Social Learning and the Accuracy-Risk Trade-off in the Wisdom of the Crowd

DHAVAL ADJODAH, YAN LENG, and SHI KAI CHONG, Media Lab, Massachusetts Institute of Technology
P. M. KRAFFT, Oxford Internet Institute, University of Oxford
ESTEBAN MORO, Universidad Carlos III de Madrid, 28911 Madrid, Spain
ALEX PENTLAND, Media Lab, Massachusetts Institute of Technology

How do we design and deploy crowdsourced prediction platforms for real-world applications where risk is an important dimension of prediction performance? To answer this question, we conducted a large online Wisdom of the Crowd study where participants predicted the prices of real financial assets (e.g. S&P 500). We observe a Pareto frontier between accuracy of prediction and risk, and find that this trade-off is mediated by social learning i.e. as social learning is increasingly leveraged, it leads to lower accuracy but also lower risk. We also observe that social learning leads to superior accuracy during one of our rounds that occurred during the high market uncertainty of the Brexit vote. Our results have implications for the design of crowdsourced prediction platforms: for example, they suggest that the performance of the crowd should be more comprehensively characterized by using *both* accuracy and risk (as is standard in financial and statistical forecasting), in contrast to prior work where risk of prediction has been overlooked.

Keywords: Crowdsourcing, Wisdom of the Crowd, Social Learning, Bayesian models, Risk

1 INTRODUCTION

Crowdsourced prediction systems through approaches such as the Wisdom of the Crowd [37, 43] or Prediction Markets [6, 49] have been successful in a range of domains such as predicting the reproducibility of scientific research [30], estimating the caloric content of food items [10], and predicting stock market prices [92].

From the perspective of platform designers who want to deploy such prediction systems at scale for real-life applications, one under-studied aspect of the performance of the crowd is the risk of the prediction of the crowd: over a collection of separate prediction tasks, it is important to measure the variance of the average accuracy — in addition to the average accuracy of the crowd in predicting the realized outcome correctly — as it provides a measure of the risk of the crowdsourced prediction system. This is because the crowd might be accurate on average (over these separate prediction tasks) but have a high variance in accuracy across tasks (i.e. for some prediction tasks, the accuracy might be quite poor), making it risky to employ this system for prediction. This view is standard in statistical [28, 40, 56] and financial [41, 58, 85] forecasting applications, but it not commonly studied in the literature on the Wisdom of Crowds and many adjacent areas of collective intelligence research.

To demonstrate the importance of measuring risk in Wisdom of Crowds platforms, we deployed a large crowdsourced prediction task in which we measure the accuracy and variance of the prediction system over 7 independent rounds of predicting financial asset prices (S&P 500, gold and WTI Oil).

1.1 Contributions

- We observe a Pareto frontier [38, 83] between accuracy and risk: as the average accuracy of the crowd over the different prediction rounds increases, so does the variance in the crowd's

Authors' addresses: Dhaval Adjodah, dval@mit.edu; Yan Leng; Shi Kai Chong, Media Lab, Massachusetts Institute of Technology; P. M. Krafft, Oxford Internet Institute, University of Oxford; Esteban Moro, Universidad Carlos III de Madrid, 28911 Madrid, Spain; Alex Pentland, Media Lab, Massachusetts Institute of Technology.

predictive accuracy. We further observe that this trade-off is mediated by the amount of social learning—i.e., the extent to which users pay attention to each other's judgments.

- We deployed one of our prediction tasks just before the Brexit vote during which there was a great deal of market uncertainty [93], and we observe that during such uncertain times social learning leads to higher accuracy.
- While modeling the belief update process of participants using Bayesian Models of Cognition [46, 47] to estimate their amount of social learning, we observe that our participants exhibit the attribute substitution heuristic of human decision-making [59], whereby a complicated problem is solved by approximating it with a simpler, less accurate model. We also observe people's preference to learn from social information rather than from non-social information.
- We are releasing our large dataset[1] which is the first dataset, to the authors' knowledge, that records not only participants pre- and post-exposure predictions, but also both the social and non-social information they were exposed to in a large-scale social Wisdom of the Crowd domain.

2 RELATED WORK

2.1 Computer-supported cooperative work (CSCW)

In the present work we study how platform design mediates collective intelligence in the context of a Wisdom of the Crowd task. The study of collective intelligence—the ability of groups to come together to solve problems collectively—has long been a key area of research in computer-supported cooperative work (CSCW). Within the CSCW community, the interest in this area of research has centered on how to improve collaborative work through the lens of collective intelligence research. Building on early work of Malone, Grosz, and colleagues [48, 81], CSCW researchers have studied factors that influence collective intelligence [64], platforms for promoting collective intelligence [4, 24, 82], frameworks for understanding collective intelligence [45], and phenomena of collective intelligence in digital settings (e.g., [55, 109]).

More recently, a new strand of work has looked into how to deliver high-quality results for complex real-world applications: for example a system inspired by distributed computing infrastructures has allowed crowds to work with thousands of workers and tasks by accounting for human factors [27], a hybrid system can overcome the issue of initial low-fidelity data [39], and new approaches has been presented that allow crowds to build datasets that approximate complex machine learning data distributions dynamically [17].

Similarly, we believe that in order to deploy Wisdom of the Crowd systems at scale — for example, as in our task, the prediction of financial asset prices — their performance must be more comprehensively characterized.

2.2 Wisdom of the Crowd

One popular domain within the collective intelligence literature is the Wisdom of the Crowd [37, 43], where participants (typically referred to as the 'crowd') are asked to make predictions of a certain quantity, such as the future price of an asset on the stock market [92] or the caloric content of food items [10]. It has been found that the central tendency of the crowd (such as the average, the median, or other aggregates) – used as a measure of collective belief – can be quite accurate [37, 43], where accuracy is defined as the error between the crowd's aggregate prediction and the "ground truth" (here, the realized future price of the asset).

There is a rich literature aiming to optimize the accuracy of the crowd, such as by recalibrating predictions against systematic biases of individuals [107], selecting participants who are resistant

[1] Data and code are available here.

to social influence [78], rewiring the network topology of information-sharing between subjects [2, 10], and optimally allocating tasks to individuals [61].

Overall, it has been hypothesized that crowds can be highly accurate in aggregate because people's individual biases are typically not correlated [92], and, therefore, cancel out on average. However, when participants in the crowd start sharing information, such as through social learning, their beliefs can become correlated and therefore degrades the accuracy of the crowd. In the next section, we discuss the impact of social learning on accuracy.

2.3 Social Learning

One of the promising avenues for advancement in the CSCW field from the science of collective intelligence is the effect of social learning—the use of information about other people's decisions to make one's own — on the collective performance of crowdsourcing systems such as the Wisdom of the Crowd.

Several threads of research in CSCW and adjacent areas examine the importance of social recommendation on engagement with web content (e.g., [13, 71, 72, 106]). From a design perspective, the relationship between social observation and collective intelligence is especially interesting because crowdsourcing platforms can often be optimized to be more social [16, 73], even though research on collective prediction problems is divided on the effect of social learning on collective performance.

On one hand, prior work has shown that exposure to social information can lead to degraded performance in aggregate guesses [75, 87, 107]: increasing the strength of social influence has been shown to increase inequality [101], selecting the predictions of people who are *resistant* to social influence has been shown to have improved collective accuracy [78], the influence of influential peers has been theoretically shown to prevent the group from converging on the true estimate [107], and exposure to the confidence levels of others has been shown to influence people to change their predictions for the worse [86].

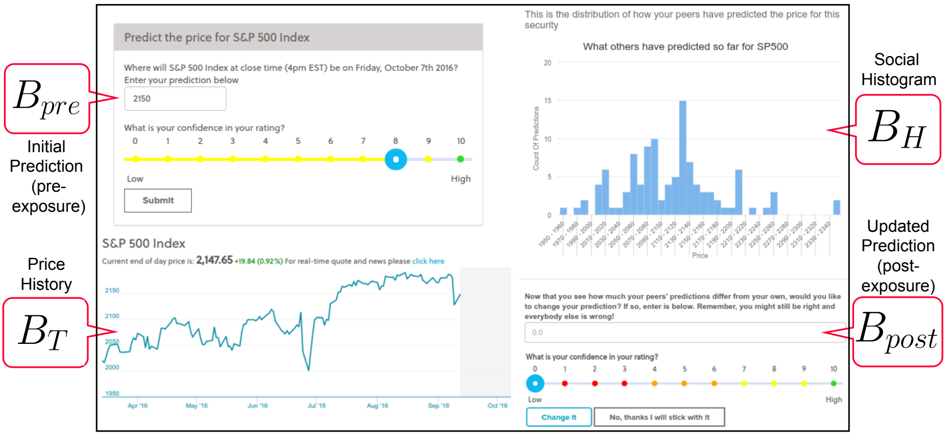
On the other hand, social learning has also been shown to lead to groups outperforming their best individuals when they work separately [2], a collective intelligence factor has been shown to predict team performance better than the maximum intelligence of members of the team [115], and human-inspired social communication between agents has been shown to improve collective performance in optimization algorithms [1, 70].

