Abstract Book

14th Annual Tinbergen Institute Conference

Bayesian Crowd

24-25 June 2019, Rotterdam

Abstracts

14th Annual Tinbergen Institute Conference

Bayesian Crowd

24-25 June 2019, Rotterdam

Organizing Committee

Aurelien Baillon, Erasmus University Rotterdam
Drazen Prelec, Massachusetts Institute of Technology
Dennie van Dolder, Vrije Universiteit Amsterdam
Tong Wang, Erasmus University Rotterdam

Time	Monday June 24	
9:00-9:30	Registration + Light breakfast	
9:30-9:50	Welcome: *How to make the most of the Bayesian Crowd conference* Room CB-5	
9:50-11:05	Session 1	
	Parallel 1 A: **Aggregation** *Chair: Benjamin Tereick* *Room C1-5* Dennie van Dolder (Vrije Universiteit Amsterdam) *The Wisdom of the Inner Crowd in Three Large Natural Experiments* John McCoy (Wharton School - University of Pennsylvania, United States) *A Possible Worlds Model: Some Theoretical Connections* Benjamin Tereick (Erasmus University Rotterdam) *The Wisdom of Self-Aggregating Crowds*	Parallel 1 B: **Truthful Reports** *Chair: Naman Goel* *Room C1-6* Yen-Lin Chiu (Johns Hopkins University, United States) *Online Product Rating with Purchase Data* Yan Xu (Erasmus University Rotterdam) *Top-Flop Betting: An Incentive Mechanism to Elicit Unverifiable Truth* Naman Goel (Swiss Federal Institute of Technology (EPFL) Lausanne) *Decentralized Oracles via Peer-Prediction in the Presence of Lying Incentives*
11:05-11:20	Coffee break	
11:20-12:35	Session 2	
	Parallel 2 A: **Weighting Judgments** *Chair: Tina Nane* *Room C1-5* Shu Huang (Carnegie Mellon University, United States) *Getting More Wisdom from the Crowd: When Weighting Individual Judgments Reliably Improves Accuracy or Just Adds Noise* Klaus Nehring (University of California, Davis, United States) *Weighing Experts, Weighing Sources: The Diversity Value* Tina Nane (Technische Universiteit Delft)	Parallel 2 B**: Game Theory** *Chair: Peter Katuscak* *Room C1-6* Leonard Wolk (Vrije Universiteit Amsterdam) *Proportional-Prize Contests to Predict Future Events: Does Group Size Matter?* Vessela Daskalova (Institute for Advanced Study in Toulouse, Toulouse School of Economics) *Categorization and Coordination* Peter Katuscak (Aachen University) *Strategy-proofness Made Simpler*

	A Cross-Validation Study of Experts' Judgments Using the Classical Model	
12:35-13:35	Lunch	
13:35-14:35	**Keynote 1**	
	Boi Faltings (École Polytechnique Fédérale de Lausanne) *Game-Theoretic Mechanisms for Eliciting Truthful Information* Room CB-5	
14:35-14:40	Break	
14:40-15:55	Session 3	
	Parallel 3 A: **Aggregating Probabilities** *Chair: Taisuke Imai* *Room C1-5* Juntao Wang (Harvard University, United States) *A Bayesian Agent Model to Explain Why We Shall Use Logit Aggregation* Robert Mislavsky (Johns Hopkins University, United States) *Combining Probability Forecasts: 60% and 60% Is 60%, but Likely and Likely Is Very Likely* Taisuke Imai (Ludwig Maximilian University of Munich) *Dynamics of Nonlinear Probability Weighting in In-play Horse Race Betting*	Parallel 3 B: **Social Learning** *Chair: Giacomo Lanzani* *Room C1-6* Ines Lindner (Vrije Universiteit Amsterdam) *Naive Learning in Social Networks with Random Communication* Ville Satopaa (INSEAD) *Shepherding the Herd* Giacomo Lanzani (Massachusetts Institute of Technology, United States) *Robust Opinion Aggregation and its Dynamics*
15:55- 16:15	Coffee break	
16:15-17:30	Session 4	
	Parallel 4 A: **Incentivized Crowds** *Chair: Robert Schmidt* *Room C1-5* Tong Wang (Erasmus University Rotterdam) *Follow the Money: Bayesian Markets to Extract Crowd Wisdom* Cem Peker (Erasmus University Rotterdam) *Extending prediction markets to elicit counterfactual beliefs*	Parallel 4 B: **Errors in Forecasting** *Chair: Emily Haisley* *Room C1-6* Florian Peters (University of Amsterdam) *Measuring Biases in Expectation Formation* Teresa Bago d'Uva (Erasmus School of Economics) *A New Decomposition of the Brier Score with an Application to Individual Subjective Survival Probabilities*

	Robert Schmidt (Heidelberg University) *Using Coordination on Probabilities as an (Incentivized) Crowd Wisdom Device*	Emily Haisley (BlackRock Inc., United States) *Reading Trade Diaries*

Time	**Tuesday June 25**	
9:00-9:30	Light breakfast	
9:30-10:30	**Keynote 2**	
	Anna Dreber Almenberg (Stockholm School of Economics) *Which Results Can We Trust? Combining Replications and Prediction Markets to Estimate the Reproducibility of Scientific Results* Room CB-5	
10:30-10:50	Coffee Break	
10:50-12:05	Session 5	
	Parallel 5 A: **Expert Selection** *Chair: Stefan M. Herzog* *Room C1-5* Ville Satopaa (INSEAD) *Expert Selection Within a Single Prediction Problem* Stefan M. Herzog (Max Planck Institute for Human Development) *How to Detect High-Performing Individuals and Groups: Decision Similarity Predicts Accuracy*	Parallel 5 B: **Biases in Beliefs** *Chair: Chen Li* *Room C1-6* Tanjim Hossain (University of Toronto, Canada) *Belief Correlation with Signal Correlation* Alexander Coutts (Nova School of Business and Economics) *No One to Blame: Biased Belief Updating without Attribution* Chen Li (Erasmus University Rotterdam) *Signal Perception and Belief Updating*
12:05-13:15	Lunch	
13:15-14:30	Session 6	
	Parallel 6 A: **Social Influence** *Chair: Mark Himmelstein* *Room C1-5* Huihui Ding (University of Cergy-Pontoise) *Does Deliberation Improve the Reliability of Epistemic Democracy?* Helge Klapper (Rotterdam School of Management) *Organizational Decision Making: When Is Social Influence Useful?*	Parallel 6 B: **Eliciting Beliefs** *Chair: Severine Toussaert* *Room C1-6* Elias Tsakas (Maastricht University) *Robust Scoring Rules* Thomas de Haan (University of Bergen) *Eliciting Entire Belief Distributions Using a Random Partitioning of the State*

	Mark Himmelstein (Fordham University, United States) *Receptiveness to Advice from Humans, Algorithmic Models and Ensembles in Forecasting*	*Space* Severine Toussaert (University of Oxford) *Measuring Sophistication about One's Future Self: A Comparison of Methods*
14:30-14:40	Break	
14:40-15:55	Session 7	
	Parallel 7 A: **Aggregation 2** *Chair: Ville Satopaa* *Room C1-5* Paul Koster (Vrije Universiteit Amsterdam) *The Economics of Participatory Value Evaluation* Juntao Wang (Harvard University, United States) *Aggregation via Peer Assessment* Ville Satopaa (INSEAD) *A Default Bayesian Aggregator of Experts' Predictions*	Parallel 7 B**: Bayesian Truth Serum** *Chair: Sonja Radas* *Room C1-6* Jens Witkowski (Frankfurt School of Finance & Management) *The Robust Bayesian Truth Serum* Yanwei Jia (National University of Singapore) *Crowd Wisdom and Prediction Markets* Sonja Radas (Institute of Economics Zagreb) *Uncovering Reliable Respondents: An Application of Bayesian Truth Serum*
15:55-16:15	Coffee break	
16:15-17:15	**Keynote 3**	
	David V. Budescu (Fordham University, United States) *The Wisdom of Forecasting Teams* Room CB-5	
17:15-17:45	Closing remarks Room CB-5	

Keynotes

Boi Faltings (École Polytechnique Fédérale de Lausanne)

Game-Theoretic Mechanisms for Eliciting Truthful Information

Several different mechanisms for eliciting truthful information from self-interested participants have been proposed in recent years, among them peer prediction, output and correlated agreement, Bayesian and peer truth serum. I will survey the different principles underlying these mechanisms and thus characterize where they might be most applicable.

