Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2021)

Forecast Aggregation via Peer Prediction

Juntao Wang,[1] **Yang Liu,**[2] **Yiling Chen**[1]
[1] Harvard University
[2] UC Santa Cruz
juntaowang@g.harvard.edu, yangliu@ucsc.edu, yiling@seas.harvard.edu

Abstract

Crowdsourcing enables the solicitation of forecasts on a variety of prediction tasks from distributed groups of people. How to aggregate the solicited forecasts, which may vary in quality, into an accurate final prediction remains a challenging yet critical question. Studies have found that weighing expert forecasts more in aggregation can improve the accuracy of the aggregated prediction. However, this approach usually requires access to the historical performance data of the forecasters, which are often not available. In this paper, we study the problem of aggregating forecasts without having historical performance data. We propose using peer prediction methods, a family of mechanisms initially designed to truthfully elicit private information in the absence of ground truth verification, to assess the expertise of forecasters, and then using this assessment to improve forecast aggregation. We evaluate our peer-prediction-aided aggregators on a diverse collection of 14 human forecast datasets. Compared with a variety of existing aggregators, our aggregators achieve a significant and consistent improvement on aggregation accuracy measured by the Brier score and the log score. Our results reveal the effectiveness of identifying experts to improve aggregation even without historical data.

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 (Atanasov et al. 2016), predicting the replicability of social science studies (Liu et al. 2020), diagnosing skin lesions (Prelec, Seung, and McCoy 2017) and labeling training sets for machine classifiers (Liu, Peng, and Ihler 2012).

Aiming to more effectively leverage collective intelligence in forecasting, we focus on improving multi-task forecast aggregation in this paper. 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 (Atanasov et al. 2016), 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 (Miller, Resnick, and Zeckhauser 2005). While they are primarily developed for the purpose of forecast elicitation, Liu, Wang, and Chen (2020) and Kong (2020) 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 (2020) 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 paper, we explore the use of five recently proposed peer prediction mechanisms (Radanovic, Faltings, and Jurca 2016; Shnayder et al. 2016; Witkowski et al. 2017; Liu, Wang, and Chen 2020; Kong 2020) 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 em-

ploy 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 (Jose and Winkler 2008; Mannes, Larrick, and Soll 2012), the logit-mean aggregator, which is based on the idea of extremization of predictions (Allard, Comunian, and Renard 2012; Satopää et al. 2014; Baron et al. 2014), a statistical-inference-based aggregator (Liu, Peng, and Ihler 2012), and the minimal pivoting aggregator, which is based on "surprising popularity." (Prelec, Seung, and McCoy 2017; Palley and Soll 2019)

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.

We present additional information about the datasets, algorithms and experimental results in the full version of this paper (Wang, Liu, and Chen 2019).

Related Work

The research of forecast aggregation with no additional information dates back to early studies about simple aggregators, such as mean, median, and their trimmed variants (Galton 1907; Clemen 1989; Jose and Winkler 2008; Mannes, Larrick, and Soll 2012). These simple aggregators have robust empirical performance and are still widely adopted in practice. They sometimes give conservative predictions (e.g. lean toward 0.5 for binary events), hence extremifying techniques that bring predictions toward 0 or 1 are explored to mitigate this issue (e.g., Satopää et al. 2014; Baron et al. 2014). More recently, several statistical-inference-based methods have been developed to use cross-task information to improve aggregation accuracy further when there exist multiple a priori similar forecasting tasks (e.g., Liu, Peng, and Ihler 2012; Lee and Danileiko 2014; McCoy and Prelec 2017). A new trend is to ask forecasters for slightly more information, i.e., their predictions about others' predictions, and use this additional information to improve accuracy (e.g., Prelec, Seung, and McCoy 2017; Palley and Soll 2019). We select representative aggregators from each of these categories as baselines for our aggregators.

