those outperforming the second best of benchmarks are highlighted in **yellow**; the worst mean scores over all aggregators on each dataset are highlighted in **red**.

those of binary events (Table 2.1 vs. Table 3.3). Second, the assumptions under which the PAS scores theoretically reflect the true accuracy of forecasters are more difficult to meet for multi-outcome events. Therefore, if we use these extended methods directly, the estimates of forecasters' performance may be noisy, leading to more noisy aggregated predictions.

A more practical alternative is to apply the PAS of forecasters estimated on binary events into the aggregation of multi-outcome events. In the GJP and HFC projects, agents face both binary events and multi-outcome events. Therefore, we can apply this approach on both GJP and HFC datasets. We present the statistics of multi-outcome forecasting questions in the GJP and HFC datasets in Table 3.3 and present the aggregation results and comparisons in Table 3.6 and 3.7. The results show a consistent and significant advantage of using the PAS-aided aggregators. The success in this approach also suggest that agents have consistent relative accuracy in making predictions on both binary events and multi-outcome events. In particular, on no dataset a benchmark outperforms a PAS-aided aggregation with statistical significance (the only exception is Logit v.s. CA-aided Mean on dataset H2).

3.7 Discussion and Future Directions

This chapter demonstrates that the PAS-aided aggregators generally have higher aggregation accuracy across datasets than the four benchmark aggregators. Among the benchmarks, the Mean, Logit, and

Base aggr.	PAS	Brier Score			Log Score		
		Mean	Logit	VI	Mean	Logit	VI
Mean	DMI	5, 0	1, 0	3, 0	3, 0	1, 0	3, 0
	CA	5, 0	1, 0	3, 0	4, 0	1, 1	3, 0
	PTS	5, 0	1, 0	3, 0	4, 0	1, 0	3, 0
	SSR	4, 0	2, 0	3, 0	5, 0	1, 0	3, 0
	PSR	5, 0	3, 0	3, 0	4, 0	3, 0	3, 0
Logit	DMI	4, 0	1, 0	3, 0	4, 0	2, 0	3, 0
	CA	4, 0	1, 0	3, 0	4, 0	3, 0	3, 0
	PTS	4, 0	2, 0	3, 0	4, 0	3, 0	3, 0
	SSR	3, 0	1, 0	3, 0	4, 0	3, 0	3, 0
	PSR	3, 0	1, 0	3, 0	4, 0	3, 0	3, 0

Table 3.7: The two-sided paired t -test for mean Brier score and mean log score of each pair of a PAS-aided aggregator and a benchmark on multi-outcome events of 6 datasets. The first integer in each cell represents the number of datasets where the PAS-aided aggregator achieves the significantly smaller mean score (with p -value <0.05), while the second integer in each cell indicates the number of datasets where the benchmark achieves the significantly smaller mean score.

MP aggregators are single-task aggregators that generate the final prediction of an event using only the forecasts on that event. However, they were the top-performing aggregators in several real-world, multi-task forecasting competitions such as in the Good Judgement project [JW08; Sat+14a]. The VI aggregator is a multi-task statistical-inference-based aggregator, which uses an inference method to infer the ground truth probability based on cross-task information. Our PAS-aided aggregators can also be viewed as a multi-task statistical-inference-based aggregator. The peer prediction methods used in the PAS-aided aggregators are inference-like methods that estimate forecasters’ underlying expertise using all forecasts collected.

Using cross-task information in aggregation gives the PAS-aided aggregators advantages over the single-task benchmark aggregator. We can see that on datasets M1b and M1c, the three single-task benchmarks perform moderately well (with a mean Brier score around 0.3), while the other benchmark aggregator using cross-task information, the VI aggregator, has almost perfect predictions (with a mean Brier score close to 0). Our PAS-aided aggregators has similarly great performance on these two datasets as the VI aggregator. On the other hand, the PAS-aided aggregators appear to have more robust performance than the statistical-inference-based VI aggregator. For example, on datasets M2, M3, and M4a, where VI has much worse performance than random guesses, the PAS-aided aggregators still have moderate performance. Intuitively, statistical inference methods are sensitive to underlying properties of the data, i.e., the extent to which the assumed probabilistic model reflects the true pattern of the data. Unlike typical statistical-inference-based aggregators, the PAS-aided aggregators do not directly infer the outcomes of the forecasting questions. Instead, they infer forecasters’ expertise from cross-task

predictions and then use the expertise information to adjust the base aggregator. This operation likely makes the PAS-aided aggregators more robust to the variation of the data.

Although the PAS-aided aggregators demonstrated significant accuracy improvement on datasets where individuals' overall performance is either good or poor and the number of forecasts collected per question is either high or low (GJP datasets and MIT datasets), we find their accuracy improvement is minimal on the HFC datasets, where the number of forecasts each forecaster made (< 40) is relatively small. This observation is consistent with the theoretical requirements for PAS scores to accurately estimate forecasters' true performance: Each forecaster has consistent accuracy across events, and each forecaster has made a sufficient number of predictions. Therefore, if an insufficient number of predictions has been made by each forecaster, the PAS scores may not reflect forecasters' factual accuracy well.

In addition, the five PAS scores that we tested in theory all rely on the assumption that the predictions of different forecasters are independent conditioned on the underlying event outcome to reflect the forecasters' true accuracy. Although the PAS-aided aggregators perform well on our 14 datasets, where the assumption is likely not hold strictly, one should still be careful about using the PAS-aided aggregators in scenarios where this assumption is saliently violated, for example, when forecasters are encouraged to discuss with each other before making predictions and when forecasters are machine predictors trained using similar data and methods.

In this chapter, we take the first step to understand the possibility of using peer prediction methods to robustly improve the collective intelligence in prediction tasks. Our approach has the advantage of only requiring a minimal amount of information to be collected and placing almost no restriction on crowdsourcing workflow. Thus, our methods have the potential of becoming a component of more interactive human-machine forecasting systems, where other techniques of boosting collective intelligence, such as teaming [CFM19], workflow design [LMW12], promoting interactions [BBA15] and AI algorithms [WLB15], are also present. From another perspective, the human-machine computation systems are now also developed for many complex tasks, such as image segmentation [Son+18] and article editing [ZVK17]. An important problem is that how we boost collective intelligence for solving these complex tasks. Our approach provides a way to potentially reduce this problem to how we can devise effective correlation metrics to capture the information quality of these responses. All above are interesting future research directions.

Chapter 4

Cursed yet Satisfied Agents

4.1 Introduction

“You win, you lose money, and you curse.”

— Kagel and Levin

Consider the following hypothetical game—Alice and Bob each have a wallet, with an amount of money that is known to them, but not to the other player. They each know that the money in the other wallet is distributed uniformly in the range between \$0 and \$100, independently of the amount of money in their own wallet. The auctioneer confiscates the two wallets, and runs a second price auction on the two wallets (the highest bidder wins the two wallets, and pays the bid of the other bidder). Say Alice has \$30 in her wallet, how should she bid? A naïve strategy Alice could take is to calculate the expected amount of money in Bob’s wallet, \$50, and add it to her amount, resulting in a bid of \$80. However, such a bidding strategy ignores the fact that if Bob invokes the same strategy, then conditioned on Alice winning the two wallets, Alice has more money in her wallet than Bob, implying that Bob’s amount is distributed uniformly between \$0 and \$30. Thus, if both agents invoke the naïve strategy, Alice’s utility conditioned on winning is the expected sum in the two wallets, $\$30 + \$15 = \$45$ minus Bob’s expected bid conditioned on him losing $\$15 + \$50 = \$65$, implying a *negative* expected utility of $-\$20$.

Of course, a rational agent should not incur a negative utility when playing a game. Klemperer [Kle98] introduced the game presented above, and named it “the wallet game”. This is an interdependent value setting (IDV) [MW82], where each agent has *private* information, termed *signal* (i.e., the amount of money in their own wallet), and a *public* valuation function that takes into account different bidders’ private information (in this case, the sum of signals). Klemperer [Kle98] analyzed the symmetric

equilibrium of rational agents in the wallet game (and introduced several asymmetric equilibria). However, in practice, the observed behavior of agents in the wallet game much resembles the naïve strategy rather than an equilibrium that rational agents end up with [AK97]. This phenomenon was first observed by Capen, Clapp, Campbell, et al. [CCC+71], three petroleum engineers who observed that oil companies experienced unexpectedly low rates of returns in oil-lease auctions since these companies “ignored the informational consequences of winning”.

This behavioral bias, where the winner fails to account for the implications of outbidding other agents is commonly referred to as *the winner’s curse*, and was consistently observed across many scenarios such as selling mineral rights [CCC+71; LD+83], book publication rights [Des+82], baseball’s free agency market [CD80; BC96], and many others (for more information about empirical evidence for the winner’s curse, see Chapter 1 in [KL09]). As standard game theory cannot account for the observed behavior, Eyster and Rabin [ER05] introduced a behavioral model that formalizes this discrepancy. They termed this model as “cursed equilibrium”.

In their model, agents correctly predict other agents’ strategies, but fail to estimate that other agents’ actions are correlated with their actual signals (or information), similar to the naïve strategy presented above. The extent of this degree of misestimation of the correlation of actions and signals is captured by a parameter χ , where the perceived utility of agents is χ times the expected utility of the agents if the actions and signals of other agents are uncorrelated *plus* $(1 - \chi)$ times the actual, correct expected utility, correlating the signals with the actions (see Section 4.3.1 for a formal definition). Having a single parameter to explain the behavioral model of the agents proved to be a very tractable modeling, as this parameter can be easily fitted using real world data to estimate χ , and better predict players behaviors [ER05].

Existing literature of interdependent values either

- (i) analyzes fully rational players’ behavior as they ‘shade’ their bid to account for the potential over-estimation of the value [Kle98; MW82; Wil69; Pri20];
- (ii) studies biased agents’ behavior when deploying known mechanisms, where the equilibrium often times implies a negative utility for the bidders (and higher revenue for the seller) [ER05; KL86; AK97; HS94]; or
- (iii) exploits the biased behavior of the agents to achieve higher revenue [BBM20].

Our work builds upon the Eyster and Rabin [ER05] “cursed equilibrium” model and views cases where agents experience *actual* negative utility as undesirable; thus, tries to avoid such scenarios.

One might wonder why a seller might care about a cost incurred by the buyer due to their own bias, especially when the outcome might increase the seller’s revenue. However, we note that this can be highly undesirable due to various reasons:

- In many real-life scenarios, such as leasing spectrum bands, violating ex-post IR can be detrimental to society at large. Perhaps a mobile company overbids on spectrum, winds up bankrupt, and then the public cannot enjoy any cellular service associated with that spectrum lease [Zhe01].
- Companies experiencing revenue loss might feel reluctant to join future auctions [HP88]. This may have the adverse affect of reducing the long-term revenue of the seller, and the long-term social welfare in the market.

In this chapter, we design mechanisms that are incentive compatible (IC) for *cursed* agents—agents maximize their *biased* utility by reporting their true private information, thus generate a predictable behavior. In order to avoid the winner’s curse, the mechanisms we introduce are *ex-post individually rational*, meaning an agent will never pay more than their *true* value. We study the quintessential objectives of revenue and welfare maximization in auction settings with interdependent values.

Our results. We focus on deterministic and anonymous mechanisms, as they are optimal for interdependent values of fully rational agents [Aus99; MS92; RT16]. We extend the cursed-equilibrium model of Eyster and Rabin [ER05] to support a strong truthfulness notion of Cursed Ex-Post Incentive-Compatible (C-EPIC), the equivalent of ex-post IC in the case of fully rational agents, which is the strongest incentive notion possible for this setting¹ (see Section 4.3.1). For interdependent values, a deterministic C-EPIC mechanism corresponds to a threshold allocation rule, which takes as input other bidders’ reports, and returns the minimum bid from which an agent starts winning.

After establishing the incentive notion we are studying in this chapter, we turn to take a closer look at the implications of ensuring the mechanisms satisfy ex-post individual rationality (Section 4.5). Our solution concept gives rise to an analogue of the payment identity of EPIC mechanisms (Proposition 4.4). As opposed to the fully rational setting, we might need to set the constant term in the payment identity ($p_i(0, \mathbf{s}_{-i})$ in Equation (4.6)) to be smaller than zero, as the mechanism might make positive transfers to compensate for the over-estimation of values due to the winner’s curse. For a fixed deterministic mechanism, we show how to optimally set this compensation term in a way that maximizes the revenue for the allocation rule, while keeping EPIR. This has the following implication—fixing the

¹The ex-post IC notion is stronger than Bayesian IC and weaker than Dominant strategy IC. In interdependent settings it is impossible to design dominant strategy IC mechanisms while obtaining good performance guarantees.

allocation rule, and letting χ grow decreases the revenue. An interesting conclusion is the following (see Propositions 4.13 and 4.14):

Proposition (Revenue and welfare monotonicity). As the cursedness-parameter χ increases: (i) the revenue of the revenue-optimal EPIC-IR and EPIR mechanism decreases. (ii) the welfare of the welfare-optimal EPIC-IR, EPIR and ex-post budget-balanced mechanism decreases.

This is in stark contrast to the case where the mechanism does not require EPIR, where cursed-agents are shown to generate more revenue in second price auctions, since the mechanisms take advantage of the possibility of agents paying more than their value [ER05].

Building upon our understanding of combining individual-rationality constraints with incentive-compatibility for agents who suffer from the winner’s curse, we turn to revenue maximization (see Section 4.6). We show that designing a revenue optimal C-EPIC-IR and EPIR deterministic mechanism decomposes into a separate problem for each agent i and other agents’ signals \mathbf{s}_{-i} . That is, the mechanism designer’s task reduces to find an optimal threshold for winning the auction for agent i for every set of signals \mathbf{s}_{-i} other agents might declare. We show how to optimally set such a threshold, resulting in a revenue-optimal mechanism (see Theorem 4.16).

Theorem (Revenue optimal mechanism). The revenue-optimal mechanism is a threshold mechanism, and the threshold rule is given via Theorem 4.16.

We discuss interesting similarities and differences of the optimal threshold rule from the case where the agents are fully rational, and the case where the seller does not require the mechanism to keep EPIR constraints at the end of Section 4.6.

Social welfare maximization is a more nuanced task, as just aiming for C-EPIC-IR and EPIR might result in the auctioneer needing to pay all the agents (even the losers). In fact, we show a scenario that the welfare-optimal C-EPIC-IR and EPIR mechanism incurs a huge revenue loss of $\Omega(n \log n)$, where the expected optimal welfare is $O(n)$. This holds even for a very simple setting where i ’s valuation is $v_i(\mathbf{s}) = s_i + \frac{1}{2} \sum_{j \neq i} s_j$ and signals are sampled independently from the $U[0,1]$ distribution (see Proposition 4.17). Therefore, when maximizing welfare, we add a requirement of ex-post budget-balance (EPBB); that is, the seller never has negative revenue.

A trivial way to ensure EPBB is to set the threshold rule such that it never sells whenever selling implies the seller will need to make positive transfers. We present a masking operation that does exactly this. Given a threshold rule, the masking of the threshold rule allocates in the maximal subset of cases where the original threshold rule made an allocation *and* the allocating induces no positive transfer from

the seller. One might wonder whether one can design a mechanism that is EPBB, while still selling at scenarios that require the mechanism to make positive transfers to some bidders, increasing the expected welfare of the mechanism. We answer this question in the negative (see Theorem 4.22).

Theorem (EPBB \equiv no positive transfers). Under natural conditions, a mechanism is EPBB if and only if the mechanism *never* makes positive transfers.

This characterization drives the following implication (Proposition 4.23).

Proposition (Welfare optimal mechanism). Under natural conditions, the welfare-optimal EPBB mechanism is a result of masking the welfare-optimal allocation.

Applying the above proposition to the wallet game example in the introduction implies that in that example, the seller should allocate the item only when the loser has at least \$50 in their wallet, since this is the case where the winner does not overestimate the amount of money in the two wallets.

We notice that for some valuation functions, as the maximum of agents' signals, this implies an EPBB mechanism can never sell the item, resulting in zero welfare. On the positive side, we introduce a new family of valuations, *Concave-Sum valuations*, which generalizes well studied classes of valuations such as weighted-sum valuations² (e.g., the wallet game) and ℓ_p -norms of signals for a finite p . We show that for this class of valuations, the optimal EPBB mechanism approximates the optimal welfare (Theorem 4.25).

Theorem (Welfare approximation). For concave-sum valuations, the welfare-optimal EPBB mechanism gets at least half of the fully efficient allocation as the number of agents grows large.

4.2 Related work

Our work investigates the problem of auction design for the agents who suffer from the winner's curse. As a behavioral anomaly [Tha92], the winner's curse has been documented and analyzed by a large literature via both field studies and lab experiments. Field evidence of the presence of the winner's curse has been discovered in auction practices across a wide array of industries, which range from the book industry [MM87] and the market of baseball players [CD80] to the offshore oil-drilling leases [CCC+71; HP88; HPB87; Por95]. Also, a large amount of lab experimental results [BS83; KL86; AK97; CL09;

²Weighted-sum valuations take the form $v_i(\mathbf{s}) = s_i + \beta \sum_{j \neq i} s_j$ for $\beta \in (0, 1]$. Note that the valuations in the wallet game example are a special case of weighted-sum valuations.

[ILN10; BDL19] show that the winner's curse occurs under various conditions, which differ in multiple dimensions, such as auction format (e.g., first-price, second-price, Dutch auction and English auction), number of participants, valuation functions and signal structures. Kagel and Levin [KL02] provided a comprehensive review for such experimental studies. Furthermore, most of these studies, such as observational ones [HP88; HPB87] and lab experimental ones [BS83; KL86; AK97; CL09], demonstrate that the bidders who suffered from the winner's curse not only experienced a reduce in the profit than anticipated, but could also be worse off upon winning, i.e., receive a negative net profit.

As discussed in the introduction, we adopt the cursed equilibrium model introduced by Eyster and Rabin [ER05] to model the bidding behaviors of agents who suffer the winner's curse. This model suggests that agents fail to fully appreciate the contingency between other bidders' bids and the auction item's value. This cause is supported by several experimental findings [BS83; KL86; AK97; CL09], and this model has been applied, generalized or analogized to different applications, including analyzing market equilibrium [ERV19; EW11], designing financial assets [KK15; EP17], unifying theoretical behavioral models [Mie09].

In addition, there is a large literature, including [Wil69; MW82; Kle98; BK02; RT16], studying the interdependent valuation auction, the type of auctions we consider in this chapter. Different from our work which designs mechanism for agents who suffer the winner's curse, these papers consider the mechanism design for fully rational agents who play (Bayesian) Nash equilibrium strategies. Milgrom and Weber [MW82] introduced the interdependent value model and analyzed the revenue of different auction formats when agents have correlated but private value over the item. Their results imply that fully rational agents who implicitly try to avoid the winner's curse bid more conservatively when there is less information (as in a second price auction) revealed in the auction than more information revealed (as in English auction). Bulow and Klemperer [BK02] and Klemperer [Kle98] showed the anomalies in certain interdependent valuation auctions that the item price may increase in supply and decrease in the number of bidders and that the item price is sensitive to even a small asymmetry of bidders. They interpreted the anomalies in terms of fully rational bidders taking the winner's curse into consideration. Roughgarden and Talgam-Cohen [RT16] developed tools to build ex-post incentive compatible mechanisms for general interdependent valuation auctions with fully rational agents. We use their tools to build our mechanisms.

In contrast to these works, Bergemann, Brooks, and Morris [BBM20] studied the auction design problem for agents suffering the winner's curse. Their work is the most similar to ours, but with several differences. The major difference is that they aim to achieve interim incentive guarantees, while we aim to achieve stronger ex-post incentive guarantees. As a result of sacrificing the ex-post incentive

guarantees, their allocation rule needs not to be monotone, i.e., they can allocate the item to an agent with a lower signal, incurring the winner’s blessing instead of the winner’s curse. In contrast, when imposing ex-post incentive guarantees, we show that the allocation rule must be monotone. They also consider a common value auction, specifically, only a single function which is the maximum of signals; Therefore, their non-monotone allocation rule will not decrease revenue and social welfare. However, this is not true in the more general valuation settings that we consider. Moreover, they consider agents who are fully cursed, while we consider agents who can be partially cursed. Finally, we study a much more general setting capturing a general family of valuation functions, while they consider a single valuation function.

The design of mechanisms when considering agents who act according to a behavioral bias, and not their objective utility, have recently gained traction. Recent examples of this line of research are finding a market equilibrium for agents who suffer from the endowment effect [EFF20; BDO18] and designing revenue-maximizing auctions for agents who are uncertainty-averse [Cha+18; LMP19], among others.

Finally, approximately optimal mechanisms in the interdependent values model recently gained attention. Works in this domain include simple and approximately optimal revenue maximizing auctions [RT16; Li17b; CFK14] and assumption-minimal welfare maximizing auctions [Ede+18; Ede+19; Ede+21; AT21; EGZ22; Gka+21]

4.3 Model

We consider *interdependent* valuation (IDV) settings commonly studied for rational agents. We consider a seller that sells a single indivisible item to a set of n agents with *interdependent* valuations. Each agent i has a signal s_i as private information. The agents’ signals are drawn from a joint distribution F with density f over the support S_i^n , where $S_i = [0, \bar{s}_i]$. We use \mathbf{s} to denote the agents’ signal profile s_1, \dots, s_n and use \mathbf{s}_{-i} to denote the signal profile of all agents except agent i . We impose the standard assumptions that f is continuous and nowhere zero on the signal space. Each agent i also has a *publicly-known* valuation function $v_i : [0, \bar{s}_i]^n \mapsto \mathbb{R}$, which represents the value received by agent i upon winning the auction as a function of all bidders’ signals. We adopt the following standard assumptions in the valuation function $v_i(\cdot)$:

- Non-negative and normalized, i.e., $\forall \mathbf{s}, v_i(\mathbf{s}) \geq 0$ and $v_i(\mathbf{0}) = 0$.
- Continuously differentiable.
- Monotone-non-decreasing in all signals and monotone-increasing in agent i ’s signal s_i .

We consider the following properties, introduced in the seminal paper of Milgrom and Weber [MW82]:

- (Signal symmetry) $S_1 = \dots = S_n = S$, where $S = [0, \bar{s}]$ for some \bar{s} , and $f(\mathbf{s}) = f(\mathbf{t})$ for any signal profile \mathbf{s} and its arbitrary permutation \mathbf{t} .
- (Valuation symmetry) For any i, j , $v_i(s_i, \mathbf{s}_{-i}) = v_j(t_j, \mathbf{t}_{-j})$ as long as $s_i = t_j$ and \mathbf{t}_{-j} is a permutation of \mathbf{s}_{-i} .
- (Signal affiliation) For any pair of signal profiles \mathbf{s} and \mathbf{s}' , it always holds that $f(\mathbf{s} \vee \mathbf{s}')f(\mathbf{s} \wedge \mathbf{s}') \geq f(\mathbf{s})f(\mathbf{s}')$, where $(\mathbf{s} \vee \mathbf{s}')$ is the component-wise maximum, and $(\mathbf{s} \wedge \mathbf{s}')$ is the component-wise minimum.³

These conditions are standard, and were considered by many papers in the literature, including [ER05; RT16]. For welfare maximization, we focus on a special form of affiliation, and consider a case where all signals are sampled i.i.d.

By the revelation principle, we consider without loss of generality direct mechanisms, in which agents directly report their private signals \mathbf{s} and then the auctioneer determines the auction outcome according to a pre-announced mechanism $M = \{(x_i, p_i)\}_{i \in [n]}$. Here, $x_i : [0, \bar{s}_i]^n \mapsto [0, 1]$ is the allocation rule, specifying agent i 's winning probability, and $p_i : [0, \bar{s}_i]^n \mapsto \mathbb{R}$ is the payment rule, specifying agent i 's payment.

We study deterministic and anonymous mechanism, as these are optimal for rational agents in IDV settings [RT16; Aus99; MS92]. This allows an easy comparison to existing results in the literature.

A mechanism is *deterministic* if $x_i(\mathbf{s}) \in \{0, 1\}$ for any i and \mathbf{s} . A mechanism is *anonymous* if for any \mathbf{s} and any permutation \mathbf{t} of \mathbf{s} , $x_i(\mathbf{s}) = x_j(\mathbf{t})$, $p_i(\mathbf{s}) = p_j(\mathbf{t})$ whenever $s_i = t_j$.

We make two technical assumptions about allocation rule x_i for simplicity of exposition: First, we do not allocate the item to an agent who reports a zero signal.

Assumption 4.1. $x_i(0, \mathbf{s}_{-i}) = 0$ for every i and \mathbf{s}_{-i} .

Second, we do not allocate when there is a tie in the highest reported signals.

Assumption 4.2. $x_i(\mathbf{s}) = 0$ for every i whenever $|\arg \max_i \{s_i\}| > 1$.

Since f is continuous, the events in both assumptions have zero probability measure, and therefore, can be ignored without affecting the expected social welfare or revenue of the mechanism. Moreover,

³Affiliation is a common form of positive correlation, which generalizes the common case of independent distributions.

these assumptions are without loss for the mechanisms we consider as implied by Lemma 4.8 in Section 4.4.2.

We use $\mathbf{b} = (b_1, \dots, b_n)$ to denote the reported signal profile (bid profile) of agents. Agents have quasilinear utilities—the utility of each agent i given private signal profile \mathbf{s} and bid profile \mathbf{b} under mechanism M is

$$u_i(\mathbf{b}, \mathbf{s}) = x_i(\mathbf{b})v_i(\mathbf{s}) - p_i(\mathbf{b}).$$

4.3.1 The Winner’s Curse—A Behavioral Model

We adopt the widely studied behavioral model, namely the *cursed equilibrium model*, introduced by Eyster and Rabin [ER05] to explain the occurrence of the winner’s curse. In this model, agents fail to incorporate the contingency between the other bidders’ actions and their signals, which determine the value of the auctioned item, but succeed in reasoning other parts of the game. To illustrate, let σ denote a bidding strategy profile of agents and f_σ denote the probability density of bids and signals under strategy profile σ , e.g., $f_\sigma(\mathbf{b}_{-i}|\mathbf{s}_{-i})$ represents the probability density of other bidders bidding \mathbf{b}_{-i} when having signals \mathbf{s}_{-i} . Given the strategy profile σ , a fully rational agent with signal s_i estimates the probability density of other agents receiving \mathbf{s}_{-i} and bidding \mathbf{b}_{-i} as

$$f_\sigma(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i) = f(\mathbf{s}_{-i}|s_i)f_\sigma(\mathbf{b}_{-i}|\mathbf{s}_{-i}).$$

Consequently, suppose the other agents follow strategy σ_{-i} , such an agent estimates their own expected utility when having signal s_i and bidding b_i as follows:

$$EU_i(b_i, s_i; \sigma_{-i}) = \int_{\mathbf{s}_{-i} \in \mathcal{S}^{n-1}} \int_{\mathbf{b}_{-i} \in \mathcal{S}^{n-1}} f(\mathbf{s}_{-i}|s_i) \cdot f_\sigma(\mathbf{b}_{-i}|\mathbf{s}_{-i}) (x_i(\mathbf{b})v_i(\mathbf{s}) - p_i(\mathbf{b})) d\mathbf{b}_{-i} d\mathbf{s}_{-i}$$

In contrast, an agent who fully neglects the contingency between other agents’ actions and signals estimates, as the naïve agent in the wallet game example, the counterpart probability density as if \mathbf{s}_{-i} and \mathbf{b}_{-i} are independent conditioned on their own signal s_i :

$$\tilde{f}_\sigma(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i) = f(\mathbf{s}_{-i}|s_i)f_\sigma(\mathbf{b}_{-i}|s_i),$$

where $f_\sigma(\mathbf{b}_{-i}|s_i) = \int_{\mathbf{s}_{-i}} f(\mathbf{s}_{-i}|s_i)f_\sigma(\mathbf{b}_{-i}|\mathbf{s}_{-i})d\mathbf{s}_{-i}$.⁴ Eyster and Rabin [ER05] further introduce a cursedness parameter χ to model the case where an agent partially neglects this contingency such that they

⁴Eyster and Rabin [ER05] suggested that the agents succeed in reasoning or perceiving all other parts of the game, except the contingency between other agents’ signals and actions. Therefore, agents get the correct $f_\sigma(\mathbf{b}_{-i}|s_i)$, a key assumption made in Eyster and Rabin’s behavioral model.

consider the counterpart probability density as

$$f_{\sigma}^{\chi}(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i) = (1 - \chi)f_{\sigma}(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i) + \chi\tilde{f}_{\sigma}(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i).$$

An agent with $\chi = 0$ is a fully rational agent, and as we will see later, an agent with $\chi > 0$ is possible to experience the winner's curse. We refer to such an agent with $\chi > 0$ as a *cursed agent* for short, and to an agent with $\chi = 1$ as a *fully cursed agent*. By analogy with a fully rational agent estimating their expected utility, an agent with parameter χ (falsely) estimates their expected utility given σ_{-i} as:

$$EU_i^{\chi}(b_i, s_i; \sigma_{-i}) = \int_{\mathbf{s}_{-i} \in S^{n-1}} \int_{\mathbf{b}_{-i} \in S^{n-1}} f_{\sigma}^{\chi}(\mathbf{b}_{-i}, \mathbf{s}_{-i}|s_i) (x_i(\mathbf{b})v_i(\mathbf{s}) - p_i(\mathbf{b})) d\mathbf{b}_{-i} d\mathbf{s}_{-i} \quad (4.1)$$

To explain and predict the winner's curse phenomenon, Eyster and Rabin [ER05] suggested that agents generally play the equilibrium strategy with respect to this misperceived utility $EU_i^{\chi}(b_i, s_i; \sigma)$ for some parameter $\chi > 0$, instead of $EU_i(b_i, s_i; \sigma)$. They referred to this equilibrium as the χ -cursed equilibrium. Formally, a strategy profile σ forms a χ -cursed equilibrium if it holds that for every agent i ,

$$\sigma(b_i|s_i) > 0 \iff b_i \in \arg \max EU_i^{\chi}(b_i, s_i; \sigma).$$

In the above definition, $\chi = 0$ gives the definition of the classic *Bayes-Nash equilibrium* (BNE). The Naïve strategy of the aforementioned wallet game is an example of a cursed equilibrium of fully cursed agents ($\chi = 1$). We refer the reader to Appendix C.1 for an illustrative example of how a χ -cursed equilibrium leads to a winner's curse in the wallet game.