Therefore, the role of social learning in collective performance is still being understood, but the question of how social learning impacts collective intelligence has great potential to impact our understanding of platform and interface design in CSCW settings.

2.4 Risk

The prior work mentioned above in both the social learning and the Wisdom of the Crowd literatures have focused on maximizing the average accuracy of groups with little regard to the variance (risk) of the predictions. It has been proven theoretically [28, 56] and observed in a variety of statistical applications [36, 40] that there is a fundamental trade-off between accuracy and risk. This means that, for any prediction system, risk will be ever-present and that maximizing accuracy will lead to increasing risks, i.e. the performance of the system will always exist within a pre-defined Pareto frontier [38, 83].

In practice, treating risk and accuracy as equally important for prediction is standard in statistical [28, 40, 56] and financial [41, 58, 85] forecasting applications and literature because it allows system designers to carefully calibrate their strategies to risk — for example to hedge for probable losses [7, 14, 15, 22, 104]. Furthermore, in a meta-study of 105 forecasting papers, 102 of them support prioritizing for lower risk [5].



Predict the price for S&P 500 Index

Where will S&P 500 Index at close time (4pm EST) be on Friday, October 7th 2016?
Enter your prediction below

2150

What is your confidence in your rating?

S&P 500 Index

Current end of day price is: 2,147.65 +19.84 (0.92%) For real-time quote and news please click here

This is the distribution of how your peers have predicted the price for this security

What others have predicted so far for SP500

Now that you see how much your peers' predictions differ from your own, would you like to change your prediction? If so, enter is below. Remember, you might still be right and everybody else is wrong!

0.0

What is your confidence in your rating?

Fig. 1. An annotated screenshot of how data was collected: the pre-exposure prediction B_{pre} is shown first, followed by the social histogram B_H and the price history B_T. Finally, the updated prediction B_{post} is collected. The ground truth of the asset's final closing price will be V (not shown here, realized at the end of the round).

At the individual level, there is strong evidence that people preferentially optimize for risk instead of accuracy in a variety of domains [54]: cognitively, people have been observed to manifest decision heuristics [60] to be conservative in the face of uncertainty [51, 116]. For example, rice farmers have been observed not to adopt significant harvest improvement technology because of the risk of it failing once and causing significant family ruin [11]. Evolutionarily, risk aversion has been shown to emerge when rare events have a large impact on individual fitness [51]. Theoretically, when considering more realistic decision-making scenarios, such as repeated trials (as opposed to one-time bets), it has been shown that accounting for risk is critical to understanding the dynamics of how people make decisions [96].

Given the importance of risk both at the individual level and as an important metric in a range of forecasting applications, it is important to study it within the context of crowdsourced prediction platforms.

3 EXPERIMENTAL DESIGN

The main goal of our study is to investigate if risk is an important dimension of crowdsourced prediction platforms that is required to fully characterize the performance of the crowd, and whether social learning affects risk. To do so, we hypothesize that a Pareto frontier exists between risk and accuracy — i.e. that there is a trade-off between risk and accuracy of prediction across several prediction rounds — and that social leaning impacts this trade-off.

To test this hypothesis, we need a dataset with the following requirements:

- Predictions made of complex and difficult to predict phenomena so that our results are applicable to the real-world platform applications being studied within the CSCW community.
- A standard against which we can compare our dataset to judge its external validity.
- A large number of predictions for statistical significance, with both pre- and post-exposure predictions collected.

- The exact social and non-social information each user was exposed to after their initial pre-exposure prediction so that we can later model how different types of information influenced them in updating their belief into their post-exposure prediction.
- Predictions over many separate independent prediction rounds.
- At least one prediction round that occurred during a period of high uncertainty to understand if our findings change in abnormal settings.

Given the above requirements, we designed the experimental procedure detailed below: we recruited a total of 2,037 participants over seven prediction rounds to predict the future prices of financial assets (the S&P 500, WTI Oil, and gold prices) during seven separate consecutive 3-week rounds over the span of 6 months, resulting in 9,268 predictions (i.e. 4,634 prediction pairs/sets). We focused on predicting financial prices as doing so is a hard prediction problem [12, 33]. Our participants were mid-career financial professionals with years of financial experience. Our participants consented to their data being used in this study and we obtained prior IRB approval.

3.1 Data Collection

As shown in the screenshot of the user interface in Fig. 1, we designed the data collection process as follows: every time a user makes a prediction of an asset's future price through our platform, the following prediction set comprising B_{pre}, B_H, B_T and B_{post} is collected:

- A "pre-exposure" belief prediction B_{pre}, which is independent of any social information. For example, a user might show-up on the platform and predict that the closing price of the S&P 500 to be $2,001 on June 24^{th}, 2016.
- The predictions B_H within the social information histogram shown to each user after each initial prediction. Additionally, we display a 6-month time-series of the asset's price B_T up to this point.
- The revised "post-exposure" prediction B_{post}. For example, after seeing the social histogram and asset price history, a user might update their belief to $2,201. Since the real price (the ground truth V) ended up being $2,037.41, this user became more accurate after information exposure (they went from $2,001 to $2,201).

We ensure that the "pre-exposure" prediction is made before any social information is shown. We present a unique histogram for every new prediction (as it is built using past predictions up to this point), as well as a unique price history time series (as it shows the 6-month price data up to the time of prediction). We require all participants to make a post-exposure prediction.

During each round, participants made a prediction of the same asset's closing price for the same final day of the round. We use the round's last day's closing market price as our measure of ground truth. We carefully instrumented the social and non-social information that our participants were exposed to, and collected their predictions before and after exposure to this information. We also deployed one of our rounds during a high uncertainty period to understand if variance reduction strategies allows the crowd to be resistant to risk.

Whenever we predict a final closing price, we only use user prediction data up to the week before the day of prediction (i.e., we don't use any data during the last week of the round) so that our predictions are not too easy. One of our rounds of prediction happened to end the day of the Brexit vote, which means that we have prediction data during a particularly volatile market period [93]. We chose the start and end dates of each round so that the expiry dates of the asset's underlying *futures* would not affect the price of both the asset and its futures. Financial data (asset and futures prices) is obtained through Barchart.com's API.

3.2 External Validity of Data Collected

As shown in Table 1, our participants are collectively quite accurate — in agreement with past Wisdom of the Crowd studies [43, 92] — indicating that their predictions are being thoughtfully elicited: we observe that the crowd is generally doing more than just linear extrapolation (we test a model where we simply extrapolate prices in time using a static slope) as the error of such a model is higher. Additionally, we observe that the crowd's mean prediction error[2] is much less than the overall price change of the asset for the 3-week prediction period.

Interestingly, the crowd's collective prediction over each round tracks (and sometimes outperforms) the futures of each asset being predicted (we calculate the futures error as the difference between the futures price and the asset price). Because futures prices are commonly used a measure of the global market's prediction of the price asset [3, 35], the fact that the crowd's performance is on-par with the futures prices indicates that our dataset is externally valid.

	Prediction Round						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Asset	**SP500**	**WTI Oil**	**Gold**	**SP500**	**SP500**	**SP500**	**SP500**
Grounth Truth ($)	2037.41	45.95	1335.80	2153.74	2126.41	2191.95	2262.53
Num. Prediction Sets	284	207	134	1174	925	1441	469
Price Change (%)	4.01	11.03	3.63	1.77	1.75	2.24	3.56
Linear Extrapolation Error (%)	6.66	16.4	1.26	1.62	2.75	0.75	3.10
Crowd Mean Error (%)	2.22	4.95	0.46	0.84	0.58	3.20	2.40
Futures Mean Error (%)	2.03	3.05	0.94	0.38	0.40	0.48	1.50

Table 1. Summary of data collected. Our crowd is accurate, and sometimes even outperforms the futures underlying the asset. Predictions made by participants are more accurate than simple linear extrapolation.

3.3 Brexit Data

We deployed one of our experiments right before the Brexit vote during which there was a lot of market uncertainty [93]: the prediction round starting on June 1st 2016 ended on June 24th 2016, the day of the Brexit vote, and participants were predicting the price of the S&P 500, an asset sensitive to global events [20, 26]. We collected 284 prediction sets during the first 2 weeks of the round, and 52 sets in the last week during which the global financial market first overestimated then underestimated the final price of the S&P 500 asset leading to a 3.7% crash, as shown in the candlestick plot in Fig. 2.

4 METHODS

Our hypothesis is that a Pareto frontier exists between risk and accuracy — i.e. that there is a trade-off between risk and accuracy of prediction across several prediction rounds — and that social learning impacts this trade-off.

In order to test this hypothesis, we need to select subsets of predictions based on how much social learning impacted the way these predictions were made. We can then calculate the risk and accuracy of these subsets and study the impact of social learning on these two aspects of performance. Given that participants were exposed to *both* social and non-social learning, we cannot directly separate predictions based on whether they were updated through social learning or not. We therefore need a way to *estimate* how much social learning was used by participants for each prediction, and then a way to select subsets of prediction based on their amount of social learning. Using these subsets,

[2] Higher relative errors in round 2 are an artifact of the fact that a few dollars' error on the lower price of WTI Oil seems like a higher error compared to same absolute error on the higher prices of the other assets (about $45 per share for WTI Oil compared to $2100 for the S&P 500 and $1300 for gold prices).