Anna Dreber Almenberg (Stockholm School of Economics)

Which Results Can We Trust? Combining Replications and Prediction Markets to Estimate the Reproducibility of Scientific Results

Why are there so many false results in the published scientific literature? And what is the actual share of results that do not replicate in different literatures? We will discuss these questions, see what prediction markets might add to the understanding of the reproducibility of science and discuss what we can do to increase the reliability of scientific results.

David V. Budescu (Fordham University, United States)

The Wisdom of Forecasting Teams

This talk is motivated by, and relies on data from, recent large-scale geopolitical forecasting tournaments (Mellers et al., 2014). Two key results of the tournaments are (1) the possibility to identify reliably expertise (Budescu & Chen, 2015; Chen et al, 2016) of individual forecasters and leverage it to improve the accuracy of the forecasts through efficient aggregation of relatively small crowds of "selected" forecasters (Mannes et al. 2014), and (2) the surprising success of small collaborative teams. This "teaming effect" is particularly intriguing because it seems to contradict the "wisdom of the crowd" hypothesis that emphasizes the importance of independence among forecasters. I will discuss both results with special attention to the "teaming effect" which I attribute to the hybrid approach implemented, which allowed forecasters to share information electronically and asynchronously, but required them to provide forecasts individually, for a statistical aggregation procedure. This hybrid approach benefits from the strengths of Computer Mediated Communication (CMC) and Statistical Aggregation.

We show that if one has access to a large number of forecasters, one can increase prediction accuracy in a novel statistical way. Place individual forecasters into teams and incentivize teams to compete against each other. Then construct new teams that combine small subsets of original teams using different original teams. This process leverages the benefits of within-team cooperative and between-team competition, two natural tendencies that can motivate people to do their best.

Parallel 1 A: Aggregation

Dennie van Dolder (Vrije Universiteit Amsterdam)

The Wisdom of the Inner Crowd in Three Large Natural Experiments

The quality of decisions depends on the accuracy of estimates of relevant quantities. According to the wisdom of crowds principle, accurate estimates can be obtained by combining the judgements of different individuals. This principle has been successfully applied to improve, for example, economic forecasts, medical judgements, and meteorological predictions. Unfortunately, there are many situations in which it is infeasible to collect judgements of others. Recent research proposes that a similar principle applies to repeated judgements from the same person. This paper tests this promising approach on a large scale in a real-world context. Using proprietary data comprising 1.2 million observations from three incentivized guessing competitions, we find that within-person aggregation indeed improves accuracy and that the method works better when there is a time delay between subsequent judgements. However, the benefit pales against that of between-person aggregation: the average of a large number of judgements from the same person is barely better than the average of two judgements from different people.
Joint work with Martijn van den Assem.

John McCoy (Wharton School - University of Pennsylvania, United States)

A Possible Worlds Model: Some Theoretical Connections

Under a possible worlds model of Bayesian respondents forming votes and meta-predictions given their private information, the surprisingly popular answer identifies the correct answer. We explore a number of theoretical results derived from this model. We examine both information aggregation and ranking respondents by expertise, and the relationship between selecting the surprisingly popular answer and the Bayesian Truth Serum. We collate multiple aggregation mechanisms under this model, identifying a particular role for the stationary distribution of the meta-prediction matrix. Lastly, we discuss how to estimate the complete model for various inputs, and thereby derive finite sample probability estimates.
Joint work with Drazen Prelec (Massachusetts Institute of Technology).

Benjamin Tereick (Erasmus University Rotterdam)

The Wisdom of Self-Aggregating Crowds

How best to form a single judgment out of many is an age-old problem in decision theory. Yet, it has so far not been studied how well a group of individuals will fare when it aggregates its own judgments. To answer this question, I propose “self-aggregation” (SELF), a method which asks individuals to vote for an option and to simultaneously provide a threshold of the number of people that would convince them of the opposite. SELF picks an option if more people vote for it than the average threshold provided in the group. I compare the performance of SELF, both theoretically and experimentally, with simple and confidence-weighted majority voting, and with the Surprisingly Popular Algorithm (SPA) recently proposed by Prelec et al. (2017). In a model in which individuals update their beliefs in a Bayesian fashion, I show that SELF is predicted to outperform alternatives. In an experimental test, respondent solve a binary decision problem in a stylized urn experiment in

which responses and aggregation results can be directly compared to the Bayesian prescription. In the experiment, SELF compares favorably to (simple and confidence-weighted) majority voting, but does not realize its theoretical advantage over the SPA. The results show that while the meta-cognitive abilities of individuals are challenged by complex methods such as SELF and the SPA, responses contain sufficient information to outperform methods based on less challenging questions.

Parallel 1 B: Truthful Reports

Yen-Lin Chiu (Johns Hopkins University, United States)

Online Product Rating with Purchase Data

We propose a mechanism to elicit user's experience in an online market where the user's purchase decision is known to the market platform and user's ratings are publicly observable. There exists an unique informative equilibrium where users tell their true experience for a product purchased on the platform. The key ingredient that induces the incentive to truthfully reveal information is the correlation between product rating and user's purchase decision: potential users will be less willing to purchase the product if they read any negative signal from the ratings. An intutive reward strucutre that is similar to second-price auction is shown to encourage users to report truthful rating according to above intuition. It is also not necessary for the platform to know the common prior between users. However, this mechanism is not without limitations: it cannot be free from babling equilibria where the users rate the product regardless of their true experience and potential users simply ignore ratings for their purchase decision. But thanks to user's purchase data, such an equilibrium will be easy to observe and how the platform's reputation alleviates the problem is discussed.

Yan Xu (Erasmus University Rotterdam)

Top-Flop Betting: An Incentive Mechanism to Elicit Unverifiable Truth

This paper introduces a simple mechanism to incentivize truth-telling even when the underlying truth is unverifiable. Most similar mechanisms in the literature rely on the existence of a common prior about the distribution of possible answers and on Bayesian arguments. In our mechanism, respondents to a question bet on the answers of others, relative to the answers given to other questions. For instance, people are asked to bet whether they think a given movie will get higher ratings than another, random movie. The bet reveals whether people themselves liked the movie. We call this method ``top-flop betting'' and show that it provides incentives to truthfully reveal private information, even in the presence of biases in the answers the bets are based on. Unlike existing methods, our method (i) relaxes assumptions on common prior; (ii) is robust to risk aversion and certainty effects, basically requiring first-order stochastic dominance only; (iii) leads to truth-telling as dominant in individual setting and a Bayesian equilibrium in game setting (not needing a Bayesian Nash equilibrium).
Joint work with Aurelien Baillon.

Naman Goel (Swiss Federal Institute of Technology (EPFL) Lausanne)

Decentralized Oracles via Peer-Prediction in the Presence of Lying Incentives

We derive conditions under which a detail-free minimal peer prediction mechanism can be used to elicit truthful data from non-trusted rational agents when an aggregate statistic of the collected data affects the amount of their incentives to lie. Furthermore, we discuss the relative saving that can be achieved by the mechanism, compared to the rational outcome, if no such mechanism was implemented. Our work is motivated by distributed platforms, where decentralized data oracles collect information about real-world events, based on the aggregate information provided by often self-interested participants. We compare our theoretical observations with numerical simulations on two publicly available real datasets.
Joint work with Aris Filos-Rastikas and Boi Faltings.