Our construction of PAS is derived from a family of mechanisms collectively called peer prediction. The term peer prediction was coined up by (Miller, Resnick, and Zeckhauser 2005) and the literature has been further developed by a series of studies (e.g., Prelec 2004; Miller, Resnick, and Zeckhauser 2005; Shnayder et al. 2016; Radanovic, Faltings, and Jurca 2016; Witkowski et al. 2017; Liu, Wang, and Chen 2020; Kong 2020).

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 $F : ([0, 1] \cup \{\emptyset\})^{|\mathcal{M}|\times|\mathcal{N}|} \to [0, 1]^{|\mathcal{M}|}$, 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 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 single-task aggregators as building blocks: the mean (Mean) and the logit-mean (Logit) (Satopää et al. 2014). Mean has empirically proved

[1] Our methods and results can be extended to multi-outcome events. Please refer to the full version of this paper for details.

robustness (Jose and Winkler 2008), while Logit extremizes the predictions of Mean and demonstrates significantly higher accuracy on some human forecast datasets (Satopää et al. 2014). 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 $(w_j)_{j\in\mathcal{N}_i}$,

- $F_i^{\text{Mean}}(\mathbf{p}_i) = \sum_{j\in\mathcal{N}_i} w_j p_{i,j}$,
- $F_i^{\text{Logit}}(\mathbf{p}_i) = \text{sigmoid}\left(\frac{\alpha}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} w_j \text{logit}(p_{ij})\right)$ with $\alpha = 2$ recommended by Satopää et al. (2014).

Logit first maps probabilistic predictions into the log-odds space using the logit function (the inverse sigmoid function). It then takes weighted mean and applies a scaling factor to further extremize the predictions. Finally, it maps the predictions back into probabilities using the sigmoid function.

Prediction accuracy metrics. The accuracy of forecasts is typically evaluated using the strictly proper scoring rules (SPSR) (Gneiting and Raftery 2007). 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 score[2]: $S^{\text{Brier}}(\hat{q}_i, Y_i) = 2(\hat{q}_i - Y_i)^2$.
- 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 score refers to a higher accuracy. The Brier score ranges from 0 to 2. The log score ranges from 0.1 to 4.61.[3] Predicting 0.5 always receives a Brier score of 0.5 and a log score of 0.69.

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 1.

In short, and in different to the true accuracy that is evaluated against the ground truth, PAS assess a prediction against only the other agents' predictions. Thus, PAS can be applied to broader crowdsourcing forecasting scenarios, requiring only knowing multiple forecasts for each task. Formally, a peer assessment score on an event set $\mathcal{M}$ and an agent set $\mathcal{N}$ is a scoring function $R : ([0,1] \cup \{\emptyset\})^{|\mathcal{M}|\times|\mathcal{N}|} \to [0,1]^{|\mathcal{N}|}$ 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.

[2]We adopt the same formula for the Brier score as in the Good Judgment Project (e.g., Atanasov et al. 2016)

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

Algorithm 1 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.

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 paper in Algorithm 1. We introduce the five peer prediction mechanisms (DMI, CA, PTS, SSR, and PSR) used in Step 1 in the next section. We choose the ranking & selection scheme rather than the softmax weight in Step 2, 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. These two schemes show similar performance with best-tuned hyper-parameters in our experiments. In Step 3, we use Mean and Logit as the base aggregator.

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 $R : ([0,1] \cup \emptyset)^{|\mathcal{M}\times\mathcal{N}|} \to [0,1]^{|\mathcal{N}|}$ that maps forecasters' prediction profile P to a reward R_j for each forecaster j. $R(\cdot)$ is carefully designed so that an agent's expected reward based on her belief about others' reports (formed by her private signal) will be maximized when she reports truthfully.

The core intuition of 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

Abbr.	Full name	Abbr.	Full name
DMI	Determinant mutual information mechanism	SPSR	Strictly proper scoring rules
CA	Correlated agreement mechanism	PAS	Peer assessment sores
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 1: The main abbreviations and the corresponding full names used in this paper

receive a better score in expectation. This property makes them ideal candidates to serve as PAS. While most peer prediction scores do not necessarily reflect prediction accuracy, we selectively review five peer prediction mechanisms and provide theoretical support for using them as PAS, i.e., the scores of these five mechanisms each correlate with accuracy of agents according to some metric.

