Crowdsourcing the Future: Predictions Made with a Social Network

Clifton Forlines[1] **Sarah Miller**[1]
[1]Draper Laboratory
Cambridge, MA
cforlines@gmail.com, smiller@draper.com

Leslie Guelcher[2] **Robert Bruzzi**[2]
[2]Mercyhurst University
Erie, PA
{lguelc84, rbruzzi}@mercyhurst.edu

ABSTRACT
Researchers have long known that aggregate estimations built from the collected opinions of a large group of people often outperform the estimations of individual experts. This phenomenon is generally described as the "Wisdom of Crowds." This approach has shown promise with respect to the task of accurately forecasting future events. Previous research has demonstrated the value of utilizing meta-forecasts (forecasts about what others in the group will predict) when aggregating group predictions. In this paper, we describe an extension to meta-forecasting and demonstrate the value of modeling the familiarity among a population's members (its social network) and applying this model to forecast aggregation. A pair of studies demonstrates the value of taking this model into account, and the described technique produces aggregate forecasts for future events that are significantly better than the standard Wisdom of Crowds approach as well as previous meta-forecasting techniques.

Author Keywords
Social Network; aggregation; forecasting; crowd-sourcing; meta-forecast; Bayesian Truth Serum.

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work, Theory and models.

INTRODUCTION
Through forecasting future events, we are able to make better decisions about actions we can take in the present. As individuals, we routinely make forecasts about the future and use them to guide our actions; activities such as deciding to bring an umbrella to work or to take the stairs rather than the elevator are all made with our future-selves in mind. The happiness and well-being of our future-selves is largely a result of our ability to forecast our future environment accurately in the present.

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

ACM978-1-4503-2473-1/14/04$15.00.
http://dx.doi.org/10.1145/2556288.2556967

Outside of our personal lives, decision makers in many fields rely on the predictions made by individual experts. Researchers have long understood that aggregate estimations built from the individual opinions of a large group of people often outperform the estimations of individual experts [8,23]. The research field of Collective Intelligence contains a growing volume of work centered on this premise. The use of the Unweighted Linear Opinion Pool (ULinOP, a fancy term for group mean) has proven to be a robust method of aggregating judgments that often outperforms more complex techniques [9]. For example, in the domain of weather forecasting, "ensemble averaging" (another domain-specific term for arithmetic mean) is currently the gold standard and is rarely if ever beaten by alternative aggregation schemes [5]. However, recent research has shown that combining forecasts through averaging can sometimes lead to poor prediction performance in the field of political judgment [24], often because the performance of the individuals themselves is low. Because knowledge and forecasting ability resides in a small subset of the population, group forecasts are often incorrect when majority voting is applied [14,21].

Indeed, group forecasting is significantly more complicated than individual forecasting as characteristics of the group come into play. Groups benefit from the diversity of the knowledge of their members; but groups are hindered by inefficient communication among group members, social cognitive biases such as groupthink, and difficulties in combining the individual forecasts into accurate group forecasts. This paper focuses on the last of these problems – the accurate aggregation of individual predictions into group forecasts. Specifically, we focus on a new method that better identifies which members of the group are in a position to provide high-quality forecasts so that their contribution can be given more weight when creating an aggregate prediction.

In this paper, we present this new method for weighting and combining forecasts and present the results of an experiment in which an objective measure of prediction performance quantifies the benefits of this approach. This new approach extends previous aggregation techniques by recognizing that the familiarity among the members of the group is highly heterogeneous and by incorporating a model of the group's social network into the aggregation process.

MOTIVATION

The Intelligence Community (IC) is tasked with collecting information from a wide variety of sources and combining this information to produce accurate forecasts about world events (for example questions, see Table 1). Clearly, the impact of forecasts made in the IC can have large consequences; as such, even small improvements to the accuracy of forecasts made by groups can have large real-world impacts. More often than not, these forecasts are not made by an individual, but rather are the output of a team.

The authors of this paper participated in the Aggregative Contingent Estimation (ACE) Program sponsored by IARPA. The goal of the ACE program was "to dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts" [**Error! Reference source not found.**]. Beyond forecasting world events, the techniques presented in this paper have applications in the areas of medicine, educational testing, politics, and economics.

PRIOR WORK ON AGGREGATE JUDGMENTS

Research into methods for combining individual judgments into accurate group judgments has been the subject of a great deal of research over the past 100 years. As such, a detailed account of this literature is beyond this paper.[1] The field is still an active research area, and the introduction of ubiquitous networking has opened the door to the exploration of systems that query and combine judgments from a large numbers of diverse individuals in real-time. This section attempts to outline some of the efforts that most closely relate to the work presented in this paper.