Parallel 2 A: Weighting Judgments

Shu Huang (Carnegie Mellon University, United States)

Getting More Wisdom from the Crowd: When Weighting Individual Judgments Reliably Improves Accuracy or Just Adds Noise

The theoretically optimal method to maximize the wisdom of a crowd (i.e., the accuracy of aggregated judgment) is a weighted average of the individual judgments where the optimal weights are determined by the accuracy (biases), reliability (variances) and dependency (correlations) of the individual judgments (Lamberson and Page, 2012; Davis-Stober et al., 2014; 2015). In practice, however, a simple average of the individual judgments (i.e., equal-weighting) often outperforms the theoretically optimal weighted average because the judgment biases, variances, and correlations are unknown and estimating them from empirical data produces unstable weights (Kang, 1986; Winkler and Clemen, 1992). We explore the conditions under which the optimal-weighting method with weights computed from a judgment covariance matrix estimated from empirical data is reliably more accurate than the simple average. Their accuracies depend on the sample size (i.e., the number of judgments from each forecaster in the empirical dataset) and on the (unknown) true covariance matrix. We find that a surprisingly large sample size is required to ensure that the weighted average is even 95% as accurate as the simple average for any true covariance matrix. However, depending on the sample covariance matrix actually observed, a moderate sample size may be enough to make us reasonably confident that the weighted average will outperform the simple average. We develop an algorithm to test whether the observed judgments in an empirical dataset are sufficient for researchers to reject using the simple average and instead trust the weighted average as reliably more accurate than the simple average. By using simulated judgments, our algorithm can display a clear diagnostic value for deciding how to combine judgments, and we find that moderate sample sizes may sometimes be sufficient to generate reliably accurate optimal weights. The measure of reliability derived from our algorithm works well with real data and should play a central role in evaluating which judgment aggregation method to use.
Joint work with Russell Golman and Stephen Broomell.

Klaus Nehring (University of California, Davis, United States)

Weighing Experts, Weighing Sources: The Diversity Value

A decision maker (DM) needs to come up with a probability judgement over a set of events based on the judgments of a set of information sources such as experts. How? There are two basic approaches. These are often referred to as (Supra-)Bayesian vs. "mechanical" or "axiomatic". More appropriate terms and conceptualizations for these distinctions may be "belief revision" vs. "belief (prior) construction", or "maximalist" vs. "minimalist". The Bayesian approach assumes that the DM already has a prior over the joint distribution of facts (states of nature) and expert beliefs. This involves substantial prior information to ground assumptions on the prior, and in this sense a fair amount of expertise on part of the DM himself. If the DM can claim (justified) confidence in possessing such information, the Bayesian approach is justified as a special instance of Bayesian rationality. However, such claims may rest on thin grounds; in particular, the task of estimating the relevant joint distribution from the past joint track record of expert judgements is often fraught with difficulty, especially since there may be a significant risk of overfitting since expert judgments are often highly correlated. So, there will be many situations when a Bayesian approach is not workable or not viewed as sufficiently reliable. To deal with such situations, we propose a constructivist approach, which assumes the least input from the DM possible. Its minimalist stance comes in two parts. Given a set of expert weights, probabilities are aggregated by some form of weighted averaging, say linear or logarithmic opinion pooling. Such schemes can be supported both on axiomatic grounds and have also proven to be strikingly successful in practice, typically using equal weights. But equal weights have significant limitations of their own, that may limit or even hurt the estimation performance. First, some experts may have something to add, but little compared to others. So adding weak experts to the pool with equal weights may increase the reliance on noisy signals and thus, on balance, diminish performance. Second, some experts may be strong by themselves, but largely duplicate the expertise of others. Thus, adding strong yet similar experts to the pool with equal weights may lead to an overreliance on certain signals compared to others, again impairing performance. Thus, to realize the potential gains from a diverse set of experts without dilution, one needs to allow for unequal weights reflecting the differential quality and/or similarity of experts. This is where some subjective input by the DM is indispensable. To allow this input to capture judgments/information of differential quality and similarity, we require the DM to specify a "reliability function" that maps sets of experts to positive real numbers measuring their "reliability". Heuristically, "reliability" can be thought of as "expected precision", and the reliability of a subset of experts measures the expected precision that can be obtained by aggregating their judgments (using optimal weights). Mathematically, reliability functions are non-additive set functions that are assumed to have the properties of diversity functions in the sense of Nehring and Puppe (2002, A Theory of Diversity, Econometrica). That is, they are monotone and totally submodular; the latter means that a given expert adds (weakly) less potential precision the more experts, especially: the more similar experts, there already are in the pool. The core task of this paper is to determine the optimal weights to be assigned on the basis of their characterization in terms of a reliability function. For this purpose, we propose and axiomatize a weighting rule called the "Diversity Value". The Diversity Value is given by a logarithmic scoring criterion and can be viewed as minimizing a generalized relative entropy. Heuristically, the Diversity Value selects those weights that best reflect the distinct marginal contributions of each expert to the overall reliability of the available set. We also show that the Diversity Value can be characterized as a weighted Shapley value in which the source weights are determined endogenously as a fixed point. In addition, we

show a number of other properties of interest. Notably, the Diversity value obeys the desideratum that larger weights should be assigned to more distinct sources (the "Similarity Principle"). In the present paper, the characterization of experts by reliability functions is taken a primitive. There are different ways how one might conceptualize and structure the reliability assessment. This defines a rich set of questions for future theoretical work and practical application.
Joint work with Ani Guerdjikova.

Tina Nane (Technische Universiteit Delft)

A Cross-Validation Study of Experts' Judgments Using the Classical Model

Structured expert judgment is routinely employed for uncertainty quantification when data are not appropriate, pose some issues or are simply lacking. "The qualifier structured means that expert judgment is treated as scientific data, albeit scientific data of a new type " (Cooke, 1991). The Classical Model (CM) or Cooke's method is arguably the most rigorous method which gathers, evaluates and mathematically combines expert opinions. CM has been used in studies sponsored by the EU, ESA, WHO, EFSA, etc. The studies have successfully used CM and many other applications span over a multitude of fields, including climate change, nuclear safety, chemical and gas industry, civil engineering, natural disasters, etc. CM advances the idea of "experts in uncertainty", which underlines the imperative need for a proper account of uncertainty, as much as for field expertise. In this regard, experts' assessments are objectively evaluated with respect to statistical accuracy (or calibration score) and informativeness. The two scores are computed from questions whose answers are known (sometimes post-hoc) to the analyst, but not known to the experts. The two scores lead to a combined score, which, in turn, gives the performance-based weight of each expert. The aggregation of experts' assessments is usually referred to as a Decision Maker (DM). DM can be regarded as any other expert in the pool of experts. Its performance can be evaluated with respect to statistical accuracy and informativeness and can be compared with scores resulting from different weighting schemes, such as equal weighting. A cross-validation analysis can be undertaken to determine DMs, using various weighting schemes, on training sets and evaluate their performance on test sets. We will perform such a cross-validation study using experts' opinions from a recent structured expert judgment study. Furthermore, we will evaluate and compare the performance of various DMs.
Joint work with Roger Cooke.

Parallel 2 B: Game Theory

Leonard Wolk (Vrije Universiteit Amsterdam)

Proportional-Prize Contests to Predict Future Events: Does Group Size Matter?