Next, we illustrate how χ -cursed equilibrium relates to the winner's curse in general. Note that Eq. (4.1) can be rewritten⁵ as follows:

$$EU_i^{\chi}(b_i, s_i; \sigma_{-i}) = \int_{\mathbf{s}_{-i} \in S^{n-1}} \int_{\mathbf{b}_{-i} \in S^{n-1}} f(\mathbf{s}_{-i}|s_i) \cdot f_{\sigma}(\mathbf{b}_{-i}|\mathbf{s}_{-i}) (x_i(\mathbf{b})v_i^{\chi}(\mathbf{s}) - p_i(\mathbf{b})) d\mathbf{b}_{-i} d\mathbf{s}_{-i}, \quad (4.2)$$

where

$$v_i^{\chi}(\mathbf{s}) := (1 - \chi)v_i(\mathbf{s}) + \chi\mathbb{E}_{\tilde{\mathbf{s}}_{-i}}[v_i(\tilde{\mathbf{s}}_{-i}, s_i)]. \quad (4.3)$$

This rewriting shows that the utility $EU_i^{\chi}(b_i, s_i; \sigma_{-i})$ optimized by an agent with valuation function v_i in the χ -cursed equilibrium is the same as the expected utility optimized by a fully rational agent with valuation function v_i^{χ} in a BNE. Thus, we have the following proposition from [ER05].

Proposition 4.1 ([ER05]). *In the IDV setting, the χ -cursed equilibrium strategy profile of agents with valuation function v_i is the same to the BNE strategy profile of agents with a modified valuation function v_i^{χ} .*

⁵This rewriting result is given by Eyster and Rabin [ER05]. We present a derivation in the Appendix for completeness.

We name the expression v_i^χ as the *cursed valuation function* of v_i . It reflects the hypothetical value of the item to the cursed agent. It contains two part—the $(1 - \chi)v_i(\mathbf{s})$ part reflects the part of the item’s value which the agent perceives through successful contingent thinking and the $\chi\mathbb{E}_{\tilde{\mathbf{s}}_{-i}}[v_i(\tilde{\mathbf{s}}_{-i}, s_i)]$ part reflects the part of the item’s value which the agent perceives when they fully ignore the contingency between other bidders’ signals and bids. Therefore, a winner i faces the winner’s curse whenever $\mathbb{E}_{\tilde{\mathbf{s}}_{-i}}[v_i(\tilde{\mathbf{s}}_{-i}, s_i)] > v_i(\mathbf{s})$; i.e., bidder i might get an item of which the value is less than i anticipated. Moreover, i might suffer a negative utility when the payment, which can be as high as $\mathbb{E}_{\tilde{\mathbf{s}}_{-i}}[v_i(\tilde{\mathbf{s}}_{-i}, s_i)]$ for a fully cursed agent, turns out larger than $v_i(\mathbf{s})$. We refer to the conceptual utilities built upon the cursed valuation functions v_i^χ as *cursed utilities*.

Eyster and Rabin [ER05] showed with empirical wallet game data that χ has a 95% confidence interval of [0.59, 0.67] with 0.63 the optimal fit. Any $\chi > 0$ predicts agents’ bids better than the BNE strategy.

We make the following key assumption:

Assumption 4.3 (Seller knows χ). *We assume the seller knows the value of χ , that is, the seller knows the extent of which the agents exhibit the cursedness bias.*

The above assumption can be justified by the following: (i) empirical studies discussed above, showing one can accurately estimate the value of χ ; (ii) moreover, the seller can observe the practical behavior of the agents, and their profit, in order to adjust the value of χ , and update the devised auction, if the estimated value of χ seems to be inaccurate. If the assumption does not hold, it is still worthy studying the problem under this assumption for following reasons. First, we show that misestimating the value of χ by ϵ still leads to an approximate C-EPIC-IR mechanism (Proposition 4.3), and thus using χ with a small estimation error still preserves some degree of incentive compatibility. Second, devising mechanisms when assuming knowing the value of χ leads to many interesting theoretical findings, which have implications on designing mechanisms for agents who suffer from winner’s curse. An example of one such implication is that there exists a tension between ensuring that the agents would not experience negative profit due to their inability to reason about their utility and the revenue of the mechanism. In other words, to ensure non-negative utility for agents who may suffer from the winner’s curse, the mechanism has to sacrifice some portion of the revenue.

4.3.2 Incentive Properties for Cursed Agents and Other Desirable Properties

Bearing above behavioral implications of agents with parameter χ in mind, a natural generalization of the interim IC concept from fully rational agents to agents with parameter χ is the following.

Definition 4.1. A mechanism $M = \{(x_i, p_i)\}_{i \in [n]}$ is interim incentive compatible for agents with parameter χ , if for all i, b_i, s_i , and for the truth-telling strategy σ^* , it holds that

$$EU^\chi(s_i, s_i; \sigma_{-i}^*) \geq EU^\chi(b_i, s_i; \sigma_{-i}^*).$$

In other words, a mechanism is interim incentive compatible for agents with parameter χ if truthful-reporting is a χ -cursed equilibrium. For fully rational agents ($\chi=0$), the above definition coincides with the standard interim IC definition [RT16], where truthful-reporting forms a BNE (that is, a 0-cursed equilibrium).

We further extend this idea to obtain a *stronger* IC notion. We consider a cursed agent's expected utility when having signal s_i , while the bid profile is $\mathbf{b} = (b_i, \mathbf{b}_{-i})$. A fully rational agent will correctly estimate the degree to which other agents' signals \mathbf{s}_{-i} are contingent on their bids \mathbf{b}_{-i} , setting this probability as $f_\sigma(\mathbf{s}_{-i} | \mathbf{b}_{-i}, s_i)$, while a fully cursed agent will think the true type is independent of agents' bids, estimating this probability as $f(\mathbf{s}_{-i} | s_i)$. Therefore, an agent with parameter χ will assess their expected utility given their signal s_i , bid b_i and others bidding \mathbf{b}_{-i} given strategy σ_{-i} as:

$$EU_i^\chi(\mathbf{b}, s_i; \sigma_{-i}) = \int_{\mathbf{s}_{-i} \in \mathcal{S}^{n-1}} ((1 - \chi)f_\sigma(\mathbf{s}_{-i} | \mathbf{b}_{-i}, s_i)u_i(\mathbf{b}, \mathbf{s}) + \chi f(\mathbf{s}_{-i} | s_i)u_i(\mathbf{b}, \mathbf{s})) d\mathbf{s}_{-i}.$$

Note that we have the following relationship between $EU_i^\chi(\mathbf{b}, s_i; \sigma_{-i})$ and $EU_i^\chi(b_i, s_i; \sigma_{-i})$:⁶

$$EU_i^\chi(b_i, s_i; \sigma_{-i}) = \int_{\mathbf{b}_{-i} \in \mathcal{S}^{n-1}} f_\sigma(\mathbf{b}_{-i} | s_i) EU_i^\chi(\mathbf{b}, s_i; \sigma_{-i}) d\mathbf{b}_{-i}. \quad (4.4)$$

Therefore, we can naturally define the ex-post incentive properties for agents with parameter χ as follows.

Definition 4.2 (Cursed ex-post incentive compatibility and individually rationality (C-EPIC-IR)). *Given a cursedness parameter χ , a mechanism is cursed ex-post incentive compatible (C-EPIC) if for every i, s_i and \mathbf{s}_{-i} , and truthfully-reporting strategy σ^* ,*

$$EU_i^\chi(\mathbf{b} = \mathbf{s}, s_i; \sigma_{-i}^*) \geq EU_i^\chi((\mathbf{b}_{-i} = \mathbf{s}_{-i}, b_i), s_i; \sigma_{-i}^*), \quad \forall b_i.$$

A mechanism is cursed ex-post individually rational (C-EPIR) if for every i, \mathbf{s} ,

$$EU_i^\chi(\mathbf{b} = \mathbf{s}, s_i) \geq 0.$$

A mechanism that is both C-EPIC and C-EPIR is denoted by C-EPIC-IR.

⁶We present the derivation of Equation (4.4) in Appendix C.2.6.

Obviously, C-EPIC implies the interim IC for agents with parameter χ .

Lemma 4.2 introduces an equivalent definition of C-EPIC-IR, which simplifies the analysis of whether a mechanism satisfies C-EPIC-IR or not.

Lemma 4.2. *A mechanism is C-EPIC if and only if for every i , s_i and \mathbf{s}_{-i} ,*

$$x_i(\mathbf{s})v_i^\chi(\mathbf{s}) - p_i(\mathbf{s}) \geq x_i(b_i, \mathbf{s}_{-i})v_i^\chi(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) \quad \forall b_i.$$

A mechanism is C-EPIR if and only if for every i , \mathbf{s} ,

$$x_i(\mathbf{s})v_i^\chi(\mathbf{s}) - p_i(\mathbf{s}) \geq 0.$$

Proof. To see this lemma holds, we only need to plug the following expression of the expected utility of bidders into Definition 4.2: $\forall \mathbf{s}, b_i$

$$\begin{aligned} & EU_i^\chi((\mathbf{b}_{-i} = \mathbf{s}_{-i}, b_i), s_i; \sigma_{-i}^*) \\ &= \int_{\tilde{\mathbf{s}}_{-i} \in \mathcal{S}^{n-1}} \left((1 - \chi) f_\sigma(\tilde{\mathbf{s}}_{-i} | \mathbf{s}_{-i}, s_i) u_i((b_i, \mathbf{s}_{-i}), (s_i, \tilde{\mathbf{s}}_{-i})) \right. \\ & \quad \left. + \chi f(\tilde{\mathbf{s}}_{-i} | s_i) u_i((b_i, \mathbf{s}_{-i}), (s_i, \tilde{\mathbf{s}}_{-i})) \right) d\tilde{\mathbf{s}}_{-i} \\ &= (1 - \chi) u_i((b_i, \mathbf{s}_{-i}), \mathbf{s}) \\ & \quad + \chi \int_{\tilde{\mathbf{s}}_{-i}} f(\tilde{\mathbf{s}}_{-i} | s_i) (v_i(\tilde{\mathbf{s}}_{-i}, s_i) x_i(b_i, \mathbf{s}_{-i}) - p_i(b_i, \mathbf{s}_{-i})) d\tilde{\mathbf{s}}_{-i} \\ &= x_i(b_i, \mathbf{s}_{-i}) \left((1 - \chi) v_i(\mathbf{s}) + \chi \int_{\tilde{\mathbf{s}}_{-i}} f(\tilde{\mathbf{s}}_{-i} | s_i) v_i(s_i, \tilde{\mathbf{s}}_{-i}) d\tilde{\mathbf{s}}_{-i} \right) - p_i(b_i, \mathbf{s}_{-i}) \\ &= x_i(b_i, \mathbf{s}_{-i}) \left((1 - \chi) v_i(\mathbf{s}) + \chi \mathbb{E}_{\tilde{\mathbf{s}}_{-i} \sim F|s_i} [v_i(\tilde{\mathbf{s}}_{-i}, s_i)] \right) - p_i(b_i, \mathbf{s}_{-i}) \\ &= x_i(b_i, \mathbf{s}_{-i}) v_i^\chi(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}), \end{aligned} \tag{4.5}$$

where $v_i^\chi(\mathbf{s})$ in the last equation is the cursed valuation function of the item, as defined in Eq (4.3). \square

Setting $\chi = 0$ in the definition of C-EPIC gives us the definition of ex-post IC (EPIC), where bidders truthfully reporting their signals forms an ex-post Nash equilibrium w.r.t. their true ex-post utilities. It is the strongest incentive guarantee one can hope for in the IDV setting. Similarly, C-EPIC is also the strongest incentive notion we can hope for with cursed agents in the IDV setting. Furthermore, Proposition 4.3 shows that C-EPIC is robust to small estimation errors of the χ parameter.

Proposition 4.3. *Let mechanism M be C-EPIC under cursedness parameter χ , and let agent i 's be a χ_i -cursed agent, where $\chi_i = \chi + \epsilon_i$. The truthful-reporting strategy σ^* forms an approximate ex-post Nash equilibrium for*

agent i with parameter χ_i in the sense that

$$EU_i^{\chi_i}(\mathbf{b} = \mathbf{s}, s_i; \sigma_{-i}^*) \geq EU_i^{\chi_i}((\mathbf{b}_{-i} = \mathbf{s}_{-i}, b_i), s_i; \sigma_{-i}^*) - \epsilon_i \cdot v_i(\bar{s}, \dots, \bar{s}) \quad \forall i, \mathbf{s}, b_i.$$

Ex-post IR (EPIR) Setting $\chi = 0$ in the definition of C-EPIR gives us the standard definition of EPIR, which ensures that no bidder will get a true negative ex-post utility at the truthful-reporting equilibrium. A mechanism that is C-EPIC-IR has the outcome that every agent bidding their true signal is a cursed equilibrium (or an ex-post equilibrium in terms of their cursed utilities) with each agent obtaining a non-negative utility based on their cursed valuation functions. However, although the agents think their utility will be non-negative for any possible realization of signals \mathbf{s} according to their belief, they might end up paying more than their value for the item leading to a negative utility, because their belief is inaccurate. Therefore, in addition to requiring C-EPIC-IR, we further consider designing mechanisms that are EPIR. Such mechanisms guarantee the agent will not experience *actual* negative utility upon receiving an item, therefore, such an agent would not regret participating in the auction in hindsight.

Ex-post budget balance (EPBB) In order to achieve the EPIR property, the mechanism might need to make positive transfers since the agents over-estimate their value for the item sold. In order to ensure the seller does not end up with negative revenue, we may also want to require that the mechanisms will satisfy the ex-post budget balance constraint.

Definition 4.3 (Ex-post budget-balance). *A mechanism $M = (x, p)$ is ex-post budget-balanced (EPBB) if for every signal profile \mathbf{s} , $\sum_i p_i(\mathbf{s}) \geq 0$.*

A more relaxed requirement is Ex-ante budget-balance, where the mechanism does not lose money *in expectation*.

When devising a mechanism that satisfies C-EPIC-IR, there is a natural tension between EPIR and budget-balance. The socially optimal mechanism might have negative revenue when satisfying EPIR (see Section 4.7.1). Moreover, while typical mechanisms usually have more revenue with cursed agents (without imposing EPIR) [ER05], when requiring the mechanism to satisfy EPIR, the revenue only decreases (see Proposition 4.13).

4.4 Preliminaries

4.4.1 C-EPIC-IR Mechanisms and Virtual Valuations

Roughgarden and Talgam-Cohen [RT16] extend Myerson's Lemma and payment identity for the IDV model. Whenever v_i^x is monotone, a simple adaptation of their results characterizes the space of C-EPIC-IR mechanisms. The proof is omitted, as it is identical to the one in [RT16] for the case of non-cursed agents.

Proposition 4.4. *A mechanism $M = (x, p)$ is C-EPIC-IR if and only if for every i, s_{-i} , the allocation rule x_i is monotone non-decreasing in the signal s_i , and the following payment identity and payment inequality hold:*

$$p_i(\mathbf{s}) = x_i(\mathbf{s})v_i^x(\mathbf{s}) - \int_{v_i^x(0, \mathbf{s}_{-i})}^{v_i^x(\mathbf{s})} x_i((v_i^x)^{-1}(t | \mathbf{s}_{-i}), \mathbf{s}_{-i}) dt - (x_i(0, \mathbf{s}_{-i})v_i^x(0, \mathbf{s}_{-i}) - p_i(0, \mathbf{s}_{-i})) \quad (4.6)$$

$$p_i(0, \mathbf{s}_{-i}) \leq x_i(0, \mathbf{s}_{-i})v_i^x(0, \mathbf{s}_{-i}) \quad (4.7)$$

We show that indeed v_i^x is monotone in our setting (affiliated signals and monotone v_i). We defer the proof to Appendix C.2.

Lemma 4.5. *The cursed valuation for agent i , $v_i^x(\mathbf{s})$, is monotone-non-decreasing in all agents' signals and monotone-increasing in s_i .*

In the setting where valuations are not cursed, setting $p_i(0, \mathbf{s}_{-i}) = 0$ maximizes the seller's revenue, and makes sure that the seller never has to pay the buyers participating in the auction, therefore ensures that the mechanism is budget-balanced. However, for cursed agents, even though $v_i^x(\mathbf{s}) \geq p_i(\mathbf{s})$, it might as well be the case that $v_i(\mathbf{s}) < p_i(\mathbf{s})$, resulting in negative utility, and breaching the EPIR property. Therefore, fixing a mechanism, one might want to set $p_i(0, \mathbf{s}_{-i})$ to be strictly smaller than zero for some values of \mathbf{s}_{-i} , which means the mechanism might pay agents for participating. Thus, in designing a mechanism to guarantee EPIR, one must take care in order not to violate budget balance.

Roughgarden and Talgam-Cohen [RT16] extend the definition of a virtual valuation to interdependent values setting. Given \mathbf{s}_{-i} , they define a function

$$\varphi_i(s_i | \mathbf{s}_{-i}) = v_i(\mathbf{s}) - v_i'(s_i, \mathbf{s}_{-i}) \frac{1 - F(s_i | \mathbf{s}_{-i})}{f(s_i | \mathbf{s}_{-i})},$$

and show that similarly to the private value setting, revenue maximization reduces to virtual welfare maximization. The definition of virtual valuations and formulating revenue maximization as virtual welfare maximization naturally to the case of cursed bidders.

Definition 4.4 (Cursed virtual value). *The cursed virtual valuation of agent i conditioned on \mathbf{s}_{-i} is defined as*

$$\varphi_i^\chi(s_i | \mathbf{s}_{-i}) = v_i^\chi(\mathbf{s}) - v_i^{\chi'}(s_i, \mathbf{s}_{-i}) \frac{1 - F(s_i | \mathbf{s}_{-i})}{f(s_i | \mathbf{s}_{-i})}.$$

The next proposition follows the exact same derivation as the one in [RT16; Mye81] for non-cursed agents.

Proposition 4.6 (Follows from [RT16; Mye81]). *For every interdependent values setting, the expected revenue of a C-EPIC-IR mechanism equals its expected conditional cursed virtual surplus, up to an additive factor:*

$$\mathbb{E}_{\mathbf{s}} \left[\sum_i p_i(\mathbf{s}) \right] = \mathbb{E}_{\mathbf{s}} \left[\sum_i x_i(\mathbf{s}) \varphi_i^\chi(s_i | \mathbf{s}_{-i}) \right] - \sum_i \mathbb{E}_{\mathbf{s}_{-i}} [x_i(0, \mathbf{s}_{-i}) v_i^\chi(0, \mathbf{s}_{-i}) - p_i(0, \mathbf{s}_{-i})]$$

4.4.2 Deterministic C-EPIC-IR Mechanisms

In this chapter we focus on deterministic mechanisms, as deterministic mechanisms are optimal for our setting whenever bidders are not cursed [RT16]. The following is a direct corollary of the monotonicity of C-EPIC-IR mechanisms.

Corollary 4.7. *Any deterministic C-EPIC-IR mechanism is a threshold mechanism. i.e., for every i , there exists a*

$$\text{function } t_i(\cdot) \text{ such that } x_i(s_i, \mathbf{s}_{-i}) = \begin{cases} 1 & s_i > t_i(\mathbf{s}_{-i}) \\ 0 & s_i \leq t_i(\mathbf{s}_{-i}) \end{cases}.$$

We refer to $t_i(\mathbf{s}_{-i})$ as the *critical bid* for agent i . Note that when $t_i(\mathbf{s}_{-i}) = \bar{s}$, we never allocate to agent i . The following lemma restricts the set of allocation rules we inspect.

Lemma 4.8. *For every deterministic, anonymous C-EPIC mechanism and for every \mathbf{s} , if the item is allocated, it is allocated to a bidder in $\arg\max_i \{s_i\}$.*

Proof. Assume the item is given to an agent j such that there exists i for which s_i is strictly bigger than s_j . By anonymity, there exists some $i \in \arg\max_\ell \{s_\ell\}$ such that if we switch i and j 's signal, i wins the item. Since j wins at \mathbf{s} , j also wins at $\mathbf{s}' = (s'_j = s_i, \mathbf{s}'_{-j} = \mathbf{s}_{-j})$ by monotonicity of C-EPIC mechanisms. Since i wins at $\mathbf{s}^* = (s_i^* = s_j, s_j^* = s_i, \mathbf{s}_{-ij}^*)$, by monotonicity, i also wins at $\mathbf{s}^* = (s_i^* = s_i, s_j^* = s_i, \mathbf{s}_{-ij}^*) = \mathbf{s}'$, a contradiction. \square

The above lemma implies that when dealing with such mechanisms, assuming Assumptions 4.1 and 4.2 are without loss. By the above lemma, a zero signal cannot win unless it is not the signal, therefore Assumption 4.1 follows from Assumption 4.2. For assumption 4.2, one can take any mechanism that violates this assumption, and set $x_i(\mathbf{s}) = 0$ for every \mathbf{s} where the highest bid is not unique. By

the above lemma, such mechanism remains monotone non-decreasing in a bidder's own bid, therefore C-EPIC. By continuity of the signal distribution, the resulting mechanism has the same expected revenue and social welfare as the original one.

4.5 Implications of Ex-post IR

In this section we discuss implications of imposing EPIR on the mechanism. We first show that in order to achieve our incentive properties, it suffices to design an ex-post IR mechanism, and cursed ex-post IR will follow.

Lemma 4.9. *For every interdependent value setting, C-EPIC and EPIR implies C-EPIR.*

Proof. For a mechanism to be EPIR, we need that the actual value an agent gets from an allocation is higher than the price they pay. That is, $x_i(\mathbf{s})v_i(\mathbf{s}) \geq p_i(\mathbf{s})$ for every \mathbf{s} . Using equation (4.6) for C-EPIC mechanisms, and rearranging, we get that for every i and every \mathbf{s}

$$p_i(0, \mathbf{s}_{-i}) \leq x_i(\mathbf{s}) (v_i(\mathbf{s}) - v_i^X(\mathbf{s})) + \int_{v_i^X(0, \mathbf{s}_{-i})}^{v_i^X(\mathbf{s})} x_i((v_i^X)^{-1}(t|\mathbf{s}_{-i}), \mathbf{s}_{-i}) dt + x_i(0, \mathbf{s}_{-i})v_i^X(0, \mathbf{s}_{-i}). \quad (4.8)$$

Specifically, fixing \mathbf{s}_{-i} and setting $s_i = 0$, we get

$$\begin{aligned} p_i(0, \mathbf{s}_{-i}) &\leq x_i(0, \mathbf{s}_{-i}) (v_i(0, \mathbf{s}_{-i}) - v_i^X(0, \mathbf{s}_{-i})) + \int_{v_i^X(0, \mathbf{s}_{-i})}^{v_i^X(0, \mathbf{s}_{-i})} x_i((v_i^X)^{-1}(t|\mathbf{s}_{-i}), \mathbf{s}_{-i}) dt + x_i(0, \mathbf{s}_{-i})v_i^X(0, \mathbf{s}_{-i}) \\ &= x_i(0, \mathbf{s}_{-i})v_i^X(0, \mathbf{s}_{-i}), \end{aligned}$$

where the inequality used Assumption 4.1, that agents aren't allocated at their lowest signal ($x_i(0, \mathbf{s}_{-i}) = 0$). This coincides with Equation (4.7), implying C-EPIC. \square

According to Corollary 4.7, every deterministic allocation rule is equivalent to a set of threshold functions $\{t_i(\cdot)\}_i$. As noted before, the only freedom one has in setting payments of C-EPIC mechanisms is by setting the term $p_i(0, \mathbf{s}_{-i})$. We show that when maximizing revenue subject to C-EPIC and EPIR constraints, there is a single optimal way to set $p_i(0, \mathbf{s}_{-i})$.

Lemma 4.10. *Fixing threshold functions $\{t_i(\cdot)\}_i$ of a deterministic anonymous mechanism, the revenue optimal C-EPIC-IR and EPIR mechanism sets*

$$p_i(0, \mathbf{s}_{-i}) = \begin{cases} \min \{0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\} & \text{if } t_i(\mathbf{s}_{-i}) < \bar{s}, \\ 0 & \text{otherwise.} \end{cases}$$

(and therefore, the payment is uniquely defined using Equation (4.6).)

Proof. Since $x_i(0, \mathbf{s}_{-i}) = 0$ by Assumption 4.1, we have to have $p_i(0, \mathbf{s}_{-i}) \leq 0$, otherwise an agent pays without getting allocated, which leads to negative utility. Consider \mathbf{s}_{-i} such that $t_i(\mathbf{s}_{-i}) < \bar{s}$. By Equation (4.8) for EPIR (which also implies C-EPIR), we get that in order to maximize revenue, one should set

$$\begin{aligned}
p_i(0, \mathbf{s}_{-i}) &= \min \left\{ 0, \min_{\bar{s}_i} \left\{ x_i(\mathbf{s}) (v_i(\mathbf{s}) - v_i^X(\mathbf{s})) + \int_{v_i^X(0, \mathbf{s}_{-i})}^{v_i^X(\mathbf{s})} x_i((v_i^X)^{-1}(t|\mathbf{s}_{-i}), \mathbf{s}_{-i}) dt \right\} \right\} \\
&= \min \left\{ 0, \min_{s_i > t_i(\mathbf{s}_{-i})} \left\{ v_i(\mathbf{s}) - v_i^X(\mathbf{s}) + \int_{v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})}^{v_i^X(\mathbf{s})} 1 dt \right\} \right\} \\
&= \min \left\{ 0, \min_{s_i > t_i(\mathbf{s}_{-i})} \left\{ v_i(\mathbf{s}) - v_i^X(\mathbf{s}) + (v_i^X(\mathbf{s}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})) \right\} \right\} \\
&= \min \left\{ 0, \min_{s_i > t_i(\mathbf{s}_{-i})} \left\{ v_i(\mathbf{s}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \right\} \right\} \\
&= \min \left\{ 0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \right\}
\end{aligned}$$

where the second equality follows from the fact that $x_i(\mathbf{s}) = 1$ only if $s_i > t_i(\mathbf{s}_{-i})$.

For $t_i(\mathbf{s}_{-i}) = \bar{s}$, agent i will never get the item and therefore, will not incur the winner's curse. Hence, we can set $p_i(0, \mathbf{s}_{-i}) = 0$. \square

We get that when agents experience the winner's curse at the "critical value" (meaning their real value is smaller than their perceived value), they pay their real value at the critical bid, while if they do not experience the winner's curse, they pay their cursed value. We get the following corollary.

Corollary 4.11. *Fixing an anonymous, deterministic, C-EPIR, EPIR mechanism with threshold function $t(\cdot)$, the revenue optimal way to set $p_i(0, \mathbf{s}_{-i})$ gives the following payment function:*

$$p_i(\mathbf{s}) = \begin{cases} p_i(0, \mathbf{s}_{-i}) & s_i \leq t_i(\mathbf{s}_{-i}) \\ v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) & s_i > t_i(\mathbf{s}_{-i}) \text{ and } v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0 \cdot \\ v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) & s_i > t_i(\mathbf{s}_{-i}) \text{ and } v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) > 0 \end{cases}$$

Proof. If $s_i \leq t_i(\mathbf{s}_{-i})$, then i does not get the item, and by Equation (4.6), $p_i(\mathbf{s}) = p_i(0, \mathbf{s}_{-i})$. If $s_i \geq t_i(\mathbf{s}_{-i})$, then i gets the item, and according to Equation (4.6) and Lemma 4.10,

$$\begin{aligned}
p_i(\mathbf{s}) &= v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + p_i(0, \mathbf{s}_{-i}) \\
&= v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \min\{0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}.
\end{aligned}$$

Therefore, if $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0$, $p_i(\mathbf{s}) = v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + 0 = v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})$, and if $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \geq 0$, then $p_i(\mathbf{s}) = v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) = v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})$. \square

An interesting implication of this corollary is that as opposed to the case where we do not require EPIR, the welfare and revenue in the case of non-cursed agents (i.e., $\chi = 0$) is at least that of the case where agents are cursed ($\chi > 0$). We show this more generally by showing that the welfare and revenue are monotonically non-increasing in χ . In order to show this, we first show that for a fixed mechanism, EPIR implies that the revenue decreases as χ increases.

Lemma 4.12. *Fix a threshold rule t . Then for every \mathbf{s} , every i , and every $0 \leq \chi \leq \chi' \leq 1$, $p_i^\chi(\mathbf{s}, t) \geq p_i^{\chi'}(\mathbf{s}, t)$, where $p_i^\chi(\mathbf{s}, t)$ is the optimal payment an agent i with cursedness parameter χ has with signals \mathbf{s} .*

Proof. We prove by the three cases of Corollary 4.11.