These mechanisms require two *assumptions* 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, 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.

Mechanisms Recovering the Strictly Proper Scoring Rules (SPSR)

SPSR are natural reward schemes to incentivize truthful reporting (Gneiting and Raftery 2007). They can be oriented in a way that a higher score corresponds to higher accuracy. But they require to know the ground truth of predicted events. Surrogate scoring rules (SSR) (Liu, Wang, and Chen 2020) and proxy scoring rules (PSR) (Witkowski et al. 2017) 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 (2020) 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

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

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

Proxy scoring rules (PSR). In constrast 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., $R^{\mathsf{PSR}}_{i,j}(p_{i,j}, \hat{Y}_i) = S(p_{i,j}, \hat{Y}_i)$. Witkowski et al. (2017) 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. (2017) recommended using an extremized mean prediction as the proxy when there is no explicit unbiased proxy of ground truth available.

Mechanisms Rewarding the Correlation

Determinant mutual information mechanism (DMI) (Kong 2020), correlated agreement (CA) (Shnayder et al. 2016), and peer truth serum (PTS) (Radanovic, Faltings, and Jurca 2016) are three mechanisms that reward agents based on their forecasts' correlation to their peers'. Their core ideas are 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. To compute the score for an agent j, all the three mechanisms first estimate 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 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 estimated empirically as

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

where $\mathcal{M}_{j,k}$ is the subset of forecasting tasks answered by both agents. We use $\hat{D}^{j,k} = \left(\hat{d}^{j,k}_{u,v}\right)_{u,v\in\{0,1\}}$ 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 (Kong 2020). 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:

$$R_j^{\mathsf{DMI}} = \eta \det(D') \cdot \det(D''), \tag{1}$$

where η is a normalization coefficient.

Correlated agreement (CA). CA rewards an agent j by

$$R_j^{\mathsf{CA}} = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} |\hat{d}^{j,k}_{u,v} - \hat{d}^{j}_{u} \cdot \hat{d}^{k}_{v}|, \tag{2}$$

where $\hat{d}^j_u = \sum_{v\in\{0,1\}} \hat{d}^{j,k}_{u,v}$ 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 $\hat{d}^{j,k}_{u,v}$) and the matching probability caused by pure chance (represented by $\hat{d}^j_u \cdot \hat{d}^k_v$).

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

$$R_j^{\mathsf{PTS}} = \hat{d}^{j,k}_{0,0}/\bar{p}_{-j,0} + \hat{d}^{j,k}_{1,1}/\bar{p}_{-j,1}. \tag{3}$$

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 1. *1. Under Assumptions A1 and A2, SSR ranks the agents in the order of their mean SPSR that SSR is built upon asymptotically ($|\mathcal{M}|, |\mathcal{N}| \to \infty$).*

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 ($|\mathcal{M}| \to \infty$).

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 2. *Under Assumptions A1 and A2, and assuming agents report truthfully, the expected rewards of DMI, CA, PTS reflect certain accuracy measures of agents. In particularly,*

1. DMI ranks the agents in the order of their reports' squared determinant mutual information (Kong 2020) w.r.t. the ground truth asymptotically ($|\mathcal{M}|, |\mathcal{N}| \to \infty$).

2. CA ranks the agents in the order of their reports' determinant mutual information w.r.t. the ground truth asymptotically ($|\mathcal{M}|, |\mathcal{N}| \to \infty$).

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 ($|\mathcal{M}|, |\mathcal{N}| \to \infty$), 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.

Item 1 in Proposition 2 follows straightforwardly from Theorem 6.4 of Kong (2020). We give the proofs for the items 2 and 3 in the full version of this paper. 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.

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 focus on binary events here and provide results for multi-outcome events in the full version.