Weighted and Unweighted Opinion Pools

The term Wisdom of Crowds refers to the general process of "taking into account the collective opinion of a group of individuals rather than a single expert [when] answer[ing] a question" [23]. It is most often used to describe the phenomenon that the mean answer of a group is often more accurate than any one individual's answer.[2] As discussed in the introduction, simple averaging or majority voting is a surprisingly robust method for aggregating individual judgments into accurate group judgments. Researchers have found that alternatives to simple averaging work well in specific domains [4,11,27], but the ULinOP, or arithmetic mean, works well in a large number of cases [9]. Indeed, many crowd-sourcing techniques assume that the average answer is correct, and focus instead on either gathering that average quickly or even estimating the group average from a subsample of the population [7].

[1] For a good overview of techniques see [3,10].

[2] Galton's original demonstration of this effect [8] used the group median rather than the mean.

With respect to combining forecasts of future events, there is ample evidence that forecast accuracy can be significantly improved through combining multiple individual forecasts, and that simple techniques (such as the ULinOP) often outperform more complex methods [3]. This said, theoretical arguments have been made that no linear combination of forecasts provides an optimal combination [20], suggesting that further work is required in the field.

Prediction Markets and Collaboration

A popular alternative to linear averaging for aggregation is the use of a prediction market. Prediction markets (as their name implies) are markets that can be used to aggregate information from market participants about the likelihood of future events (e.g., [26]). Generally, these markets are highly accurate and compare favorably with the ULinOP.

Recent work has focused on leveraging the combined benefits of ULinOP and Prediction Markets [10].These hybrid approaches, which average the prediction of several methods (so called "poll of polls"), have shown promise, and have performed well in recent U.S. election forecasting.

Prediction Markets rely on a large number of participants and are inherently "zero sum" in their reward structure. As such, they do not foster collaboration among individual forecasters who are seen as competing, although there is inherent information sharing in the market due to buy/sell pricing. Recent work on the role of collaborative approaches to combining expert predictions have shown very good results [25]. In this work, small teams working collaboratively to come up with consensus forecasts were able to outperform linear averaging techniques for a large percentage of forecasting questions examined.

Another related area of collaboration that has received a good deal of attention in recent years is collaborative, or social, search [2,6,13]. Researchers have identified that searchers often query their social network for answers [6], and have demonstrated that there is value in leveraging people's beliefs about the knowledge of other individuals when reading responses to crowd-sourced answers [13].

Forecaster Knowledge and Quality Assurance

Much work has been completed on methods to improve aggregate forecasts through applying quality assurance methods to the answers given by individuals. Because of the popularity of on-line crowd-sourcing tools like Amazon Mechanical Turk, many groups used to provide aggregate answers include a number of individuals who provide unreliable answers in order to quickly receive their compensation. Research into automatically identifying these poor and malicious performers [18] and into filtering out low-quality answers through making multiple comparisons among judgments [21] can help improve aggregate answers in many cases.

One particular method of filtering out poor answers/ forecasters is through the use of a "gold question" [15,17].

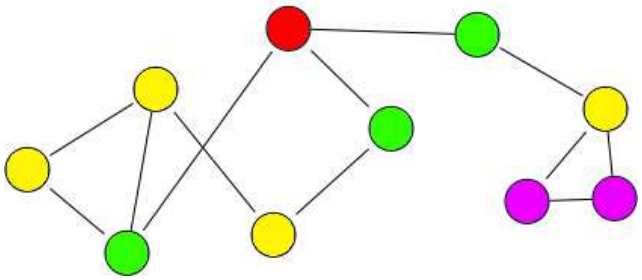


Figure 1. An illustrative social network.

A gold question has a known answer and is a question that the experimenter has determined has a high level of diagnostic value. Those forecasters who can answer the gold question correctly have a high probability of answering the question of interest correctly. While the value of such questions has been demonstrated, it is difficult to generate gold questions quickly and reliably for many topics of interest [16].

Prelec [19] introduced the intriguing notion that a meta-prediction could act as a sort of common gold question with a high level of diagnostic value for factual and forecasting questions. A meta-prediction is an individual's expectations concerning other people's forecasts, and can be elicited at the same time as an individual's personal forecast.

Prelec's Bayesian Truth Serum (BTS) [19] is the fundamental aggregation method that we base our aggregation approach on. In a simple variation of the technique, each individual forecaster provides not only their own forecast about a question, but also their meta-forecast about the average answer from the group as a whole. Forecasters who can predict what others will say are generally better forecasters themselves. The BTS algorithm produces a BTS score, which is itself an indicator of expertise. Through using BTS scores to weight individual forecasts in the aggregation process, one can assign greater importance to the "experts" identified by BTS and thus outperform the traditional ULinOP Wisdom of Crowds technique. For reference, we have found improvements in aggregate forecasting performance in the range of 5-10% when using the BTS technique.