Case 1: $s_i \leq t(\mathbf{s}_{-i})$. Then, $p_i^\chi(\mathbf{s}, t) = \min\{0, v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}$, and $p_i^{\chi'}(\mathbf{s}, t) = \min\{0, v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^{\chi'}(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}$. Notice that if $p_i^\chi(\mathbf{s}, t) < 0$, then

$$\begin{aligned}
v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) &< v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \\
&= (1 - \chi)v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \chi \mathbb{E}_{\mathbf{t}_{-i} \sim F_{t(\mathbf{s}_{-i})}}[v_i(t(\mathbf{s}_{-i}), \mathbf{t}_{-i})] \\
\iff v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) &< \mathbb{E}_{\mathbf{t}_{-i} \sim F_{t(\mathbf{s}_{-i})}}[v_i(t(\mathbf{s}_{-i}), \mathbf{t}_{-i})] \\
\iff v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) &= (1 - \chi)v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \chi \mathbb{E}_{\mathbf{t}_{-i} \sim F_{t(\mathbf{s}_{-i})}}[v_i(t(\mathbf{s}_{-i}), \mathbf{t}_{-i})] \\
&\leq (1 - \chi')v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \chi' \mathbb{E}_{\mathbf{t}_{-i} \sim F_{t(\mathbf{s}_{-i})}}[v_i(t(\mathbf{s}_{-i}), \mathbf{t}_{-i})] \\
&= v_i^{\chi'}(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \\
\iff p_i^\chi(\mathbf{s}, t) &= v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \\
&\geq v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^{\chi'}(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \\
&= p_i^{\chi'}(\mathbf{s}, t), \tag{4.9}
\end{aligned}$$

which means that either $p_i^\chi(\mathbf{s}, t) = p_i^{\chi'}(\mathbf{s}, t) = 0$, or $p_i^\chi(\mathbf{s}, t) \geq p_i^{\chi'}(\mathbf{s}, t)$.

Case 2: $s_i > t_i(\mathbf{s}_{-i})$ and $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0$. In this case, according to Corollary 4.11, we have $p_i^\chi(\mathbf{s}, t) = v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})$. According to Equation (4.9), $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0$ implies $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^{\chi'}(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0$, and therefore, $p_i^{\chi'}(\mathbf{s}, t) = v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})$ as well.

Case 3: $s_i > t_i(\mathbf{s}_{-i})$ and $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \geq 0$. According to Equation (4.9), we also have $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^{\chi'}(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \geq 0$, which implies that $p_i^\chi(\mathbf{s}, t) = v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})$ and $p_i^{\chi'}(\mathbf{s}, t) = v_i^{\chi'}(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})$. Since by Equation (4.9), $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \geq 0$ implies $v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) > v_i^{\chi'}(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})$, the lemma follows. \square

We get the following.

Proposition 4.13 (Revenue monotonicity). *For any $0 \leq \chi \leq \chi' \leq 1$ the revenue optimal anonymous deterministic C-EPIC-IR EPIR mechanism for χ -cursed agents has revenue at least as high as the revenue optimal anonymous deterministic C-EPIC-IR EPIR mechanism for χ' -cursed agents.*

Proof. Let $t^{\chi'}$ be the threshold rule of the revenue optimal deterministic C-EPIC-IR EPIR mechanism for χ -cursed agents, and let $\text{REV}^\chi(t)$ be the optimal revenue of χ -cursed agents with threshold rule t . Then

$$\text{REV}^\chi(t^{\chi'}) = \int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^\chi(\mathbf{s}, t^{\chi'}) d\mathbf{s} \geq \int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^{\chi'}(\mathbf{s}, t^{\chi'}) = \text{REV}^{\chi'}(t^{\chi'}),$$

where the inequality follows Lemma 4.12. Therefore, for the optimal deterministic mechanism for agents with cursedness parameter χ , the revenue can only be larger. \square

Proposition 4.14 (Welfare monotonicity). *For any $0 \leq \chi \leq \chi' \leq 1$, (a) the welfare optimal deterministic C-EPIC-IR EPIR mechanism for χ -cursed agents that satisfies ex-post (ex-ante) budget-balance has welfare at least as high as the welfare optimal deterministic C-EPIC-IR EPIR mechanism for χ' -cursed agents that satisfies ex-post (ex-ante) budget-balance.*

Proof. We show that every threshold rule that is ex-post (ex-ante) budget-balanced for χ' -cursed agents is also ex-post (ex-ante) budget-balanced for χ -cursed agents. Since the space of mechanisms is larger for χ -cursed agents, the proposition follows. Fix a threshold rule t and signal profile \mathbf{s} . By Lemma 4.12, we have that $\sum_i p_i^\chi(\mathbf{s}, t) \geq \sum_i p_i^{\chi'}(\mathbf{s}, t)$ and $\int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^\chi(\mathbf{s}, t) d\mathbf{s} \geq \int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^{\chi'}(\mathbf{s}, t) d\mathbf{s}$. Therefore, if $\sum_i p_i^{\chi'}(\mathbf{s}, t)$ or $\int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^{\chi'}(\mathbf{s}, t) d\mathbf{s}$ are non-negative, so are $\sum_i p_i^\chi(\mathbf{s}, t)$ or $\int_{\mathbf{s}} f(\mathbf{s}) \sum_i p_i^\chi(\mathbf{s}, t) d\mathbf{s}$. \square

4.6 Revenue Maximization

In this section, we devise a mechanism that maximizes revenue among all deterministic, anonymous, C-EPIC-IR and EPIR mechanisms. According to Lemma 4.8, we only consider rules for which $t_i(\mathbf{s}_{-i}) \geq \max_{j \neq i} s_j$. We will use $s_{-i}^* = \max_{j \neq i} s_j$ to denote the second highest signal. We notice that every such mechanism is feasible by definition.

Observation 4.15. *Every deterministic mechanism such that $x_i(\mathbf{s}) = 1$ implies that $s_i > s_{-i}^*$ is feasible, meaning that for every \mathbf{s} , $\sum_i x_i(\mathbf{s}) \leq 1$.*

Proof. This is simply because there cannot be two bidders with a signal strictly larger than all other signals. \square

Therefore, when designing a revenue optimal deterministic anonymous mechanism, one needs to only care about a threshold rule $t(\cdot)$ satisfying $t(\mathbf{s}_{-i}) > s_{-i}^*$ (the same for all bidders due to anonymity),

and not worry about feasibility or how to set $p_i(\mathbf{s}_{-i})$, as those are set according to Lemma 4.10. We now show how to devise the optimal anonymous deterministic mechanism.

Theorem 4.16. *Let*

$$r_i(t, \mathbf{s}_{-i}) = \begin{cases} \min\{v_i^\chi(t, \mathbf{s}_{-i}), v_i(t, \mathbf{s}_{-i})\} - v_i^\chi(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}) & t \in [s_{-i}^*, \bar{s}] \\ 0 & t = \bar{s} \end{cases} \quad (4.10)$$

The optimal deterministic anonymous mechanism sets a threshold

$$t(\mathbf{s}_{-i}) \in \arg \max_{t \in [s_{-i}^*, \bar{s}]} r_i(t, \mathbf{s}_{-i}).$$

Proof. Using Assumption 4.1 to simplify the expression in Proposition 4.6, our objective is to maximize the following expected revenue.

$$\mathbb{E}_{\mathbf{s}} \left[\sum_i x_i(\mathbf{s}) \varphi_i^\chi(s_i|\mathbf{s}_{-i}) \right] + \sum_i \mathbb{E}_{\mathbf{s}_{-i}} [p_i(0, \mathbf{s}_{-i})].$$

Since we are interested in anonymous mechanisms, this is equivalent to maximizing

$$\mathbb{E}_{\mathbf{s}} [x_i(\mathbf{s}) \varphi_i^\chi(s_i|\mathbf{s}_{-i})] + \mathbb{E}_{\mathbf{s}_{-i}} [p_i(0, \mathbf{s}_{-i})]$$

for a given agent i .

Applying Lemma 4.10, we first consider the case $t_i(\mathbf{s}_{-i}) < \bar{s}$ and aim to find a threshold function $t(\cdot)$ such that the following is maximized, and then consider the case that $t_i(\mathbf{s}_{-i}) = \bar{s}$.

$$\begin{aligned} & \int_{\mathbf{s}} f(\mathbf{s}) x_i(\mathbf{s}) \varphi_i^\chi(s_i|\mathbf{s}_{-i}) d\mathbf{s} + \int_{\mathbf{s}_{-i}} f(\mathbf{s}_{-i}) \min\{0, v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})\} d\mathbf{s}_{-i} \\ &= \int_{\mathbf{s}_{-i}} f(\mathbf{s}_{-i}) \left(\int_{t(\mathbf{s}_{-i})}^1 f(s_i|\mathbf{s}_{-i}) \varphi_i^\chi(s_i|\mathbf{s}_{-i}) ds_i + \min\{0, v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})\} \right) d\mathbf{s}_{-i}. \end{aligned} \quad (4.11)$$

We use the following expansion.

$$\begin{aligned} \int_{t(\mathbf{s}_{-i})}^{\bar{s}} f(s_i|\mathbf{s}_{-i}) \varphi_i^\chi(s_i|\mathbf{s}_{-i}) ds_i &= \int_{t(\mathbf{s}_{-i})}^{\bar{s}} f(s_i|\mathbf{s}_{-i}) \left(v_i^\chi(\mathbf{s}) - v_i^{\chi'}(\mathbf{s}) \frac{1 - F(s_i|\mathbf{s}_{-i})}{f(s_i|\mathbf{s}_{-i})} \right) ds_i \\ &= \int_{t(\mathbf{s}_{-i})}^{\bar{s}} f(s_i|\mathbf{s}_{-i}) v_i^\chi(\mathbf{s}) + F(s_i|\mathbf{s}_{-i}) v_i^{\chi'}(\mathbf{s}) ds_i - \int_{t(\mathbf{s}_{-i})}^{\bar{s}} v_i^{\chi'}(\mathbf{s}) ds_i \\ &= (F(s_i|\mathbf{s}_{-i}) - 1) v_i^\chi(\mathbf{s}) \Big|_{t(\mathbf{s}_{-i})}^{\bar{s}} \\ &= (F(\bar{s}|\mathbf{s}_{-i}) - 1) v_i^\chi(\bar{s}, \mathbf{s}_{-i}) - (F(t(\mathbf{s}_{-i})|\mathbf{s}_{-i}) - 1) v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \\ &= (1 - F(t(\mathbf{s}_{-i})|\mathbf{s}_{-i})) v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}), \end{aligned} \quad (4.12)$$

where the third equality uses integration by parts, and the last inequality is due to $F(\bar{s}|\mathbf{s}_{-i}) = 1$. Plugging

Equation (4.12) into Equation (4.11), we get:

$$\int_{\mathbf{s}_{-i}} f(\mathbf{s}_{-i}) ((1 - F(t(\mathbf{s}_{-i})|\mathbf{s}_{-i}))v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \min\{0, v_i(t(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}) d\mathbf{s}_{-i} \quad (4.13)$$

Notice that for each \mathbf{s}_{-i} , the choice of threshold $t(\mathbf{s}_{-i})$ is independent of a different \mathbf{s}'_{-i} 's threshold (we are guaranteed to be feasible by Observation 4.15). Let $t = t(\mathbf{s}_{-i})$. We wish to choose $t \geq s_{-i}^*$ that maximizes:

$$(1 - F(t|\mathbf{s}_{-i}))v_i^\chi(t, \mathbf{s}_{-i}) + \min\{0, v_i(t, \mathbf{s}_{-i}) - v_i^\chi(t, \mathbf{s}_{-i})\}$$

For t such that $v_i(t, \mathbf{s}_{-i}) \geq v_i^\chi(t, \mathbf{s}_{-i})$, we get

$$(1 - F(t|\mathbf{s}_{-i}))v_i^\chi(t, \mathbf{s}_{-i}) + v_i(t, \mathbf{s}_{-i}) - v_i^\chi(t, \mathbf{s}_{-i}) = v_i(t, \mathbf{s}_{-i}) - v_i^\chi(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}),$$

while for t satisfying $v_i(t, \mathbf{s}_{-i}) < v_i^\chi(t, \mathbf{s}_{-i})$, we have

$$(1 - F(t|\mathbf{s}_{-i}))v_i^\chi(t, \mathbf{s}_{-i}) + 0 = v_i^\chi(t, \mathbf{s}_{-i}) - v_i^\chi(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}).$$

Hence, for every \mathbf{s}_{-i} , we should choose a $t > s_{-i}^*$ that maximizes

$$\min\{v_i^\chi(t, \mathbf{s}_{-i}), v_i(t, \mathbf{s}_{-i})\} - v_i^\chi(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}). \quad (4.14)$$

Second, we consider the case $t_i(\mathbf{s}_{-i}) = \bar{s}$. In this case, we never allocate to agent i whichever s_i is and $p_i(\mathbf{s}) = p_i(0, \mathbf{s}_{-i}) = 0$, leading to zero expected revenue. Therefore, Theorem 4.16 gives the optimal threshold.

□

Note that in the case we only require C-EPIC-IR without EPIR, we can set $p_i(0, \mathbf{s}_{-i}) = 0$, and the optimal mechanism chooses a $t > s_{-i}^*$ that maximizes

$$v_i^\chi(t, \mathbf{s}_{-i}) - v_i^\chi(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}),$$

which coincides with the mechanism in [RT16] for cursed valuation, while the optimal EPIC-IR mechanism in [RT16] for non-cursed valuations will choose $t > s_{-i}^*$ that maximizes

$$v_i(t, \mathbf{s}_{-i}) - v_i(t, \mathbf{s}_{-i})F(t|\mathbf{s}_{-i}),$$

which coincides with our mechanism when $\chi = 0$.

4.7 Welfare Maximization

We consider the objective of maximizing welfare in an EPIR and EPBB way. Unless stated otherwise, we assume that the bidders' signals are sampled i.i.d. In Section 4.7.1 we demonstrate the tension that arises between devising a truthful mechanism for cursed agents, and the requirement that the agents would not experience a negative utility. We show a simple example where in the fully efficient mechanism, the mechanism pays the agents much more than it earns. In Section 4.7.2 we present an operation that takes any deterministic mechanism, and makes it an EPBB mechanism by not allocating in scenarios where the mechanisms would be required to make positive transfers. In Section 4.7.3 we show that under natural assumptions on the valuation functions, every EPBB mechanism does not make positive transfers. This implies that the masking operation on the socially optimal mechanism yields a welfare optimal EPBB mechanism (see Proposition 4.23). We then present two interesting scenarios: (i) when agents' valuation is the max function, any EPBB mechanism obtains zero welfare, and (Section 4.7.4) (ii) a family of valuations including weighted-sum valuations and ℓ_p -norm of signals, where as the number of agents grows large, the expected EPBB welfare approaches 1/2 of the expected optimal allocation (Section 4.7.4). Missing proofs can be found in the Appendix C.2.

4.7.1 Welfare Optimal Mechanism is not Budget Balanced

Consider n bidders with valuations⁷ $v_i(\mathbf{s}) = s_i + \frac{1}{2} \sum_{j \neq i} s_j$ and signals drawn independently from $U[0, 1]$. Suppose that $\chi = 1$, that is, the agents are fully cursed. The welfare optimal C-EPIC-IR mechanism gives the item to the agent with the highest value/signal, and charges payments according to Equations (4.6) and (4.8). We show that such mechanism must incur a negative revenue of order $\Theta(n\sqrt{n})$. We note that the mechanism does not even satisfy the less restrictive requirement of ex-ante budget-balance.

Proposition 4.17. *There exists a setting where the welfare optimal mechanism has an expected revenue loss of $\Theta(n\sqrt{n})$.*

Proof. For a signal profile \mathbf{s} , assuming that 1 is the highest, agent 1 pays $v_1^\chi(s_{-1}^*, \mathbf{s}_{-1})$, and the seller pays each bidder $i - p_i(0, \mathbf{s}_{-i})$ subject to the EPIR constraints. To minimize their payment, Lemma 4.10 implies that the seller sets $p_i(0, \mathbf{s}_{-i}) = \min \{0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}$. While since the signals

⁷Note these are weighted-sum valuations with $\beta = 1/2$.

are bounded by 1, it is clear that the expected payment of the highest bidder is $O(n)$. We will show that

$$\begin{aligned}\mathbb{E}_{\mathbf{s}}\left[\sum_i -p_i(0, \mathbf{s}_{-i})\right] &= \mathbb{E}_{\mathbf{s}}\left[\sum_i -p_i(0, \mathbf{s}_{-i})\right] = \mathbb{E}_{\mathbf{s}}\left[\sum_i \max\{0, v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\}\right] \\ &= \Omega(n\sqrt{n}).\end{aligned}$$

Notice that for a signal profile \mathbf{s} and a bidder i with $\chi = 1$,

$$v_i^1(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) = s_{-i}^* + \mathbb{E}_{\tilde{\mathbf{s}}}\left[\frac{1}{2}\sum_{j \neq i} \tilde{s}_j\right] - \left(s_{-i}^* + \frac{1}{2}\sum_{j \neq i} s_j\right) = \frac{n-1}{4} - \frac{1}{2}\sum_{j \neq i} s_j.$$

We get

$$\begin{aligned}\mathbb{E}_{\mathbf{s}}\left[\sum_i -p_i(0, \mathbf{s}_{-i})\right] &= \sum_i \mathbb{E}_{\mathbf{s}}\left[\max\left\{0, \frac{n-1}{4} - \frac{1}{2}\sum_{j \neq i} s_j\right\}\right] \\ &= \sum_i \frac{1}{2} \mathbb{E}_{\mathbf{s}}\left[\max\left\{0, \frac{n-1}{2} - \sum_{j \neq i} s_j\right\}\right] \\ &\approx \sum_i \frac{1}{2} \mathbb{E}_{x \sim N(\frac{n-1}{2}, (n-1)/12)}\left[\max\left\{0, \frac{n-1}{2} - x\right\}\right] \\ &= \sum_i \frac{1}{2} \mathbb{E}_{x \sim N(0, (n-1)/12)}\left[\max\{0, x\}\right] \\ &= \sum_i \frac{1}{2} \mathbb{E}_{x \sim N(0, (n-1)/12)}\left[x \mid x \geq 0\right] \Pr_x[x \geq 0] \\ &= \frac{n}{4} \sqrt{\frac{(n-1)}{24\pi}} = \Theta(n\sqrt{n}).\end{aligned}$$

Here, the approximation follows the central limit theorem, the third equality follows by symmetry of the normal distribution, and the fourth equality follows by taking the expected value of a half-normal distribution.

Since the seller collects $O(n)$ from the buyers, but pays them $\Theta(n\sqrt{n})$, the proof follows. \square

4.7.2 Masked Mechanisms

We define an operation that takes a deterministic mechanism, and imposes no positive transfers ($p_i(0, \mathbf{s}_{-i}) = 0$), therefore, the masking operation outputs a mechanism that is trivially EPBB.

Definition 4.5. *Given a deterministic mechanism with threshold function $t_i(\mathbf{s}_{-i})$ for all i, \mathbf{s}_{-i} , let $\mathcal{NC}(\mathbf{s}_{-i}) = \{t \mid t \geq s_i^*, \text{ and } v_i(t, \mathbf{s}_{-i}) \geq \mathbb{E}_{\tilde{\mathbf{s}}_{-i} | s_i = st} [v_i(t, \tilde{\mathbf{s}}_{-i})]\}$. A masking of the mechanism is a threshold mechanism with a*

new threshold function

$$t'_i(\mathbf{s}_{-i}) = \begin{cases} \inf\{t | t \geq t_i(\mathbf{s}_{-i}), t \in \mathcal{NC}(\mathbf{s}_{-i})\} & \text{if } \{t | t \geq t_i(\mathbf{s}_{-i}), t \in \mathcal{NC}(\mathbf{s}_{-i})\} \neq \emptyset \\ 1 & \text{otherwise} \end{cases}.$$

The following lemma shows that indeed masking a mechanism results in a mechanism that is EPBB. We note that this implication does not assume that the signals are independent (as opposed to the rest of this section).

Lemma 4.18. *For any deterministic mechanism that can be implemented in a C-EPIC-IR, EPIR, its masking can be implemented in a C-EPIC-IR, EPIR, and EPBB.*

Proof. The masking of a deterministic mechanism is still a threshold mechanism, therefore, it is can be implemented in a C-EPIC-IR and EPIR given payments that satisfy Equations (4.7) and (4.8). To show that the mechanism can be implemented in an EPBB manner, we show that for every i , \mathbf{s}_{-i} and χ ,

$$v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \geq \mathbb{E}_{\tilde{\mathbf{s}}_{-i} \sim F_{|t_i(\mathbf{s}_{-i})}}[v_i(t_i(\mathbf{s}_{-i}), \tilde{\mathbf{s}}_{-i})] \quad (4.15)$$

implies $\min\{0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\} = 0$; Therefore, according to Lemma 4.10, we can set $p_i(0, \mathbf{s}_{-i}) = 0$.

We have that

$$\begin{aligned} v_i^\chi(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) &= (1 - \chi)v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) + \chi \mathbb{E}_{\tilde{\mathbf{s}}_{-i} \sim F_{|t_i(\mathbf{s}_{-i})}}[v_i(t_i(\mathbf{s}_{-i}), \tilde{\mathbf{s}}_{-i})] \\ &\leq v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}), \end{aligned}$$

where the equality follows Equation (4.3), and the first inequality follows from the condition in Equation (4.15). The lemma follows. \square

4.7.3 Ex-post Budget-Balance Implies No Positive Transfers

In the following we show that under natural conditions, every mechanism that is deterministic, anonymous, C-EPIC-IR EPIR and ex-post budget-balance has no positive transfers. This will imply that the optimal deterministic, anonymous, C-EPIC-IR EPIR and ex-post budget-balance mechanism is a masking of the generalized Vickrey auction [MS92; Aus99]. We then show that for the max function, every masked mechanism that allocates the item with zero probability must have positive transfers, implying that such a mechanism will never sell in order to impose ex-post budget-balance. This gives an unbounded gap between the optimal welfare for non-cursed agents and for cursed agents. Finally, we show some

interesting families of valuations, where one can provably show that the masking of the generalized Vickrey auction approximate the optimal mechanism for non-cursed agents.

We first introduce lemmas that will be useful in proving the main result of this section.

Lemma 4.19. *For every anonymous, deterministic, C-EPIC-IR EPIR, and EPBB mechanism, for every i and \mathbf{s}_{-i} , $p_i(0, \mathbf{s}_{-i}) < 0$ implies $t_i(\mathbf{s}_{-i}) = s_{-i}^*$.*

Proof. We prove by contradiction. Suppose there exists some \mathbf{s}_{-i} such that $p_i(0, \mathbf{s}_{-i}) < 0$ but $t_i(\mathbf{s}_{-i}) > s_{-i}^*$. Then, for $s_i = \frac{1}{2}(s_{-i}^* + t_i(\mathbf{s}_{-i}))$, we have that for the signal profile \mathbf{s} there are no winners — i has the highest signal, but it is still lower than i 's threshold, and by Lemma 4.8, only the highest signal can win. Therefore, we have that $p_j(\mathbf{s}) = p_j(0, \mathbf{s}_{-j})$ for every bidder j , and $\sum_j p_j(\mathbf{s}) = \sum_j p_j(0, \mathbf{s}_{-j}) \leq p_i(0, \mathbf{s}_{-i}) < 0$, which violates the ex-post budget-balance property. \square

Lemma 4.20. *For every anonymous, deterministic, C-EPIC-IR EPIR, and EPBB mechanism, if there exist i and \mathbf{s}_{-i} such that $p_i(0, \mathbf{s}_{-i}) < 0$, then there's a unique bidder j with maximum signal in \mathbf{s}_{-i} , and for every $s_i \in [0, s_j)$, we have $t_j(\mathbf{s}_{-ij}, s_i) < s_j$.*

Proof. We prove by contradiction. Suppose there are two bidders with the same highest signal in \mathbf{s}_{-i} . For signal profile $\mathbf{s}' = (0, \mathbf{s}_{-i})$, according to Assumption 4.2, no one will be allocated with item and $p_k(\mathbf{s}') = p_k(0, \mathbf{s}'_{-k}) \leq 0$ for k . As $p_i(0, \mathbf{s}'_{-i}) = p_i(0, \mathbf{s}_{-i}) < 0$, therefore, $\sum_k p_k(\mathbf{s}') < 0$, violating EPBB. Thus, there is a unique bidder j with the highest signal in \mathbf{s}_{-i} .

Fix $s_i < s_j$, and suppose $t_j(\mathbf{s}_{-ij}, s_i) \geq s_j$, then j cannot win the auction since j 's signal is no larger than the threshold j faces. By Lemma 4.8, no other agents can win the item, as their signals are not the highest one. Therefore, we have $p_{j'}(\mathbf{s}) = p_{j'}(0, \mathbf{s}_{-j'})$ for all j' , and $\sum_{j'} p_{j'}(\mathbf{s}) = \sum_{j'} p_{j'}(0, \mathbf{s}_{-j'}) \leq p_i(0, \mathbf{s}_{-i}) < 0$, violating the ex-post budget-balance property. \square

We now prove the main result of this section, that under a natural conditions, then every EPBB mechanism that satisfies our desired incentive properties makes no positive transfers. The condition fits the basic intuition that as the actual signals of all other bidders but some bidder i get smaller, i 's cursedness increases (as typically, i will overestimate others' signal according to original distribution of signals).

Definition 4.6 (Cursedness-monotonicity condition). *A valuation function satisfies the cursedness-monotonicity condition if for every i , \mathbf{s}_{-i} for which there exists $s_i > \max_{j \neq i} \{s_j\}$ such that $v_i(\mathbf{s}) - v_i^X(\mathbf{s}) < 0$, then for any*

$\mathbf{s}'_{-i} \leq \mathbf{s}_{-i}$ ⁸ and any $s'_i \in (\max_{j \neq i} \{s'_j\}, \bar{s})$ ⁹, we also have $v_i(\mathbf{s}') - v_i^X(\mathbf{s}') < 0$.

Proposition 4.21 below shows that cursedness-monotonicity condition holds for many widely studied valuation functions such as weighted sum valuations [RT16; Ede+19; Ede+21; Mye81] and max of signals [BBM20; BK02].

Proposition 4.21. *The following valuation functions satisfy the cursedness-monotonicity condition:*

1. $v_i(\mathbf{s}) = s_i + \beta \sum_{j \neq i} s_j$. (Weighted-sum valuations.)
2. $v_i(\mathbf{s}) = \max_i \{s_i\}$. (Maximum of signals.)

Theorem 4.22. *For every anonymous, deterministic, C-EPIC-IR, EPIR mechanisms, if the valuation function v_i satisfies cursedness-monotonicity, then a mechanism is ex-post-budget balanced if and only if for every i , \mathbf{s}_{-i} , $p_i(0, \mathbf{s}_{-i}) = 0$.*

Proof. As the “if” direction is immediate, we focus on proving the “only if” direction. Assume that we use the optimal way to set $p_i(0, \mathbf{s}_{-i})$ as described in Lemma 4.10. This is without loss since if the mechanism is not budget-balanced using optimal setting of $p_i(0, \mathbf{s}_{-i})$, it is not budget-balance for every setting of $p_i(0, \mathbf{s}_{-i})$.

We prove by contradiction. Suppose that there exists \mathbf{s}_{-i} such that $p_i(0, \mathbf{s}_{-i}) < 0$, therefore, by Lemma 4.10, $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) < 0$. Let z be the smallest non-zero signal in the set of all signals but s_i . If there is no such signal, we have $p_i(0) < 0$, and by anonymity, the revenue of the all zero signal profile is $\sum_j p_j(0) = np_i(0) < 0$.

Assume, z is not the only non-zero signal in \mathbf{s}_{-i} . By Lemma 4.20, let j be the agent with the highest signal in \mathbf{s}_{-i} and for any $s_i < s_{-i}^*$, we have $t_j(\mathbf{s}_{-j}) < s_j \leq \bar{s}$. Thus, we have for any $s_i < s_{-i}^*$, $p_j(0, \mathbf{s}_{-j}) = \min\{0, v_j(t_j(\mathbf{s}_{-j}), \mathbf{s}_{-j}) - v_j^X(t_j(\mathbf{s}_{-j}), \mathbf{s}_{-j})\} < 0$, where the inequality is due to the cursedness-monotonicity condition, since $\mathbf{s}_{-j} \leq \mathbf{s}_{-i}$ and $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) < 0$, as stated above. Therefore, we can set $s_i = 0$, and have $p_j(0, 0, \mathbf{s}_{-ij}) < 0$. By anonymity, we also have $p_i(0, 0, \mathbf{s}_{-ij}) < 0$.

We continue the process iteratively until we reach the signal profile $\mathbf{s}_{-i} = (0, \dots, 0, z)$, we have $p_i(0, \mathbf{s}_{-i}) < 0$. Moreover, by Lemma 4.20, $t_i(0_{-i}) < z$. Continuing with the process for another step, we also get that $p_i(0) < 0$, which implies $t_i(0_{-i}) = 0$ by Lemma 4.19.

Now consider the final signal profile we get, $\mathbf{s} = (s_1 = z, 0, \dots, 0)$, where agent 1 denotes the agent with the smallest non-zero signal z in the original signal profile. By the above argument, agent 1 wins

⁸For two vectors \mathbf{s}, \mathbf{t} , $\mathbf{s} \leq \mathbf{t}$ if \mathbf{s} is coordinate-wise smaller than or equal to \mathbf{t} when we sort the entries in decreasing order.

⁹Recall the support of each signal s_i is denoted as $[0, \bar{s}]$.

the auction (as $t_i(\mathbf{0}_{-i}) = 0$). Moreover, by Corollary 4.11, $p_1(\mathbf{s}) = v_1(t_i(\mathbf{0}_{-i}), \mathbf{0}_{-i}) = v_1(\mathbf{0}) = 0$, while for all other agents $p_i(\mathbf{s}) = p_i(0, \mathbf{s}_{-i}) < 0$, therefore, $\sum_i p_i(\mathbf{s}) < 0$, contradicting ex-post budget-balance. \square

Next, we apply the above characterization to determine the welfare-optimal EPBB mechanism. Then, we show that for some valuation function (e.g., the max function), the welfare-optimal EPBB mechanism attains zero welfare; while for some other valuation functions, including the weighted sum valuations, the welfare-optimal mechanisms are simple and can approximate the full efficiency attained by an EPBB mechanism for fully rational agents (i.e., $\chi = 0$).

4.7.4 Optimal Mechanism

As standard in the interdependent literature, when devising the welfare-optimal mechanism, we assume that the valuations satisfy the single-crossing condition, presented here in the context of symmetric valuation functions.