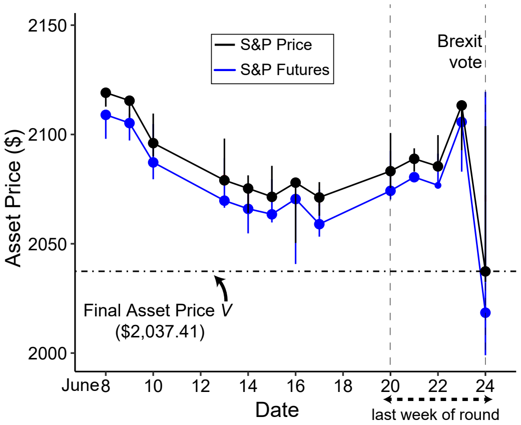


Fig. 2. The close, low, and high price of the asset and its underlying futures are shown as candlestick plots. The asset and futures overestimated the price and then crashed during the last week.

we then want to measure how much social learning impacted accuracy in this subset *compared* to the full set of predictions (namely, the Wisdom of the Crowd) - we calculate not just the accuracy of a subset but its improvement relative to the original Wisdom of the Crowd (average of the full set of predictions).

From the perspective of platform designers who want to be able to select predictions based on required levels of accuracy or risk (e.g. to fit a certain portfolio of risk), it is important to measure improvement of subsets relative to the full collection of predictions. This is because, currently, platform designers only have access to one global measure of risk and accuracy — that of the whole set of predictions (when there is no subset filtering). To demonstrate that selecting subsets of predictions can lead to significant *improvements* in accuracy and risk, we therefore need to calculate these improvements.

For our results to support our hypothesis, we need to find a statistically significant trade-off between risk and accuracy over subsets with varying amounts of social learning.

The structure of this section is as follows:

- In section 4.1, we describe how we model individual belief update — how a participant updates their prediction from their pre- to post-exposure prediction — using either numerical Monte Carlo methods or approximate methods based on prior work by [46, 47].
- In section 4.2, we detail how we evaluate the error (residual) between the *modeled* updated belief and the actual updated belief of a participants. This error will allow us to introduce a parameter which allows us to estimate the relative amount of social learning for each prediction set. Creating such parameters is standard in the Wisdom of the Crowd literature. [63, 105].
- In section 4.3, we describe our methodology for subsetting predictions based on their estimated amount of social learning. Subsetting predictions based on estimated quantities is common [78].
- In section 4.4, we explain how the accuracy and risk of subsets are measured, and how they are used to create a Pareto curve [83].

4.1 Modeling Belief Updates

Using formalism inspired by Bayesian models of cognition [46], we can model the 4,634 prediction sets collected over many rounds, at a high level, as a Bayesian update. To use this formalism, we need to select a prior distribution for each individual's belief, a likelihood (evidence) distribution, and a way to use them to compute the posterior (updated belief) distribution. Here, we describe modeling at a high level, and describe more thoroughly our two modeling approaches in the next sections.

As we are interested in how individuals update their belief regarding the asset's future price (ground truth) V based on the information we expose them to, the choice of the prior distribution is straightforward: $P_{prior}(V) \approx P(B_{pre})$, the distribution over an individual's pre-exposure belief of V. There are two main likelihood (evidence) distributions participants employ: the assets' price history B_T participants are shown, giving us $P_{likelihood}(V) \approx P(B_T)$, or analogously, the social histogram B_H, giving us $P_{likelihood}(V) \approx P(B_H)$.

Given the prior and likelihood, the *modeled* posterior prediction $P_{posterior}(V)$, can, therefore, be approximated as $P_{posterior}(V) \propto P(B_H) \cdot P(B_{pre})$ in the case of the social histogram, and $P_{posterior}(V) \propto P(B_T) \cdot P(B_{pre})$ when participants learn from the past price history. When using the social histogram, we can simply bin the prices shown to a participant and obtain a distribution over prices. When using price history B_T, the time-series of prices needs to first be transformed into a 'rates' histogram as is standard in financial technical analysis [88, 94]. To do so, the daily rate in price change is calculated, and this histogram of price change per day is used to extrapolate and predict asset prices. Specifically, a daily rate, r_t, of asset price change is calculated for each day during the 6-month interval that a user is shown, $r_t = \frac{B_t - B_{t-1}}{B_t}$. These rates are then used to create a histogram of prices similar to when using the social histogram.

We do not make any other assumptions in terms of what data to use to approximate the likelihood and prior distributions. Given these distributions, the question is then how to compute the posterior (updated) belief of an individual.

Although the space of possible distributions and posterior computation approaches is very large, we focus here on using two simple, interpretable, and theoretically-motivated approaches from prior work [47], namely using Gaussian (normal) distributions to approximate priors and likelihoods, and using a Monte Carlo numerical sampling approach to calculate the posterior from the actual distributions of prices that participants were exposed to. We leave to future work the exploration of richer distributions and approaches to modeling belief update as it is beyond the scope of this study.

4.1.1 Approximate Approach: In this approach, we assume both the prior and likelihood to be normally distributed such that the *modeled* posterior — which models the belief update of an individual and therefore allows us to predict their belief after exposure to information — is also normally distributed. When using the social histogram B_T as evidence, the posterior is $P_{posterior}(V) = (B_{pre} + \overline{B_H})/2$. We call this model `GaussianSocial`[3]w. When modeling belief update from the price history B_T we obtain `GaussianPrice`, where $P_{posterior}(V) = (B_{pre} + \overline{B_T})/2$. We include the derivation of these models in the supplementary.

4.1.2 Numerical Approach: Instead of using an approximated distribution for the likelihood, following the formalism of [47], we can use a numerical approach by binning the likelihood distributions to estimate the posterior distribution using Monte Carlo methods. Because we do not have access to the distribution of the prior belief of each individual (as we only have an individual point estimate for each prediction set), we still have to approximate the prior. We model the prior to be Gaussian,

[3] We can sum the scalar B_{pre} to the average of B_T, $\overline{B_T}$, as the average is also a scalar.

with the mean set as the pre-exposure prediction of an individual, B_{pre}, and the standard deviation set as the standard deviation of the social histogram B_H or the standard deviation of the price history B_T, depending on which likelihood distribution was used.

Specifically, we calculate the posterior distribution $P_{posterior}(b)$ of an individual's post-exposure prediction b in the following way: let b_j be a unique value in $\mathbf{B_H}$, and $P_{B_H}(b_h)$ be the probability density of b_h in $\mathbf{B_H}$. Let $P_{prior}(b)$ be the density of b in the parametrized prior distribution. The posterior distribution for the numerical model is defined as $P_{posterior}(b) = \frac{P_{B_H}(b)\times P_{prior}(b)}{\sum_{b_j \in B_H} P_{B_H}(b_j)\times P_{prior}(b_j)}$ when using the social information B_H. After computing this posterior distribution using rejection sampling [42] — our data and distribution are small enough that rejection sampling was fast enough —, we use the mean of the distribution as the *modeled* updated belief of a participant.

4.2 Evaluating Model Error

For all models, we compute the relative residual error between the model's prediction of the posterior ($\mu_{\sim P_{posterior}(V)}$) and the actual post-exposure prediction (B_{post}) as: $(\mu_{\sim P_{posterior}(V)} - B_{post})/B_{post}$. For the approximate approach, $\mu_{\sim P_{posterior}(V)}$ is simply the mean of the normal distribution representing the posterior, while in the numerical approach, the mean is estimated through averaging over all bins of the empirical distribution (the distribution is small enough that sampling was not needed).

For all models, the 95% confidence intervals are calculated as follows: we assume the data follows Student's t-distribution since the variance of the true distribution is unknown and, therefore, we estimate it from the sample data. Let s_e be the estimated standard error of the sample mean and t_e be the t-value for the 95% confidence interval desired, which can be computed via inverse t-distribution. The lower and upper limits for the 95% confidence interval are $[\mu_e - t_e s_e, \mu_e + t_e s_e]$, where μ_e is the estimated sample mean.

4.3 Subsetting Predictions

Using these models, we can *estimate* which information source — social information or price history — each participant used to update their belief by comparing the residual errors of models using either social information or price history as likelihood. This will allow us to select subsets of prediction based on whether they were more likely updated using social or non-social information.

Our approach is illustrated in Fig. 3. This approach of estimating characteristics of how predictions are revised is standard in the Wisdom of the Crowd literature. For example, prior work has estimated resistance to social influence [78] and influenceability in revising judgements after seeing the opinion of others [63, 105], and used them to improve collective performance.