Experiment Setup

Datasets. Our 14 test datasets consist of 4 datasets from the Good Judgement Projects (GJP) collected from 2011 to 2014 (Good Judgment Project 2016), 3 datasets from the Hybrid Forecasting Competition (HFC) of varied populations (IARPA 2019), and 7 MIT datasets (Prelec, Seung, and McCoy 2017). These datasets vary in several dimensions, including dataset size, sparsity, topics, collecting environment, and participants' performance. Together they offer a rich environment 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

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

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.

Our paper focuses 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 of the 14 datasets after filtering in Table 2. More details about datasets can be found in the full version of this paper.

Benchmarks. In addition to the two base aggregators, Mean and Logit, which are widely-used in the minimal-information aggregation setting (Satopää et al. 2014; Jose and Winkler 2008), 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 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 the full version. Other sophisticated methods in this category include the cultural consensus model (Oravecz, Vandekerckhove, and Batchelder 2014), the cognitive hierarchy model (Lee and Danileiko 2014), and the multi-task statistical surprising popularity method (McCoy and Prelec 2017)[5]. We will also compare to the performance these aggregators reported by McCoy and Prelec (2017) on the MIT datasets.
- *Surprising-popularity-based methods* are not minimal-information aggregators, but they represent a new trend of forecast aggregation (Prelec, Seung, and McCoy 2017; Palley and Soll 2019). 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 (Prelec, Seung, and McCoy 2017). 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 hyperparameter—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 hyperparameter 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 VI aggregator as the proxy used in PSR.[6] All these aggregators are described in Algorithm 1.

Smaller but Smarter Crowd

Before we delve 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.

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

[5]This aggregator combines both inference and surprising-popularity.

[6]We also tested using proxies (e.g, the mean of agents' predictions and the extremized mean (Witkowski et al. 2017)) in PSR, while using VI as the proxy gives us the best result.

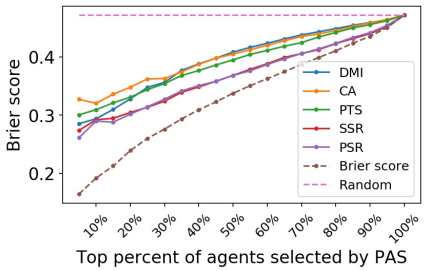


Figure 1: The averages of the true mean Brier score of top forecasters selected by the five PAS and by the true Brier score.

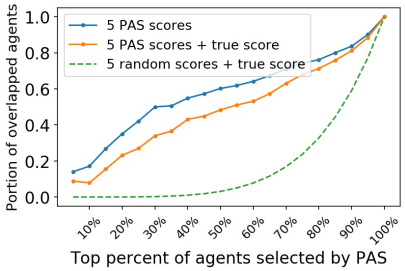


Figure 2: The portions of overlapped agents, who are simultaneously selected by all of the five PAS and the true score.

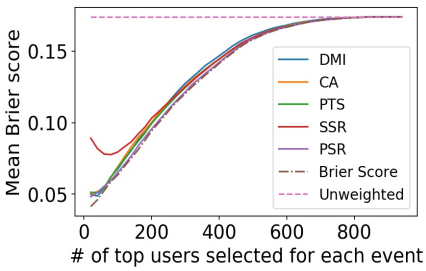


Figure 3: The Brier scores of the mean-based PAS-aided aggregators varying with the numbers of top agents selected on dataset G2.

Fig. 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. 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. 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 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, 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).

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" (Mellers et al. 2015) and (Goldstein, McAfee, and Suri 2014)). Our contribution is to demonstrate that we can identify this set of smarter forecasters using only their prediction information.

Forecast Aggregation Performance

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 minimal-information forecast aggregation.
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 and Table 4. Table 3 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.

Table 4 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 4 and more in the full version of this paper). Next, we give a more detailed review of the experimental results.

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: The mean Brier scores (with range [0, 2]) 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**.