Definition 4.7 (Single-Crossing for symmetric valuation functions). *Symmetric valuation functions $\{v_i\}_{i \in [n]}$ satisfy the single crossing condition if for any signal profile \mathbf{s} and agents i, j , $s_i \geq s_j$ if and only if $v_i(\mathbf{s}) \geq v_j(\mathbf{s})$.*

Theorem 4.22 implies that for valuations that satisfy the single-crossing condition, the following masked version of the generalized Vickrey auction (GVA) for cursed valuations is the welfare optimal mechanism that satisfies EPBB. GVA assigns the item to the bidder with the highest valuation given all reported signals, and charges the winner the valuation of the item at the minimum winning signal for the winner fixing others' reported signals. With the symmetry settings and single-crossing condition, GVA allocates to the bidder with the highest signal. The masking of GVA is defined as follows.

Definition 4.8 (Masked Generalized Vickrey Auction (M-GVA)). *The masked generalized Vickrey auction considers the threshold rule $t'(\mathbf{s}_{-i})$ which is the result of taking the the threshold rule $t(\mathbf{s}_{-i}) = \max_{j \neq i} s_j$, and masking it to ensure no positive transfers (Definition 4.5). The payments are set using the payment identity (Equation (4.6)), where $p_i(0, \mathbf{s}_{-i})$ is set according to Lemma 4.10.*

Proposition 4.23. *M-GVA is the welfare-optimal mechanism among deterministic, anonymous, C-EPIC-IR, EPIR and EPBB mechanisms, for continuous valuation functions v_i that satisfies cursedness-monotonicity.*

Proof. By Theorem 4.22, we have the socially optimal mechanism must be a masked mechanism. Also, notice that the allocation rule of a GVA mechanism can be written as a threshold allocation rule, with threshold function $t_i^{\text{GVA}}(\mathbf{s}_{-i}) = s_i^*$. Therefore, we only need to prove that the threshold allocation rule $t_i^{\text{M}}(\cdot)$ masking over $t_i^{\text{GVA}}(\cdot)$ maximizes the social welfare over all valid threshold allocation rules $t_i(\cdot)$. To

see this, if we lower the threshold $t_i(\mathbf{s}_{-i})$ from $t_i^M(\mathbf{s}_{-i})$, then we either violate the feasible constraint that $t_i(\mathbf{s}_{-i}) \geq s_{-i}^*$ or violate no positive transfer, i.e., $p_i(0, \mathbf{s}_{-i}) = v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) < 0$. If we increase the threshold $t_i(\mathbf{s}_{-i})$ from $t_i^M(\mathbf{s}_{-i})$, then the social welfare is decreased as $\mathbb{E}_{\mathbf{s}}[SW(\mathbf{s})] = \sum_i \mathbb{E}_{\mathbf{s}_{-i}}[\mathbb{E}_{s_i|\mathbf{s}_{-i}}[SW(\mathbf{s})]] = \sum_i \mathbb{E}_{\mathbf{s}_{-i}}[\int_{t_i(\mathbf{s}_{-i})}^{\bar{s}} v_i(t, \mathbf{s}_{-i}) f(t|\mathbf{s}_{-i}) dt]$.

□

max Function Has Zero Welfare

For the max function, Theorem 4.22 implies that EPBB mechanisms allocate the item with zero probability, because as long as the winner's signal is not \bar{s} , the winner is cursed, i.e., $v_i(\mathbf{s}) < v_i^X(\mathbf{s})$ for winner i and $s_i < \bar{s}$.

Corollary 4.24. *Consider $v_i(\mathbf{s}) = \max_i\{s_i\}$ with signals drawn i.i.d. from $U[0, 1]$ and χ -cursed agents for $\chi > 0$. Then any deterministic, anonymous, C-EPIC-IR, EPIR and EPBB mechanism allocates with 0 probability (and thus has 0 revenue and welfare).*

Approximate Efficiency

In contrast to the max function, many other valuation functions achieve good efficiency guarantees. To demonstrate this, we define a family of valuations functions including well studied functions such as weighted-sums valuations [RT16; Kle98; Wil69; Mye81; Ede+19; Ede+21], and ℓ_p norms for a finite p , for which the M-GVA mechanism approximates the fully efficient mechanism.

Definition 4.9 (Concave-Sum valuations). *Concave-Sum valuations are valuations that can be expressed in the form of $v_i(\mathbf{s}) = l(g(s_i) + \sum_{j \neq i} h(s_j))$, where g, h, l are strictly increasing and bounded functions on the support of s_i , and l is concave.*

Theorem 4.25. *For agents with Concave-Sum valuations, if the valuation function v_i satisfies the single-crossing condition, the M-GVA mechanism has welfare that approaches $\frac{1}{2}$ of the optimal social welfare as the number of agents grows large.*

Proof. Let $h_i = h(s_i)$ for any $i \in [n]$, and $\mathbf{h} = (h_1, \dots, h_n)$. Let $\bar{h} = \frac{\sum_i h_i}{n}$ and $\bar{h}_{-i} = \frac{\sum_{j \neq i} h_j}{n-1}$. Let $\lambda = \mathbb{E}_{s_i}[h_i]$, and let b be the supremum of $h(\cdot)$ on the support of s_i . Note that as $s_i, i \in [n]$ are i.i.d., $h_i, i \in [n]$ are also i.i.d. Let i^* denote the bidder with highest signals among all agents with tie breaking arbitrarily. Note that because of the single-crossing condition, $v_{i^*}(\mathbf{s})$ is also no less than any other value $v_j(\mathbf{s})$ for $j \neq i^*$.

Consider the signal profile \mathbf{s} such that $\bar{h} \geq \lambda + \frac{b}{n}$. We know that

$$\mathbb{E}_{\mathbf{s}}[SW^{\text{M-GVA}}(\mathbf{s})] \geq \mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] \cdot \Pr \left[\bar{h} \geq \lambda + \frac{b}{n} \right].$$

So to prove our theorem, we only need to prove that

$$\lim_{n \rightarrow +\infty} \Pr \left[\bar{h} \geq \lambda + \frac{b}{n} \right] = \frac{1}{2} \text{ and } \mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] \geq \mathbb{E} \left[SW^{\text{OPT}}(\mathbf{s}) \right].$$

First, we prove that $\lim_{n \rightarrow +\infty} \Pr \left[\bar{h} \geq \lambda + \frac{b}{n} \right] = \frac{1}{2}$. To see this, as $h_i, i \in [n]$ are i.i.d., by the central limit theorem, we have when $n \rightarrow +\infty$, $\sqrt{n}(\bar{h} - \lambda) \rightarrow \mathcal{N}(0, \sigma^2)$ for some fixed σ . Thus, we have $\lim_{n \rightarrow +\infty} \Pr \left[\bar{h} \geq \lambda + \frac{b}{n} \right] = \lim_{n \rightarrow +\infty} 1 - \Phi\left(\frac{b}{\sigma\sqrt{n}}\right) = \frac{1}{2}$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Second, we prove that $\mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] \geq \mathbb{E} \left[SW^{\text{OPT}}(\mathbf{s}) \right]$. We also divide the proof into two parts. In the first part, we show that $\mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] = \mathbb{E} \left[v_{i^*}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right]$: note that for any \mathbf{s} such that $\bar{h} \geq \lambda + \frac{b}{n}$, we have $\forall i \in [n]$, $\bar{h}_{-i} \geq \lambda$ (because $\forall i, h_i \leq b$), and thus, we have that $\forall i \in N$,

$$\begin{aligned} v_i(t_i^{\text{M-GVA}}(\mathbf{s}_{-i}), \mathbf{s}_{-i}) &= l \left(g(t_i^{\text{M-GVA}}(\mathbf{s}_{-i})) + \sum_{j \neq i} h_j \right) \\ &\geq l \left(g(t_i^{\text{M-GVA}}(\mathbf{s}_{-i})) + (n-1)\lambda \right) \\ &= l \left(\mathbb{E}_{\tilde{\mathbf{s}}_{-i}} \left[g(t_i^{\text{M-GVA}}(\mathbf{s}_{-i})) + \sum_{j \neq i} h(\tilde{s}_j) \right] \right) \\ &\geq \mathbb{E}_{\tilde{\mathbf{s}}_{-i}} \left[l \left(g(t_i^{\text{M-GVA}}(\mathbf{s}_{-i})) + \sum_{j \neq i} h(\tilde{s}_j) \right) \right] \\ &= \mathbb{E}_{\tilde{\mathbf{s}}_{-i}} [v_i(t_i^{\text{M-GVA}}(\mathbf{s}_{-i}), \tilde{\mathbf{s}}_{-i})]. \end{aligned}$$

The last inequality is due to the concavity of l and the Jensen's inequality. Then, according to the definition of M-GVA mechanism, we have that that if \mathbf{s} satisfies $\bar{h} \geq \lambda + \frac{b}{n}$, the M-GVA will allocate the item to agent i^* and produce social welfare $v_{i^*}(\mathbf{s})$. Thus, we have $\mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] = \mathbb{E} \left[v_{i^*}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right]$.

In the second part, we prove that $\mathbb{E} \left[v_{i^*}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] \geq \mathbb{E} \left[SW^{\text{OPT}}(\mathbf{s}) \right]$. This proof relies on Lemma 4.26 below, whose proof is given in Appendix C.2.

Lemma 4.26. *If \mathbf{z} is a vector of (possibly correlated) random variables, each with the same support $[a, b]$, and $q(\mathbf{z})$ is a non-decreasing function in z_i for any i given any \mathbf{z}_{-i} , then for any i , and a constant d ,*

$$\mathbb{E}_{\mathbf{z}} \left[q(\mathbf{z}) | \sum z_j \geq d \right] \geq \mathbb{E}_{\mathbf{z}_{-i}} \left[\mathbb{E}_{z_i | \mathbf{z}_{-i}} [q(\mathbf{z})] | \sum_{j \neq i} z_j \geq d - b \right].$$

Let $\tilde{v}_i(\mathbf{h}) = l(g(h^{-1}(h_i)) + \sum_{j \neq i} h_j)$. Let $h_{(i)}$ be the order statistics of \mathbf{h} in the order that $h_{(1)} \geq \dots \geq h_{(n)}$. Let $\mathbf{h}_{(i)+} = (h_{(i)}, \dots, h_{(n)})$. Let $\tilde{\xi}_{(i)}(\mathbf{h}_{(i)+}) = \mathbb{E}_{\mathbf{h}_{(1), \dots, (i-1)} | \mathbf{h}_{(i)+}} [l_{(1)}(g(h^{-1}(h_i)) + \sum_{j \neq i} h_j))]$. As $h_i, i \in [n]$

are i.i.d., we have for any $i \in [n - 1]$ and for any $j > i$,

$$F(h_{(i)} | h_{(i+1)}, \dots, h_{(j-1)}, h'_{(j)}, h_{(j+1)}, \dots, h_{(n)})$$

weakly FOSD¹⁰ $F(h_{(j)} | \mathbf{h}_{(j+1)+})$ if $h'_{(j)} > h_{(j)}$. As $\tilde{v}_i(\mathbf{h})$ is non-decreasing with h_i for any i given any $\mathbf{h}_{-(i)}$, by mathematical induction and the first of stochastic dominance, we get that for any $i \in [n]$, $\zeta_{(i)}(\mathbf{h}_{(i)+})$ is non-decreasing in $h_{(j)}$ for any $j \geq i$ given any $\mathbf{h}_{-(j)}$. Therefore, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{s}} \left[v_{i^*}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] &= \mathbb{E}_{\mathbf{h}} \left[\tilde{v}_{(1)}(\mathbf{h}) | \sum_j h_{(j)} \geq n\lambda + b \right] \\ &\geq \mathbb{E}_{\mathbf{h}_{(2)+}} \left[\zeta_{(2)+}(\mathbf{h}_{(2)+}) | \sum_{j=2}^n h_{(j)} \geq n\lambda \right] \\ &\geq \dots \\ &\geq \mathbb{E}_{\mathbf{h}_{(n)+}} \left[\zeta_{(n)+}(\mathbf{h}_{(n)+}) | h_{(n)} \geq n\lambda - (n-1)b \right] \\ &\geq \mathbb{E}_{\mathbf{h}_{(n)+}} \left[\zeta_{(n)+}(\mathbf{h}_{(n)+}) \right] \\ &= \mathbb{E}_{\mathbf{h}}[\tilde{v}_{(1)}(\mathbf{h})] = \mathbb{E}_{\mathbf{s}}[v_{i^*}(\mathbf{s})] = \mathbb{E}_{\mathbf{s}}[SW^{\text{OPT}}(\mathbf{s})]. \end{aligned}$$

The inequalities are because of the monotone-increasing property of $\zeta_{(i)}(\cdot)$ and Lemma 4.26. Combining the two parts, we finally have $\mathbb{E} \left[SW^{\text{M-GVA}}(\mathbf{s}) | \bar{h} \geq \lambda + \frac{b}{n} \right] \geq \mathbb{E} \left[SW^{\text{OPT}}(\mathbf{s}) \right]$, which concludes our entire proof. □

4.8 Conclusion and Future Directions

In this work, we studied the design of mechanisms for agents who suffer from overestimating their value, where the seller tries to avoid agents from suffering from the winner's curse. We designed mechanisms that are deterministic and anonymous, while maximizing the revenue and the welfare of the seller without allowing buyers to have a negative utility. For welfare maximization, we added a requirement that the seller would never have a negative revenue *ex-post*. While we devised optimal mechanisms for these settings, there are many dimensions to the problem, where relaxing any one of these dimensions might lead to a new and interesting design problem. For instance, one can ask the question of what happens if we allow for randomized mechanisms? Mechanisms that can discriminate against bidders? Mechanisms that satisfy the budget-balance constraint only *ex-ante*? Relaxing each of

¹⁰Random variable A weakly FOSD random variable B if for any outcome x , $\Pr[A \geq x] \geq \Pr[B \geq x]$. If A weakly FOSD B , then we have for any non-decreasing function $q(\cdot)$, $\mathbb{E}_A[q(A)] \geq \mathbb{E}_B[q(B)]$.

these dimensions, or a combination, will lead to a new and intricate design problem.

Another interesting question one may ask is how much revenue or welfare exactly do we lose by the fact the agents are biased? Can we relate this loss to the cursedness parameter χ ? How much do we lose by being 'nice' and helping the buyers not lose money, although they are not playing rationally?

We hope our work opens the way for other studies answering these, and other interesting and related questions.

Chapter 5

Can Laypeople Predict the Replicability of Social Science Studies without Expert Intervention: an Exploratory Study

5.1 Introduction

The replicability of scientific studies has played a critical role in the development of science [FW21; Sch16]. Scientific results that fail to replicate will misguide the research advancement, and impair the credibility of the research community. Concerns of the replicability of social science studies have been raised long ago [Ioa05; ID13; MTL14], followed by several large-scale replication projects conducted to systematically examine the replicability of published studies across various fields of social science. Four notable such projects conducted in the recent decade are the Reproducibility Project: Psychology (RPP) [Col15]; the Social Science Replication Project (SSRP) [Cam+18]; the Many Labs 2 Project (ML2) [Kle+18]; and the Experimental Economics Replication Project (EERP) [Cam+16]. They found a low replication rate ranging from 36% (RPP) to 62% (SSRP), which further heated the debate of the replication crisis in social science [Bak16; CM16; Fan18].

Besides these replication projects, efforts have also been made to develop scientific methods to forecast the replication probability of social science studies, aiming to provide a fast and more economical alternative to indicate the reliability of studies. Accompanying with the four replication projects, four forecasting projects were conducted to explore the potential of using the collective intelligence of the

research community to predict the replication probabilities of social science studies [Dre+15; Cam+18; For+19; Cam+16]. In these forecasting projects, hundreds of experts were recruited from relevant research communities to predict the success probabilities of these replication experiments via either surveys or prediction markets. All these projects achieved above-chance prediction accuracy, ranging from 58% to 86%, with an average of 66% for surveys and 73% for prediction markets [Gor+21], showing the effectiveness of this approach. Inspired by the success of this crowd forecasting approach, the Defense Advanced Research Projects Agency (DARPA)'s program "Systematizing Confidence in Open Research and Evidence (SCORE)" further used this approach to generate confidence scores for thousands of social science studies and investigate the replicability differences across fields [Gor+20].

However, expert resources are usually scarce and expensive. The development of online crowdsourcing markets like the Amazon Mechanical Turk platform has enabled us to access laypeople's intelligence much more flexibly, inexpensively, and scalable than accessing expert resources. Using laypeople to make collective forecasts has been proven to be surprisingly accurate in various applications, such as predicting the outcomes of geopolitics, economics, or sports events [Mel+14; GMS14]. This phenomenon is referred to as the *wisdom of crowds*, which has been extensively researched [PSM17], with the earliest research dating back to a hundred years ago [Gal07]. Recently, Hoogeveen et al. [HSW20] have conducted the pioneering work of using laypeople's wisdom to predict the replicability of social science studies. They presented the participants with the materials of 27 selected studies from SSRP and ML2 and asked them to provide a replication prediction for each study. They achieved an accuracy of 59% when presenting the participants with a short description of the study and 67% when additionally presenting with the Bayes factor and its verbal interpretation of the study. However, they still required experts to compose these short descriptions that are comprehensible to laypeople, which may be a potential bottleneck to the scalability of their approach.

In this work, we explored the potential of using laypeople to make predictions about the replicability of social science studies without expert intervention. We investigated to what extent we could elicit useful information when we presented laypeople with raw materials truncated from the published papers of the studies. In particular, we had three objectives: i) evaluating laypeople's engagement in such technical tasks, ii) knowing their perceptions about social science studies from the perspectives of the surprisingness of the findings and the accessibility of the raw materials, and iii) predicting the replicability of social science studies using the solicited information. In the following, we provide a methods section introducing details about our experiment design and data collection process. This is followed by a results section describing our findings and related statistical analysis and a discussion section discussing our results in light of Hoogeveen et al.'s results.

5.2 Methods

5.2.1 Materials

We selected 89 studies out of the 97 published studies investigated in RPP. These 89 studies have both a known replication outcome and an abstract section. We released the surveys about these studies using Human Intelligence Task (HIT) on the Amazon Mechanical Turk (Mturk) platform. Each HIT contained two surveys about two studies uniformly randomly selected from the 89 studies and an additional exit survey about participant's demographic information and user experience.

Presentation of studies and survey questions

In each survey, the reading material and the survey questions were presented using two web pages to reduce participants' cognitive burden. The participants could proceed to the second page only after completing the questions on the first page. The first page presented the title and the abstract of the study and contained four selection and rating questions. Each numerical rating was associated with a brief description of this rating. The four questions are listed below.

- (Q1) Please select a single sentence in the abstract that best describes the main findings/claims of the paper.
- (Q2) In the abstract, how many phrases or terms are NOT familiar to you? Rate between 1 to 4.¹
- (Q3) How much do you understand the main findings/claims of the study after reading the abstract? Rate between 1 to 4.²
- (Q4) Do you find the main findings/claims surprising? Rate between 1 to 4.³

Participants were also asked to enumerate some unfamiliar terms and phrases and to describe the main findings in their own language in text boxes.

The second page presented the section of one of the experiments in the study.⁴ Our goal was to

¹1="0"; 2="1 or 2"; 3="3 or 4"; 4="5 or more".

²1="It is very clear to me what the study has found"; 2="I have some general ideas about what the study has found but still find some places unclear"; 3="I am not so sure about what the study has found, but I can make a rough guess"; 4="I have no idea what the study attempts to achieve".

³1="Completely unsurprising"; 2="Somewhat unsurprising"; 3="Somewhat surprising"; 4="Completely surprising"

⁴RPP ran replication experiments on the last experiment of each study. Therefore, we aimed to present the section of the last experiment of each study. However, in some studies, the methods of the last experiment referred

provide the participants with a taste of how the experiment was designed and conducted to support the study's findings. We excluded the results subsection from the presented material for two reasons. First, we wanted to keep a reasonable length of the reading material for laypeople. Second, the result subsection usually contains many statistical terminologies, which are not accessible to laypeople. The reading material was followed by five selection and rating questions listed below.

- (Q5) Please select a single sentence that best describes what the experiment/study does. Such a sentence usually appears in a paragraph (if exists) before the "Method" section. If you think that there is no such a sentence, please check the box below.
- (Q6) Please select three sentences that best describe the most important steps of the experiment/study.
- (Q7) How many participants were recruited in the experimental study?
- (Q8) If the same type of experiments are re-performed, what probability do you assign that the findings presented in the abstract will be observed? Make your best prediction from 0% (not likely at all) to 100% (with certainty the same finding).
- (Q9) How do you feel that the findings/claims of the paper may hold in other scenarios besides the scenario tested in the above experiment? Rate between 1 to 3.⁵

The materials of the studies were presented in the same format as they were presented in the web version (Html full-text version) on sagepub.com and ebscohost.com. In sentence selection questions, the participant could directly click to select the sentence in the given reading material. Brief instructions and examples were given to help participants understand what type of sentences were good selections for each question.

Design of survey questions

The survey questions Q1 to Q9 were designed to explore three dimensions of information we wanted to collect from participants: participants' engagement in our HITs, participants' perception about the studies, and participants' predictions about the studies' replicability.

to those of previous experiments. In these cases, we presented the participants with the section of the experiment referred to.

⁵1="The claims/findings will likely NOT hold in other scenarios"; 2="The claims/findings may partially hold in some similar scenarios"; 3="The claims/findings will hold to be true in many other similar scenarios".

Participants' engagement: To laypeople, reading social science studies and answering related questions might be a tedious and non-trivial task. Therefore, we designed the sentence selection questions Q1, Q5, Q6, and the factual question Q7 to help us evaluate participants' engagement in our HITs. We compared the responses from different participants in the sentence selections, Q1, Q5, and Q6, to evaluate whether participants provided random answers. Q7 asked about the number of participants in the given study. We compared the responses to the correct answer to evaluate whether participants answered this question authentically. Moreover, Q2 and Q3 are both related to the accessibility of the study, and Q8 and Q9 are related to the replicability of the study. Intuitively, if participants completed the surveys in good faith, the answers to each pair of questions should demonstrate positive correlations to some extent.

Participants' perceptions about the studies: We designed Q2, Q3, and Q4 to evaluate participants' perceptions of the studies from the perspectives of accessibility and surprisingness. We wanted to investigate how these perceptions correlate with participants' replication predictions. Q2 and Q3 both aimed to evaluate the accessibility of the studies to laypeople. While Q2 asked participants to report the number of unfamiliar terms and phrases in the abstract, Q3 asked participants to rate their understanding of the abstract directly. Q4 asked participants to rate the surprisingness of the main claims/findings presented in the abstracts of the studies.

Participants' replication predictions about the studies: Q8 and Q9 were designed to elicit participants' predictions about the replicability of the studies. While Q8 asked concretely about a replication probability, Q9 asked about the generalizability of the main claims/findings.

5.2.2 Procedure and incentive

We conducted the surveys using HITs on the Amazon Mturk platform. Mturk workers could see a preview page of our experiments and then determine whether to take the HITs. The preview page presented the motivations and the purpose of our experiments. It stated that the HIT was about reading social science papers and answering related questions about the accessibility, the plausibility, and the replicability about the studies. The preview page also stated that the estimated completion time of the HIT was 30 minutes, with a fixed \$6 compensation upon completion. After the workers accepted the HITs, we showed an instruction page describing that the HIT consisted of reading materials about two published social science studies and answering about ten questions. Meanwhile, in this page, we also stated that "Your good-faith effort to understand the paper materials and answer the questions is crucial

to our experiment and to the development of modern social sciences. We appreciate your contributions!” Participants could move forward only if they checked the box “I will put forward my good-faith effort in completing the task.” The next page showed the consent of participation. After that, participants were presented with two surveys about two different studies uniformly randomly selected from the 89 RPP studies, followed by an exit survey that ended the whole HIT.

Participants could quit the HITs at any time during their participation, but if they did not complete the entire HITs, they would not receive the fixed \$6 payment. If a participant completed a HIT within 10 minutes, they would be blocked from taking more HITs during the next 12 hours. They would be shown a clear message that they were blocked for 12 hours, but the reason was not given to them.

5.2.3 Participants

Participants were recruited from the Amazon Mturk platform. Each participant could take at most 3 HITs per day and 20 HITs in total. Each completed HIT was paid with a fixed \$6 upon completion. 405 Mturk workers completed at least one entire HIT. The median HITs completed by participants is 2 ($M=5.50$, $SD=6.56$). According to the exit survey, among the 405 participants, 1.48% had a Ph.D. or equivalent degree, 10.62% had a master’s degree or were pursuing a Ph.D. degree, 45.19% had a Bachelor’s degree, and 42.72% were undergraduate students or below. Only 3.21% indicated that they had previously heard about RPP or similar replication projects.

5.3 Results

We received 2229 complete HITs, corresponding to 4458 complete survey responses for individual papers. Each of the 89 papers received either 50 or 51 responses with a mean of 50.09. The median time spent on a single HIT was 29 minutes ($M=35$, $SD=19$). The median of the HIT experience ratings, which ranges from 1 (poor) to 5 (excellent), is 4 ($M=3.84$, $SD=0.92$).

5.3.1 Participants’ Engagement

We analyzed the responses from the questions related to participants’ engagement and observed that the the participants i) correctly answered the factual question Q7, ii) formed consensuses on the sentence selections, and iii) demonstrated anticipated correlation in correlated questions.

In the factual question Q7, 4089 (91.72%) out of 4458 responses correctly answered the number of participants in the given study. The majority was correct on 87 out of 89 studies. These results suggested

that the participants answered this factual question authentically, and laypeople can identify the number of participants given the corresponding experiment materials truncated from the social science papers.

In the sentence selection questions Q1, Q5, and Q6, we observed a salient concentration in the selection, differing from the pattern generated by random selection. Here we presented our detailed analysis of Q1 and Q5. Q1 asked participants to select the main claim sentence in the abstract section, while Q5 asked participants to select a single sentence in the method section that best summarizes the experiment method. In Q1, the average votes received by the top 5 most frequently selected sentences over 89 studies were 26.2 (SD=6.4), 14.6 (SD=5.46), 5.2 (SD=2.48), 2.6 (SD=1.62), 1.2 (SD=0.98) respectively. In comparison, if participants uniformly randomly selected a sentence in the abstract, the average votes of the top 5 most frequently selected answers would be 7.26 (SD=1.29), 7.23 (SD=1.27), 7.10 (SD=1.24), 7.04 (SD=1.23), 6.96 (SD=1.2), significantly differing from the pattern we observed. In Q5, 28.24% of the responses indicated that there was no single sentence summarizing the experiment method. These responses were associated with specific studies. On average, each study received 14.15 (SD=14.42) such responses, while the median was only 5. The top 25% (22) studies received 62.75% of these responses. Meanwhile, for the responses that selected a single sentence, the top 5 most frequently selected sentences in each study received an average of 23 (SD=10.14), 7.6 (SD=4.22), 2.8 (SD=0.98), 2.2 (SD=0.98), and 1.4 (SD=0.6) votes, respectively. In comparison, the most frequently selected sentence would receive only 1.43 (SD=0.51) votes if participants uniformly randomly selected a sentence in Q5.

We further investigated the locations of these selected sentences to examine whether participants tended to select the first or the last sentence regardless of the context. We found that both questions had a considerable number of responses selecting a sentence in the middle, and the distributions of the locations of the selected sentence in the two questions differ significantly. In particular, in Q1, 12.1% of the responses selected the first sentence, 50.98% selected a sentence in the middle, and 36.95% selected the last sentence. In Q5, 45.14% of the responses selected the first sentence, 49.98% selected a sentence in the middle, and only 4.88% selected the last sentence. These results suggested that participants did not select the sentence purely based on the sentence location regardless of the context.

We also observed a moderate negative correlation in Spearman's correlation test ($\rho = -0.50$, $p < 0.0001$) between participants' ratings of the number of unfamiliar terms and phrases in the abstract (Q2) and their ratings of the accessibility of the abstract (Q3). This result followed the intuition that the more unfamiliar terms and phrases in the material, the harder it was to understand the material. We also observed a strong positive correlation in Spearman's correlation test ($\rho = 0.68$, $p < 0.0001$) between participants' ratings of the generalizability of the results of the given study (Q9) and their replication predictions (Q8). These results suggested that the participants' responses were self-contained.

Overall, these results suggested that the participants in our experiments completed the surveys with effort and good faith. They also suggested that laypeople can read published social science papers, understand the materials to some extent, and answer related questions authentically.

5.3.2 Participants' perceptions about studies

We focused on the perceptions of the accessibility (Q3) and the surprisingness (Q4) of the studies. For the accessibility, most responses indicated that they either clearly understood or had a general idea about the main claims/findings in the abstract. In particular, 26.02% of the responses had a rating of 4 in Q3, referring to that the participant thought that they clearly understood the main claims. 40.51% had a rating of 3, indicating that the participant had a general idea about the main claims. 23.62% gave a rating of 2, meaning that the participant could make a rough guess, and the rest, 9.85%, gave a rating of 1, indicating that the participant had no idea about the main claims/findings. Each study's mean accessibility rating was concentrated around 3 (M=2.83, SD=0.54, Median=2.72).

For the surprisingness, most responses found the main claims/findings unsurprising. In particular, 23.53% gave a rating 1 of completely unsurprising, 43.16% gave a rating 2 of somewhat unsurprising, 27.16% gave a rating 3 of somewhat surprising, and the rest, 6.15%, gave a rating 4 of completely surprising. Each study's mean rating of surprisingness was concentrated around rating 2 (M=2.16, SD=0.30, Median=2.18). There existed a weak negative correlation between participants' surprisingness and accessibility ratings in Spearman's correlation test ($\rho = -0.22, p < 0.0001$).