Although we explored many models of belief update (as detailed in Result section 5.3), we choose to focus on the `GaussianSocial` and `GaussianPrice` models (which assume the prior and likelihood to be Gaussian) due to their superior modeling accuracy, and because they are simple, interpretable, and theoretically-motivated models from prior work [47]. We leave for future work the interesting question of searching the large space of parametric and non-parametric models and distributions to best predict people's belief update process.

Therefore, using `GaussianSocial` and `GaussianPrice`, we calculate a residual ϵ_H for when learning from social information B_H and a residual ϵ_T when learning from the price history B_T, as $\epsilon_H = \frac{|\texttt{GaussianSocial}-B_{post}|}{B_{post}}$ and $\epsilon_T = \frac{|\texttt{GaussianPrice}-B_{post}|}{B_{post}}$ respectively. We define $\alpha = \epsilon_T - \epsilon_H$, and we use it to measure how likely a participant used each source of information to update their prediction. For example, for a prediction set $[B_{pre}, B_H, B_T, B_{post}]$ if $\alpha > 0$ (i.e., $\epsilon_T > \epsilon_H$), this means that this prediction set is better modeled using the social histogram of peer's belief B_H instead of the price history B_T.

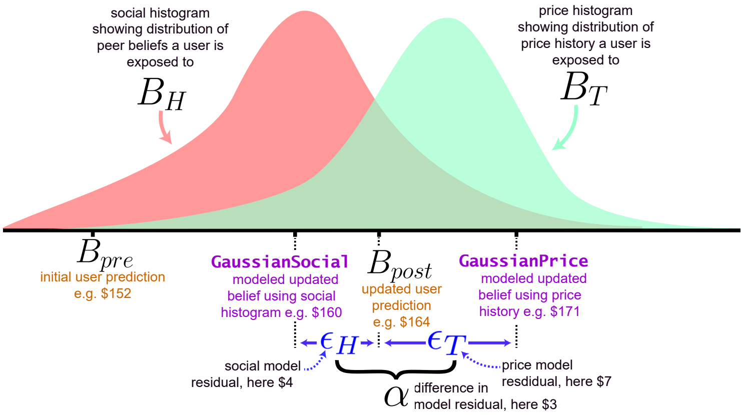

social histogram showing distribution of peer beliefs a user is exposed to
B_H
price histogram showing distribution of price history a user is exposed to
B_T
B_{pre}
initial user prediction e.g. $152
GaussianSocial
modeled updated belief using social histogram e.g. $160
B_{post}
updated user prediction e.g. $164
GaussianPrice
modeled updated belief using price history e.g. $171
ϵ_H
ϵ_T
social model residual, here $4
price model resdidual, here $7
α
difference in model residual, here $3

Fig. 3. An example belief update: for each prediction set, a user updates their belief from the pre-exposure prediction B_{pre} to the updated prediction B_{post} by either learning from the social histogram B_H and/or the price history B_T. ϵ_H is the residual between the *modeled* updated prediction GaussianSocial and the participant's updated prediction B_{post}; ϵ_T is the residual between GaussianPrice and B_{post}. α is the difference between ϵ_T and ϵ_H.

Using α, which we re-scale to be in the interval [-1,1] for each round, we can select a subset S_{α_s} of the prediction sets such that the α of these prediction sets lie in the range $0 \leq \alpha < \alpha_s$ (or $\alpha_s < \alpha \leq 0$ when $\alpha_s < 0$), where α_s is the one-sided boundary we will vary in order to measure how much more likely a participant updated their belief from the social information instead of the price history. For example, the higher α_s is, the more likely a prediction set is better modeled using the social histogram of peer's belief B_H instead of the price history B_T.

4.4 Evaluating Improvement of Subsets

Each prediction set is now associated with a measure of the relative amount of social vs non-social learning, the parameter α. To subset predictions based on α, we bin the α's from all 4,634 prediction sets into 15 groups of equal size, and compute the improvement in prediction error of each subset S_{α_s} and its variance compared to when all the crowd's predictions are used (i.e. compared to *the* Wisdom of the Crowd). To select a subset S_{α_s} of the prediction sets, we select them based on whether their the α of these prediction sets lie in the range $0 \leq \alpha < \alpha_s$ (or $\alpha_s < \alpha \leq 0$ when $\alpha_s < 0$). In order to measure the **improvement** in accuracy after exposure to information, we select only the post-exposure (updated) predictions π_j^{post} within the subset of predictions S_{α_s}, calculate the average prediction of this subset $\mathbb{E}_{j \sim S_{\alpha_s}}[\pi_j^{post}]$, and then the (absolute) error of this average within each round i, with respect to the ground truth V, as $\frac{|\mathbb{E}_{j \sim S_{\alpha_s}}[\pi_j^{post}] - V|}{V}$

Similarly, we then calculate the error of the whole crowd's post-exposure predictions S_{all}, through the same computation as above, using $-1 \leq \alpha \leq 1$ for the set S_{all} instead of S_{α_s}. We define improvement $I^{S_{\alpha_s}}$ as the absolute difference between these two errors, as it measures the improvement in accuracy of using a subset S_{α_s} over using the full set of predictions — the Wisdom of the Crowd — within S_{all}. Critically, this will allow us to measure if exposure to varying degrees of social vs non-social information improves or worsen the performance of the crowd, and is an important metric for platform designers looking to improve the performance of a crowdsourced prediction system.

Note that the improvement defined so far is for each bootstrap b for each round i, and is more clearly denoted as $I_{i,b}^{S_{\alpha_s}}$. Our reported value of improvement (the one in Fig. 4) is over 100 random bootstraps with replacement and is thereby calculated as such: over all rounds i, we first calculate the average improvement, $\mathbb{E}_i[I_{b,i}^{S_{\alpha_s}}]$ for each bootstrap b, and then, over all bootstraps, we calculate $\mathbb{E}_b[\mathbb{E}_i[I_{b,i}^{S_{\alpha_s}}]]$. We use boostrapping [31] in order to have a more robust estimate of the average accuracy and its variance (described in the next section).

In a Pareto curve we are interested in not only measuring the average accuracy described above but also the risk of the crowd in predicting the wrong collective prediction **over the different rounds**. This is because we are interested in estimating the spread of the distribution of accuracy — risk — of the Wisdom of the Crowd over all prediction rounds. Therefore, in Fig. 4, we first calculate, for each bootstrap b the standard deviation (our measure of risk) across the seven rounds i of prediction, i.e. $\sqrt{\mathbb{E}_i[(I_{b,i}^{S_{\alpha_s}} - \mathbb{E}_i[I_{b,i}^{S_{\alpha_s}}])^2]}$. We then compute the average of this risk over 100 bootstraps and report this value. We use standard deviation instead of variance as it is the more popular measure of risk in practice [85].

4.5 Summary of Methods

We first model the belief update of individuals after they have been exposed to either the social information or the price history, using either an approximate approach or a numerical approach. We observe the residual (error) between the *modeled* and the real updated belief of a participant using these various models and find that the `GaussianSocial` and `GaussianPrice` models (which assume the prior and likelihood to be Gaussian) outperform all other models (as detailed in Result section 5.3). Using these models, we compute a parameter α which allows us to estimate whether each prediction set was more likely updated using social information or the price history. Using α, we can select subsets of predictions that were more likely to have been made using one of the information sources. The accuracy this subset can then be compared to the full set of predictions (referred to as the Wisdom of the Crowd) to calculate the improvement of this subset. Similarly, we calculate the variance (risk) of this improvement.

5 RESULTS

5.1 Accuracy-Risk Trade-off

Using a Pareto curve [83], we compare the improvement in prediction accuracy and risk of each subset S_{α_s}. As shown in the Pareto plot in Fig. 4, we observe that although people who learn more from price history are more accurate, there is increased variance—and therefore risk—in their predictions. This suggests that there is a risk-return trade-off between learning from one's peers versus looking at the price history: as social learning is increasingly leveraged, it leads to lower accuracy but also lower risk (replicating prior findings that exposure to social information decreases the variance of the crowd [75]). Note that the social histogram is quite often non-unimodal (as detailed in the next result section 5.3), which means that participants are intentionally collapsing multiple distribution modes to decrease variance.

Such a Pareto trade-off between risk and accuracy is common in financial forecasting [41, 85] and statistical prediction [28, 36, 40, 56], but has not been typically observed in the literature on the Wisdom of Crowds. This has implications for the design of crowdsourced prediction platforms as described in the Discussion section 6.1.

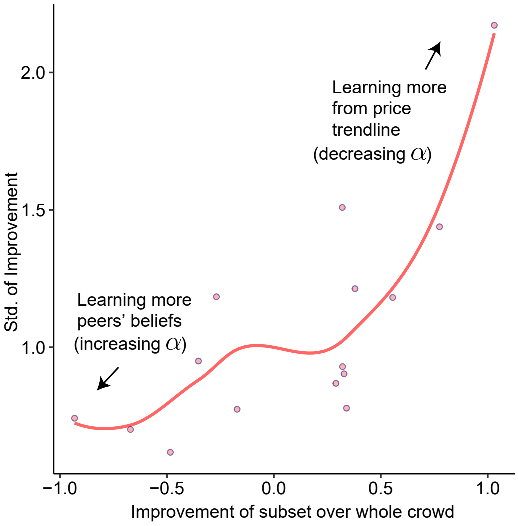


Fig. 4. In this Pareto curve, we plot the improvement of each subset vs. the standard deviation in improvement within this subset. We see a risk-return trade-off: predictions made with price history are more accurate, but with higher risk (standard deviation). Smoothed curved is generated using LOESS [18].