In this paper we present a forecasting technique based on a Blotto game with n players and proportional payoffs. We show that this mechanism possesses perfect forecasting ability: With common knowledge of realization probabilities, for any group size, in the unique symmetric equilibrium, players' strategies fully reveal these realization probabilities. Findings in our laboratory experiment confirm the invariance of forecasting performance to group size when realization probabilities are common knowledge. When these probabilities are not common knowledge, we find

that groups of size three do better than groups of size two; a further increase in group size beyond three does not lead to a further improvement in performance. The evidence supports the use of the Blotto game with proportional payoffs as a viable method to elicit forecasts about future events that requires only a limited number of participants.
Joint work with Fan Rao (Maastricht University), Ronald Peeters (University of Otago).

Vessela Daskalova (Institute for Advanced Study in Toulouse, Toulouse School of Economics)

Categorization and Coordination

This paper considers the use of categories to make predictions. It presents a framework to examine when decision makers may be better off using fewer rather than more categories, even without exogenous costs of using more. We study three cases: individual prediction, coordination of predictions, and the convex combination of the two. The analysis focuses on how the attempt to coordinate predictions with others affects incentives for coarse categorization in different environments. We show that while a coordination motive does not provide incentives for coarse categorization in deterministic environments, it could provide such incentives in stochastic environments.
Joint work with Nicolaas J. Vriend.

Peter Katuscak (Aachen University)

Strategy-proofness Made Simpler

There is a growing evidence that many people do not report their preferences truthfully in strategy-proof mechanisms. As examples, consider student-school matching under a strategy-proof mechanism or the second-price auction. A leading explanation of this observation are cognitive limitations of players. We propose a novel way of framing any strategy-proof mechanism that aims to reduce the ensuing cognitive load on players. First, a player's reported type is used to determine opportunity sets of the other players, whereas reported types of the other players are used to determine the opportunity set of the player in question. Second, a player's private allocation is determined as (one of) the most preferred allocation(s) in her opportunity set according to her reported preference type. We then experimentally test whether this framing increases truthful reporting relative to ``traditional'' frames using top trading cycles. We find that the proposed framing increases the rate of truthful reporting by almost one half. Moreover, there is no effect for low numeracy subjects, whereas the rate of truthful reporting more than doubles for high-numeracy subjects.
Joint work with Thomas Kittsteiner.

Parallel 3A: Aggregating Probabilities

Juntao Wang (Harvard University, United States)

A Bayesian Agent Model to Explain Why We Shall Use Logit Aggregation

Aggregation using the logit model has been proved to be extremely successful on Good Judgment Project datasets~[Satopa ??a ?? et al., 2014]. Satopa ??a ?? et al derived the logit aggregator as the maximum likelihood estimation when the log-odds of agents' predictions are normally distributed around an extremized log-odd of the true probability, where the extremization models the system biases in human forecasts. In this work, we show that when all agents are Bayesian and have a common prior with observations of signals independently drawn from a fixed set of signals of the world, the extremized logit model also approximates the posterior probability given the common prior and observations. Our model further illustrates the connection between the logit aggregator and the flat Bayesian aggregator. The latter approximates the posterior prediction when these observations are fully independent.
Joint work with Yang Liu (UCSC), Yiling Chen (Harvard University).

Robert Mislavsky (Johns Hopkins University, United States)

Combining Probability Forecasts: 60% and 60% Is 60%, but Likely and Likely Is Very Likely

To make optimal decisions, people must accurately estimate the likelihood of uncertain events (e.g., "Will this price increase?"). As such, they may solicit opinions from multiple advisors, who may provide verbal ("prices will likely increase") or numeric ("60% chance that prices will increase") forecasts. Although existing research documents differences in how we process verbal and numeric probabilities in isolation, less is known about how we aggregate forecasts of each type. In four studies, we find that people typically average numeric probabilities but count verbal probabilities (i.e., making forecasts that are more extreme than each advisor's). In Study 1 (N=205), participants saw a stock, its most recent price, and two advisor forecasts about whether the price would increase in a year. The advisor forecasts were either numeric ("60-69%") or verbal ("Rather Likely"), with both advisors giving the same advice. Participants then provided their own forecasts on scales that matched the advice (numeric: 1="0-9%", 10="90-100%"; verbal: 1="Nearly Impossible", 10="Nearly Certain"). Critically, all advisor forecasts corresponded to the 7th point on their respective scales. More participants gave extreme forecasts (i.e., greater than 7) in the verbal condition (30.1%) than in the numeric condition (11.8%), p=.001. In Study 2 (N=806), we replicate this effect manipulating the number of advisors within-subjects and using probabilities both above and below the scale midpoint. Participants were assigned to one of four between-subjects conditions in a 2 (numeric vs. verbal) x 2 (above vs. below midpoint) design using stimuli similar to Study 1. However, instead of seeing advisor forecasts simultaneously, participants saw the first advisor's forecast, made a prediction, then saw the second advisor's forecast and could revise their prediction. For probabilities above the midpoint, the proportion of extreme forecasts increased from 18.3% to 29.7%, p<.001, after participants saw the second advisor in the verbal condition, but decreased from 11.4% to 9.0% in the numeric condition, p=.21. For probabilities below the midpoint, the proportion of extreme forecasts increased from 13.1% to 23.1% in the verbal condition, p=.001, but decreased from 18.3% to 13.4% in the numeric condition, p=.01. Study 3 (N=626) replicates these findings using real expert forecasts. Participants were assigned to one of four between-subjects conditions in a 2 (numeric vs. verbal) x 2 (one vs. two advisors) design. Participants predicted the outcomes of 10 baseball games

by providing the likelihood that the favored team would win. For each game, participants saw one or two real expert forecasts on a 0%-100% scale (e.g., “53%”) or a 0-100 verbal probability scale (e.g., “53 “ Somewhat Likely”). Participants then made their own forecasts on a 0-100 scale with verbal or numeric labels, depending on condition. When seeing one advisor, participants in the verbal and numeric conditions are equally likely to make an extreme forecast (56% vs. 50%), p=.17. However, when seeing two advisors, participants were much more likely to make an extreme forecast in the verbal condition (47% vs. 30%, p<.001; interaction: p=.005). In Study 4 (N=809), we test how combination strategies carry over into decision-making. Participants read a scenario where they were making a purchase involving uncertainty. They saw an advisor’s forecast, given numerically or verbally and suggesting participants wait for the uncertainty to be resolved. Participants then indicated whether they would buy the item or wait (1=definitely buy; 7=definitely wait). They then saw a second forecast, qualitatively identical to the first, and again indicated their choice. More participants increased their likelihood of waiting in the verbal condition (33.8%) than in the numeric condition (20.5%; p<.001), suggesting that participants update their beliefs more after seeing a second verbal prediction than a second numeric condition, which then influences decisions.
Joint work with Celia Gaertig (University of Pennsylvania).

TaisukeImai (Ludwig Maximilian University of Munich)

Dynamics of Nonlinear Probability Weighting in In-play Horse Race Betting

There is evidence from many experiments that low probabilities of risky outcomes are overweighted, and high probabilities are underweighted. In the context of gambling such as horse racing, the over- and under-weighting pattern is manifested by the favorite-longshot bias (FLB). Specifically, bettors value longshots (horses with a relatively small chance of winning) more than expected given how rarely they win, and they value favorites too little given how often they actually win. As a result, the expected return from any bet increases with the probability that the event will occur. The bias is considered as an important deviation from the market efficiency hypothesis, which argues that the betting odds for an event provide the best forecast of its probability of occurrence and that the expected return at all odds will be the same. In this paper, we use a high-frequency limit-order book dataset on horse race betting from an online betting exchange market. It is one type of "prediction market," in which probabilities derived from market prices prove to be useful in forecasting. A unique aspect of this market is that traders can bet not only before races start but also after the races start. This feature allows us to estimate calibration curves from the betting odds that the crowd establishes at each point in time, and to investigate the emergence and evolution of non-linear probability weighting over time. We obtain two main findings. First, unlike standard findings in parimutuel betting markets which show evidence supporting FLB, we find that betting odds taken from a 10 minute time window prior to races are well-calibrated (that means, there is no FLB). Second, when we look at time window just before (40 seconds to 5 seconds) races finish, market odds exhibit systematic FLB. Furthermore, the degree of bias gets larger as the races approach to the finish line.
Joint work with Colin F. Camerer.