5.3.3 Forecasting replicability

Participants' prediction accuracy

Among the 89 RPP studies, 36 studies replicated successfully, resulting in a replication rate of 40.45%. We used the participants' mean replication prediction from Q8 as the final prediction of the replicability of each study. We observed a salient overestimation of the replication probability in laypeople's predictions. The median of these 89 final predictions was 0.69 with a minimum of 0.58, greater than 0.5 (M=0.69, SD=0.05, Max=0.81). If we threshold final predictions at 0.5, all these final predictions forecasted that the corresponding study could replicate successfully, resulting in an accuracy score of 40.45%, lower than random guesses. At the response level, only 595 out of 4458 (13%) responses gave a replication probability lower than 0.5 with a median of 0.71 (M=0.69, SD=0.19, Min=0.00, Max=1.00).⁶ These results

⁶We provided participants a sliding bar ranging from 0 to 100 (%) with step size 1 (%) to indicate their prediction. The sliding button was initialized at the center position 50. Participants were allowed to submit their answers only

Accessibility	1	2	3	4
# responses	439	1053	1806	1160
Mean replication prediction	0.61 (0.21)	0.65 (0.19)	0.70 (0.17)	0.74 (0.17)
Actual replication rate	0.52 (0.50)	0.44 (0.50)	0.41 (0.49)	0.32 (0.47)
Spearman's ρ	0.02 (p=0.61)	-0.04 (p=0.24)	-0.03 (p=0.25)	-0.02 (p=0.43)

Table 5.1: Statistics of the responses with different accessibility ratings. Spearman's rank correlation coefficient ρ is calculated between participants' replication predictions and actual replication outcomes.

suggested that laypeople tend to believe that a published social science study can replicate successfully. The overestimation of social science studies' replicability has also been found in other studies with participants recruited from either the research community [Gor+20; Gor+21] or the laypeople [HSW20]. To remove the effect of overestimation on the accuracy, we investigated the discriminatory power of the participants' mean predictions and conducted a rank correlation test. However, Spearman's correlation test showed no significant rank correlation ($\rho = -0.18$, $p = 0.0764$) between the mean predictions and the replication outcomes. This result suggested that laypeople's replication predictions collected in our experiments contained very limited signals about whether a social study can replicate or not.

We further investigated whether the participants' predictions and their prediction accuracy were influenced by the accessibility of the studies. Table 5.1 shows the statistics of the responses with different accessibility ratings. We observed that the mean replication prediction increased with the accessibility rating. In fact, there existed a weak positive Spearman's rank correlation between the reported accessibility and the replication prediction at the response level ($\rho = 0.22$, $p < 0.0001$) and a moderate positive correlation between them at the study level ($\rho = 0.51$, $p < 0.0001$). In contrast, the actual replication rate decreased with the reported accessibility. There existed a very weak negative rank correlation between the reported accessibility rating and the actual replication outcome at the response level ($\rho = -0.11$, $p < 0.0001$), and no significant correlation at the study level ($\rho = -0.18$, $p = 0.08$). Meanwhile, at each of the four accessibility levels, there was no discriminatory power found between the participants' replication predictions and the actual replication outcomes, as the Spearman's rank correlation coefficients were all close to zero with a p-value greater than 0.10 (last row, Table 5.1). These results suggested that the participants tended to give a higher replication prediction to the more accessible studies. However, the reported accessibility of the studies had poor discriminatory power (a weak negative correlation) in predicting the replication outcome. This might explain why the participants' replication predictions also had poor predictive power.

if a movement of the sliding button was detected. 2.4% responses predicted exactly 50 (%).

We also observed similar but smaller correlation between the participants' self-rated surprisingness and the replication prediction performance. The participants' self-rated surprisingness had a very weak negative Spearman's rank correlation to their replication predictions at the response level ($\rho = -0.15$, $p < 0.0001$), and a weak negative rank correlation at the study level ($\rho = -0.32$, $p = 0.002$). This suggested that the participants tended to give a lower replication prediction when they found the main claims/findings in the abstract surprising to them. In contrast, the self-rated surprisingness had a very weak positive rank correlation to the actual replication outcome at the response level ($\rho = 0.07$, $p < 0.0001$) and no significant correlation at the study level ($\rho = -0.19$, $p = 0.08$).

5.3.4 Forecasting using machine learning

Machine learning is a technique to learn a pattern from historical data to make predictions on new coming data. Machine learning has been applied to predict the replicability of social science studies in various settings [Alt+19; YYU20; Sal+18]. We investigated whether we can use machine learning to improve the prediction accuracy of people's predictions. We used the responses collected on the 89 RPP studies as our dataset to evaluate the machine learning approach. We divided these 89 studies into a training set and a validation set using the cross-validation method. This setup simulated the situation where we have access to participants' historical prediction data to help us make final predictions.

As we only have a limited 89 samples, we use only three features to predict the replicability of each study: the mean replication prediction, the mean accessibility, and the mean surprisingness received by each study. We use the classic logistic regression as the classifier to avoid over-fitting the data. To investigate the predictive power of each feature, we also evaluated the performance of using every single feature to make forecasts. We focused on two accuracy metrics, the accuracy score and the AUC-ROC [DG06]. The latter is a common accuracy metric used in the machine learning community to evaluate the discriminatory power of predictions. We ran 2000 times 5-fold cross-validation on the 89 studies and collected 10000 accuracy scores and AUC-ROC on both the training and validation sets.

Table 5.2 shows the mean accuracy and the mean AUC-ROC and their confidence intervals on both the training set and the validation set, when the mean replication prediction (Q8), the mean accessibility (Q3), and the mean surprisingness (Q4) and all of them were used as learning features, respectively. All four sets of features showed similar prediction performance in the accuracy score and AUC-ROC. We observed an improvement in the accuracy score (around 0.60) compared to the participants' raw predictions (0.41). This result demonstrated the potential of using machine learning to correct the bias in laypeople's replication prediction data via learning from historical data. However, this improvement

Features	Training		Validation	
	Accuracy	AUC-ROC	Accuracy	AUC-ROC
Replication prediction (Q8)	0.62 [0.59, 0.65]	0.61 [0.59, 0.66]	0.61 [0.5, 0.72]	0.60 [0.43, 0.69]
Accessibility (Q3)	0.60 [0.58, 0.62]	0.61 [0.59, 0.63]	0.60 [0.53, 0.72]	0.61 [0.53, 0.71]
Surprisingness (Q4)	0.62 [0.59, 0.65]	0.61 [0.58, 0.66]	0.62 [0.44, 0.72]	0.60 [0.40, 0.75]
All (Q3, Q4, Q8)	0.64 [0.61, 0.66]	0.65 [0.64, 0.68]	0.63 [0.50, 0.72]	0.60 [0.44, 0.73]

Table 5.2: Mean accuracy scores and AUC-ROCs on the training set and validation set when the mean replication prediction (Q8), mean accessibility (Q3) and mean surprisingness (Q4) and all of them are used as features respectively. Brackets show the 95% confidence intervals of the corresponding values.

was limited, as there was no significant difference in the accuracy score from always predicting that the study could not replicate (which obtained an accuracy score of 0.59 on the 89 RPP studies). Meanwhile, these machine learning predictions achieved an AUC-ROC around 0.6., better than random guesses (AUC-ROC=0.5). This improvement was significant ($p < 0.0001$) when the classifier used the mean accessibility rating as the only feature to predict the replicability. We also observed no improvement in using all features together to predict the replicability compared to using a single feature. This might be because these three features turned out to be correlated with each other, and each feature had limited discriminatory power.

5.4 Discussion

In this work, we explored whether laypeople can predict the replicability of social science studies without expert intervention. We carefully designed surveys and collected responses via releasing HITs on the Amazon Mturk platform. Our experiments revealed several interesting findings.

First, Amazon Mturk workers engaged in our very technical HITs, which involved reading raw material truncated from published social science papers and answering related questions. They devoted considerable time and effort to the HITs and provided reasonable and self-contained answers. This showed the potential of using Amazon Mturk workers to extract information from social science papers that might be difficult to extract via a pure machine approach.

Second, we found that these social science studies in the RPP projects were accessible to laypeople to some extent, as most responses indicated that they either had a general idea about or clearly understood the main findings of the studies. Participants also formed consensus about the main sentences that summarized the abstract and the experimental method and that described the main experimental steps.

Third, we found that laypeople’s replication predictions or perceptions about the studies in our

experiments had limited predictive power in predicting actual replication outcomes. Without expert intervention, laypeople demonstrated a prediction accuracy (40%) lower than chance in our experiments. In contrast, both the researcher forecasters and laypeople with expert intervention achieved above-chance prediction accuracy in similar survey-based experiments. Dreber et al. [Dre+15] reported a prediction accuracy of 58% achieved by researcher forecasters on the same RPP paper set. Hoogeveen et al. [HSW20] reported prediction accuracy of 59% and 67% achieved by laypeople with varied conditions on 27 selected social science papers when experts interpreted the main findings of the studies into more accessible languages.

Although we and Hoogeveen et al. [HSW20] both focused on the prediction performance of laypeople, our experiments had four main differences from theirs, which may explain the prediction performance drop we found.

- **Expert intervention.** For each study, Hoogeveen et al. presented participants with a short description of the research question, the operationalization, and the key finding of the given study and then asked participants about the replication probability. These materials were composed and rephrased by experts to be comprehensible to laypeople. Thus, the participants might be more clear about the main findings of the studies. In fact, 72% of participants indicated that they understood the descriptions of all the 27 studies used in Hoogeveen et al.'s experiments. In contrast, to reduce expert participation, for each study, we presented participants with the raw abstract and one experiment section directly truncated from the published paper of the study. This increased the cognitive burden of laypeople and raised the difficulty of understanding the studies' main findings. In our experiments, only 26% of the responses indicated that they clearly understood the given abstract, and 41% indicated that they had a general idea. This accessibility issue might further introduce vagueness in making their predictions about the replicability of the studies.
- **Study set.** Hoogeveen et al. selected 27 studies from SSRP and ML2 replication projects. These two projects had a higher replication rate overall, 62% for SSRP and 50% for ML2, and the selected studies had a replication rate of 52%. In contrast, we selected 89 studies from RPP due to its larger sample size (97 studies in RPP vs. 21 for SSRP and 24 for ML2). However, the studies in RPP had a much lower replication rate, 37.5% overall and 40% for our 89 studies. This low replication rate creates a disadvantage, as people (both researcher forecasters and laypeople) tend to overestimate the replication rate. Moreover, the replicability of studies in RPP was more difficult to predict than SSRP and ML2. To see this point, the prediction accuracy score of researcher forecasters via

surveys was 0.58 on RPP [Dre+15], compared to 0.86 on SSRP [Cam+18] and 0.67 on ML2 [For+19]. These features of RPP may partially explain why laypeople's replication predictions had a salient overestimation and limited predictive power in our experiments.

- **Participant population.** In Hooegeveen et al.'s experiments, most participants (54%) were first-year students at the University of Amsterdam, 32% were Amazon Mturk workers, and the rest were recruited via social media. In contrast, all of our participants were Amazon Mturk workers. We conjectured that a student admitted by the psychology major of the world's renowned universities might have better skills in reading psychology papers and conducting related reasoning than an average Amazon Mturk worker. This advantage might contribute to more informative replication predictions in Hooegeveen et al.'s experiments.
- **Presentation of material.** Instead of presenting an interpreted description of the studies in Hooegeveen et al.'s experiments, we presented the participants with the raw material truncated from the published papers, which might create an impression to laypeople that these studies were designed and conducted thoughtfully and with rigorous examinations, potentially driving laypeople to make a higher replication prediction. This might also contribute to the salient overestimation of the replication probability.

Given our results, we still think that there is a potential to rely on laypeople to make predictions about the replicability of social science studies, because we did find that the laypeople were willing to devote time and effort to complete our tasks. They did show an understanding of the materials to some extent and provided self-contained answers with good faith. However, some adjustments to the information elicitation procedure might be necessary to increase the chance of success and will be exciting for future research. For example, we can provide participants with more refined material such as the main claim sentences and the main result sentences to reduce participants' cognitive burden and mitigate the vagueness in identifying the main findings. Moreover, some probabilistic training and replication judgment training may be required to help participants carry out necessary reasoning and reduce the overestimate bias. Furthermore, iterative and cooperative elicitation processes can be explored besides using one-shot surveys. For example, we can ask laypeople to articulate the main claims into more comprehensible languages iteratively and then make predictions.

Chapter 6

Conclusion

In this thesis, I have investigated four sub-domains of information elicitation and aggregation: information elicitation with verification, information elicitation without verification, information aggregation without verification, and the human behavior aspect. It is clear that these four sub-domains have varied challenges and require significantly different techniques and methodologies to address these challenges. I summarize the main results of this thesis as follows.

In Chapter 1, I studied the wagering mechanism design in the information elicitation with verification setting. Wagering mechanisms have four desirable properties, individual rationality, incentive compatibility (truthfulness), budget balance, and Pareto optimality. However, it has been shown that these four properties cannot be achieved simultaneously with deterministic mechanisms. To address this obstacle, I extended the mechanism design space into randomized mechanisms and proposed two families of randomized wagering mechanisms that obtain these four properties simultaneously.

In Chapter 2, I investigated probabilistic prediction elicitation problem in the without-verification setting. I considered the scenarios where the principal aims to elicit a set of homogeneous prediction questions. I devised a data-driven approach to recover the strictly proper scoring rules in the without-verification setting, borrowing the techniques from learning with noisy labels. The resulting surrogate scoring rules inherit the capability to reflect participants' prediction accuracy from the strictly proper scoring rules while achieving the strongest truthfulness notion, the dominant uniform strategy truthfulness.

In Chapter 3, I studied the information aggregation without verification problem. It has been shown that identifying expert forecasters using historical data in the with-verification setting and aggregating accordingly can consistently improve the aggregation accuracy. I extended this approach into the

without-verification setting. The obstacle is that without verification, it seems impossible to identify the prediction accuracy of participants. However, I identified an empirical correlation between the rewards of several peer prediction mechanisms and the prediction accuracy of participants. With the help of the rewards, I can still select the potential expert forecasters from the participation population. I further demonstrated on 14 real-world forecasting datasets that we could consistently improve the aggregation accuracy over existing aggregation methods by aggregating over these selected forecasters using peer prediction mechanisms. This result provides a new effective approach to do aggregation without verification.

Chapters 4 and 5 consider the human behavior aspect of information elicitation and aggregation. In Chapter 4, I investigated the interdependent valuation auction design with human bidders who demonstrate the winner's curse behavioral bias. Such a bias not only leads to non-truthful report from the bidders but also leads to a negative utility of the winner, harming the long-term revenue and social welfare of the auction. I provided a complete characterization of all ex-post incentive compatible and individual rational mechanisms for cursed bidders, which also guarantee a non-negative utility for the winner. I further presented the optimal revenue and social welfare mechanisms within this space. The result shows that in order to adapt to the behavioral bias of agents to achieve truthful reporting, the auctioneer has to sacrifice the revenue and social welfare. In Chapter 5, I conducted a human-subject elicitation and aggregation experiment of using laypeople to predict the replicability of social science studies. This experiment aims to investigate whether we can elicit useful information via rich elicitation, eliciting information in addition to a direct prediction, to improve aggregation performance. The results showed that the direct replication predictions of laypeople correlate with the understandability of the study materials of laypeople but do not correlate with the actual replication outcomes.

The broadening application of the wisdom of crowds keeps raising new challenges for information elicitation and aggregation. I hope the results of this thesis could shed some light on following future directions in this extensive area.

Application-oriented design. Most of the existing works in information elicitation have focused on achieving truthfulness in various settings. However, to design a practical elicitation mechanism, it is crucial to consider other desirable properties beyond truthfulness. These properties are usually application-specific. For example, in the betting/wagering scenarios, the Pareto optimality might be a more important property than the truthfulness in practice, as the most widely used wagering mechanism is the pari-mutuel wagering mechanism, a mechanism achieving the Pareto optimality but sacrificing the strict truthfulness [FP18]. In short-term elicitation scenarios such as the tasks using Amazon Mturkers as participants, the simplicity of the elicitation scheme might also be a very important property, because

in these tasks, the participants usually only spend no more than an hour to complete the tasks and are not likely to spend time to understand and reason about a complex elicitation scheme. In our experiments of using laypeople to predict the replicability of social science studies, we provided a simple fixed payment scheme with intrinsic incentive and successfully obtained authentic and self-contained responses from the participants. Other properties to pursue beyond the truthfulness include the richness of the elicited information form, the capability to characterize the information quality (like in the strictly proper scoring rules and the surrogate scoring rules), and the budget efficiency.

On the other hand, the truthfulness property could be relaxed to some extent in practice. As demonstrated in our experiments in Chapter 5, although our elicitation mechanism is not strictly truthful (fixed payment), we still found the participants were engaging in our tasks, providing authentic and self-contained responses. Recent behavioral studies also showed that people have a truthful-telling preference, i.e., people tend to lie only if the utility of lying significantly outweighs the utility of truthful reporting [ANR19; EG12; FF13]. This behavioral evidence suggests that we might relax the strict truthfulness property to approximate truthfulness without harming the elicitation performance in practice too much. The relaxation of the truthfulness property and the pursuit of other application-specific properties have not been extensively explored in the literature, leaving ample space for future research.

Data-driven design. In the era of machine learning, algorithms that dig out useful information from noisy data are emerging. There is a growing literature on applying these data-driven algorithms to design economic mechanisms [Fen21]. The multi-task information elicitation and aggregation is an ideal scenario to apply this data-driven approach. Both the surrogate scoring rules and the forecast aggregation via peer prediction framework in this thesis are examples of using the data-driven approach to solve elicitation and aggregation obstacles in the without-verification setting. The information collected from multiple tasks provides the principal leverages to infer the ground truth and design mechanisms and algorithms accordingly. In information elicitation, the data-driven approach usually provides approximate truthfulness, just as in the surrogate scoring rules, as inference errors are inevitable and hardly align with incentive properties. However, as discussed in the previous paragraph, achieving strict truthfulness may not be very important. It will be fascinating to see what benefits we can obtain using a data-driven approach and how we shall trade-off between the benefits of learning from data and the rigorousness of the truthfulness property in real practices.

Behavior-robust design. Human behavior is a super important but overlooked domain in information elicitation and aggregation. The challenge of investigating human behavior stems from the fact that humans' behavior varies with individual subjects and with subtle changes in the environment. In this

thesis, I took both theoretical and empirical approaches to study human behavior. In the theoretical approach, I designed a truthful mechanism based on an existing theoretical model of the bounded rationality of humans. Such a method can reveal insights into the what-if questions, e.g., what the performance of the mechanisms is if the participants follow certain behavioral models. This method helps us understand the boundary of what can or cannot be achieved by designing mechanisms. However, it is still far from designing a practical mechanism adapted to human behavior because no model can perfectly predict human behaviors, which vary from individual to individual and case to case. It is important to consider that in designing practical mechanisms, we have to face mixed groups of crowds, each having its own behavioral patterns. Therefore, a systematic way of evaluating the robustness of elicitation and aggregation mechanisms and developing behavior-robust mechanisms is urgently needed. Using elicitation mechanisms as an example, there exist several promising directions to approach the robustness. One direction is to design mechanisms such that truthful reporting is always the optimal strategy under different behavior models. An example is the obviously strategy-proof mechanisms proposed by [Li17a]. Such mechanisms have been explored in the other economic settings, such as auctions and matching games, but not in the forecast elicitation problems. Another direction is to acknowledge a parametric behavior model (like in our auction mechanism design in Chapter 4) and conduct analysis and design considering some reasonable worst-case distribution of the parameters. Third, we can focus on simple mechanisms like the fixed payment mechanism. These mechanisms usually have less strategic space and are more robust to the difference between different behavioral patterns. They are also more friendly for conducting experiments to understand humans' behavioral reactions to them and make corresponding adjustments.

References

- [ACR12] D. Allard, A. Comunian, and P. Renard. “Probability aggregation methods in geoscience”. In: *Mathematical Geosciences* 44.5 (2012), pp. 545–581 (cit. on pp. 30, 74, 75).
- [Aga+17] A. Agarwal, D. Mandal, D. C. Parkes, and N. Shah. “Peer prediction with heterogeneous users”. In: *ACM EC*. ACM. 2017, pp. 81–98 (cit. on p. 76).
- [AK97] C. Avery and J. H. Kagel. “Second-Price Auctions with Asymmetric Payoffs: An Experimental Investigation”. In: *Journal of Economics & Management Strategy* 6.3 (1997), pp. 573–603 (cit. on pp. 96, 99, 100, 182).
- [AL88] D. Angluin and P. Laird. “Learning from noisy examples”. In: *Machine Learning* 2.4 (1988), pp. 343–370 (cit. on p. 36).
- [Alt+19] A. Altmejd, A. Dreber, E. Forsell, J. Huber, T. Imai, M. Johannesson, M. Kirchler, G. Nave, and C. Camerer. “Predicting the replicability of social science lab experiments”. In: *PloS one* 14.12 (2019) (cit. on pp. 1, 35, 136).
- [ANR19] J. Abeler, D. Nosenzo, and C. Raymond. “Preferences for truth-telling”. In: *Econometrica* 87.4 (2019), pp. 1115–1153 (cit. on p. 142).
- [Arm01] J. S. Armstrong. “Combining forecasts”. In: *Principles of forecasting*. Springer, 2001, pp. 417–439 (cit. on p. 6).
- [Arr+08] K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, et al. *The promise of prediction markets*. 2008 (cit. on p. 3).

- [Asp10] W. Aspinnall. “A route to more tractable expert advice”. In: *Nature* 463.7279 (2010), pp. 294–295 (cit. on pp. 4, 75).
- [AT21] A. Amer and I. Talgam-Cohen. “Auctions with Interdependence and SOS: Improved Approximation”. In: *Algorithmic Game Theory - 14th International Symposium, SAGT 2021, Aarhus, Denmark, September 21-24, 2021, Proceedings*. Ed. by I. Caragiannis and K. A. Hansen. Vol. 12885. Lecture Notes in Computer Science. Springer, 2021, pp. 34–48. DOI: [10.1007/978-3-030-85947-3_3](https://doi.org/10.1007/978-3-030-85947-3_3). URL: https://doi.org/10.1007/978-3-030-85947-3%5C_3 (cit. on p. 101).
- [Ata+16] P. Atanasov, P. Rescober, E. Stone, S. A. Swift, E. Servan-Schreiber, P. Tetlock, L. Ungar, and B. Mellers. “Distilling the wisdom of crowds: Prediction markets vs. prediction polls”. In: *Management science* 63.3 (2016), pp. 691–706 (cit. on pp. 63, 73, 77, 176).
- [Aus99] L. M. Ausubel. “A Generalized Vickrey Auction”. In: (1999) (cit. on pp. 97, 102, 119).
- [Bak16] M. Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604 (2016) (cit. on p. 127).
- [Bar+14] J. Baron, B. A. Mellers, P. E. Tetlock, E. Stone, and L. H. Ungar. “Two reasons to make aggregated probability forecasts more extreme”. In: *Decision Analysis* 11.2 (2014), pp. 133–145 (cit. on pp. 6, 74, 75).
- [BBA15] J. P. Bigham, M. S. Bernstein, and E. Adar. “Human-computer interaction and collective intelligence”. In: *Handbook of collective intelligence* 57 (2015) (cit. on p. 94).
- [BBM20] D. Bergemann, B. Brooks, and S. Morris. “Countering the winner’s curse: optimal auction design in a common value model”. In: *Theoretical Economics* 15.4 (2020), pp. 1399–1434 (cit. on pp. 96, 100, 121).
- [BC15] D. V. Budescu and E. Chen. “Identifying expertise to extract the wisdom of crowds”. In: *Management Science* 61.2 (2015), pp. 267–280 (cit. on pp. 4, 75).

- [BC96] B. Blecherman and C. F. Camerer. “Is there a winner’s curse in the market for baseball players? Evidence from the field”. In: (1996) (cit. on p. 96).
- [BDL19] B. D. Bernheim, S. DellaVigna, and D. Laibson. *Handbook of Behavioral Economics-Foundations and Applications 2*. Elsevier, 2019 (cit. on pp. 7, 100).
- [BDO18] M. Babaioff, S. Dobzinski, and S. Oren. “Combinatorial Auctions with Endowment Effect”. In: *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*. Ed. by É. Tardos, E. Elkind, and R. Vohra. ACM, 2018, pp. 73–90. DOI: [10.1145/3219166.3219197](https://doi.org/10.1145/3219166.3219197). URL: <https://doi.org/10.1145/3219166.3219197> (cit. on p. 101).
- [BK02] J. Bulow and P. Klemperer. “Prices and the Winner’s Curse”. In: *RAND journal of Economics* (2002), pp. 1–21 (cit. on pp. 100, 121).
- [Bri50a] G. W. Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1 (1950), pp. 1–3 (cit. on pp. 35, 37).
- [Bri50b] G. W. Brier. “Verification of Forecasts Expressed in Terms of Probability”. In: *Monthly Weather Review* 78.1 (1950), pp. 1–3 (cit. on p. 11).
- [BS83] M. H. Bazerman and W. F. Samuelson. “I won the auction but don’t want the prize”. In: *Journal of conflict resolution* 27.4 (1983), pp. 618–634 (cit. on pp. 99, 100).
- [Byl94] T. Bylander. “Learning linear threshold functions in the presence of classification noise”. In: *Proceedings of the seventh annual conference on Computational learning theory*. ACM, 1994, pp. 340–347 (cit. on pp. 12, 36, 48).
- [Cam+16] C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. “Evaluating replicability of laboratory experiments in economics”. In: *Science* 351.6280 (2016), pp. 1433–1436 (cit. on pp. 127, 128).
- [Cam+18] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. “Evaluating the replicability

of social science experiments in Nature and Science between 2010 and 2015". In: *Nature Human Behaviour* 2.9 (2018), p. 637 (cit. on pp. 127, 128, 139).

- [CCC+71] E. C. Capen, R. V. Clapp, W. M. Campbell, et al. "Competitive bidding in high-risk situations". In: *Journal of petroleum technology* 23.06 (1971), pp. 641–653 (cit. on pp. 96, 99).
- [CD80] J. Cassing and R. W. Douglas. "Implications of the auction mechanism in baseball's free agent draft". In: *Southern Economic Journal* (1980), pp. 110–121 (cit. on pp. 96, 99).
- [CFK14] S. Chawla, H. Fu, and A. R. Karlin. "Approximate revenue maximization in interdependent value settings". In: *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*. Ed. by M. Babaioff, V. Conitzer, and D. A. Easley. ACM, 2014, pp. 277–294. DOI: [10.1145/2600057.2602858](https://doi.org/10.1145/2600057.2602858). URL: <https://doi.org/10.1145/2600057.2602858> (cit. on p. 101).
- [CFM19] L. B. Canonico, C. Flathmann, and N. McNeese. "Collectively intelligent teams: Integrating team cognition, collective intelligence, and AI for future Teaming". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 63. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, pp. 1466–1470 (cit. on p. 94).
- [Cha+18] S. Chawla, K. Goldner, J. B. Miller, and E. Pountourakis. "Revenue Maximization with an Uncertainty-Averse Buyer". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*. Ed. by A. Czumaj. SIAM, 2018, pp. 2050–2068. DOI: [10.1137/1.9781611975031.134](https://doi.org/10.1137/1.9781611975031.134). URL: <https://doi.org/10.1137/1.9781611975031.134> (cit. on p. 101).
- [Che+11] Y. Chen, I. Kash, M. Ruberry, and V. Shnayder. "Decision markets with good incentives". In: *International Workshop on Internet and Network Economics*. Springer. 2011, pp. 72–83 (cit. on p. 3).

- [Che+14] Y. Chen, N. R. Devanur, D. M. Pennock, and J. W. Vaughan. “Removing arbitrage from wagering mechanisms”. In: *Proceedings of the 15th ACM EC*. ACM. 2014, pp. 377–394 (cit. on pp. 3, 9, 11, 15, 17, 27).
- [CL09] G. Charness and D. Levin. “The origin of the winner’s curse: a laboratory study”. In: *American Economic Journal: Microeconomics* 1.1 (2009), pp. 207–36 (cit. on pp. 99, 100).
- [Cla+16] G. Claeskens, J. R. Magnus, A. L. Vasnev, and W. Wang. “The forecast combination puzzle: A simple theoretical explanation”. In: *International Journal of Forecasting* 32.3 (2016), pp. 754–762 (cit. on p. 4).
- [Cle89] R. T. Clemen. “Combining forecasts: A review and annotated bibliography”. In: *International journal of forecasting* 5.4 (1989), pp. 559–583 (cit. on p. 75).
- [CM16] G. Christensen and E. Miguel. “Transparency, reproducibility, and the credibility of economics research. Pt NBER WP 22989. Forthcoming in the”. In: *Journal of Economic Literature* (2016) (cit. on p. 127).
- [Col15] O. S. Collaboration. “Estimating the reproducibility of psychological science”. In: *Science* 349.6251 (2015), aac4716 (cit. on p. 127).
- [CP10] Y. Chen and D. M. Pennock. “Designing markets for prediction”. In: *AI Magazine* 31.4 (2010), pp. 42–52 (cit. on p. 3).
- [CPW16] R. Cummings, D. M. Pennock, and J. Wortman Vaughan. “The possibilities and limitations of private prediction markets”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM. 2016, pp. 143–160 (cit. on pp. 11, 12, 17, 18).
- [CW86] R. T. Clemen and R. L. Winkler. “Combining economic forecasts”. In: *Journal of Business & Economic Statistics* 4.1 (1986), pp. 39–46 (cit. on pp. 1, 4, 75).
- [Des+82] J. P. Dessauer, M. Dunbar, D. M. Brownstone, and I. Franck. *Book Publishing*. 1982 (cit. on p. 96).
- [DG06] J. Davis and M. Goadrich. “The relationship between Precision-Recall and ROC curves”. In: *ICML*. ACM. 2006, pp. 233–240 (cit. on p. 136).