5.2 Performance under High Uncertainty

An additional result of our study is the investigation of the crowd's performance during high uncertainty using the data from the prediction round that happened during the Brexit vote. We ran the analysis described earlier, but only for predictions made during the last week. Note that in all previous results, we took care not to use the last week of data to calculate collective accuracy so that prediction was not too easy, but we do so here as the high uncertainty makes prediction quite hard. This last week of data that we use is a disjoint subset of data from the data we previously used.

Again, we bin all α's from the prediction sets during this week and investigate the improvements of subsets of predictions compared to the whole crowd. We use a smaller number of bins due to the smaller number of predictions during the last week: 52 prediction sets in the last week compared to 284 during the open period of prediction that we previously used for predictions.

As can be seen in Fig. 5, as α decreases (i.e. we select predictions that were more likely updated using the price history instead of the social information), improvement in accuracy of subsets compared to the Wisdom of the Crowd (all predictions) severely decays down to -3.14% (95% confidence interval [2.49, 3.79]). Conversely, as subsets of predictions updated using the social histogram ($\alpha_s > 0$) are selected, the improvement in their accuracy is fairly stable (although negative).

Note that although a smaller number of predictions were made during the last week before Brexit (52 prediction sets compared to 284 during the open period of prediction as discussed earlier), we have sufficient data to afford statistically significant results as shown by the 95% confidence intervals of our findings. The improvement values, confidence intervals, and their accompanying α_s are included in Table 4 in the supplementary. Unfortunately, given that such high market uncertainty only occurred during one round, we do not have enough data to produce a Pareto curve over multiple rounds.

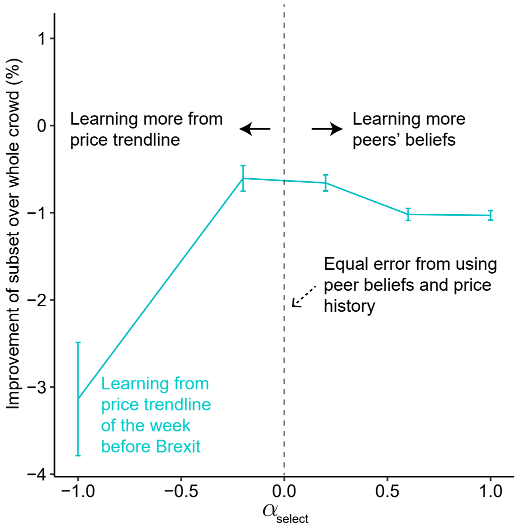


Fig. 5. Improvement when selecting predictions based on how much more they were likely made using social information ($\alpha_s > 0$) vs. price history ($\alpha_s < 0$). 95% Confidence intervals obtained through 100 bootstraps. When the price history is itself very uncertain, participants who learn from their peers do better than when they learned from the price history.

Our results suggest that although social learning generally leads to lower accuracy as shown in the Pareto curve of Fig. 4, during periods of high uncertainly, social learning leads to higher accuracy. This suggests that social learning could be leveraged by platform designers as a valuable tool that minimizes catastrophic performance.

5.3 Belief Update Models

Although our goal is not to search for the best model of individual belief update, we highlight our observations from fitting simple, interpretable, and theoretically-motivated models and distributions from prior work. It is important to note that `GaussianSocial` and `GaussianPrice` are both parameter-less models and did not require any parameter fitting, making their success in modeling belief update even more interesting.

As can be seen in Fig. 6, models that use social information for modeling the belief update of participants (`GaussianSocial`,`GaussianSocialModes`, `NumericalSocial`) perform better than models that use the price history (`GaussianPrice`, `NumericalPrice`). This suggests that our participants predominantly use social information instead of the price history to update their belief.

More specifically, `GaussianSocial`, our simple Gaussian model that assumes the data follows a single-mode Gaussian distribution, outperforms `GaussianSocialModes`, a model that measures when the social histogram is non-unimodal (which we estimate using the Hartigan's dip test of unimodality [50]) and uses the largest mode as the mean of the distribution in the same belief update procedure as `GaussianSocial`. This suggests that people assume the data they learn from to be unimodal even when it is non-unimodal, in line with prior work [29, 91].

Additionally, `GaussianSocial` outperforms the more precise numerical model `NumericalSocial` which makes no parametric assumption on the data distributions and uses a Monte Carlo procedure to estimate the posterior distribution. This suggests that people use simple heuristics when

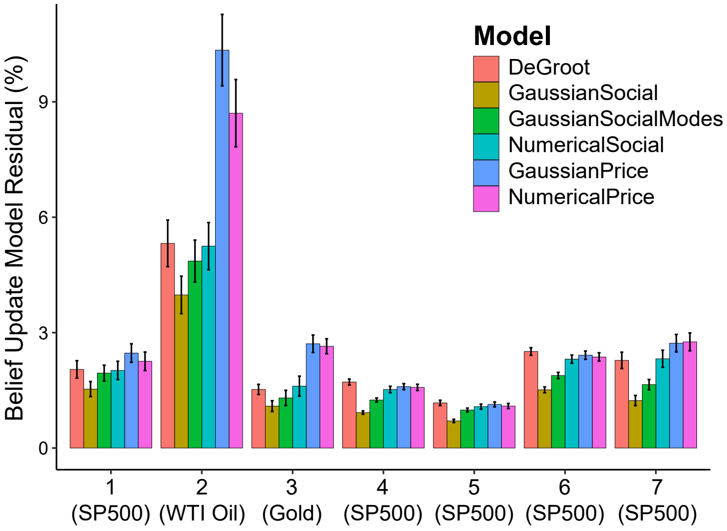


Fig. 6. The y-axis shows the relative residual between *modeled* belief update and *actual* updated belief. Simple approximated models do better at modeling belief update than numerical models, and models using social histograms as likelihood perform better than models using the price history. Error bars represent 95% CI.

learning from their peers. However, when people are learning from the price history, the dominance of simpler models is not as clear as the performance of the simple `GaussianPrice` model is indistinguishable from that of the numerical model (`NumericalPrice`).

`GaussianSocial` also outperforms the popular `DeGroot` model commonly used as a benchmark in the literature [25], where an individual updates their belief as the weighted average belief of their peers. Here we set the weights (trust values) equal for all peers, as we have no data to estimate these weights, and therefore assume a uniform prior. It is interesting to note that `GaussianSocial` is equivalent to the `DeGroot` model when a participant's weight on their own prior belief is equal to the total of the weights of all other participants. This is in agreement with previous work showing that people put a disproportionately larger weight on their own prior belief [23, 89].

Overall, the superiority of `GaussianSocial` in predicting belief update suggests that participants use a heuristic, uni-modal, and simple belief update procedure when updating their beliefs, and that they predominantly update their predictions using social information instead of price history.

6 DISCUSSION

Our hypothesis was that a Pareto frontier exists between risk and accuracy — i.e. that there is a trade-off between risk and accuracy of prediction — and that social learning impacts this trade-off. Our results support this hypothesis: Fig. 4 demonstrates that a Pareto frontier exists — similarly to statistical [28, 40, 56] and financial [41, 58, 85] forecasting systems — and that it is mediated by social learning. Additionally, our belief update models (Fig. 6) suggest that participants rely on social learning (and other heuristics) to update their belief.

6.1 Design Implications

Currently, crowdsourced prediction systems using a Wisdom of the Crowd approach focus on measuring and optimizing the average accuracy of their participants with little regard to the variance (risk) of the predictions. Practically, this means that platform designers and operators deploy a task to be predicted (e.g. predicting future prices) and end up with a single measure of performance for the task - the average accuracy of the crowd (or some similar measure of central tendency).

Our results suggest that the performance of such systems can be more comprehensively characterized by using both the average accuracy of the crowd and its variance. This is especially important because crowdsourced systems are being increasingly deployed for applications where uncertainty and risk can be quite harmful, such as in the medical information discovery domain [44, 69, 114] where incorrect predictions are extremely costly. More generally, the modeling of risk supports more powerful and versatile applications of crowdsourced predictions such as hedging risks over portfolios of tasks which is standard in financial and statistical forecasting.

Additionally, given that the performance of a crowdsourced prediction system lies along a Pareto frontier, a practical question for designers is how to *tune* the platform to reach a desired value of risk and accuracy. Our result that social learning can mediate the accuracy-risk trade-off provides a practical means to attain performances along this frontier. Practically, our results suggest that social learning within a crowdsourcing platform could be more purposefully leveraged to fit the task at hand: for example, platform designers could vary the social learning between users by incentivizing it to have lower risk — especially during highly uncertain times, as our results from the Brexit prediction (Fig. 5) round showed.