Parallel 3 B: Social Learning

Ines Lindner (Vrije Universiteit Amsterdam)

Naive Learning in Social Networks with Random Communication

We study social learning in a social network setting where agents receive independent noisy signals about the truth. Agents na??vely update beliefs by repeatedly taking weighted averages of neighbors' opinions. The weights are fixed in the sense of representing average frequency and intensity of social interaction. However, the way people communicate in real life is random such that agents do not update their belief in exactly the same way at every point in time. Our main finding suggests the following. Even if the social network does not privilege any agent in terms of influence, a large society almost always fails to converge to the truth. We conclude that wisdom of crowds seems an illusive concept and bares the danger of mistaking consensus for truth.
Joint work with Bernd Heidergott, Jia-Ping Huang.

Ville Satopaa (INSEAD)

Shepherding the Herd

This article analyzes multiple experts who forecast an underlying dynamic state based on a stream of public and private signals. Each expert minimizes a convex combination of her forecasting error and deviation from the other experts' forecasts. As a result, the experts exhibit herding behavior -- a bias that has been well-recognized in the economics and psychology literature. Our first contribution derives and analyzes the experts' optimal forecast under different levels of herding. This extends the Kalman filter and smoothing to applications where the underlying dynamics can be non-linear and herding is an important part of the process. Our second contribution is a welfare analysis where we show that, on average, the precision of public information affects welfare more than the level of herding among the experts. However, on average, the level of herding decreases the heterogeneity in the experts' forecasts more than the precision of public information. We also show that the negative effects of a sudden drop in public information are greater and last longer at higher levels of herding. Our third contribution shows how the model can be estimated in practice and used in a simple compensation scheme that minimizes the negative effects of herding.
Joint work with Jussi Keppo (NUS).

Giacomo Lanzani (Massachusetts Institute of Technology, United States)

Robust Opinion Aggregation and its Dynamics

We consider a generalization of DeGroot's linear model of social learning. By relaxing the assumption of a quadratic utility for the agents, we obtain an opinion aggregator that is normalized, monotone, and translation invariant. We directly link these properties to the natural conditions of the micro-foundation. In addition to the less demanding assumptions on the payoff function of the agents, the opinion aggregator allows for several economically relevant patterns ruled out by the linear model. For instance, agents can feature homophily, dislike (or attraction) for extreme opinions as well as discard information obtained from sources that are perceived as redundant. We also show that under this weaker assumptions is still possible to explore the standard questions addressed by the linear model, such as convergence of limit opinions, and the properties of consensus and

wisdom for this limit.
Joint work with Simone Cerreia-Vioglio, Roberto Corrao.

Parallel 4A: Incentivized Crowds

Tong Wang (Erasmus University Rotterdam)

Follow the Money: Bayesian Markets to Extract Crowd Wisdom

The answer to many questions, such as whether extraterrestrial life exists, is unknown. It is also uncertain when and how the answer will be known. In such situations, asking experts for their opinion seems a reasonable thing to do but we face two problems. First, if experts disagree, should we trust the majority? Second, how can we incentivize their truth-telling if we do not know whether or when the correct answer will be known? In this paper, we solve both problems simultaneously. We design a market in which experts report their opinion about a statement (endorse it or not) and bet on each other's opinion. Those who endorse the statement are offered to buy a bet that more than p% of others will also endorse it, where p is randomly drawn. Those not endorsing the statement can sell the bet (equivalently, buy the opposite bet, that more than (100-p)% of others will not endorse it). We then ''follow the money'', selecting the opinion of those who got a positive payoff on average. We demonstrate theoretically and illustrate empirically that our market elicits truthful answers and that ''following the money'' outperforms selecting the majority opinion.
Joint work with Aurelien Baillon and Benjamin Tereick.

Cem Peker (Erasmus University Rotterdam)

Extending prediction markets to elicit counterfactual beliefs

How do we evaluate causal effects of an investment or a public policy? Standard approach is Rubin causal model, which considers a counterfactual world where the project is not implemented and compares outcomes in actual and counterfactual worlds. In practice, counterfactual outcomes are not observable. Researchers either use natural interventions or controlled experiments, relying on assumptions that restrict how intervention/treatment may affect treatment and control groups. We propose an extended prediction market to elicit beliefs on both actual and counterfactual outcomes. Agents first participate in a standard prediction market for outcomes in the actual world. Then, they bet on counterfactual beliefs of others in a secondary market. The standard prediction market aggregates beliefs on outcomes in the actual world. The secondary market incentivizes agents to reveal if they consider favorable outcomes more likely in the counterfactual world. Such information can be used to form wisdom of crowds estimates for signs of hypothesized causal effects.
Joint work with Aurelien Baillon.

Robert Schmidt (Heidelberg University)

Using Coordination on Probabilities as an (Incentivized) Crowd Wisdom Device

We examine whether coordination on probabilities serves as a crowd wisdom device. That is, subjects are asked about the probability of a particular event, and they then have to coordinate on a percentage number by stating an integer between 0 and 100. We tested the proposed mechanism by asking subjects to estimate probabilities for events in an ultimatum game conducted by Trautmann and van de Kuilen (2014). In the main treatment, subjects had to estimate probabilities of the proposers?? and responders?? actions in the ultimatum game, and they were paid based on the precision with which they anticipated the average probability stated by the "crowd" (i.e. all subjects within a session). In the control treatment, subjects did not coordinate, but they were incentivized to state their actual beliefs about the factual probabilities of the proposers?? and responders?? actions. Our data strongly supports the hypothesis that coordination on probabilities serves as a crowd wisdom device, as we do not find any differences between the main treatment (coordination) and the control treatment (beliefs). That is, the subjects in the coordination treatment stated the same probabilities as the subjects in the belief treatment. Also, we compare the performance of the proposed mechanism with Trautmann and van de Kuilen (2014), who elicited beliefs about the same ultimatum game using six different methods: introspection, probability matching, outcome matching (corrected and uncorrected) and quadratic scoring rule (corrected and uncorrected). We find that the proposed mechanism, i.e. coordination on probabilities, is significantly more accurate in terms of Brier scores (Brier, 1950) than either of the approaches examined by Trautmann and van de Kuilen (2014). We therefore conclude that the proposed mechanism is suited as an (incentivized) crowd wisdom device. The mechanism appears particularly appealing with regard to the elicitation of probabilities that are unverifiable or of events that that did not yet occur, such as upcoming elections, financial market events or sport matches.

Parallel 4B: Errors in Forecasting

Florian Peters (University of Amsterdam)

Measuring Biases in Expectation Formation

We develop a general framework for measuring biases in expectation formation. The basic insight is that under- and overreaction to new information is identified by the impulse response function of forecast errors. This insight leads to a simple and widely applicable measurement procedure. The procedure yields estimates of under- and overreaction to new information at different horizons. Our framework encompasses all major models of expectations, sheds light on existing approaches to measuring biases, and provides new empirical predictions. In an application to inflation expectations, we find that forecasters underreact to aggregate shocks but overreact to idiosyncratic shocks.
Joint work with Simas Kucinskas.