- [DG13] A. Dasgupta and A. Ghosh. “Crowdsourced judgement elicitation with endogenous proficiency”. In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 319–330 (cit. on pp. 5, 38, 42, 72).
- [DP97] D. Duffie and J. Pan. “An overview of value at risk”. In: *Journal of derivatives* 4.3 (1997), pp. 7–49 (cit. on p. 35).
- [Dre+15] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. “Using prediction markets to estimate the reproducibility of scientific research”. In: *Proceedings of the National Academy of Sciences* 112.50 (2015), pp. 15343–15347 (cit. on pp. 128, 138, 139).
- [Ede+18] A. Eden, M. Feldman, A. Fiat, and K. Goldner. “Interdependent Values without Single-Crossing”. In: *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*. Ed. by É. Tardos, E. Elkind, and R. Vohra. ACM, 2018, p. 369. DOI: [10.1145/3219166.3219173](https://doi.org/10.1145/3219166.3219173). URL: <https://doi.org/10.1145/3219166.3219173> (cit. on p. 101).
- [Ede+19] A. Eden, M. Feldman, A. Fiat, K. Goldner, and A. R. Karlin. “Combinatorial Auctions with Interdependent Valuations: SOS to the Rescue”. In: *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*. Ed. by A. Karlin, N. Immorlica, and R. Johari. ACM, 2019, pp. 19–20. DOI: [10.1145/3328526.3329759](https://doi.org/10.1145/3328526.3329759). URL: <https://doi.org/10.1145/3328526.3329759> (cit. on pp. 101, 121, 123).
- [Ede+21] A. Eden, M. Feldman, I. Talgam-Cohen, and O. Zviran. “PoA of Simple Auctions with Interdependent Values”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 5321–5329. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/16671> (cit. on pp. 101, 121, 123).

- [EFF20] T. Ezra, M. Feldman, and O. Friedler. “A General Framework for Endowment Effects in Combinatorial Markets”. In: *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*. Ed. by P. Biró, J. D. Hartline, M. Ostrovsky, and A. D. Procaccia. ACM, 2020, pp. 499–500. doi: [10.1145/3391403.3399516](https://doi.org/10.1145/3391403.3399516). URL: <https://doi.org/10.1145/3391403.3399516> (cit. on p. 101).
- [EG12] S. Erat and U. Gneezy. “White lies”. In: *Management Science* 58.4 (2012), pp. 723–733 (cit. on p. 142).
- [EGZ22] A. Eden, K. Goldner, and S. Zheng. “Private Interdependent Valuations”. In: *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, to appear*. 2022. URL: <https://arxiv.org/abs/2111.01851> (cit. on p. 101).
- [EP17] A. Ellis and M. Piccione. “Correlation misperception in choice”. In: *American Economic Review* 107.4 (2017), pp. 1264–92 (cit. on p. 100).
- [ER05] E. Eyster and M. Rabin. “Cursed equilibrium”. In: *Econometrica* 73.5 (2005), pp. 1623–1672 (cit. on pp. 96–98, 100, 102–105, 108, 182).
- [ERV19] E. Eyster, M. Rabin, and D. Vayanos. “Financial markets where traders neglect the informational content of prices”. In: *The Journal of Finance* 74.1 (2019), pp. 371–399 (cit. on p. 100).
- [EW11] E. Eyster and G. Weizsäcker. *Correlation neglect in financial decision-making*. Tech. rep. DIW Discussion Papers, 2011 (cit. on p. 100).
- [Fan18] D. Fanelli. “Opinion: Is science really facing a reproducibility crisis, and do we need it to?” In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2628–2631 (cit. on p. 127).
- [Fen21] Z. Feng. “Machine Learning-Aided Economic Design”. PhD thesis. Harvard University, 2021 (cit. on p. 142).

- [FF13] U. Fischbacher and F. Föllmi-Heusi. “Lies in disguise—an experimental study on cheating”. In: *Journal of the European Economic Association* 11.3 (2013), pp. 525–547 (cit. on p. 142).
- [FK19] A. Frankel and E. Kamenica. “Quantifying information and uncertainty”. In: *American Economic Review* 109.10 (2019), pp. 3650–80 (cit. on p. 39).
- [For+19] E. Forsell, D. Viganola, T. Pfeiffer, J. Almenberg, B. Wilson, Y. Chen, B. A. Nosek, M. Johannesson, and A. Dreber. “Predicting replication outcomes in the Many Labs 2 study”. In: *Journal of Economic Psychology* 75 (2019), p. 102117 (cit. on pp. 128, 139).
- [FP18] R. Freeman and D. M. Pennock. “An Axiomatic View of the Parimutuel Consensus Wagering Mechanism”. In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2018, pp. 1936–1938 (cit. on pp. 9–11, 17, 141).
- [FPW17] R. Freeman, D. M. Pennock, and J. Wortman Vaughan. “The double clinching auction for wagering”. In: *Proceedings of the 18th ACM EC*. ACM. 2017, pp. 43–60 (cit. on pp. 3, 9–11, 15, 17, 23, 30).
- [Fri+18] J. A. Friedman, J. D. Baker, B. A. Mellers, P. E. Tetlock, and R. Zeckhauser. “The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament”. In: *International Studies Quarterly* 62.2 (2018), pp. 410–422 (cit. on pp. 1, 35).
- [FV14] B. Frénay and M. Verleysen. “Classification in the presence of label noise: a survey”. In: *IEEE transactions on neural networks and learning systems* 25.5 (2014), pp. 845–869 (cit. on p. 39).
- [FW16] R. Frongillo and J. Witkowski. “A geometric method to construct minimal peer prediction mechanisms”. In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016 (cit. on p. 38).

- [FW21] F. Fidler and J. Wilcox. “Reproducibility of Scientific Results”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021 (cit. on p. 127).
- [Gal07] F. Galton. *Vox populi*. 1907 (cit. on pp. 1, 75, 128).
- [Gao+14] X. A. Gao, A. Mao, Y. Chen, and R. P. Adams. “Trick or treat: putting peer prediction to the test”. In: *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM. 2014, pp. 507–524 (cit. on p. 7).
- [GF19a] N. Goel and B. Faltings. *Deep Bayesian Trust : A Dominant and Fair Incentive Mechanism for Crowd*. 2019 (cit. on p. 76).
- [GF19b] N. Goel and B. Faltings. “Deep bayesian trust: A dominant and fair incentive mechanism for crowd”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 1996–2003 (cit. on pp. 5, 36, 39, 45).
- [GJP16] G. J. P. GJP. *GJP Data*. Version V1. 2016. DOI: [10 . 7910 / DVN / BPCDH5](https://doi.org/10.7910/DVN/BPCDH5). URL: <https://doi.org/10.7910/DVN/BPCDH5> (cit. on pp. 83, 176).
- [Gka+21] V. Gkatzelis, R. Patel, E. Pountourakis, and D. Schoepflin. “Prior-Free Clock Auctions for Bidders with Interdependent Values”. In: *Algorithmic Game Theory - 14th International Symposium, SAGT 2021, Aarhus, Denmark, September 21-24, 2021, Proceedings*. Ed. by I. Caragiannis and K. A. Hansen. Vol. 12885. Lecture Notes in Computer Science. Springer, 2021, pp. 64–78. DOI: [10 . 1007 / 978 - 3 - 030 - 85947 - 3 _ 5](https://doi.org/10.1007/978-3-030-85947-3_5). URL: https://doi.org/10.1007/978-3-030-85947-3%5C_5 (cit. on p. 101).
- [GMS14] D. G. Goldstein, R. P. McAfee, and S. Suri. “The wisdom of smaller, smarter crowds”. In: *ACM EC*. ACM. 2014, pp. 471–488 (cit. on pp. 75, 87, 128).
- [Gor+20] M. Gordon, D. Viganola, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, C. Twardy, et al. “Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme”. In: *Royal Society open science* (2020) (cit. on pp. 128, 135).

- [Gor+21] M. Gordon, D. Viganola, A. Dreber, M. Johannesson, and T. Pfeiffer. “Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects”. In: *Plos one* 16.4 (2021), e0248780 (cit. on pp. 128, 135).
- [GR05] T. Gneiting and A. E. Raftery. “Weather forecasting with ensemble methods”. In: *Science* 310.5746 (2005), pp. 248–249 (cit. on p. 35).
- [GR07a] T. Gneiting and A. E. Raftery. “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378 (cit. on pp. 11, 13, 77, 80).
- [GR07b] T. Gneiting and A. E. Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378 (cit. on pp. 3, 35, 38, 40).
- [Gun92] S. I. Gunter. “Nonnegativity restricted least squares combinations”. In: *International Journal of Forecasting* 8.1 (1992), pp. 45–59 (cit. on p. 4).
- [HP88] K. Hendricks and R. H. Porter. “An empirical study of an auction with asymmetric information”. In: *The American Economic Review* (1988), pp. 865–883 (cit. on pp. 97, 99, 100).
- [HPB87] K. Hendricks, R. H. Porter, and B. Boudreau. “Information, returns, and bidding behavior in OCS auctions: 1954-1969”. In: *The Journal of Industrial Economics* (1987), pp. 517–542 (cit. on pp. 99, 100).
- [HRS16] A. Hassidim, A. Romm, and R. I. Shorrer. ““ Strategic” Behavior in a Strategy-Proof Environment”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. 2016, pp. 763–764 (cit. on p. 7).
- [HS94] C. A. Holt and R. Sherman. “The loser’s curse”. In: *The American Economic Review* 84.3 (1994), pp. 642–652 (cit. on p. 96).
- [HSW19] S. Hoogeveen, A. Sarafoglou, and E.-J. Wagenmakers. “Laypeople Can Predict Which Social Science Studies Replicate”. In: (2019) (cit. on pp. 1, 8, 35).

- [HSW20] S. Hoogeveen, A. Sarafoglou, and E.-J. Wagenmakers. “Laypeople can predict which social-science studies will be replicated successfully”. In: *Advances in Methods and Practices in Psychological Science* 3.3 (2020), pp. 267–285 (cit. on pp. 128, 135, 138).
- [IAR19] IARPA. *Hybrid Forecasting Competition*. <https://www.iarpa.gov/index.php/research-programs/hfc?id=661>. 2019 (cit. on pp. 64, 83, 176).
- [ID13] J. Ioannidis and C. Doucouliagos. “What’s to know about the credibility of empirical economics?” In: *Journal of Economic Surveys* 27.5 (2013), pp. 997–1004 (cit. on p. 127).
- [ILN10] A. Ivanov, D. Levin, and M. Niederle. “Can relaxation of beliefs rationalize the winner’s curse?: an experimental study”. In: *Econometrica* 78.4 (2010), pp. 1435–1452 (cit. on p. 100).
- [Ioa05] J. P. Ioannidis. “Why most published research findings are false”. In: *PLoS medicine* 2.8 (2005), e124 (cit. on p. 127).
- [JF07] R. Jurca and B. Faltings. “Collusion-resistant, incentive-compatible feedback payments”. In: *Proceedings of the 8th ACM conference on Electronic commerce*. ACM. 2007, pp. 200–209 (cit. on p. 38).
- [JF09] R. Jurca and B. Faltings. “Mechanisms for making crowds truthful”. In: *Journal of Artificial Intelligence Research* 34 (2009), pp. 209–253 (cit. on p. 38).
- [JNW06] V. R. Jose, R. F. Nau, and R. L. Winkler. “Scoring Rules, Generalized Entropy and utility maximization”. Working Paper, Fuqua School of Business, Duke University. 2006 (cit. on pp. 3, 11, 35, 38).
- [Joh07] D. J. Johnstone. “The Parimutuel Kelly Probability Scoring Rule”. In: *Decision Analysis* 4.2 (June 2007), pp. 66–75. ISSN: 1545-8490. DOI: [10.1287/deca.1070.0091](https://doi.org/10.1287/deca.1070.0091). URL: <http://dx.doi.org/10.1287/deca.1070.0091> (cit. on p. 11).

- [JW08] V. R. R. Jose and R. L. Winkler. “Simple robust averages of forecasts: Some empirical results”. In: *International journal of forecasting* 24.1 (2008), pp. 163–169 (cit. on pp. 6, 74, 75, 77, 84, 93).
- [KG04] D. M. Kilgour and Y. Gerchak. “Elicitation of Probabilities Using Competitive Scoring Rules”. In: *Decision Analysis* 1.2 (2004), pp. 108–113. ISSN: 1545-8490. DOI: <http://dx.doi.org/10.1287/deca.1030.0003> (cit. on p. 11).
- [KK15] P. Kondor and B. Köszegi. *Cursed financial innovation*. Tech. rep. WZB Discussion Paper, 2015 (cit. on p. 100).
- [KL02] J. H. Kagel and D. Levin. “Bidding in common-value auctions: A survey of experimental research”. In: *Common value auctions and the winner’s curse* 1 (2002), pp. 1–84 (cit. on p. 100).
- [KL09] J. H. Kagel and D. Levin. *Common value auctions and the winner’s curse*. Princeton University Press, 2009 (cit. on p. 96).
- [KL86] J. H. Kagel and D. Levin. “The winner’s curse and public information in common value auctions”. In: *The American economic review* (1986), pp. 894–920 (cit. on pp. 96, 99, 100).
- [Kle+18] R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, et al. “Many Labs 2: Investigating variation in replicability across samples and settings”. In: *Advances in Methods and Practices in Psychological Science* 1.4 (2018), pp. 443–490 (cit. on p. 127).
- [Kle98] P. Klemperer. “Auctions with almost common values: The Wallet Game and its applications”. In: *European Economic Review* 42.3-5 (1998), pp. 757–769 (cit. on pp. 95, 96, 100, 123).
- [KLS16] Y. Kong, K. Ligett, and G. Schoenebeck. “Putting peer prediction under the micro (economic) scope and making truth-telling focal”. In: *International Conference on Web and Internet Economics*. Springer. 2016, pp. 251–264 (cit. on pp. 5, 38, 76).

- [Kon+20] Y. Kong, G. Schoenebeck, B. Tao, and F.-Y. Yu. “Information elicitation mechanisms for statistical estimation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02. 2020, pp. 2095–2102 (cit. on pp. 5, 39).
- [Kon20] Y. Kong. “Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks”. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2020, pp. 2398–2411 (cit. on pp. 5, 39, 42, 45, 57, 66, 72, 74, 81–83, 91, 177).
- [KS18] Y. Kong and G. Schoenebeck. “Water from two rocks: Maximizing the mutual information”. In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 177–194 (cit. on pp. 5, 39, 42).
- [KS19] Y. Kong and G. Schoenebeck. “An information theoretic framework for designing information elicitation mechanisms that reward truth-telling”. In: *ACM Transactions on Economics and Computation (TEAC)* 7.1 (2019), p. 2 (cit. on pp. 5, 38, 42, 45, 57).
- [Kur+19] R. H. Kurvers, S. M. Herzog, R. Hertwig, J. Krause, M. Moussaid, G. Argenziano, I. Zalaudek, P. A. Carney, and M. Wolf. “How to detect high-performing individuals and groups: Decision similarity predicts accuracy”. In: *Science advances* 5.11 (2019), eaaw9011 (cit. on p. 76).
- [Lam+08] N. S. Lambert, J. Langford, J. Wortman, Y. Chen, D. Reeves, Y. Shoham, et al. “Self-financed wagering mechanisms for forecasting”. In: *Proceedings of the 9th ACM EC*. ACM. 2008, pp. 170–179 (cit. on pp. 3, 9, 11–13, 15, 17, 18, 21, 25).
- [Lam+15] N. S. Lambert, J. Langford, J. W. Vaughan, Y. Chen, D. M. Reeves, Y. Shoham, and D. M. Pennock. “An axiomatic characterization of wagering mechanisms”. In: *Journal of Economic Theory* 156 (2015), pp. 389–416 (cit. on pp. 3, 9, 11, 12).
- [LBD16] M. Lang, N. Bharadwaj, and C. A. Di Benedetto. “How crowdsourcing improves prediction of market-oriented outcomes”. In: *Journal of Business Research* 69.10 (2016), pp. 4168–4176 (cit. on p. 1).

- [LC18] Y. Liu and Y. Chen. “Surrogate Scoring Rules and a Dominant Truth Serum for Information Elicitation”. In: *CoRR* abs/1802.09158 (2018). arXiv: [1802.09158](https://arxiv.org/abs/1802.09158). URL: <http://arxiv.org/abs/1802.09158> (cit. on pp. 12, 20).
- [LD+83] J. Lohrenz, E. Dougherty, et al. “Bonus bidding and bottom lines: federal offshore oil and gas”. In: *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers. 1983 (cit. on p. 96).
- [LD14] M. D. Lee and I. Danileiko. “Using cognitive models to combine probability estimates”. In: *Judgment and Decision Making* 9.3 (2014), p. 259 (cit. on pp. 6, 75, 85, 174, 181).
- [Li17a] S. Li. “Obviously strategy-proof mechanisms”. In: *American Economic Review* 107.11 (2017), pp. 3257–87 (cit. on p. 143).
- [Li17b] Y. Li. “Approximation in mechanism design with interdependent values”. In: *Games and Economic Behavior* 103 (2017), pp. 225–253 (cit. on p. 101).
- [Liu+20] Y. Liu, M. Gordon, J. Wang, M. Bishop, Y. Chen, T. Pfeiffer, C. Twardy, and D. Viganola. “Replication Markets: Results, Lessons, Challenges and Opportunities in AI Replication”. In: *arXiv preprint arXiv:2005.04543* (2020) (cit. on p. 73).
- [LL15] Y. Liu and M. Liu. “An Online Learning Approach to Improving the Quality of Crowd-Sourcing”. In: *Proceedings of the 2015 ACM SIGMETRICS*. Portland, Oregon, USA: ACM, 2015, pp. 217–230. ISBN: 978-1-4503-3486-0 (cit. on p. 58).
- [LMP19] S. Liu, J. B. Miller, and A. Psomas. “Risk Robust Mechanism Design for a Prospect Theoretic Buyer”. In: *Algorithmic Game Theory - 12th International Symposium, SAGT 2019, Athens, Greece, September 30 - October 3, 2019, Proceedings*. Ed. by D. Fotakis and E. Markakis. Vol. 11801. Lecture Notes in Computer Science. Springer, 2019, pp. 95–108. DOI: [10.1007/978-3-030-30473-7_7](https://doi.org/10.1007/978-3-030-30473-7_7). URL: https://doi.org/10.1007/978-3-030-30473-7_7 (cit. on p. 101).
- [LMW12] C. Lin, M. Mausam, and D. Weld. “Dynamically Switching between Synergistic Workflows for Crowdsourcing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1. 2012 (cit. on p. 94).

- [LPI12] Q. Liu, J. Peng, and A. T. Ihler. “Variational inference for crowdsourcing”. In: *Advances in neural information processing systems*. 2012, pp. 692–700 (cit. on pp. 6, 73–75, 179, 180).
- [LWC20] Y. Liu, J. Wang, and Y. Chen. “Surrogate scoring rules”. In: *Proceedings of the 21st ACM Conference on Economics and Computation*. 2020, pp. 853–871 (cit. on pp. 74, 76, 80, 91).
- [McC56] J. McCarthy. “Measures of the Value of Information”. In: *PNAS: Proceedings of the National Academy of Sciences of the United States of America* 42.9 (1956), pp. 654–655 (cit. on p. 40).
- [Mel+14] B. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. A. Swift, et al. “Psychological strategies for winning a geopolitical forecasting tournament”. In: *Psychological science* 25.5 (2014), pp. 1106–1115 (cit. on p. 128).
- [Mel+15] B. Mellers, E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, E. Chen, J. Baker, Y. Hou, M. Horowitz, et al. “Identifying and cultivating superforecasters as a method of improving probabilistic predictions”. In: *Perspectives on Psychological Science* 10.3 (2015), pp. 267–281 (cit. on pp. 87, 176).
- [Men+15] A. Menon, B. Van Rooyen, C. S. Ong, and B. Williamson. “Learning from corrupted binary labels via class-probability estimation”. In: *International Conference on Machine Learning*. 2015, pp. 125–134 (cit. on p. 48).
- [Mie09] T. Miettinen. “The partially cursed and the analogy-based expectation equilibrium”. In: *Economics Letters* 105.2 (2009), pp. 162–164 (cit. on p. 100).
- [MLS12] A. E. Mannes, R. P. Larrick, and J. B. Soll. “The social psychology of the wisdom of crowds.” In: (2012) (cit. on pp. 6, 74, 75).
- [MM87] R. P. McAfee and J. McMillan. “Auctions and bidding”. In: *Journal of economic literature* 25.2 (1987), pp. 699–738 (cit. on p. 99).

- [MP17] J. McCoy and D. Prelec. “A statistical model for aggregating judgments by incorporating peer predictions”. In: *arXiv preprint arXiv:1703.04778* (2017) (cit. on pp. 6, 75, 85, 90, 174).
- [MRZ05a] N. Miller, P. Resnick, and R. Zeckhauser. “Eliciting informative feedback: The peer-prediction method”. In: *Management Science* 51.9 (2005), pp. 1359–1373 (cit. on pp. 74, 76).
- [MRZ05b] N. Miller, P. Resnick, and R. Zeckhauser. “Eliciting Informative Feedback: The Peer-Prediction Method”. In: *Management Science* 51.9 (2005), pp. 1359–1373 (cit. on pp. 5, 38, 46).
- [MS92] E. Maskin and H. Siebert. “Auctions and privatization”. In: *Privatization* (1992) (cit. on pp. 97, 102, 119).
- [MT00] S. Mullainathan and R. H. Thaler. *Behavioral economics*. 2000 (cit. on p. 7).
- [MTL14] Z. Maniadis, F. Tufano, and J. A. List. “One swallow doesn’t make a summer: New evidence on anchoring effects”. In: *American Economic Review* 104.1 (2014), pp. 277–90 (cit. on p. 127).
- [MW76] J. E. Matheson and R. L. Winkler. “Scoring Rules for Continuous Probability Distributions”. In: *Management Science* 22.10 (1976), pp. 1087–1096 (cit. on p. 11).
- [MW82] P. R. Milgrom and R. J. Weber. “A theory of auctions and competitive bidding”. In: *Econometrica: Journal of the Econometric Society* (1982), pp. 1089–1122 (cit. on pp. 95, 96, 100, 102).
- [Mye81] R. B. Myerson. “Optimal auction design”. In: *Mathematics of operations research* 6.1 (1981), pp. 58–73 (cit. on pp. 110, 121, 123).
- [Nat+13] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. “Learning with noisy labels”. In: *Advances in neural information processing systems*. 2013, pp. 1196–1204 (cit. on pp. 10, 12, 36, 39, 48).

- [OAB15] Z. Oravecz, R. Anders, and W. H. Batchelder. “Hierarchical Bayesian modeling for test theory without an answer key”. In: *Psychometrika* 80.2 (2015), pp. 341–364 (cit. on p. 174).
- [OS10] A. Othman and T. Sandholm. “Decision rules and decision markets.” In: *AAMAS*. Citeseer. 2010, pp. 625–632 (cit. on p. 3).
- [OVB14] Z. Oravecz, J. Vandekerckhove, and W. H. Batchelder. “Bayesian cultural consensus theory”. In: *Field Methods* 26.3 (2014), pp. 207–222 (cit. on pp. 6, 75, 85, 180).
- [Par+16] M. Parry et al. “Linear scoring rules for probabilistic binary classification”. In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1596–1607 (cit. on pp. 65, 66).
- [Por95] R. H. Porter. “The role of information in US offshore oil and gas lease auction”. In: *Econometrika: Journal of the Econometric Society* (1995), pp. 1–27 (cit. on p. 99).
- [Pre04a] D. Prelec. “A Bayesian Truth Serum for Subjective Data”. In: *Science* 306.5695 (2004), pp. 462–466 (cit. on pp. 5, 76).
- [Pre04b] D. Prelec. “A Bayesian truth serum for subjective data”. In: *science* 306.5695 (2004), pp. 462–466 (cit. on pp. 6, 38, 39).
- [Pri20] T. C. for the Prize in Economic Sciences in Memory of Alfred Nobel. *Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2020*. 2020. URL: <https://www.nobelprize.org/uploads/2020/09/advanced-economicsciencesprize2020.pdf> (cit. on p. 96).
- [PS19] A. B. Palley and J. B. Soll. “Extracting the Wisdom of Crowds When Information is Shared”. In: *Management Science* 65.5 (2019), pp. 2291–2309 (cit. on pp. 6, 74, 75, 85).
- [PS20] A. Palley and V. Satopää. “Boosting the wisdom of crowds within a single judgment problem: Selective averaging based on peer predictions”. In: *Available at SSRN 3504286* (2020) (cit. on p. 75).
- [PSM17] D. Prelec, H. S. Seung, and J. McCoy. “A solution to the single-question crowd wisdom problem”. In: *Nature* 541.7638 (2017), p. 532 (cit. on pp. 6, 38, 64, 73–75, 83, 85, 128, 176, 180).

- [Rab13] M. Rabin. “An approach to incorporating psychology into economics”. In: *American Economic Review* 103.3 (2013), pp. 617–22 (cit. on p. 7).
- [Ree18] A. Rees-Jones. “Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match”. In: *Games and Economic Behavior* 108 (2018), pp. 317–330 (cit. on p. 7).
- [RF13a] G. Radanovic and B. Faltings. “A Robust Bayesian Truth Serum for Non-Binary Signals”. In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence*. AAAI ’13. 2013 (cit. on p. 38).
- [RF13b] G. Radanovic and B. Faltings. “A robust bayesian truth serum for non-binary signals”. In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013 (cit. on p. 5).
- [RF14] G. Radanovic and B. Faltings. “Incentives for truthful information elicitation of continuous signals”. In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014 (cit. on p. 39).
- [RFJ16] G. Radanovic, B. Faltings, and R. Jurca. “Incentives for effort in crowdsourcing using the peer truth serum”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.4 (2016), p. 48 (cit. on pp. 5, 38, 42, 66, 71, 72, 74, 76, 81, 91).
- [RG10] R. Ranjan and T. Gneiting. “Combining probability forecasts”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.1 (2010), pp. 71–91 (cit. on pp. 30, 75).
- [Ril14] B. Riley. “Minimum truth serums with optional predictions”. In: *Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC14)*. 2014 (cit. on p. 38).
- [RT16] T. Roughgarden and I. Talgam-Cohen. “Optimal and Robust Mechanism Design with Interdependent Values”. In: *ACM Trans. Economics and Comput.* 4.3 (2016), 18:1–18:34. DOI: [10.1145/2910577](https://doi.org/10.1145/2910577). URL: <https://doi.org/10.1145/2910577> (cit. on pp. 97, 100–102, 106, 109, 110, 116, 121, 123).

- [RW15] B. van Rooyen and R. C. Williamson. “Learning in the Presence of Corruption”. In: *arXiv preprint:1504.00091* (2015) (cit. on pp. 39, 48).
- [Sal+18] L. Salsabil, J. Wu, M. H. Choudhury, W. A. Ingram, E. A. Fox, S. J. Rajtmajer, and C. L. Giles. “A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software”. In: (2018) (cit. on p. 136).
- [Sat+14a] V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. “Combining multiple probability predictions using a simple logit model”. In: *International Journal of Forecasting* 30.2 (2014), pp. 344–356 (cit. on pp. 6, 30, 74, 75, 77, 84, 91, 93).
- [Sat+14b] V. A. Satopää, S. T. Jensen, B. A. Mellers, P. E. Tetlock, L. H. Ungar, et al. “Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs”. In: *Annals of Applied Statistics* 8.2 (2014), pp. 1256–1280 (cit. on p. 75).
- [Sat+21] V. A. Satopää, M. Salikhov, P. E. Tetlock, and B. Mellers. “Bias, information, noise: The BIN model of forecasting”. In: *Management Science* 67.12 (2021), pp. 7599–7618 (cit. on p. 6).
- [Sav71] L. J. Savage. “Elicitation of Personal Probabilities and Expectations”. In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801 (cit. on pp. 3, 35, 38, 40).
- [Sch16] S. Schmidt. “Shall we really do it again? The powerful concept of replication is neglected in the social sciences.” In: (2016) (cit. on p. 127).
- [Sco+13] C. Scott, G. Blanchard, G. Handy, S. Pozzi, and M. Flaska. “Classification with Asymmetric Label Noise: Consistency and Maximal Denoising.” In: *COLT*. 2013, pp. 489–511 (cit. on p. 36).
- [Sco15] C. Scott. “A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels.” In: *AISTATS*. 2015 (cit. on pp. 10, 12, 36, 39, 48).

- [Shn+16] V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. “Informed truthfulness in multi-task peer prediction”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM. 2016, pp. 179–196 (cit. on pp. 5, 38, 39, 42, 66, 72, 74, 81, 91).
- [Son+18] J. Y. Song, R. Fok, A. Lundgard, F. Yang, J. Kim, and W. S. Lasecki. “Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance”. In: *23rd International Conference on Intelligent User Interfaces*. 2018, pp. 559–570 (cit. on p. 94).
- [Sur05] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005 (cit. on p. 1).
- [SY20a] G. Schoenebeck and F.-Y. Yu. “Learning and Strongly Truthful Multi-Task Peer Prediction: A Variational Approach”. In: *arXiv preprint arXiv:2009.14730* (2020) (cit. on pp. 5, 39).
- [SY20b] G. Schoenebeck and F.-Y. Yu. “Two Strongly Truthful Mechanisms for Three Heterogeneous Agents Answering One Question”. In: *International Conference on Web and Internet Economics*. Springer. 2020, pp. 119–132 (cit. on p. 38).
- [Tet+14] P. E. Tetlock, B. A. Mellers, N. Rohrbaugh, and E. Chen. “Forecasting tournaments: Tools for increasing transparency and improving the quality of debate”. In: *Current Directions in Psychological Science* 23.4 (2014), pp. 290–295 (cit. on pp. 1, 35).
- [Tha92] R. H. Thaler. “The winner’s curse”. In: *Across the Board* 29 (1992), pp. 30–30 (cit. on p. 99).
- [Ung+12] L. Ungar, B. Mellers, V. Satopää, P. Tetlock, and J. Baron. “The good judgment project: A large scale test of different methods of combining expert predictions”. In: *2012 AAAI Fall Symposium Series*. 2012 (cit. on p. 176).
- [Wan+11] G. Wang, S. R. Kulkarni, H. V. Poor, and D. N. Osherson. “Aggregating large sets of probabilistic forecasts by weighted coherent adjustment”. In: *Decision Analysis* 8.2 (2011), pp. 128–144 (cit. on p. 76).