Additionally our results provide evidence that there exists competing effects of exposure to social information versus non-social information on both accuracy and risk. Prior work had separately investigated exposure to social information [75] or to non-social information [53, 95]. By designing our experimental procedure such that people were freely able to learn from either social or non-social information and then estimating how much more of each source of information a person learned from, we are able to show that each type of learning causes opposite effects in terms of accuracy and risk: learning from the price history encourages higher accuracy, while learning from one's peers minimizes risk. This provides direct insights as to the design of crowdsourced prediction platforms as it indicates that there is an important balance between providing social and non-social information.

6.2 Heuristics and Biases

Our results also have implications for the literature on decision heuristics and biases [91, 108]. Through the modeling of belief update, we observe that our subjects exhibit the attribute substitution heuristic of human decision-making [59], whereby a complicated problem is solved by approximating it with a simpler, less accurate model. We observe this heuristic as our participants simplify the data they are using to update their belief. This is evidenced by the fact that our participants' updated beliefs are better modeled by the `GaussianSocial` model (which assumes the data to be unimodal) than by the multi-modal belief update model `GaussianSocialModes`, indicating that our participants generally wrongly assume the data to be unimodal even when it is not (measured using the Hartigan's dip test of unimodality [50]). This is in line with previous studies that have shown that people wrongly assume data to be unimodal [29, 74, 90] due to the fact that using multi-modal data is cognitively costly [52]. Additional evidence of this substitution heuristic is from the fact that approximate models better predict the updated beliefs of participants than the more complicated (numerical) models: the `GaussianSocial` model outperforms the more precise Monte Carlo numerical models (as shown in Fig. 6).

Another decision heuristic that we observe is that people prefer to use social information rather than the underlying price history of an asset to update their belief as models which use social information (`GaussianSocial`,`GaussianSocialModes`, `NumericalSocial`) outperform models that use price history (`GaussianPrice`, `NumericalPrice`) as shown in Fig. 6. However, this collective preference for social learning comes at the price of lower accuracy (Fig. 4). It is especially surprising that our participants preferred to use social information instead of prices to update their belief given that they were mid-career finance professionals with strong financial experience who should know that price information is generally better to predict future prices [32, 79]. Such behavior has been observed in prior work where even experts performing a familiar task demonstrate sub-optimal decision heuristics [66, 103], and often over-rely on social information [34, 97]. However, instead of seeing such behavior as irrational, our results suggests that perhaps participants are preferentially aiming for lower risk instead of higher accuracy. This preference for social information especially pays off during the high uncertainty period before the Brexit vote.

Therefore, our results support growing evidence that heuristics and biases are not merely *defects* of human decision-making, but that perhaps they optimize for richer objectives or are optimized for more time- or data-constrained decision-making [21, 57, 62, 68, 80, 84, 102]. For example, when individual decision-making is viewed within the lens of more realistic requirements such as limited time [8, 19] or attention [110], heuristics and biases—such as people assuming that the environment around them undergoes strong abrupt changes even when it is quite stable [77, 100]—act as priors that facilitate fast decision-making [76, 99], and are quite helpful in practice.

6.3 Future Work

Our work demonstrates that crowdsourced prediction platforms behave similarly to financial and statistical forecasting systems in that they exhibit an accuracy-risk Pareto frontier, and that this trade-off is mediated by social learning. This observation opens a number of interesting avenues for future work within the CSCW community. One interesting next step would be to investigate if different modalities of social communication and learning have have a similar accuracy-risk trade-off such as different types of discussions on forums [67] or the diversity of backgrounds of people interacting [111]. In our work, we restricted each round to have a static population of participants whose predictions were shared: an interesting direction for future work would be to embed participants in social networks given the importance and popularity of recent work on the effect of communication topologies [1, 2, 9, 10] on group performance. Another interesting avenue for future work would be to utilize established metrics of risk aversion [98] and investigate how subsetting predictions using these metrics affects collective accuracy and risk minimization. We also leave to future work the exploration of the large space of parametric and non-parametric models that best model people's belief update process. All these directions of future work pave the way for improving the design and deployment of crowdsourced prediction platforms.

REFERENCES

[1] Dhaval Adjodah, Dan Calacci, Abhimanyu Dubey, Anirudh Goyal, Peter Krafft, Esteban Moro, and Alex Pentland. 2020. Leveraging Communication Topologies Between Learning Agents in Deep Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

[2] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, PM Krafft, Mehdi Moussaid, and Alex Pentland. 2020. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences* (2020).

[3] Ron Alquist and Lutz Kilian. 2010. What do we learn from the price of crude oil futures? *Journal of Applied econometrics* 25, 4 (2010), 539–573.

[4] Ofra Amir, Barbara J Grosz, Krzysztof Z Gajos, Sonja M Swenson, and Lee M Sanders. 2015. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1419–1428.

[5] J Scott Armstrong, Kesten C Green, and Andreas Graefe. 2015. Golden rule of forecasting: Be conservative. *Journal of Business Research* 68, 8 (2015), 1717–1731.

[6] Kenneth J Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D Nelson, et al. 2008. The promise of prediction markets.

[7] Søren Asmussen and Dirk P Kroese. 2006. Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability* 38, 2 (2006), 545–558.

[8] Ron Azuma, Mike Daily, and Chris Furmanski. 2006. A review of time critical decision making models and human cognitive processes. In *2006 IEEE aerospace conference*. IEEE, 9–pp.

[9] Daniel Barkoczi and Mirta Galesic. 2016. Social learning strategies modify the effect of network structure on group performance. *Nature communications* 7, 1 (2016), 1–8.

[10] Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences* (2017), 201615978. https://doi.org/10.1073/pnas.1615978114

[11] Hans P Binswanger and Donald A Sillers. 1983. Risk aversion and credit constraints in farmers' decision-making: A reinterpretation. *The Journal of Development Studies* 20, 1 (1983), 5–21.

[12] John Y Campbell and Robert J Shiller. 1988. Stock prices, earnings, and expected dividends. *The Journal of Finance* 43, 3 (1988), 661–676.

[13] L Elisa Celis, Peter M Krafft, and Nathan Kobe. 2016. Sequential voting promotes collective discovery in social recommendation systems. In *Tenth International AAAI Conference on Web and Social Media*.

[14] Ariane Chapelle, Yves Crama, Georges Hübner, and Jean-Philippe Peters. 2008. Practical methods for measuring and managing operational risk in the financial sector: A clinical study. *Journal of Banking & Finance* 32, 6 (2008), 1049–1061.

[15] Valérie Chavez-Demoulin, Paul Embrechts, and Johanna Nešlehová. 2006. Quantitative models for operational risk: extremes, dependence and aggregation. *Journal of Banking & Finance* 30, 10 (2006), 2635–2658.

[16] Pin-Yu Chen, Shin-Ming Cheng, Pai-Shun Ting, Chia-Wei Lien, and Fu-Jen Chu. 2015. When crowdsourcing meets mobile sensing: A social network perspective. *IEEE Communications Magazine* 53, 10 (2015), 157–163.

[17] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[18] William S Cleveland and Susan J Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association* 83, 403 (1988), 596–610.

[19] Izack Cohen. 2008. Improving time-critical decision making in life-threatening situations: Observations and insights. *Decision Analysis* 5, 2 (2008), 100–110.

[20] Justin Cox and Todd Griffith. 2018. Political Uncertainty and Market Liquidity: Evidence from the Brexit Referendum and the 2016 US Presidential Election. *Available at SSRN 3092335* (2018).

[21] Henrik Cronqvist and Stephan Siegel. 2014. The genetics of investment biases. *Journal of Financial Economics* 113, 2 (2014), 215–234.

[22] Marcelo G Cruz. 2002. *Modeling, measuring and hedging operational risk*. Vol. 346. Wiley New York.

[23] Chetan Dave and Katherine W Wolfe. 2003. On confirmation bias and deviations from Bayesian updating. *Retrieved on* 24, 02 (2003), 2011.

[24] Anna De Liddo, Ágnes Sándor, and Simon Buckingham Shum. 2012. Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work (CSCW)* 21, 4-5 (2012), 417–448.

[25] Morris H DeGroot. 1974. Reaching a consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.

[26] Shubhada Deshpande. 2020. Brexit Myth on FTSE and DAX Companies: A Review. *Available at SSRN 3517139* (2020).

[27] Djellel Difallah, Alessandro Checco, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. Deadline-Aware Fair Scheduling for Multi-Tenant Crowd-Powered Systems. *ACM Transactions on Social Computing* 2, 1 (2019), 1–29.

[28] Pedro Domingos. 2000. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*. 231–238.

[29] N Donnelly, KR Cave, M Welland, and T Menneer. 2006. Breast screening, chicken sexing and the search for oil: Challenges for visual cognition. *Geological Society, London, Special Publications* 254, 1 (2006), 43–55.

[30] Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112, 50 (2015), 15343–15347.

[31] Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*. Springer, 569–593.