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), and the Logit aggregator, which extremizes the mean prediciton, 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 4, the PAS-aided aggregators outperform the Mean and the Logit aggregators with statistical significance on most datasets. Dataset H2 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 number[7] of predictions are collected from each forecaster to compute the PAS).

[7]We will discuss this number in the next section.

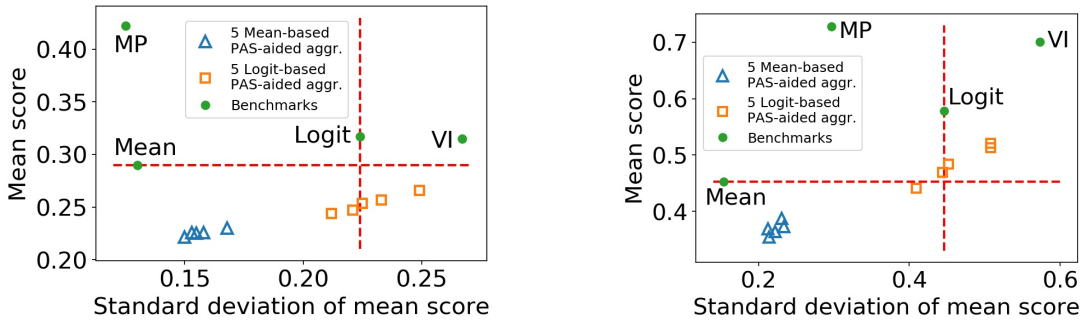
PAS-aided aggregators vs. VI and other inference-based methods. We notice that although VI ranks the worst in many datasets, the number of datasets on which VI statistically underperforms each PAS-aided aggregator is smaller than those numbers of the other benchmarks (Table 4). This is because VI tends 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). 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 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 (2017) 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 (See the full version of this paper for concrete data). 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 4 still shows a salient advantage of PAS-aided Mean aggregators over MP. This result implies that when forecasters make predictions on multiple events,

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 4: 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**.



(a) Brier score

(b) Log scoring rule

Figure 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.

the cross-task information leveraged by the PAS scores may be more powerful in facilitating aggregation than the additionally solicited information used in MP.

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. 4 (Concrete data can be found in the full version of this paper). 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 fore-

[8]The 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.

cast aggregation settings, where only a small set of forecasts is available with no ground truth yet revealed. We present the details of this experimentin the full version of this paper.

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. 2.

Discussion and Future Directions

This paper 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 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 (Jose and Winkler 2008; Satopää et al. 2014). 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 paper, 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 crowsourcing 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 (Canonico, Flathmann, and McNeese 2019), workflow design (Lin, Mausam, and Weld 2012), promoting interactions (Bigham, Bernstein, and Adar 2015) and AI algorithms (Weld, Lin, and Bragg 2015), are also present. From another perspective, the human-machine computation systems are now also developed for many complex tasks, such as image segmentation (Song et al. 2018) and article editing (Zhang, Verou, and Karger 2017). 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.

Acknowledgements

This research is supported in part by National Science Foundation (NSF) under grants CCF-1718549, IIS-2007951, and IIS-2007887, and the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific) under Contract No. N66001-19-C-4014. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, DARPA, SSC Pacific or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Allard, D.; Comunian, A.; and Renard, P. 2012. Probability aggregation methods in geoscience. *Mathematical Geosciences* 44(5): 545–581.

Atanasov, P.; Rescober, P.; Stone, E.; Swift, S. A.; Servan-Schreiber, E.; Tetlock, P.; Ungar, L.; and Mellers, B. 2016. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science* 63(3): 691–706.

Baron, J.; Mellers, B. A.; Tetlock, P. E.; Stone, E.; and Ungar, L. H. 2014. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis* 11(2): 133–145.

Bigham, J. P.; Bernstein, M. S.; and Adar, E. 2015. Human-computer interaction and collective intelligence. *Handbook of collective intelligence* 57.