- [WC92] R. L. Winkler and R. T. Clemen. “Sensitivity of weights in combining forecasts”. In: *Operations research* 40.3 (1992), pp. 609–614 (cit. on p. 4).
- [Wil69] R. B. Wilson. “Communications to the editor—competitive bidding with disparate information”. In: *Management science* 15.7 (1969), pp. 446–452 (cit. on pp. 96, 100, 123).
- [Win69] R. L. Winkler. “Scoring rules and the evaluation of probability assessors”. In: *Journal of the American Statistical Association* 64.327 (1969), pp. 1073–1078 (cit. on pp. 3, 11, 35, 38).
- [Wit+17] J. Witkowski, P. Atanasov, L. H. Ungar, and A. Krause. “Proper proxy scoring rules”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (cit. on pp. 36, 39, 48, 66, 74, 76, 80, 86, 91).
- [Wit+18] J. Witkowski, R. Freeman, J. W. Vaughan, D. M. Pennock, and A. Krause. “Incentive-compatible forecasting competitions”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. 2018 (cit. on pp. 11, 12).
- [WLB15] D. S. Weld, C. H. Lin, and J. Bragg. “Artificial intelligence and collective intelligence”. In: *Handbook of Collective Intelligence* (2015), pp. 89–114 (cit. on p. 94).
- [WLC19] J. Wang, Y. Liu, and Y. Chen. “Forecast aggregation via peer prediction”. In: *arXiv preprint arXiv:1910.03779* (2019) (cit. on p. 70).
- [WP12a] J. Witkowski and D. Parkes. “Peer prediction without a common prior”. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC ’12. ACM. 2012, pp. 964–981 (cit. on p. 39).
- [WP12b] J. Witkowski and D. C. Parkes. “A Robust Bayesian Truth Serum for Small Populations”. In: *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. AAAI ’12. 2012 (cit. on pp. 5, 38, 76).
- [WP13] J. Witkowski and D. C. Parkes. “Learning the prior in minimal peer prediction”. In: *Proceedings of the 3rd Workshop on Social Computing and User Generated Content at the ACM Conference on Electronic Commerce*. Vol. 14. 2013 (cit. on pp. 5, 38).

- [WZ04] J. Wolfers and E. Zitzewitz. “Prediction markets”. In: *Journal of economic perspectives* 18.2 (2004), pp. 107–126 (cit. on p. 3).
- [YYU20] Y. Yang, W. Youyou, and B. Uzzi. “Estimating the deep replicability of scientific findings using human and artificial intelligence”. In: *Proceedings of the National Academy of Sciences* 117.20 (2020), pp. 10762–10768 (cit. on p. 136).
- [Zhe01] C. Z. Zheng. “High bids and broke winners”. In: *Journal of Economic theory* 100.1 (2001), pp. 129–171 (cit. on p. 97).
- [ZVK17] A. X. Zhang, L. Verou, and D. Karger. “Wikum: Bridging discussion forums and wikis using recursive summarization”. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 2082–2096 (cit. on p. 94).

Appendix A

Appendix to Chapter 2

A.1 Missing Figures

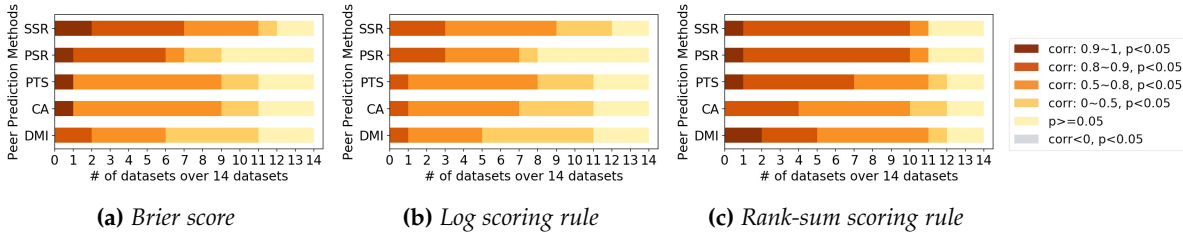


Figure A.1: The number of datasets in each level of correlation (measured by Pearson’s correlation coefficient) between individuals’ peer prediction scores and different SPSR when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.

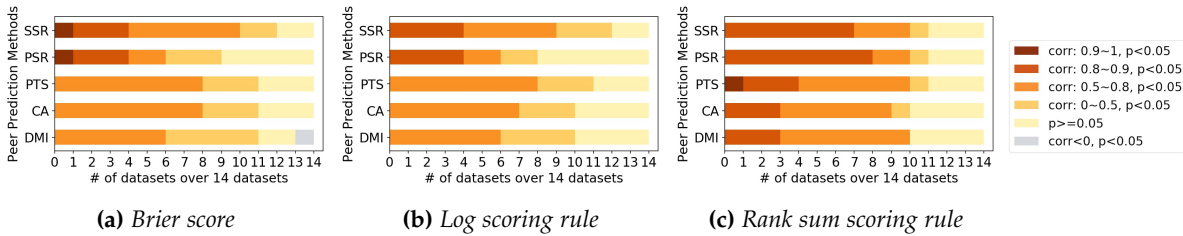


Figure A.2: The number of datasets in each level of correlation (measured by Spearman’s correlation coefficient) between individuals’ peer prediction scores and different SPSR when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.

A.2 Missing Proofs

A.2.1 Proof of Lemma 2.2

Proof. Suppose z and y are not stochastically relevant, we have

$$\Pr[y = 0|z = 0] = \Pr[y = 0|z = 1], \quad (\text{A.1})$$

$$\Pr[y = 1|z = 0] = \Pr[y = 1|z = 1]. \quad (\text{A.2})$$

From Eqn. (A.1) we know that

$$\frac{\Pr[y = 0, z = 1]}{\Pr[z = 1]} = \frac{\Pr[y = 0, z = 0]}{\Pr[z = 0]} \Leftrightarrow \frac{\Pr[y = 0]e_z^-}{\Pr[z = 1]} = \frac{\Pr[y = 0](1 - e_z^-)}{\Pr[z = 0]},$$

When $\Pr[y = 0] \neq 0$, we have $\frac{\Pr[z=1]}{\Pr[z=0]} = \frac{e_z^-}{1 - e_z^-}$. Similarly from Eqn. (A.2), we have $\frac{\Pr[z=1]}{\Pr[z=0]} = \frac{1 - e_z^+}{e_z^+}$., when $\Pr[y = 1] \neq 0$. Therefore we, obtained

$$\frac{e_z^-}{1 - e_z^-} = \frac{1 - e_z^+}{e_z^+},$$

from which we have $e_z^- + e_z^+ = 1$. Contradiction. The other direction follows similarly. \square

A.2.2 Proof of Lemma 2.12

Proof. We consider the estimation of the error rates e_z^+, e_z^- of an agent i , and we consider a generic task as tasks are a priori similar. Thus, in the proof, we drop the subscript k , which indexes the tasks. There are two layers of estimation error is solving the system of equations Eqn. (2.4, 2.5, 2.6):

- **1. Estimation error due to heterogeneous agents:** the higher order equations doesn't capture the true matching probability with heterogeneous agents. As we draw Z_2 and Z_3 in a task without replacement, with finite number of agents, Z_2 and Z_3 are dependent with Z_1 , and the error rates of Z_2 and Z_3 are not exactly the same to the error rates of Z_1 (z).
- **2. Estimation errors due to finite estimation samples:** The last sources of errors come from the estimation errors of $\widetilde{\beta}_{-i}, \widetilde{\gamma}_{-i}$ and $\widetilde{\alpha}_{-i}$.

Next we bound the two errors separately.

1. Estimation error due to heterogeneous agents: The challenge lies in the fact that the higher order equations doesn't capture the true matching probability with heterogeneous agents.

We first consider Eqn. (2.5). (2.5) is not precise– randomly picking a prediction signal from all agents without replacement leads to a different error rates. This will complicate the solution for the system of equations. We show that our estimation, though being ignoring the above bias, will not affect our results

by too much: Let k_1 be the agent whose prediction signal is picked to be Z_1 . Conditioned on agent k_1 being picked and on reports q_1, \dots, q_N , we have $\Pr[Z_1 = Z_2 = 1 | q_1, \dots, q_N, k_1] = q_{k_1} \cdot \left(\frac{\sum_{j \neq i, k_1} q_j}{N-2} \right)$. Recall that q_{k_1} is a random variable because of the private signal c_{k_1} received by agent k_1 and the randomness in σ_{k_1} , and that $e_z^+ = \mathbb{E}_{q_1, \dots, q_N | y=1}[\bar{q}_{-i}]$. We have that

$$\begin{aligned}
\Pr[Z_1 = Z_2 = 1 | y = 1] &= \mathbb{E}_{k_1} [\mathbb{E}_{q_1, \dots, q_N | y=1} [\Pr[Z_1 = Z_2 = 1 | k_1, q_1, \dots, q_N]]] \\
&= \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1} \left[q_{k_1} \cdot \left(\frac{\sum_{j \neq i, k_1} q_j}{N-2} \right) \right] \right] \\
&= \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1} [q_{k_1}] \cdot \mathbb{E}_{q_1, \dots, q_N | y=1} \left[\frac{\sum_{j \neq i, k_1} q_j}{N-2} \right] \right] \\
&= \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1} [q_{k_1}] \cdot \mathbb{E}_{q_1, \dots, q_N | y=1} \left[\frac{(N-1)\bar{q}_{-i}}{N-2} - \frac{q_{k_1}}{N-2} \right] \right] \\
&= \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1} [q_{k_1}] \cdot \left(\frac{N-1}{N-2} e_z^+ - \frac{1}{N-2} \mathbb{E}_{q_1, \dots, q_N | y=1} [q_{k_1}] \right) \right] \\
&= \frac{N-1}{N-2} e_z^+ \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1} [q_{k_1}] \right] - \frac{1}{N-2} \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1}^2 [q_{k_1}] \right] \\
&= \frac{N-1}{N-2} (e_z^+)^2 - \frac{1}{N-2} \omega,
\end{aligned}$$

where $\omega := \mathbb{E}_{k_1} \left[\mathbb{E}_{q_1, \dots, q_N | y=1}^2 [q_{k_1}] \right]$.

Note both e_z^+ and ω are no more than 1. Then ,

$$\left| \frac{N-1}{N-2} (e_z^+)^2 - \frac{1}{N-2} \omega - (e_z^+)^2 \right| \leq \frac{(e_z^+)^2}{N-2} + \frac{1}{N-2} \omega \leq \frac{2}{N-2}$$

This adds $\frac{2}{N-2}$ error bias in the step where we replace $\Pr[z_1 = z_2 = 1 | y = 1]$ with $(e_z^+)^2$ in the deduction of Eqn. (2.5). And, it finally adds $\frac{2}{N-2}$ error bias in estimating β_{-i} (through both $(e_z^+)^2$ and $(1 - e_z^+)^2$) in Eqn. (2.5).

Similarly for the matching among three agents (Eqn. (2.6)) we have

$$\left| \Pr[Z_1 = Z_2 = Z_3 = 1 | y = 1] - (e_z^+)^3 \right| \leq \frac{3}{N-3}.$$

And this adds $\frac{3}{N-3}$ error bias in estimating γ_{-i} .

2. Estimation errors due to finite estimation samples: The last sources of errors come from the estimation errors of $\widetilde{\beta}_{-i}$, $\widetilde{\gamma}_{-i}$ and $\widetilde{\alpha}_{-i}$. Direct application of the Chernoff bound gives us the following lemma:

Lemma A.1. *When there are M samples for estimating $\widetilde{\beta}_{-i}$, $\widetilde{\gamma}_{-i}$ and $\widetilde{\alpha}_{-i}$ respectively (total budgeting $3M$), we*

have with probability at least $1 - \delta$ that

$$|\widetilde{\beta}_{-i} - \beta_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}}, |\widetilde{\gamma}_{-i} - \gamma_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}}, |\widetilde{\alpha}_{-i} - \alpha_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}}.$$

The error analysis in **1** and **2** jointly imply that with probability at least $1 - \delta$

$$|\widetilde{\beta}_{-i} - \beta_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2}, |\widetilde{\gamma}_{-i} - \gamma_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{3}{N-3}, |\widetilde{\alpha}_{-i} - \alpha_{-i}| \leq \sqrt{\frac{\ln \frac{6}{\delta}}{2M}}.$$

Now we are ready to prove Lemma 2.12. First of all, from Algorithm 6, we can easily derive that

$$|\widetilde{e}_z^- - e_z^-| \leq \frac{|\tilde{a} - a|}{2} + \frac{|\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}|}{2} \quad (\text{A.3})$$

$$|\widetilde{e}_z^+ - e_z^+| \leq \frac{|\tilde{a} - a|}{2} + \frac{|\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}|}{2} \quad (\text{A.4})$$

For the latter term in Eqn. (A.3) and (A.4), we have

$$\begin{aligned} \frac{|\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}|}{2} &= \frac{|(\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}) \cdot (\sqrt{\tilde{a}^2 - 4\tilde{b}} + \sqrt{a^2 - 4b})|}{2(\sqrt{\tilde{a}^2 - 4\tilde{b}} + \sqrt{a^2 - 4b})} \\ &\leq \frac{|\tilde{a}^2 - a^2|}{2\sqrt{a^2 - 4b}} + \frac{4|\tilde{b} - b|}{2\sqrt{a^2 - 4b}} \\ &\leq \frac{|\tilde{a} - a|^2}{2\sqrt{a^2 - 4b}} + \frac{a \cdot |\tilde{a} - a|}{\sqrt{a^2 - 4b}} + \frac{2|\tilde{b} - b|}{\sqrt{a^2 - 4b}} \end{aligned}$$

The first inequality is due to that we drop the positive $2\sqrt{\tilde{a}^2 - 4\tilde{b}}$ in the denominator. For the second inequality, note that a is non-negative as essentially, $a = 1 - e_z^+ + e_z^-$ shown in proof for Theorem 2.9.

To summarize, we have

$$|\widetilde{e}_z^- - e_z^-| \leq \left(\frac{1}{2} + \frac{a}{\sqrt{a^2 - 4b}} \right) |\tilde{a} - a| + \frac{2|\tilde{b} - b|}{\sqrt{a^2 - 4b}} + \frac{1}{2\sqrt{a^2 - 4b}} |\tilde{a} - a|^2 \quad (\text{A.5})$$

$$|\widetilde{e}_z^+ - e_z^+| \leq \left(\frac{1}{2} + \frac{a}{\sqrt{a^2 - 4b}} \right) |\tilde{a} - a| + \frac{2|\tilde{b} - b|}{\sqrt{a^2 - 4b}} + \frac{1}{2\sqrt{a^2 - 4b}} |\tilde{a} - a|^2 \quad (\text{A.6})$$

The key tasks here reduce to bounding $|\tilde{a} - a|$ and $|\tilde{b} - b|$. Recall

$$\begin{aligned} a &:= \frac{\gamma_{-i} - \alpha_{-i}\beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2} \\ b &:= \frac{\alpha_{-i}\gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2} \end{aligned}$$

We know the following facts

$$\begin{aligned} |(\widetilde{\beta}_{-i} - (\widetilde{\alpha}_{-i})^2) - (\beta_{-i} - (\alpha_{-i})^2)| &\leq |(\widetilde{\alpha}_{-i})^2 - (\alpha_{-i})^2| + |\widetilde{\beta}_{-i} - \beta_{-i}| \\ &\leq 2|\widetilde{\alpha}_{-i} - \alpha_{-i}| + |\widetilde{\beta}_{-i} - \beta_{-i}| \leq 3\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2}, \end{aligned}$$

$$\begin{aligned} |(\widetilde{\gamma}_{-i} - \widetilde{\beta}_{-i}\widetilde{\alpha}_{-i}) - (\gamma_{-i} - \beta_{-i}\alpha_{-i})| &\leq |\widetilde{\gamma}_{-i} - \gamma_{-i}| + |\widetilde{\beta}_{-i}\widetilde{\alpha}_{-i} - \beta_{-i}\alpha_{-i}| \\ &\leq |\widetilde{\gamma}_{-i} - \gamma_{-i}| + |\widetilde{\beta}_{-i} - \beta_{-i}| + |\widetilde{\alpha}_{-i} - \alpha_{-i}| \\ &\leq 3\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2} + \frac{3}{N-3}, \end{aligned}$$

$$\begin{aligned} |(\widetilde{\alpha}_{-i}\widetilde{\gamma}_{-i} - (\widetilde{\beta}_{-i})^2) - (\alpha_{-i}\gamma_{-i} - (\beta_{-i})^2)| &\leq |\widetilde{\alpha}_{-i} - \alpha_{-i}| + |\widetilde{\gamma}_{-i} - \gamma_{-i}| + 2|\widetilde{\beta}_{-i} - \beta_{-i}| \\ &\leq 4\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2} + \frac{3}{N-3}, \end{aligned}$$

Next, denoting $\eta = p(1-p)(1 - e_z^+ - e_z^-)^2$ (which also means $\Delta = p(1-p)(x^- - x^+)^2$), we have

$$\begin{aligned} \beta_{-i} - (\alpha_{-i})^2 &= (1-p) \cdot (x^-)^2 + p \cdot (x^+)^2 - ((1-p) \cdot x^- + p \cdot x^+)^2 \\ &= (1-p) \cdot p \cdot (x^- - x^+)^2 \\ &= \eta \end{aligned}$$

Let N be sufficiently large such that

$$3\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2} < \eta \tag{A.7}$$

then

$$\widetilde{\beta}_{-i} - (\widetilde{\alpha}_{-i})^2 \geq \frac{p(1-p)}{2} \cdot \frac{\eta}{2}$$

Therefore,

$$\begin{aligned} |\tilde{a} - a| &= \left| \frac{\widetilde{\gamma}_{-i} - \widetilde{\alpha}_{-i}\widetilde{\beta}_{-i}}{\widetilde{\beta}_{-i} - (\widetilde{\alpha}_{-i})^2} - \frac{\gamma_{-i} - \alpha_{-i}\beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2} \right| \\ &\leq \frac{|(\widetilde{\beta}_{-i} - (\widetilde{\alpha}_{-i})^2) - (\beta_{-i} - (\alpha_{-i})^2)| + |(\widetilde{\gamma}_{-i} - \widetilde{\beta}_{-i}\widetilde{\alpha}_{-i}) - (\gamma_{-i} - \beta_{-i}\alpha_{-i})|}{|\widetilde{\beta}_{-i} - (\widetilde{\alpha}_{-i})^2| \cdot |\beta_{-i} - (\alpha_{-i})^2|} \\ &\leq \frac{2}{\eta^2} \left(6\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{4}{N-2} + \frac{3}{N-3} \right) \end{aligned}$$

Note that the first inequality uses Lemma 2.13.

Similarly for b , we have

$$|\tilde{b} - b| \leq \frac{2}{\eta^2} \left(7\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{4}{N-2} + \frac{3}{N-3} \right)$$

Together, we proved that when M and N are sufficiently large such that Eqn. (A.7) holds, i.e., $3\sqrt{\frac{\ln \frac{6}{\delta}}{2M}} + \frac{2}{N-2} < \frac{\eta}{2}$, we have

$$|\tilde{e}_z^- - e_z^-| \leq O \left(\sqrt{\frac{\ln \frac{1}{\delta}}{M}} + \frac{1}{N} \right)$$

$$|\tilde{e}_z^+ - e_z^+| \leq O \left(\sqrt{\frac{\ln \frac{1}{\delta}}{M}} + \frac{1}{N} \right)$$

□

Appendix B

Appendix to Chapter 3

B.1 Forecast aggregation performance on small datasets

This section examines the performance of our PAS-aided aggregators and benchmark aggregators over smaller datasets. Specifically, for each of the 14 original datasets, we uniformly randomly sample without replacement 20 binary events and 30 or 50 participants to generate a smaller dataset. We keep the original participant set for those MIT datasets with less than 30 or 50 participants (Table 2.1). Meanwhile, we still maintain that each event receives at least 10 responses and that each participant forecasts on at least 15 events. The HFC datasets are too sparse to generate such small datasets with this forecast density requirement. Therefore, we remove them from the examination. For each of the remaining 11 datasets, we run random sampling 30 times and report the average aggregation performance over these 30 runs under the Brier score in Table B.1a (50 participants sampled for each run) and Table B.1b (30 participants sampled for each run). Both tables demonstrate a consistent improvement of using the Mean-based PAS-aided aggregators, with better relative performance (compared to the benchmarks) achieved on the datasets with 50 participants sampled. This result indicates that our PAS-aided aggregators can also be applied to relatively small prediction datasets (e.g., the forecasts collected at the cold-start stage of long-term forecast competitions, where no ground truth information has yet been resolved) and improve the aggregation performance.

Base aggr.	Score	G1	G2	G3	G4	M1a	M1b	M1c	M2	M3	M4a	M4b
Mean	DMI	.124	.070	.087	.047	.389	.175	.143	.496	.407	.468	.273
	CA	.115	.066	.076	.040	.371	.161	.134	.490	.406	.500	.269
	PTS	.117	.066	.076	.041	.423	.177	.135	.496	.407	.500	.269
	SSR	.121	.073	.076	.050	.492	.200	.134	.483	.406	.505	.277
	PSR	.120	.070	.076	.051	.524	.183	.170	.498	.413	.504	.283
Logit	DMI	.115	.067	.093	.032	.540	.172	.091	.563	.507	.607	.348
	CA	.112	.060	.087	.027	.524	.156	.091	.557	.508	.662	.330
	PTS	.112	.060	.087	.029	.597	.190	.085	.566	.506	.668	.339
	SSR	.106	.064	.088	.043	.660	.232	.073	.545	.505	.674	.368
	PSR	.113	.067	.085	.049	.691	.199	.132	.588	.515	.652	.359
Mean (benchmark)		.193	.166	.106	.135	.453	.347	.345	.480	.399	.436	.310
Logit (benchmark)		.115	.084	.076	.055	.683	.438	.340	.497	.497	.599	.458
VI (benchmark)		.213	.110	.093	.070	.673	.265	.308	.862	.577	.721	.353
SP (benchmark)		N/A	N/A	N/A	N/A	.507	.190	.310	.890	.487	.637	.543

(a) 20 binary events and 50 participants sampled for each run

Base aggr.	Score	G1	G2	G3	G4	M1a	M1b	M1c	M2	M3	M4a	M4b
Mean	DMI	.166	.096	.090	.080	.442	.160	.160	.473	.390	.512	.313
	CA	.154	.083	.059	.061	.440	.153	.151	.479	.386	.534	.296
	PTS	.154	.085	.061	.062	.465	.155	.154	.472	.388	.547	.299
	SSR	.156	.085	.061	.064	.480	.150	.152	.481	.393	.542	.321
	PSR	.158	.082	.062	.064	.528	.170	.179	.482	.397	.540	.332
Logit	DMI	.158	.080	.077	.054	.611	.153	.133	.515	.500	.692	.397
	CA	.149	.068	.062	.048	.640	.140	.112	.539	.494	.713	.363
	PTS	.148	.069	.063	.046	.652	.151	.120	.519	.496	.712	.380
	SSR	.141	.069	.066	.049	.697	.136	.093	.527	.500	.696	.416
	PSR	.152	.072	.069	.057	.720	.176	.160	.551	.507	.704	.412
Mean (benchmark)		.208	.161	.091	.135	.473	.327	.358	.475	.387	.475	.354
Logit (benchmark)		.134	.084	.054	.058	.720	.381	.380	.491	.493	.665	.512
VI (benchmark)		.239	.113	.080	.077	.724	.224	.274	.869	.550	.773	.411
SP (benchmark)		nan	nan	nan	nan	.573	.230	.313	.903	.440	.687	.647

(b) 20 binary events and 30 participants sampled for each run

Table B.1: The mean Brier scores (range [0, 2], the lower the better) of different aggregators on randomly sampled sub-datasets of 4 GJP datasets and 7 MIT datasets. The best mean Brier score among benchmarks on each dataset is marked by bold font. The mean Brier scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in **green**; those outperforming the second best of benchmarks are highlighted in **yellow**; the worst mean Brier scores over all aggregators on each dataset are highlighted in **red**.

B.2 Missing tables

	Mean-based					Logit-based					Benchmarks			
	DMI	CA	PTS	SSR	PSR	DMI	CA	PTS	SSR	PSR	Mean	Logit	VI	MP ¹
Mean (Brier)	.221	.226	.226	.225	.230	.244	.247	.254	.257	.266	.290	.317	.315	.423
Std. (Brier)	.150	.153	.158	.155	.168	.212	.221	.225	.233	.249	.130	.224	.267	.125
Mean (Log)	.354	.369	.364	.388	.373	.441	.470	.484	.521	.513	.453	.578	.701	.728
Std. (Log)	.214	.213	.222	.231	.234	.409	.444	.452	.508	.508	.154	.446	.573	.297

Table B.2: The mean and the standard deviation of the mean Brier scores and the mean log scores of the 10 PAS-aided aggregators and the benchmarks over 14 datasets. The bold font means that the data is significantly better than the counterparts of all benchmarks with p -value < 0.05 .

Aggregators	M1a	M1b	M1c	M2	M3	M4a	M4b
Cultural consensus model [OAB15]	0.55	0.02	0.00	0.76	0.56	0.64	0.31
Cognitive hierarchy model [LD14]	-	-	0.32	0.48	0.46	-	-
Statistical surprising popularity method [MP17]	0.24	0.06	0.02	0.60	0.51	0.65	0.35

Table B.3: The mean Brier scores of three statistical-inference-based aggregators on MIT datasets reported by McCoy and Prelec [MP17]. The Brier score has been re-scaled to the range $[0,2]$ to align with ours. The bold font marks the five cases where theirs outperform the worst of our five mean-based PAS aggregators.

¹As MP only applies to 7 MIT datasets, the data of MP in this table should not be compared directly to that of the others.

Base aggr.	PAS	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
Mean	DMI	.236	.141	.143	.148	.370	.324	.187	.377	.242	.230	.625	.643	.880	.414
	CA	.241	.146	.156	.162	.351	.323	.235	.477	.238	.230	.642	.640	.880	.450
	PTS	.231	.142	.141	.147	.326	.317	.194	.499	.236	.230	.666	.640	.880	.450
	SSR	.246	.188	.148	.143	.314	.309	.212	.632	.226	.291	.669	.643	.911	.502
	PSR	.261	.134	.139	.126	.314	.310	.198	.642	.221	.236	.678	.644	.880	.441
Logit	DMI	.176	.115	.137	.084	.344	.327	.260	.583	.125	.094	.643	1.097	1.495	.691
	CA	.168	.114	.128	.073	.244	.330	.271	1.040	.100	.094	.734	1.093	1.495	.689
	PTS	.167	.114	.135	.082	.280	.329	.280	1.132	.111	.094	.776	1.093	1.495	.689
	SSR	.155	.110	.135	.093	.209	.318	.282	1.542	.086	.138	.746	1.125	1.431	.920
	PSR	.164	.115	.136	.091	.272	.334	.267	1.517	.075	.054	.805	1.097	1.495	.766
Mean (benchmark)	.365	.323	.242	.296	.373	.313	.268	.633	.520	.521	.672	.634	.686	.497	
Logit (benchmark)	.185	.138	.131	.119	.205	.267	.257	1.338	.782	.524	.718	1.047	1.380	1.003	
VI (benchmark)	.548	.176	.198	.206	.712	.699	.384	1.356	.073	.010	1.859	1.385	1.464	.741	
MP (benchmark)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	.597	.384	.373	.671	.804	1.226	1.042	

Table B.4: The mean log scores (the lower the better) of different aggregators on binary events of 14 datasets. The best mean score among benchmarks on each dataset is marked by bold font. The mean scores of 10 PAS-aided aggregators that outperform the best of benchmarks on each dataset are highlighted in **green**; those outperforming the second best of benchmarks are highlighted in **yellow**; the worst mean scores over all aggregators on each dataset are highlighted in **red**.

B.3 More details about the datasets

GJP datasets. GJP datasets [Ung+12; Ata+16; GJP16] contain four datasets about forecasts on geopolitical questions collected from 2011 to 2014. The dataset of each year differs in both the forecasting questions and the participant pools, and is denoted by G1 to G4 in our paper correspondingly. When collecting the forecasts, the participants were given different treatments: some were given probabilistic training, some were teamed up and allowed to discuss with each other before giving their own predictions, and some made predictions solely. Participants who demonstrated consistently high prediction accuracy across different forecasting questions in previous years were identified as “superforecasters” and were teamed up to participate in the forecast tournament in the following year [Mel+15]. The participants’ prediction accuracy has also been shown to be influenced by different treatments [Ata+16].

HFC datasets. HFC datasets [IAR19] contain three datasets collected in 2018 with forecasting questions ranging from geopolitics to economics and environments. We use H1 to denote the dataset collected by the Hughes Research Laboratories (HRL), with participants recruited from Amazon Mechanical Turk (AMT) as H1. We use H2 to denote the dataset collected by IRAPA, with participants recruited from Amazon Mechanical Turk (AMT). Moreover, we use H3 to denote the dataset collected by IRAPA, with participants recruited via invitation and recommendation.