[32] Eugene F Fama. 1965. The behavior of stock-market prices. *The journal of Business* 38, 1 (1965), 34–105.

[33] Eugene F Fama. 1995. Random walks in stock market prices. *Financial analysts journal* 51, 1 (1995), 75–80.

[34] F Douglas Foster and S Viswanathan. 1996. Strategic trading when agents forecast the forecasts of others. *The Journal of Finance* 51, 4 (1996), 1437–1478.

[35] Kenneth R French. 1986. Detecting spot price forecasts in futures prices. *Journal of Business* (1986), S39–S54.

[36] Francesco Gagliardi. 2011. Instance-based classifiers applied to medical databases: diagnosis and knowledge extraction. *Artificial Intelligence in Medicine* 52, 3 (2011), 123–139.

[37] Francis Galton. 1907. Vox populi (The wisdom of crowds). *Nature* 75, 7 (1907), 450–451.

[38] Alexander Gammerman and Vladimir Vovk. 2007. Hedging predictions in machine learning. *Comput. J.* 50, 2 (2007), 151–163.

[39] Kapil Garg, Yongsung Kim, Darren Gergle, and Haoqi Zhang. 2019. 4X: A Hybrid Approach for Scaffolding Data Collection and Interest in Low-Effort Participatory Sensing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[40] Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural computation* 4, 1 (1992), 1–58.

[41] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. 2005. There is a risk-return trade-off after all. *Journal of Financial Economics* 76, 3 (2005), 509–548.

[42] Wally R Gilks, Nicky G Best, and KKC Tan. 1995. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 44, 4 (1995), 455–472.

[43] Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–49.

[44] Assaf Gottlieb, Robert Hoehndorf, Michel Dumontier, and Russ B Altman. 2015. Ranking adverse drug reactions with crowdsourcing. *Journal of medical Internet research* 17, 3 (2015), e80.

[45] Antonietta Grasso and Gregorio Convertino. 2012. Collective intelligence in organizations: Tools and studies. *Computer Supported Cooperative Work (CSCW)* 21, 4-5 (2012), 357–369.

[46] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition. *The Cambridge Handbook of Computational Psychology, Ron Sun (ed.), Cambridge University Press.* (2008), 1–49.

[47] Thomas L. Griffiths and Joshua B. Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological Science* 17, 9 (2006), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

[48] Barbara Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence* (1996).

[49] Robin Hanson. 2003. Combinatorial information market design. *Information Systems Frontiers* 5, 1 (2003), 107–119.

[50] John A Hartigan, Pamela M Hartigan, et al. 1985. The dip test of unimodality. *The annals of Statistics* 13, 1 (1985), 70–84.

[51] Arend Hintze, Randal S Olson, Christoph Adami, and Ralph Hertwig. 2015. Risk sensitivity as an evolutionary adaptation. *Scientific reports* 5 (2015), 8242.

[52] Aaron B Hoffman and Bob Rehder. 2010. The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General* 139, 2 (2010), 319.

[53] Robin M Hogarth and Hillel J Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive psychology* 24, 1 (1992), 1–55.

[54] Charles A Holt and Susan K Laury. 2002. Risk aversion and incentive effects. *American economic review* 92, 5 (2002), 1644–1655.

[55] Eaman Jahani, Peter M Krafft, Yoshihiko Suhara, Esteban Moro, and Alex Sandy Pentland. 2018. Scamcoins, s*** posters, and the search for the next bitcoinTM: Collective sensemaking in cryptocurrency discussions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[56] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning.* Vol. 112. Springer.

[57] Anika K Josef, David Richter, Gregory R Samanez-Larkin, Gert G Wagner, Ralph Hertwig, and Rui Mata. 2016. Stability and change in risk-taking propensity across the adult life span. *Journal of personality and social psychology* 111, 3 (2016), 430.

[58] Jon M Joyce and Robert C Vogel. 1970. The uncertainty in risk: Is variance unambiguous? *The Journal of Finance* 25, 1 (1970), 127–134.

[59] Daniel Kahneman and Shane Frederick. 2014. *Representativeness Revisited: Attribute Substitution in Intuitive Judgment.* Number January 2002. 49–81 pages. https://doi.org/10.1017/CBO9780511808098.004 arXiv:arXiv:1011.1669v3

[60] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I.* World Scientific, 99–127.

[61] David R Karger, Sewoong Oh, and Devavrat Shah. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research* 62, 1 (2014), 1–24.

[62] Douglas T Kenrick and Vladas Griskevicius. 2013. *The rational animal: How evolution made us smarter than we think.* Basic Books (AZ).

[63] Corentin Vande Kerckhove, Samuel Martin, Pascal Gend, Peter J Rentfrow, Julien M Hendrickx, and Vincent D Blondel. 2016. Modelling influence and opinion evolution in online collective behaviour. *PloS one* 11, 6 (2016).

[64] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2316–2329.

[65] Yea-Seul Kim, Logan A Walls, Peter Krafft, and Jessica Hullman. 2019. A Bayesian Cognition Approach to Improve Data Visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 682.

[66] Derek J Koehler, Lyle Brenner, and Dale Griffin. 2002. The calibration of expert judgment: Heuristics and biases beyond the laboratory. *Heuristics and biases: The psychology of intuitive judgment* (2002), 686–715.

[67] Peter M Krafft, Nicolás Della Penna, and Alex Sandy Pentland. 2018. An experimental study of cryptocurrency market dynamics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[68] Venkat R Lakshminarayanan, M Keith Chen, and Laurie R Santos. 2011. The evolution of decision-making under risk: framing effects in monkey risk preferences. *Journal of Experimental Social Psychology* 47, 3 (2011), 689–693.

[69] Jérémy Lardon, Redhouane Abdellaoui, Florelle Bellet, Hadyl Asfari, Julien Souvignet, Nathalie Texier, Marie-Christine Jaulent, Marie-Noëlle Beyens, Anita Burgun, and Cédric Bousquet. 2015. Adverse drug reaction identification and extraction in social media: a scoping review. *Journal of medical Internet research* 17, 7 (2015), e171.

[70] David Lazer and Allan Friedman. 2007. The network structure of exploration and exploitation. *Administrative science quarterly* 52, 4 (2007), 667–694.

[71] Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[72] Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*. 621–630.

[73] Soo Ling Lim, Daniele Quercia, and Anthony Finkelstein. 2010. StakeSource: harnessing the power of crowdsourcing and social networks in stakeholder analysis. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, Vol. 2. IEEE, 239–242.

[74] Marcus Lindskog. 2013. *Is the Intuitive Statistician Eager or Lazy?: Exploring the Cognitive Processes of Intuitive Statistical Judgments*. Ph.D. Dissertation. Acta Universitatis Upsaliensis.

[75] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025. https://doi.org/10.1073/pnas.1008636108

[76] Mark Lubell and John T Scholz. 2001. Cooperation, reciprocity, and the collective-action heuristic. *American Journal of Political Science* (2001), 160–178.

[77] Ning Ma and Angela J Yu. 2015. Statistical learning and adaptive decision-making underlie human response time variability in inhibitory control. *Frontiers in psychology* 6 (2015), 1046.

[78] Gabriel Madirolas and Gonzalo G de Polavieja. 2015. Improving collective estimations using resistance to social influence. *PLoS computational biology* 11, 11 (2015), e1004594.

[79] Burton G Malkiel and Eugene F Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance* 25, 2 (1970), 383–417.

[80] Dave EW Mallpress, Tim W Fawcett, Alasdair I Houston, and John M McNamara. 2015. Risk attitudes in a changing environment: An evolutionary model of the fourfold pattern of risk preferences. *Psychological Review* 122, 2 (2015), 364.

[81] Thomas W Malone and Kevin Crowston. 1990. What is coordination theory and how can it help design cooperative work systems?. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*. 357–370.

[82] Thomas W Malone, Jeffrey V Nickerson, Robert J Laubacher, Laur Hesse Fisher, Patrick De Boer, Yue Han, and W Ben Towne. 2017. Putting the pieces back together again: Contest webs for large-scale problem solving. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1661–1674.

[83] Harry Markowitz. 1952. Portfolio selection. *The journal of finance* 7, 1 (1952), 77–91.

[84] Sandeep Mishra. 2014. Decision-making under risk: Integrating perspectives from biology, economics, and psychology. *Personality and Social Psychology Review* 18, 3 (2014), 280–307.

[85] Franco Modigliani and Modigliani Leah. 1997. Risk-adjusted performance. *Journal of portfolio management* 23, 2 (1997), 45.

[86] Mehdi Moussaïd, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. 2013. Social influence and the collective dynamics of opinion formation. *PloS one* 8, 11 (2013), e78433. https://doi.org/10.1371/journal.pone.0078433 arXiv:1311.3475

[87] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[88] Salih N Neftci. 1991. Naive trading rules in financial markets and wiener-kolmogorov prediction theory: a study of" technical analysis". *Journal of Business* (1991), 549–571.

[89] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[90] Richard E Nisbett and Ziva Kunda. 1985. Perception of social distributions. *Journal of Personality and Social Psychology* 48, 2 (1985), 297.