Canonico, L. B.; Flathmann, C.; and McNeese, N. 2019. Collectively intelligent teams: Integrating team cognition, collective intelligence, and AI for future Teaming. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, 1466–1470. SAGE Publications Sage CA: Los Angeles, CA.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5(4): 559–583.

Galton, F. 1907. Vox populi. *Nature* 75: 450–451.

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

Goldstein, D. G.; McAfee, R. P.; and Suri, S. 2014. The wisdom of smaller, smarter crowds. In *ACM EC*, 471–488. ACM.

Good Judgment Project. 2016. GJP Data. https://doi.org/10.7910/DVN/BPCDH5. doi:10.7910/DVN/BPCDH5. Accessed: 2021-09-07.

IARPA. 2019. Hybrid Forecasting Competition. https://www.iarpa.gov/index.php/research-programs/hfc. Accessed: 2021-09-07.

Jose, V. R. R.; and Winkler, R. L. 2008. Simple robust averages of forecasts: Some empirical results. *International journal of forecasting* 24(1): 163–169.

Kong, Y. 2020. Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks. In *SODA*, 2398–2411. SIAM.

Lee, M. D.; and Danileiko, I. 2014. Using cognitive models to combine probability estimates. *Judgment and Decision Making* 9(3): 259.

Lin, C.; Mausam, M.; and Weld, D. 2012. Dynamically Switching between Synergistic Workflows for Crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Liu, Q.; Peng, J.; and Ihler, A. T. 2012. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, 692–700.

Liu, Y.; Gordon, M.; Wang, J.; Bishop, M.; Chen, Y.; Pfeiffer, T.; Twardy, C.; and Viganola, D. 2020. Replication Markets: Results, Lessons, Challenges and Opportunities in AI Replication. *arXiv preprint arXiv:2005.04543* .

Liu, Y.; Wang, J.; and Chen, Y. 2020. Surrogate scoring rules. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 853–871.

Mannes, A. E.; Larrick, R. P.; and Soll, J. B. 2012. The social psychology of the wisdom of crowds. In *J. I. Krueger (Ed.), Social Judgment and Decision Making*, 227–242. Psychology Press.

McCoy, J.; and Prelec, D. 2017. A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778* .

Mellers, B.; Stone, E.; Murray, T.; Minster, A.; Rohrbaugh, N.; Bishop, M.; Chen, E.; Baker, J.; Hou, Y.; Horowitz, M.; et al. 2015. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3): 267–281.

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

Oravecz, Z.; Vandekerckhove, J.; and Batchelder, W. H. 2014. Bayesian cultural consensus theory. *Field Methods* 26(3): 207–222.

Palley, A. B.; and Soll, J. B. 2019. Extracting the Wisdom of Crowds When Information is Shared. *Management Science* 65(5): 2291–2309.

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

Prelec, D.; Seung, H. S.; and McCoy, J. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541(7638): 532.

Radanovic, G.; Faltings, B.; and Jurca, R. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM TIST* 7(4): 48.

Satopää, V. A.; Baron, J.; Foster, D. P.; Mellers, B. A.; Tetlock, P. E.; and Ungar, L. H. 2014. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting* 30(2): 344–356.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed truthfulness in multi-task peer prediction. In *ACM EC*, 179–196. ACM.

Song, J. Y.; Fok, R.; Lundgard, A.; Yang, F.; Kim, J.; and Lasecki, W. S. 2018. Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance. In *23rd International Conference on Intelligent User Interfaces*, 559–570.

Wang, J.; Liu, Y.; and Chen, Y. 2019. Forecast aggregation via peer prediction. *arXiv preprint arXiv:1910.03779* .

Weld, D. S.; Lin, C. H.; and Bragg, J. 2015. Artificial intelligence and collective intelligence. *Handbook of Collective Intelligence* 89–114.

Witkowski, J.; Atanasov, P.; Ungar, L. H.; and Krause, A. 2017. Proper proxy scoring rules. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhang, A. X.; Verou, L.; and Karger, D. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2082–2096.