Teresa Bago d'Uva (Erasmus School of Economics)

A New Decomposition of the Brier Score with an Application to Individual Subjective Survival Probabilities

Subjective probabilities about person-specific events, such as survival to a particular age, are now elicited in many surveys of the general population, such as the Health and Retirement Study (HRS).

These data are intended to help model individual behaviour with respect to saving, retirement and insurance that is presumed to depend on subjective distributions of future life events. Very little is known about the accuracy of individuals' subjective expectations of events that are specific to them. Early analyses of HRS data on subjective probabilities of survival to a particular age show a positive correlation with true survival and negative correlations with objective risks, including health problems and parental mortality. Such correlations, however, provide an incomplete assessment of prediction accuracy. In a previous paper, we used a scoring rule (Brier score) to measure accuracy in a formal and complete way. We show these subjective probabilities are highly inaccurate, particularly among the lower educated and cognitively less able, which potentially leads to grave mistakes in household decision making. In this paper, we introduce a new method to decompose prediction accuracy that is suited to explaining the sources of the inaccuracy of subjective probabilities of person-specific events. The method requires data on subjective probabilities, outcomes of the event that is predicted and on cues that are relevant to the outcome and potentially influence formation of the subjective probabilities. This new decomposition draws on the vast literature on probability judgement in the fields of meteorology and psychology. It is based on the Yates decomposition of the Brier score into bias, discrimination, outcome uncertainty and pure noise. Making use of the lens model framework, we extend the decomposition to identify the contributions of i) inappropriate weighting of the cues and ii) private information that comes from knowledge of person-specific risk factors that are not captured by the measured cues. We use this new decomposition to explain differences in the accuracy of subjective survival probabilities (SSP) by educational attainment in the US using data from the HRS. This spans 22 years and enables comparison of respondents' predictions of their chances of living to 75 with their true survival to that age. We find that SSP are very inaccurate - Brier score larger than 0.25, worse than if everyone had reported a fifty-fifty chance of surviving to 75. This inaccuracy is greatest for the least educated. The decomposition reveals that, in part, that is due to low educated facing greater uncertainty about longevity because of higher mortality rates. But it is also because their predictions are much noisier, consistent with lower ability to form beliefs about longevity. SSP have some, albeit low, discriminatory power: on average, the probability of survival to 75 reported by those who reach that age is 10 pp higher than the probability reported by those who do not. Our decomposition further unveils that the low discriminatory power is partly due to inappropriate weighting of (insufficient responsiveness to) risk factors, including onset of disease (e.g. cancer, lung, stroke, etc.), smoking and body mass index. Even the higher educated underestimate mortality risks but the lower educated do so by about twice as much. While SSP are inaccurate, the decomposition reveals that they do contain private information that predicts longevity. However, they are much less accurate than predictions obtained by regressing them on the cues ' ie, from aggregation of SSP over individuals with the same cues. This is because the resulting gain from reduced individual noise in SSP more than compensates the loss of private information. This dominance of noise over private information in SSP holds for all groups, but especially for the lower educated. Our conclusions shed new light on determinants of (in)accuracy of subjective survival probabilities. Our method can also be applied to an array of subjective probabilities collected routinely in population surveys. Their accuracy has hitherto not been assessed and is potentially influenced in varying degrees by factors such as uncertainty, inappropriate weighting and private information. It also holds promise for evaluation of a variety of forecasts by professionals and experts, as well as in experimental settings. Owen O'Donnell (Erasmus School of Economics; University of Macedonia).

Emily Haisley (BlackRock Inc., United States)

Reading Trade Diaries

Markets are often held up as the quintessential illustration of the “wisdom of the crowds”. They are an aggregation of incentivized forecasts of future value that distil this wisdom into price. We examine an antecedent of this market mechanism, using a “trade diary” dataset from a sample of 15 fund managers at Blackrock. The dataset contains the forecasts and logic behind roughly 4,000 trades. Immediately prior to ordering a trade, fund managers recorded their rationale, the expected time horizon, and their conviction. Conviction was measured as the probability that the bought (sold) security will out (under) perform the fund’s benchmark over the expected time horizon. This captured verifiable probabilistic predictions. Comparison of predicted and actual hit rates suggest overconfidence in estimations of uncertainty. Improvement with longer, more detailed rationales suggest a value for slow thinking. Natural language processing (NLP) is used to make the text of the rationales analysable and suggests alignment between ex-ante and ex-post factor attribution. Most interestingly, the NLP analysis reveals which fund managers have well-calibrated risk sensors, the value of emotion, and who is accurately able to harness second order beliefs.
Joint work with Shweta Agarwal, Nicky Lai.

Parallel 5A: Expert Selection

Ville Satopaa (INSEAD)

Expert Selection Within a Single Prediction Problem

Averaging predictions from multiple judges often provides a more accurate estimate than using the prediction from a single individual, a phenomenon called ``the wisdom of crowds''. This prediction can often be improved further by averaging predictions from a subset of judges, known as a select crowd. However, previously proposed approaches for deciding which and how many judges to include in the select crowd rely on past performance data from similar questions. This paper develops methodology to identify a high-performing subset of judges within a single prediction task. Judges are asked to estimate both the quantity of interest and the average prediction of all other judges. Predictions of others are then used as part of a customized criterion for identifying a select crowd. Although testing every possible subgroup is combinatorially intractable and prone to overfitting, we propose a sequential selection procedure that is robust to noise and can be solved quickly. The merits of this method are illustrated with data from several studies.
Joint work with Asa Palley and Jack Soll.

Stefan M. Herzog (Max Planck Institute for Human Development)

How to Detect High-Performing Individuals and Groups: Decision Similarity Predicts Accuracy

Distinguishing between high- and low-performing individuals and groups is of prime importance in a wide range of high-stakes contexts. While this is straightforward when accurate records of past performance exist, in most real-world contexts, such records are unavailable. Focusing on the class of binary decision problems, we use a combined theoretical and empirical approach to develop and

test a novel approach to this important problem. First, we employ a general mathematical argument and numerical simulations to show that the similarity of an individual's decisions to others is a powerful predictor of that individual's decision accuracy. Second, testing this prediction with several large data sets on breast and skin cancer diagnostics, geopolitical forecasting and a general knowledge task, we find, as predicted, that decision similarity robustly permits the identification of high-performing individuals and groups. Our findings offer an intriguingly simple, yet broadly applicable, heuristic of improving real-world decision-making systems.
Joint work with Ralf H. J. M. Kurvers, Ralf Hertwig, Jens Krause, Mehdi Moussaid, Giuseppe Argenziano, Iris Zalaudek, Patricia A. Carney, Max Wolf.

Parallel 5 B: Biases in Beliefs

Tanjim Hossain (University of Toronto)

Belief Correlation with Signal Correlation

Using a set of incentivized laboratory experiments, we characterize how people form beliefs about a random variable based on independent and correlated signals. First, we theoretically show that while pure correlation neglect always leads to overvaluing correlated signals, this depends on the exact structure of signal generation process if people misperceive precision. Specifically, they may sometimes undervalue correlated signals depending on the correlation structure. Our experimental results support this theoretical finding. Subjects form their beliefs in an unbiased, yet suboptimal, way. They do not completely neglect precision level or correlation structure of signals. While they do overvalue moderately or strongly correlated signals, we find that they undervalue weakly correlated signals. Estimated parameters of our model suggest that subjects show a high level of correlation neglect and also suffer from overprecision --- believing the signals are more precise than they actually are. Additionally, we find that subjects do not fully benefit from wisdom of the crowd --- they undervalue aggregated information about others' actions in favor of their private information. This is consistent with the models of overprecision where people do not properly incorporate the variance reducing power of averages.
Joint work with Ryo Okui (Seoul National University).