[91] Richard E Nisbett and Lee Ross. 1980. Human inference: Strategies and shortcomings of social judgment. (1980).

[92] Michael Nofer and Oliver Hinz. 2014. Are crowds on the internet wiser than experts? The case of a stock prediction community. *Journal of Business Economics* 84, 3 (2014), 303–338.

[93] Andreas Oehler, Matthias Horn, and Stefan Wendt. 2017. Brexit: Short-term stock price effects and the impact of firm-level internationalization. *Finance Research Letters* 22 (2017), 175–181.

[94] Cheol-Ho Park and Scott H Irwin. 2007. What do we know about the profitability of technical analysis? *Journal of Economic Surveys* 21, 4 (2007), 786–826.

[95] John W Payne, John William Payne, James R Bettman, and Eric J Johnson. 1993. *The adaptive decision maker.* Cambridge university press.

[96] Ole Peters. 2019. The ergodicity problem in economics. *Nature Physics* 15, 12 (2019), 1216–1221.

[97] Marta Posada, Cesareo Hernandez, and Adolfo Lopez-Paredes. 2006. Learning in continuous double auction market. In *Artificial Economics.* Springer, 41–51.

[98] John W Pratt. 1978. Risk aversion in the small and in the large. In *Uncertainty in Economics.* Elsevier, 59–79.

[99] David G Rand, Victoria L Brescoll, Jim AC Everett, Valerio Capraro, and Hélène Barcelo. 2016. Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General* 145, 4 (2016), 389.

[100] Chaitanya Ryali, Gautam Reddy, and J Yu Angela. 2018. Demystifying excessively volatile human learning: A Bayesian persistent prior and a neural approximation. In *Advances in Neural Information Processing Systems.* 2781–2790.

[101] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

[102] Laurie R Santos and Alexandra G Rosati. 2015. The evolutionary roots of human decision making. *Annual review of psychology* 66 (2015), 321–347.

[103] James Shanteau. 1988. Psychological characteristics and strategies of expert decision makers. *Acta psychologica* 68, 1-3 (1988), 203–215.

[104] Pavel V Shevchenko and Mario V Wuthrich. 2006. The structural modelling of operational risk via Bayesian inference: Combining loss data with expert opinions. *The Journal of Operational Risk* 1, 3 (2006), 3–26.

[105] Jack B Soll and Richard P Larrick. 2009. Strategies for revising judgment: How (and how well) people use othersâĂŹ opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 3 (2009), 780.

[106] Greg Stoddard. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *Ninth International AAAI Conference on Web and Social Media (ICWSM).*

[107] Brandon M Turner, Mark Steyvers, Edgar C Merkle, David V Budescu, and Thomas S Wallsten. 2014. Forecast aggregation via recalibration. *Machine learning* 95, 3 (2014), 261–289.

[108] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[109] Marlon Twyman, Brian C Keegan, and Aaron Shaw. 2017. Black Lives Matter in Wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* 1400–1412.

[110] Daan Van Knippenberg, Linus Dahlander, Martine R Haas, and Gerard George. 2015. Information, attention, and decision making.

[111] Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. 2004. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology* 89, 6 (2004), 1008.

[112] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive science* 38, 4 (2014), 599–637.

[113] Edward Vul and Harold Pashler. 2008. Measuring the crowd within probabilistic representations within individuals. *Psychological Science* 19, 7 (2008), 645–647.

[114] Zichen Wang, Caroline D Monteiro, Kathleen M Jagodnik, Nicolas F Fernandez, Gregory W Gundersen, Andrew D Rouillard, Sherry L Jenkins, Axel S Feldmann, Kevin S Hu, Michael G McDermott, et al. 2016. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications* 7, 1 (2016), 1–11.

[115] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[116] Ruixun Zhang, Thomas J Brennan, and Andrew W Lo. 2014. The origin of risk aversion. *Proceedings of the National Academy of Sciences* 111, 50 (2014), 17777–17782.

SUPPLEMENTARY MATERIAL

DERIVATION OF GAUSSIAN MODEL

We describe `GaussianSocial` here. `GaussianPrice` follows the same derivation, substituting the social histogram B_H with the price history B_T.

Our notation follows that of [65]. We assume that people's estimate of the future price before information exposure, B_{pre}, is being sampled from an internal prior distribution [113], and that the sample we obtain is indicative of the mean of the prior distribution following the results of [112].

We suppose that people think each asset has a true value, V^*, which people are trying to estimate to predict the future asset value, V (the ground truth); that prior beliefs about V^* follow a Normal (Gaussian) prior distribution, $V^* \sim Normal(\mu_{prior}, \sigma_{prior})$; and that evidence about V^* can be understood as being generated from a Normal distribution, $Normal(V^*, \sigma_{data})$. In this case the posterior beliefs people have follows a simple form. Letting information content be defined as the inverse of the Normal distribution's variance $I = \frac{1}{\sigma}$, we have that

$$\mu_{posterior} = \frac{\mu_{prior} \cdot I_{prior} + \mu_{data} \cdot I_{data}}{I_{prior} + I_{data}}. \tag{1}$$

Additionally, the social histogram is treated as representing the information content of data about V^*, then we have:

$$\mu_{posterior} = \frac{B_{pre} \cdot I_{prior} + \overline{B_H} \cdot I_{data}}{I_{prior} + I_{data}}. \tag{2}$$

The `GaussianSocial` rule therefore can be viewed as reflecting an assumption of a Normal distribution as a mental model, and assuming private information and social information have the same information content ($I_{prior} = I_{data}$), which gives:

$$\mu_{posterior} = \frac{B_{pre} + \overline{B_H}}{2}. \tag{3}$$

PERFORMANCE OVER ALL ROUNDS

Here we report the values of the residual for each round for all models. We can observe that `GaussianSocial` does best.

	ROUND						
MODEL	**1** **(S&P 500)**	**2** **WTI Oil**	**3** **Gold**	**7** **(S&P 500)**	**8** **(S&P 500)**	**9** **(S&P 500)**	**12** **(S&P 500)**
`GaussianSocial`	1.53 (0.19)	3.97 (0.48)	1.08 (0.13)	0.92 (0.04)	0.70 (0.04)	1.51 (0.07)	1.23 (0.13)
`GaussianSocialModes`	1.94 (0.20)	4.85 (0.54)	1.30 (0.19)	1.24 (0.05)	0.98 (0.04)	1.88 (0.08)	1.64 (0.13)
`NumericalSocial`	2.01 (0.23)	5.24 (0.61)	1.60 (0.25)	1.52 (0.08)	1.07 (0.06)	2.31 (0.10)	2.31 (0.22)
`NumericalPrice`	2.25 (0.23)	8.70 (0.87)	2.64 (0.19)	1.57 (0.08)	1.09 (0.06)	2.36 (0.10)	2.75 (0.23)
`GaussianPrice`	2.46 (0.24)	10.3 (0.92)	2.70 (0.22)	1.59 (0.07)	1.13 (0.06)	2.41 (0.10)	2.72 (0.22)
`DeGroot`	2.04 (0.22)	5.32 (0.60)	1.52 (0.13)	1.71 (0.07)	1.17 (0.06)	2.51 (0.09)	2.27 (0.21)

Table 2. Values of the residual for each round for all models. Numbers in parentheses show the 95% error.

TABLE OF SUBSETTING

In this section, we report the improvement when selecting a subset of participants.

α_s	**Improvement (%)**	**95% CI**
-1.0	1.03	0.02
-0.9	0.77	0.05
-0.7	0.33	0.07
-0.6	0.32	0.07
-0.4	0.29	0.07
-0.3	0.34	0.06
-0.1	-0.17	0.02
0.0	-0.48	0.06
0.1	0.56	0.03
0.3	0.38	0.03
0.4	0.32	0.03
0.6	-0.27	0.08
0.7	-0.35	0.06
0.9	-0.67	0.04
1.0	-0.93	0.02

Table 3. Improvements achieved by subsetting predictions via α_s for all rounds. Confidence intervals are calculated through 100 bootstraps.

α_s	**Improvement (%)**	**95% CI**
-1.0	-3.14	0.65
-0.2	-0.61	0.15
0.2	-0.66	0.09
0.6	-1.02	0.07
1.0	-1.03	0.05

Table 4. Improvements achieved by subsetting predictions via α_s only for predictions the week before Brexit. Confidence intervals are calculated through 100 bootstraps.

α_s	**Improvement (%)**	**Standard Deviation**
-1.0	1.03	2.17
-0.9	0.77	1.44
-0.7	0.33	0.90
-0.6	0.32	0.93
-0.4	0.29	0.87
-0.3	0.34	0.78
-0.1	-0.17	0.77
0.0	-0.48	0.62
0.1	0.56	1.18
0.3	0.38	1.21
0.4	0.32	1.51
0.6	-0.27	1.18
0.7	-0.35	0.95
0.9	-0.67	0.70
1.0	-0.93	0.74

Table 5. Improvement and bootstrapped Standard Deviation used in Pareto curve.