MIT datasets. MIT datasets contain seven datasets (denoted as M1a, M1b, M1c, M2, M3, M4a, M4b [PSM17]) collected for seven forecast behavior studies and for testing forecast aggregation methods. The forecasting questions range from the capital of states to the price interval of some artworks and some trivial knowledge. In the datasets, participants were asked to give binary (yes-or-no) answers to the forecasting questions instead of probabilistic predictions. Datasets M1c, M2, M3 also contain the confidence for the binary answers, which we directly interpret into probabilistic predictions of the favored binary answers when we aggregate the predictions. Moreover, all of the seven datasets contain participants’ answers to an additional question for each forecasting question. This additional question asks the participants to estimate the percentage of other forecasters who choose the same binary answer as theirs. This information will be used by one of the benchmark aggregators we test. In particular, these seven datasets were collected to develop and evaluate information elicitation and aggregation methods on questions where the majority is likely to be wrong [PSM17]. Therefore, these datasets have a relatively low participants’ performance.

B.4 Missing Proofs

B.4.1 Proof of Theorem 3.2

Proof. The result about DMI is implied by Theorem 6.4 in [Kon20]. The result about CA can be proved in a similar way by observing that CA is asymptotically equivalent to determinant mutual information for binary events. For completeness, we present the proof for CA. We also present the proof for PTS below.

For CA: First, we introduce the determinant mutual information [Kon20]. Consider two discrete random variable X and W with the same support \mathcal{V} . Let $d(X, W) = (d_{u,v})_{u,v \in \mathcal{V}}$ be the joint distribution of X and W , where $d_{u,v} = \Pr(X = u \text{ and } W = v)$. Let $d(X|W) = (d_{u,v})_{u,v \in \mathcal{V}}$ be the conditional probability matrix, where $d_{u,v} = \Pr(X = u | W = v)$.

Definition B.1. *The determinant mutual information of two binary random variables X and W is $|\det(d(X, W))|$.*

We denote the determinant mutual information of X, W as $DM(X, W) = |\det(d(X, W))|$. We will involve the use of its two properties introduced below.

Proposition B.1. *Let X, X', W be three discrete random variables with the same support, and X' is less informative than X w.r.t. W , i.e., X' is independent of W conditioning X .*

- (Information monotonicity) $DM(X', W) \leq DM(X, W)$. The inequality is strict when $|\det(d(X, W))| \neq 0$ and $d(X'|X)$ is not a permutation matrix.
- (Relatively invariance) $DM(X', W) = DM(X, W) |\det(d(X'|X))|$.

The information monotonicity is the key property for being a mutual information.

Now, by Assumption A1 and the truthfulness assumption, we can consider the reported signal of agent j on a generic task as a binary random variable p_j ($p_j \in \{0, 1\}$). We denote the ground truth of the generic task as y and denote the joint distribution of agent j 's reports and the ground truth as $D^{j,*} = (d_{u,v}^{j,*})_{u,v \in \{0,1\}}$, where $d_{u,v}^{j,*} = \Pr(p_j = u \text{ and } y = v)$. Similarly, let $D^{j,k}$ be the joint distribution of agent j 's and agent k 's reports. The empirical joint distribution $\hat{D}^{j,k}$ is an unbiased and asymptotically consistent estimator of the true joint distribution $D^{j,k}$. So asymptotically ($|M| \rightarrow \infty$), we have $\hat{D}^{j,k} = D^{j,k}$.² Recall that CA compute the reward of agent j given a reference peer k as:

$$R_j^{\text{CA}} = \Delta \cdot \text{Sgn}(\Delta),$$

²For simplicity of exposition, we abuse the use of “=” here.

where $\Delta = (\delta_{u,v})_{u,v \in \{0,1\}}$, and $\delta_{u,v} = \hat{d}_{u,v}^{j,k} - \hat{d}_u^j \cdot \hat{d}_v^k$. By trivial math, we have $\delta_{0,0} = \delta_{1,1} = -\delta_{0,1} = -\delta_{1,0}$ and $R_j^{\text{DMI}} = 2|\delta_{0,0}|$. Further, asymptotically ($|\mathcal{M}| \rightarrow \infty$), $|\delta_{0,0}| = d_{0,0}^{j,k} - d_0^j \cdot d_0^k = |\det(D^{j,k})| = DM(p_j, p_k)$. Thus, we get that asymptotically ($|\mathcal{M}| \rightarrow \infty$), $R_j^{\text{DMI}} = DM(p_j, p_k)$.

Now, for another agent $j' \neq j$, when k is also selected as her reference peer, we have asymptotically ($|\mathcal{M}| \rightarrow \infty$),

$$\begin{aligned} R_j^{\text{DMI}} - R_{j'}^{\text{DMI}} &= DM(p_j, p_k) - DM(p_{j'}, p_k) \\ &= |DM(p_k|y)|(DM(p_j, y) - DM(p_{j'}, y)) \\ &\propto DM(p_j, y) - DM(p_{j'}, y) \end{aligned} \tag{B.1}$$

This equation holds due to the relatively invariance of the determinant mutual information and the Assumption A2. This equation holds for any reference agent $k \neq j, j'$. Thus, when agent j has a higher mutual information w.r.t. ground truth, she gets a higher reward than agent j' for any reference peer $k \neq j, j'$.

Asymptotically ($|\mathcal{N}| \rightarrow \infty$), with sufficient number of agents, the probability that agent j' (j) are selected as agent j 's (j' 's) reference peer can be neglected. Therefore, the expected reward, with expectation taken over the reference peer selection, of CA rank the agents in the order of the determinant mutual information of agents' reports w.r.t. ground truth.

For PTS: By the counterpart argument in the proof for CA, under Assumption A1, we can treat the report p_j as a random variable for a generic task with ground truth variable denoted as y . Let $\bar{d}_{u,v} = \sum_{j \in \mathcal{M}} d_{u,v}^{j,*} / |\mathcal{M}|$ representing the joint distribution of a uniformly randomly picked report on a task w.r.t. the ground truth. Let $\bar{d}_u = \bar{d}_{u,0} + \bar{d}_{u,1}$, $u \in \{0,1\}$ be the marginal probability that an average agent reporting $p_j = 1$. Further, let q_v be the marginal distribution of $y = v$. We have $q_v = \bar{d}_{0,v}^{j,*} + \bar{d}_{1,v}^{j,*}$, $\forall v \in \{0,1\}$. Let $\mathbb{E}[R_j^{\text{PTS}}]$ be the expected reward of agent j under PTS.

$$\begin{aligned}
\mathbb{E}[R_j^{\text{PTS}}] &= \frac{1}{|\mathcal{N}| - 1} \sum_{k \neq j} \frac{d_{0,0}^{j,k}}{\bar{p}_{-j,0}} + \frac{d_{1,1}^{j,k}}{\bar{p}_{-j,1}} && (|\mathcal{M}| \rightarrow \infty) \\
&= \frac{1}{|\mathcal{N}| - 1} \sum_{k \neq j} \frac{q_0 d_{0,0}^{j,*} d_{0,0}^{k,*} + q_1 d_{0,1}^{j,*} d_{0,1}^{k,*}}{\bar{p}_{-j,0}} + \frac{q_0 d_{1,0}^{j,*} d_{1,0}^{k,*} + q_1 d_{1,1}^{j,*} d_{1,1}^{k,*}}{\bar{p}_{-j,1}} && (\text{Assumption A2}) \\
&= \frac{q_0 d_{0,0}^{j,*} \bar{d}_{0,0} + q_1 d_{0,1}^{j,*} \bar{d}_{0,1}}{\bar{d}_0} + \frac{q_0 d_{1,0}^{j,*} \bar{d}_{1,0} + q_1 d_{1,1}^{j,*} \bar{d}_{1,1}}{\bar{d}_1} && (|\mathcal{N}| \rightarrow \infty) \\
&= \frac{q_0 d_{0,0}^{j,*} \bar{d}_{0,0} + q_1 (q_1 - d_{1,1}^{j,*}) \bar{d}_{0,1}}{\bar{d}_0} + \frac{q_0 (q_0 - d_{0,0}^{j,*}) \bar{d}_{1,0} + q_1 d_{1,1}^{j,*} \bar{d}_{1,1}}{\bar{d}_1} \\
&= q_0 \left(\frac{\bar{d}_{0,0}}{\bar{d}_0} - \frac{\bar{d}_{1,0}}{\bar{d}_1} \right) d_{0,0}^{j,*} + q_0 \left(\frac{\bar{d}_{1,1}}{\bar{d}_1} - \frac{\bar{d}_{1,0}}{\bar{d}_0} \right) d_{1,1}^{j,*} + \text{constant},
\end{aligned}$$

where $\text{constant} = (q_1)^2 \frac{\bar{d}_{0,1}}{\bar{d}_0} + (q_0)^2 \frac{\bar{d}_{1,0}}{\bar{d}_1}$. As with sufficient number of agents, q_0, q_1 and $\bar{d}_{u,v}, \bar{d}_u, (u, v \in \{0, 1\})$ are all constant to each agent, therefore, for each agent $j \in \mathcal{N}$, $\mathbb{E}[R_j^{\text{PTS}}]$ is the same weighted function of the matching probability $d_{0,0}^{j,*}$ and $d_{1,1}^{j,*}$ of the agent. Note that $\frac{\bar{d}_{0,0}}{\bar{d}_0}$ and $\frac{\bar{d}_{1,1}}{\bar{d}_1}$ are the precision of the mean prediction of agents for $y = 0$ and $y = 1$. If $\frac{\bar{d}_{0,0}}{\bar{d}_0} > 0.5$ and $\frac{\bar{d}_{1,1}}{\bar{d}_1} > 0.5$, we have $\frac{\bar{d}_{0,0}}{\bar{d}_0} - \frac{\bar{d}_{1,0}}{\bar{d}_1} > 0.5$ and $\frac{\bar{d}_{1,1}}{\bar{d}_1} - \frac{\bar{d}_{1,0}}{\bar{d}_0} > 0.5$, then $\mathbb{E}[R_j^{\text{PTS}}]$ is a negative function of a an expected weighted 0-1 loss of agent j . (Note that an expected weighted 0-1 loss of agent j is expressed by $\alpha d_{0,1}^{j,*} + \beta d_{1,0}^{j,*} (\alpha, \beta > 0)$.)

□

B.5 Variational inference for crowdsourcing

Variational inference for crowdsourcing (VI), proposed in [LPI12], is a computationally efficient inference method that builds a statistical model on agents' predictions over multiple questions to infer the ground truths of these questions. To make our paper self-contained, we present a sketch of VI, which mainly follows Section 3.2 of [LPI12].

VI consider the following statistical settings (assumptions): Agents provide binary predictions, i.e., $p_{ij} \in \{0, 1\}$ and have heterogeneous prediction abilities. Each agent j 's prediction ability is characterized by a parameter c_j , which is the correct probability of its predictions, i.e., $c_j = \mathbb{P}(p_{ij} = y_i), \forall i \in \mathcal{M}_j$. Moreover, $c_j, \forall j$ are i.i.d. drawn from some beta distribution $\text{Beta}(\alpha, \beta)$ with an expectation no less than 0.5, i.e., $\mathbb{E}_{c_j \sim \text{Beta}(\alpha, \beta)} \geq 0.5, \forall j$.

The goal of VI is to compute the marginal distribution of y_i under the above statistical assumptions. The marginal distribution is then used as the aggregated prediction \hat{q}_i for event i . Let $\delta_{ij} = \mathbb{1}\{p_{ij} = y_i\}$. The joint posterior distribution of the agents' abilities $\mathbf{c} := (c_1, \dots, c_{|\mathcal{N}|})$ and the ground truth outcomes

$\mathbf{y} := (y_1, \dots, y_{|\mathcal{M}|})$ conditioned on the predictions and hyper-parameter α, β is

$$\mathbb{P}(\mathbf{c}, \mathbf{y} | \{p_{ij}\}_{ij}, \alpha, \beta) \propto \prod_{j \in \mathcal{N}} \left(\mathbb{P}(c_j | \alpha, \beta) \prod_{i \in \mathcal{M}_j} c_j^{\delta_{ij}} (1 - c_j)^{(1 - \delta_{ij})} \right). \quad (\text{B.2})$$

Therefore, the marginal distribution of y_i is $\mathbb{P}(y_i | \{p_{ij}\}_{ij}, \alpha, \beta) = \sum_{y_i=0,1, i \in \mathcal{M} \setminus \{i\}} \int_{\mathbf{c}} \mathbb{P}(\mathbf{c}, \mathbf{y} | \{p_{ij}\}_{ij}, \alpha, \beta) d\mathbf{c}$.

$\mathbb{P}(y_i | \{p_{ij}\}_{ij}, \alpha, \beta)$ is computationally hard due to the summation of all $y_i, i \in \mathcal{M}$ and the integration of $c_j, j \in \mathcal{N}$. To solve this obstacle, VI adopts the mean field method. It approximates $\mathbb{P}(\mathbf{c}, \mathbf{y} | \{p_{ij}\}_{ij}, \alpha, \beta)$ with a fully factorized distribution $d(\mathbf{c}, \mathbf{y}) = \prod_{i \in \mathcal{M}} \mu_i(y_i) \prod_{j \in \mathcal{N}} \nu_j(c_j)$ for some probability distribution function $\mu_i, i \in \mathcal{M}$ and $\nu_j, j \in \mathcal{N}$, and determines the best $d(\mathbf{c}, \mathbf{y})$ by minimizing the the KL divergence:

$$\text{KL}[d(\mathbf{c}, \mathbf{y}) | \mathbb{P}(\mathbf{c}, \mathbf{y} | \{p_{ij}\}_{ij}, \alpha, \beta)] = -\mathbb{E}_{(\mathbf{c}, \mathbf{y}) \sim d(\mathbf{c}, \mathbf{y})} [\log(\mathbb{P}(\mathbf{c}, \mathbf{y} | \{p_{ij}\}_{ij}, \alpha, \beta))] - \sum_{i \in \mathcal{M}} H(\mu_i) - \sum_{j \in \mathcal{N}} H(\nu_j) \quad (\text{B.3})$$

$H(\cdot)$ is the entropy function. Noting the prior distribution of $q_j, j \in \mathcal{N}$ is a Beta distribution, we could derive the following mean field update using the block coordinate descent method:

$$\text{Updating } \mu_i : \mu_i(y_i) \propto \prod_{j \in \mathcal{N}_i} a_j^{\delta_{ij}} b_j^{1 - \delta_{ij}}, \quad (\text{B.4})$$

$$\text{Updating } \nu_j : \nu_j(c_j) \propto \text{Beta} \left(\sum_{i \in \mathcal{M}_j} \mu_i(p_{ij}) + \alpha, \sum_{i \in \mathcal{M}_j} \mu_i(1 - p_{ij}) + \beta \right), \quad (\text{B.5})$$

where $a_j = \exp(\mathbb{E}_{c_j \sim \nu_j}[\ln c_j])$ and $b_j = \exp(\mathbb{E}_{c_j \sim \nu_j}[\ln(1 - c_j)])$. Let $\bar{c}_j = \mathbb{E}_{c_j \sim \nu_j}[c_j]$. Applying the first order approximation $\ln(1 + x) \approx x$ with $x = \frac{c_j - \bar{c}_j}{\bar{c}_j}$ on a_j and b_j , we can get $a_j \approx \bar{c}_j$ and $b_j \approx 1 - \bar{c}_j$ and an approximate mean field update,

$$\text{Updating } \mu_i : \mu_i(y_i) \propto \prod_{j \in \mathcal{N}_i} \bar{c}_j^{\delta_{ij}} (1 - \bar{c}_j)^{1 - \delta_{ij}}, \quad (\text{B.6})$$

$$\text{Updating } \nu_j : \bar{c}_j = \frac{\sum_{i \in \mathcal{M}_j} \mu_i(p_{ij}) + \alpha}{|\mathcal{M}_j| + \alpha + \beta}. \quad (\text{B.7})$$

In our experiments, we used the two-coin model extension of VI [LPI12], where the prediction ability of an agent j is characterized by two parameters $c_{j,0}$ and $c_{j,1}$ with $c_{j,0} := \mathbb{P}(p_{ij} = 0 | y_i = 0)$ and $c_{j,1} := \mathbb{P}(p_{ij} = 1 | y_i = 1)$. Consequently, the approximate mean field update is

$$\text{Updating } \mu_i : \mu_i(y_i) \propto \prod_{j \in \mathcal{N}_i} \bar{c}_{j,y_i}^{\delta_{ij}} (1 - \bar{c}_{j,y_i})^{1 - \delta_{ij}}, y_i \in \{0, 1\}, \quad (\text{B.8})$$

$$\text{Updating } \nu_j : \bar{c}_{j,k} = \frac{\sum_{i \in \mathcal{M}_j} \mu_i(k) + \alpha}{\sum_{i \in \mathcal{M}_j} \mathbb{1}\{p_{ij} = k\} + \alpha + \beta}, k \in \{0, 1\}. \quad (\text{B.9})$$

[PSM17] has tested the performance of the culture consensus model (CCM) [OVb14] and the cognitive

hierarchy model (CHM) [LD14] on MIT datasets, while CCM has a slightly better performance. VI has the similar performance on MIT datasets compared to CCM. Therefore, we choose to test VI as a representative for multi-task aggregators.

Appendix C

Appendix to Chapter 4

C.1 Cursed Equilibrium in the Wallet-Game

In this section, we demonstrate how the winner's curse naturally arises from considering the cursed-equilibrium model.

Recall the wallet game example and suppose that Alice and Bob have $\chi = 1$. Then in the χ -cursed equilibrium, Alice bids as if Alice is a fully rational agent who values the item by $\mathbb{E}_{s_{Bob} \sim U[0,100]}[s_{Alice} + s_{Bob} | s_{Alice} = \$30] = \$80$. Thus, under the second price auction, Alice bids \$80 and experiences the winner's curse upon winning. Avery and Kagel [AK97] conducted a lab experiment about this wallet game with each agent's wallet money drawn from $U[1,4]$. The best linear regressor of agents' strategy shows that agents bid by $2.64 + 1.13s_i$, close to the expected valuation $\mathbb{E}_{s_{-i} \sim U[1,4]}[s_i + s_{-i}] = 2.5 + s_i$, instead of the BNE strategy $2s_i$, indicating the agents have a $\chi > 0$. Eyster and Rabin [ER05] further showed that any $\chi > 0$ fits data better than the fully rational case ($\chi = 0$) and with a 95% confidence interval of $[0.59, 0.67]$.

C.2 Missing proofs

C.2.1 Proof of Proposition 4.3

Proof. As the mechanism is C-EPIC under parameter χ , we have

$$x_i(\mathbf{s})v_i(\mathbf{s}) - p_i(\mathbf{s}) \geq x_i(b_i, \mathbf{s}_{-i})v_i(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) \quad \forall i, \mathbf{s}, b_i,$$

Therefore, we have $\forall i, \mathbf{s}, b_i$:

$$\begin{aligned}
EU_i^{\chi_i}(\mathbf{b} = \mathbf{s}, s_i; \sigma_{-i}^*) &= x_i(\mathbf{s})v_i^{\chi'}(\mathbf{s}) - p_i(\mathbf{s}) \\
&= x_i(\mathbf{s})((1 - \chi - \epsilon)v_i(\mathbf{s}) - (\chi + \epsilon)\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})]) - p_i(\mathbf{s}) \\
&= x_i(\mathbf{s})v_i^{\chi}(\mathbf{s}) - p_i(\mathbf{s}) + \epsilon \cdot x_i(\mathbf{s}) \left(\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})] - v_i(\mathbf{s}) \right) \\
&\geq x_i(b_i, \mathbf{s}_{-i})v_i^{\chi}(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) + \epsilon \cdot x_i(\mathbf{s}) \left(\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})] - v_i(\mathbf{s}) \right) \\
&= x_i(b_i, \mathbf{s}_{-i})v_i^{\chi'}(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) \\
&\quad + \epsilon \cdot x_i(\mathbf{s}) \left(\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})] - v_i(\mathbf{s}) \right) - \epsilon \cdot x_i(b_i, \mathbf{s}_{-i}) \left(\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})] - v_i(\mathbf{s}) \right) \\
&= x_i(b_i, \mathbf{s}_{-i})v_i^{\chi'}(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) + \epsilon(x_i(\mathbf{s}) - x_i(b_i, \mathbf{s}_{-i})) \left(\mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i}[v_i(s_i, \bar{\mathbf{s}}_{-i})] - v_i(\mathbf{s}) \right) \\
&\geq x_i(b_i, \mathbf{s}_{-i})v_i^{\chi'}(\mathbf{s}) - p_i(b_i, \mathbf{s}_{-i}) - \epsilon \cdot v_i(\bar{s}, \dots, \bar{s}) \\
&= EU_i^{\chi_i}(\mathbf{b}_{-i} = \mathbf{s}_{-i}, b_i, s_i; \sigma_{-i}^*) - \epsilon_i \cdot v_i(\bar{s}, \dots, \bar{s})
\end{aligned}$$

□

C.2.2 Proof of Lemma 4.5

Proof. We only need to prove that for agent i , $\mathbb{E}_{\mathbf{s}_{-i}|s_i}[v_i(\mathbf{s})]$ is non-decreasing in s_i . We first prove that for any function $g^{(n)}(\mathbf{s})$ non-decreasing in each signal, and for affiliated signals $\mathbf{s} = (s_1, \dots, s_n)$, $g^{(n-1)}(\mathbf{s}_{-j}) = \mathbb{E}_{s_j|\mathbf{s}_{-j}}[g^{(n)}(\mathbf{s})]$ is also non-decreasing in signal s_i for all $i \neq j$. This is because signal affiliation of \mathbf{s} implies that for any pair of i and j , fixing \mathbf{s}_{-ij} , s_i and s_j are also affiliated, which further implies that $s_j|s_i = x$ weakly first-order stochastically dominates (FOSD) $s_j|s_i = y$ for any $x > y$. As a result of the FOSD and the non-decreasing property of $g^{(n)}(\mathbf{s})$, we have

$$\begin{aligned}
g^{(n-1)}(s_i = x, \mathbf{s}_{-ij}) &= \mathbb{E}_{s_j|\mathbf{s}_{-ij}, s_i=x} \left[g^{(n)}(s_j, \mathbf{s}_{-ij}, s_i = x) \right] \\
&\geq \mathbb{E}_{s_j|\mathbf{s}_{-ij}, s_i=y} \left[g^{(n)}(s_j, \mathbf{s}_{-ij}, s_i = y) \right] \\
&\geq \mathbb{E}_{s_j|\mathbf{s}_{-ij}, s_i=y} \left[g^{(n)}(s_j, \mathbf{s}_{-ij}, s_i = y) \right] = g^{(n-1)}(s_i = y, \mathbf{s}_{-ij})
\end{aligned}$$

Therefore, $g^{(n-1)}(\mathbf{s}_{-j})$ is non decreasing in s_i for any $i \neq j$. By induction starting with $g^{(n)}(\mathbf{s}) = v_i(\mathbf{s})$, we can get $g^{(1)}(s_i) = \mathbb{E}_{\mathbf{s}_{-i}|s_i}[v_i(\mathbf{s})]$ is non-decreasing in s_i . □

C.2.3 Proof of Proposition 4.21

Proof. We first prove that $v_i(\mathbf{s}) = s_i + \beta \sum_{j \neq i} s_j$ satisfies the cursedness-monotonicity condition. We have

$$\begin{aligned} v_i(\mathbf{s}) - v_i^X(\mathbf{s}) &= \chi \left(v_i(\mathbf{s}) - \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [v_i^X(s_i, \bar{\mathbf{s}}_{-i})] \right) \\ &= \chi \left(\sum_{j \neq i} s_j - \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} \left[\sum_{j \neq i} \tilde{s}_j \right] \right) \\ &= \chi \left(\sum_{j \neq i} s_j - \mathbb{E}_{\bar{\mathbf{s}}_{-i}} \left[\sum_{j \neq i} \tilde{s}_j \right] \right) \end{aligned}$$

The last equation is due to signals are assumed to be independent in this section. Therefore, if for some \mathbf{s} , $v_i(\mathbf{s}) < v_i^X(\mathbf{s})$, then we have $\sum_{j \neq i} s_j < \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [\sum_{j \neq i} \tilde{s}_j]$ and thus for any $\mathbf{s}'_{-i} \leq \mathbf{s}_{-i}$, $\sum_{j \neq i} s'_j \leq \sum_{j \neq i} s_j < \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [\sum_{j \neq i} \tilde{s}_j]$, implying for any s'_i , $v_i(\mathbf{s}') - v_i^X(\mathbf{s}') < 0$, which completes the proof. Also, note that if the signals are not independent but positively affiliated in the sense that $\forall s'_i > s_i, \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s'_i} [\sum_{j \neq i} \tilde{s}_j] \geq \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [\sum_{j \neq i} \tilde{s}_j]$, the cursedness monotonicity still holds.

Second, we prove that $v_i(\mathbf{s}) = \max_{\mathbf{s}} \{s_i\}$ satisfies the cursedness-monotonicity condition. This is simply because that $\forall \mathbf{s}_{-i}$ and $s_i \in (\max_{j \neq i} \{s_j\}, \bar{s})$, $v_i(\mathbf{s}) - v_i^X(\mathbf{s}) < 0$. To see this, we have $v_i(\mathbf{s}) = s_i < \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [\max\{s_i, \max_{j \neq i} \{\tilde{s}_j\}\}] = \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [v_i(s_i, \bar{\mathbf{s}}_{-i})]$, implying $v_i(\mathbf{s}) - v_i^X(\mathbf{s}) < 0$. \square

C.2.4 Proof of Corollary 4.24

Proof. Because $v_i(\mathbf{s}) = \max_{\mathbf{s}} \{s_i\}$, we have $\forall \mathbf{s}_{-i}$ and $s_i \in (\max_{j \neq i} \{s_j\}, \bar{s})$, $v_i(\mathbf{s}) - v_i^X(\mathbf{s}) < 0$. To see this, we have $v_i(\mathbf{s}) = s_i < \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [\max\{s_i, \max_{j \neq i} \{\tilde{s}_j\}\}] = \mathbb{E}_{\bar{\mathbf{s}}_{-i}|s_i} [v_i(s_i, \bar{\mathbf{s}}_{-i})]$, implying $v_i(\mathbf{s}) - v_i^X(\mathbf{s}) < 0$. Therefore, for any threshold function $t_i(\cdot)$, which satisfies $t_i(\mathbf{s}_{-i}) \geq \max_{j \neq i} \{s_j\}$, $\forall \mathbf{s}_{-i}$ according to Lemma 4.8, we have $v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) \leq 0$, where the equality holds only if $t_i(\mathbf{s}_{-i}) = \bar{s}$. Therefore, for any \mathbf{s}_{-i} , if $t_i(\mathbf{s}_{-i}) < \bar{s}$, then we have $p_i(0, \mathbf{s}_{-i}) = \min\{0, v_i(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i}) - v_i^X(t_i(\mathbf{s}_{-i}), \mathbf{s}_{-i})\} < 0$. Since the max function satisfies the cursedness-monotonicity, Theorem 4.22 implies such $t_i(\cdot)$ cannot be supported by any deterministic, anonymous, C-EPIC-IR, and EPBB mechanism. Consequently, the only deterministic, anonymous, C-EPIC-IR, and EPBB mechanism is to either allocate to a bidder with $s_i = \bar{s}$ or never allocate, leading to zero allocation probability and thus zero social welfare and revenue. \square

C.2.5 Proof of Lemma 4.26

Proof.

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} \left[q(\mathbf{z}) \mid \sum z_j \geq d \right] &= \mathbb{E}_{\mathbf{z}_{-i}} \left[\mathbb{E}_{z_i | \mathbf{z}_{-i}} \left[q(\mathbf{z}) \mid z_i \geq d - \sum_{j \neq i} z_j \right] \mid \sum_{j \neq i} z_j \geq d - b \right] \\ &\geq \mathbb{E}_{\mathbf{z}_{-i}} \left[\mathbb{E}_{z_i | \mathbf{s}_{-i}} [q(\mathbf{z})] \mid \sum_{j \neq i} z_j \geq d - b \right] \end{aligned}$$

The equality is because when the supremum of the support of any z_i is b , then $\sum z_j \geq d$ if and only if $\sum_{j \neq i} z_j \geq d - b$ and $z_i \geq c - \sum_{j \neq i} z_j$. The inequality is because $v(\mathbf{z})$ is non-decreasing in z_i given any \mathbf{z}_{-i} . \square

C.2.6 Derivation of Equation 4.4

$$\begin{aligned} &\int_{\mathbf{b}_{-i}} f_{\sigma}(\mathbf{b}_{-i} | s_i) EU_i^{\chi}(\mathbf{b}, s_i; \sigma_{-i}) d\mathbf{b}_{-i} \\ &= \int_{\mathbf{b}_{-i}} \int_{\mathbf{s}_{-i}} f_{\sigma}(\mathbf{b}_{-i} | s_i) \left((1 - \chi) f_{\sigma}(\mathbf{s}_{-i} | \mathbf{b}_{-i}, s_i) u_i(\mathbf{b}, \mathbf{s}) + \chi f(\mathbf{s}_{-i} | s_i) u_i(\mathbf{b}, \mathbf{s}) \right) d\mathbf{s}_{-i} d\mathbf{b}_{-i} \\ &= \int_{\mathbf{b}_{-i}} \int_{\mathbf{s}_{-i}} \left((1 - \chi) f_{\sigma}(\mathbf{s}_{-i}, \mathbf{b}_{-i} | s_i) u_i(\mathbf{b}, \mathbf{s}) + \chi \tilde{f}(\mathbf{b}_{-i}, \mathbf{s}_{-i} | s_i) u_i(\mathbf{b}, \mathbf{s}) \right) d\mathbf{s}_{-i} d\mathbf{b}_{-i} \\ &= \int_{\mathbf{b}_{-i}} \int_{\mathbf{s}_{-i}} f_{\sigma}^{\chi}(\mathbf{b}_{-i}, \mathbf{s}_{-i} | s_i) u_i(\mathbf{b}, \mathbf{s}) d\mathbf{s}_{-i} d\mathbf{b}_{-i} = EU_i^{\chi}(b_i, s_i; \sigma_{-i}) \end{aligned}$$