Alexander Coutts (Nova School of Business and Economics)

No one to blame: Biased belief updating without attribution

Evidence suggests that individuals are on average overconfident about their ability, affecting career and financial decisions, among others. We investigate how overconfidence may persist in the face of objective feedback. Self-attribution biases are said to exist when we take credit for good outcomes, but blame poor outcomes on external factors. While heavily studied in social psychology, and often referenced in economics, rigorous evidence is scarce. We present a modified Bayesian model of self-attribution bias, which distinguishes biases in attribution towards idiosyncratic noise versus a stable

fundamental factor, thus defining two types of attribution bias: (1) noisy, and (2) fundamental. Using an experiment where individuals receive noisy performance feedback that also depends on a teammate, we elicit beliefs about both own and the teammate's performance to identify precise patterns in attribution. Our novel elicitation procedure operates indirectly: subjects are incentivized to report true beliefs in order to maximize their payoffs in the experiment, but are not paid directly for accuracy. In this way our procedure has all the theoretical properties of "matching probabilities" and the binarized scoring rule but remains intuitive and relevant for subjects. Our main results are that individuals are overconfident, and take too much credit for positive feedback. However, they significantly under-weight negative feedback, in a way that leads us to reject both types of self-attribution bias. We show biased information processing using both a structural model of modified Bayesian updating as well as non-parametric matching on prior beliefs with a fully powered control experiment where subjects report beliefs about two strangers matched in a team.
Joint work with Leonie Gerhards, Zahra Murad.

Chen Li (Erasmus University Rotterdam)

Signal Perception and Belief Updating

This paper introduces a theory of signal perception to study how people update their beliefs. By allowing perceived signals to deviate from actual signals, we identify the probability that people miss or misread signals, giving indices of conservatism and confirmatory bias. In an experiment, we elicited perceived signals from choices and obtained a structural estimation of the indices. The subjects were conservative and acted as if they missed 43% of the signals they received. Also they exhibited confirmatory bias by misreading 19% of the signals contradicting their prior beliefs.
Joint work with Ilke Aydogan, Aurelien Baillon, Emmanuel Kemel.

Parallel 6A: Social Influence

Huihui Ding (University of Cergy-Pontoise)

Does deliberation improve the reliability of epistemic democracy?

We study the effects of deliberation on epistemic social choice, in two settings. In the first setting, the group faces a binary epistemic decision analogous to the Condorcet Jury Theorem. In the second setting, group members have probabilistic beliefs arising from their private information, and the group wants to aggregate these beliefs in a way that makes optimal use of this information. We assume that each agent wants other agents to agree with her, and discloses her private information to persuade the other agents for this purpose; this is how we model deliberation. We find that deliberation is guaranteed to improve the performance of the group only under certain conditions; these involve the nature of the social decision rule, the group size, and also the presence of ‘neutral observers’ whom the other agents try to persuade.
Joint work with Marcus Pivato.

Helge Klapper (Rotterdam School of Management)

Organizational Decision Making: When Is Social Influence Useful?

We present a formal computational model to analyze how social influence relationships shape organizational decision-making involving information aggregation. Our model reproduces several important patterns from experiments and empirical observations, namely that small groups may fail to surface privately held information, that the quality of decision making is sensitive to order of speech effects, that groups may fall victim to the illusion of consensus (such that public and private votes diverge sharply), as well as sometimes succumb to "groupthink" or "pluralistic ignorance". Critically, we also identify the conditions when groups, through social influence, improve decision quality. Our analysis suggests that these different outcomes can all arise under different settings of the same basic mechanism: a tradeoff between certification and censoring of complementary information, which is shaped by the distribution of influence patterns and preferences in the group. Joint work with Boris Maciejovsky, Phanish Puranam, Markus Reitzig.

Mark Himmelstein (Fordham University, United States)

Receptiveness to Advice from Humans, Algorithmic Models and Ensembles in Forecasting

When making forecasts, people often have various resources at their disposal to aid their judgments. Past research has suggested that under many conditions, people show a preference for probabilistic advice generated by a human rather than by an algorithm; while in other cases people may prefer the algorithmic advice. This study sought to understand under what conditions people might prefer one source to the other. Two hundred and fifty Amazon Mechanical Turk users were asked to make forecasts on 20 different items about actual geopolitical and economic outcomes. Items varied with regard to their temporal distance to resolution (time horizon): half of all items were set to resolve within approximately two weeks and the other half within approximately six weeks. After making a preliminary forecast, participants were then shown an advisory reference forecast and given an opportunity to revise their original judgment. The reference forecasts were described as having been obtained from either human experts or statistical models, but were obtained by averaging the results of a previous pilot study. Participants were also assigned to different conditions such that advice was either described as obtained from a single expert (or algorithm; single condition) or averaged across multiple experts (or multiple algorithms; ensemble condition). In a secondary experimental task, the same participants were shown an additional 10 items, and asked whether they thought statistical models or human experts would be better suited to forecast them. Results indicated no overall preference between algorithmic and human advice across all items, but most participants showed a preference for algorithmic advice for longer time horizons and human advice for shorter ones. This contrasted with results from the secondary task, in which people expressed belief that algorithms would be better suited to forecast economic items and humans would be better suited to forecast political items, but did not expect there would be a difference with regard to time horizon. Overall, people adjusted their forecasts in about 50% of the cases in the direction suggested by the advice. These adjustments were beneficial, as they systematically increased the accuracy of the forecasts. No differences were found between single and ensemble conditions for updating behavior, but participants in the single conditions were more accurate than participants in the ensemble conditions in both their preliminary and revised forecasts for both human and algorithmic advice. A follow-up experiment will replicate this procedure with the addition of a third condition, in which human and algorithmic advice is aggregated together to form 'hybridized' reference forecasts.
Joint work with David V Budescu.

Parallel 6B: Eliciting Beliefs

Elias Tsakas (Maastricht University)

Robust Scoring Rules

Does the mere exposure of a subject to a belief elicitation task affect the very same beliefs that we are trying to elicit? Is it theoretically possible to guarantee that this will not be the case? In this paper, we introduce mechanisms that make it simultaneously strictly dominant for the subject to (a) not update his beliefs as a response to the incentives provided by the mechanism itself, and (b) report his beliefs truthfully. Such non-invasive mechanisms are called robust scoring rules, and they are useful in a number of settings. First, their existence guarantees that the usual assumption of stationary beliefs (that we often explicitly or implicitly impose, e.g., in revealed preference tests or in experimental designs) is at least theoretically plausible. Second, robust scoring rules are needed for eliciting unbiased estimates of population beliefs. We prove that robust scoring rules exist under mild assumptions. Our existence proof is constructive, thus identifying an entire class of robust scoring rules. Subsequently, we show that well-known scoring rules (viz., the quadratic and the discrete) are approximately robust in the sense that they can arbitrarily approximate the subject's beliefs. Hence, robustness does not force us to use some exotic elicitation mechanism. Instead, it merely places restrictions on the magnitude of incentives that the scoring rule provides.

Thomas de Haan (University of Bergen)

Eliciting Entire Belief Distributions Using a Random Partitioning of the State Space

I introduce a new method to incentivize the elicitation of belief distributions, the Random Partitions Method. With this method, an agent's payoff not only depends on the realized state and the elicited distribution, but also on a random two-level partitioning of the state-space. The method creates a binary lottery payoff structure where reports closer to an agent's true belief distribution generate a higher probability to earn a high payout. The randomization of the state-space partitioning ensures that the agent is incentivized to report correctly across the entire distribution. I compare the introduced Random Partitions method with both the well known Quadratic Scoring Rule (Brier 1950, Savage 1971, and e.g. Selten 1998), and a method based on the Becker-DeGroot-Marschak procedure (see e.g. Karni, 2009) and argue that the Random Partitions method gives substantially stronger truth-telling incentives to agents in situations where there are many states/bins.

Severine Toussaert (University of Oxford)

Measuring Sophistication about One's Future Self: A Comparison of Methods

Measuring individuals' beliefs about what they will accomplish in the future is a very challenging task. In the presence of self-control problems and signaling concerns, the belief elicitation mechanism is likely to distort both the distribution of beliefs about a given outcome and the distribution of realized outcomes. First, individuals likely to face a self-control problem might use the elicitation mechanism as a commitment device to implement their plan. Second, individuals might

use the elicitation procedure as a way to signal their "type" (to themselves or to others) so as to preserve their image or maintain optimism in the face of challenges. The objective of this talk is to discuss (i) how various elicitation methods can distort the joint distribution of beliefs and outcomes; (ii) propose ways of measuring bias and correcting for distortions.

Parallel 7A
Aggregation 2

Paul Koster (Vrije Universiteit Amsterdam)

The Economics of Participatory Value Evaluation

This paper develops a novel approach to the economic evaluation of public policies: participatory value evaluation (PVE). PVE involves citizens directly in decisions of the government, taking into account governmental and individual budget constraints. Citizens receive reliable information on social impacts and can choose the best portfolio of projects according to their social preferences. This paper develops the economic and econometric theoretical framework for fixed budget and flexible budget PVE experiments which allows us to directly measure the change in social welfare for investments in water infrastructure in The Netherlands.
Joint work with Thijs Dekker (University of Leeds), Niek Mouter (TU Delft).

Juntao Wang (Harvard University, United States)

Aggregation via Peer Assessment

In this paper, we explore the possibility of using peer information to do peer assessment so that we are able to i) calibrate the accuracy of each individual without assessing ground truth, and ii) make more accurate aggregated predictions. We derive the peer assessment methods from the peer prediction and scoring rule literature and show by experiments over 14 real-world datasets that these assessment methods can effectively achieve the above two objectives.

In particular, we used surrogate scoring rules [Liu and Chen, 2018] and proxy scoring rules [Witkowski et al., 2017] with proxy generated by different aggregators, which include the extremized mean [Witkowski et al., 2017], variational inference EM algorithm [Liu et al., 2012], surprisingly popular algorithm [Prelec et al., 2017], to evaluate the mean accuracy of individuals across multiple forecasting questions. We call these the peer assessment scores (PA scores) for each individual. Then, we used the mean or the logit model [Satopa ??a ?? et al., 2014] to aggregate the predictions from individuals with higher accuracy based on the PA scores.

We ran experiments on the Good Judgment Project [Atanasov et al., 2016] datasets, the datasets collected by Prelec et al. [Prelec et al., 2017], and the Hybrid Forecast Competition datasets (www.iarpa.gov/index.php/research-programs/ hfc?id=661) and showed that these PA scores of an individual have an extremely high positive correlation with the mean Brier score of the individual (with correlation coefficient above 0.8 and p-value???0.05 on half of the datasets) and the aggregators applied on a subset of individuals with better PA scores outperformed the mean, median, logit model, variational inference algorithm and surprisingly popular algorithm on these

datasets.
Joint work with Yang Liu (UC Santa Cruz), Yiling Chen (Harvard University).

Ville Satopaa (INSEAD)

A Default Bayesian Aggregator of Experts' Predictions

This work is about forecasting future events. In particular, we develop a Bayesian aggregator that inputs a decision-maker's (DM) prior probability of the event, a.k.a., the base rate and then updates it based on experts' probability predictions. The underlying probability model is parametric, describes heterogeneity in terms of information asymmetry and noise, and captures the experts' bias, precision, and dependence. For cases where the base rate cannot be chosen from past data or by the DM, we motivate a default choice based on the experts' predictions alone. This way our aggregator becomes automatic and can be applied directly to any set of predictions. We illustrate our aggregator on a real-world dataset with thousands of experts and millions of predictions about 500 future events in finance, politics, and other areas. Overall, our aggregator improves the mean squared error of the average prediction by 23% and other state-of-the-art aggregators by 13-28%.

Parallel 7B: BTS

Jens Witkowski (Frankfurt School of Finance &Management)

The Robust Bayesian Truth Serum

The Bayesian Truth Serum (BTS) is the first peer prediction mechanism that truthfully elicits private information from respondents without requiring knowledge of the respondents' belief models. It asks for two reports: a report about the information itself (the opinion, rating, or experience, henceforth referred to as signal) and a prediction report corresponding to a respondent's belief about the distribution of signals in the population. The mechanism's major drawback is that it is truthful only for a large number of respondents, where this number depends on the respondents' belief model and is thus unknown to the mechanism.

We design the Robust Bayesian Truth Serum (RBTS), which alleviates this problem. As in BTS, RBTS takes a signal and a prediction report, and does not require knowledge of the respondents' belief model. However, RBTS is already truthful for two or more respondents. (An earlier analysis of RBTS required at least three respondents and was restricted to binary responses.) RBTS has additional advantages over BTS: first, the payments computed by RBTS are bounded for all reports, and these bounds can be set to any values chosen by the designer. For example, the designer can ensure that all payments will lie in between $0 and $1. Second, RBTS is ex post individual rationality, meaning that no respondent incurs a negative payment in any outcome. This is important for crowdsourcing applications, where it is often infeasible for the mechanism to receive payments from respondents. Moreover, analogous to proper scoring rules in forecasting, RBTS also identifies experts among the respondents. In contrast to proper scoring rules, however, RBTS does so without needing to know the correct answers or outcomes. Interestingly, RBTS also has tight technical connections to recently proposed algorithms for aggregating crowd responses with meta beliefs.
Joint work with David C. Parkes (Harvard University).

Janwei Jia (National University of Singapore)

Crowd Wisdom and Prediction Markets

Thanks to digital innovation, the concept of crowd wisdom, which aims at gathering information (e.g. Wikipedia) and making a prediction (e.g. using prediction markets) from a group's aggregated inputs, has been widely appreciated. An innovative survey design, based on a Bayesian learning framework, called the Bayesian truth serum (BTS), was proposed previously to reduce the bias in the simple majority rule, so as to get a consistent estimator, by asking additional survey questions. A natural question is whether we can extend the BTS framework to prediction markets (not just polls). To do so, this paper proposes two estimators, one based on a prediction market alone and the other based on both the market and a poll question. We show that both estimators are consistent within the BTS framework, under different sets of regularity conditions. Numerical results are given to illustrate the convergence of different estimators.
Joint work with Min Dai (National University of Singapore), Steven Kou (Boston University).

Sonja Radas (Institute of Economics Zagreb) and Drazen Prelec (MIT)

Uncovering Reliable Respondents: An Application of Bayesian Truth Serum

Many areas of economics use subjective data, although it had been known to present problems regarding its reliability. To improve data quality, researchers may use scoring rules that reward respondents so that it is most profitable for them to tell the truth. However, if the subjects are not well informed about the topic or if they do not pay sufficient attention, they will produce data that could not be reliably used for decision-making even though subjects gave their honest answer. The problem is compounded by the fact that it is usually not possible to a priori differentiate “bad” respondents from the “good” ones. In this paper we develop a model based on Bayesian Truth Serum, a scoring rule that incentivizes truth telling. A crucial property of BTS is to reward meta-knowledge, which is measured by accuracy of respondents’ predictions about choices of their peers. We show how these peer-predictions can be used for identification of reliable respondents, which allows researchers to discard the unreliable data. We use purchase intention survey, a popular method to elicit early adoption forecasts for a new concept, as a test bed for our approach. We present results from three online experiments, demonstrating that corrected purchase intentions are closer to the real outcomes.
Joint work with Drazen Prelec (Massachusetts Institute of Technology).

https://bayesiancrowd.com/