The research described in this paper stems from the observation that the current BTS approach to eliciting and using the meta-forecast from an individual assumes that the individual's knowledge of the group is uniform. Intuitively, one knows that this knowledge is *not* uniform, and in fact can be highly disparate. For example, an individual has varying degrees of knowledge about different subpopulations and even individuals in a larger group. Furthermore, this knowledge is often directional (for example, the authors of this paper know a lot more about the views and beliefs of the President of the United States than vice versa) and is dependent on topic areas (for example, the equipment manager of a sport team probably know a lot more about players' equipment preferences than their political leanings).

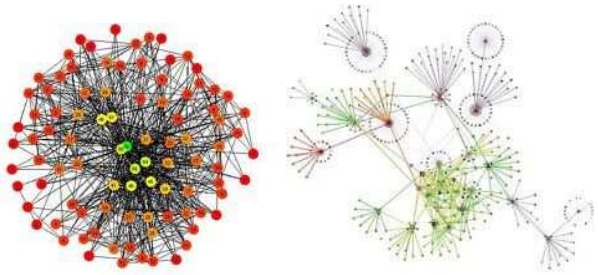


Figure 2. Two social network topologies that should greatly influence the value of meta-forecasts.

The motivation behind this investigation is to start to understand if an appreciation for this heterogeneity can be useful when generating aggregate forecasts. Toward this end, we investigate the usefulness of meta-forecasts made about familiar and unfamiliar group members, and describe an elicitation and aggregation technique that produces superior aggregate forecasts for a diverse set of forecasting questions.

ILLUSTRATIVE EXAMPLE

The baseline idea is to amend the BTS technique so that BTS scores are influenced by our understanding of the population's social network. The BTS approach to aggregation works by suppressing the answers from certain (highly predictable) people when it comes to averaging the groups' overall answer. Each individual is asked to model the distribution of answers given by the group, and as discussed previously, one's knowledge of the group is decidedly nonuniform.

One's view of the group as a whole is greatly influenced by the social network of that group, such as the one depicted in Figure 1. In this example, Alice (shown in red) is more closely familiar with the opinions and knowledge of her immediate neighbors (green) in this network than with those people further away from her (yellow, and then purple). In this example, Alice's opinion of the group is greener than the group actually is; thus, her meta-forecasts should be influenced more by the green members than those farther removed from Alice in the graph.

One can hypothesize that groups with a highly connected social network will result in inferior traditional BTS aggregations because each individual "knows" quite a lot about the knowledge of the group as a whole, which is what they're being asked to estimate. With this knowledge of the opinions of others shared among the group, the BTS technique would do little to identify the experts in the population.

Consider the two network topologies shown in Figure 2. The network on the left represents a highly connected network, such as you might find residing in a

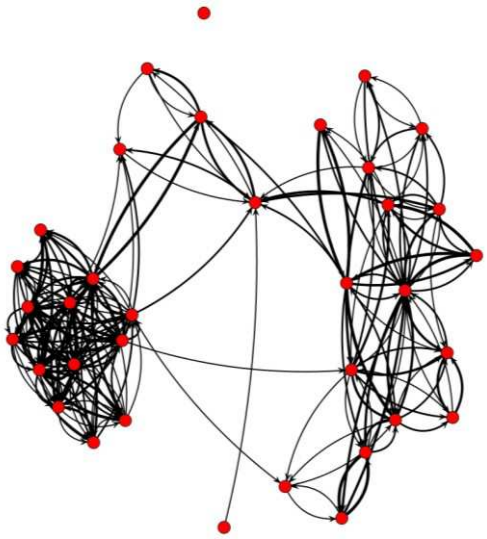


Figure 3. The social network of the pilot study investigation.

single location and working closely with one another. The network on the right would represent an organization with several remote locations that each includes a highly connected team connected to the overall organization by a small number of individuals. As we hypothesize that the value of the meta-forecast depends greatly on the familiarity among group members, understanding the social ties among the group may prove to be an important task when creating aggregate forecasts.

MEASURING FORECAST PERFORMANCE

In the two studies discussed in this paper, individual and group performance on forecast problems was measured using the Brier score [1]. The Brier score quantifies the accuracy of a probability forecast by computing the average squared deviation between predicted probabilities for a set of events and the (eventual) outcomes (detailed in Eq. 1).

$$BrierScore = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{r}(f_{ti} - o_{ti})^2$$

where:

- $\mathbf{f_{ti}}$ is the forecast probability
- $\mathbf{o_{ti}}$ is the binary indicator of the event outcome
- **r** is the number of possible outcomes
- **t** is the number of forecast instances

Equation 1. Measuring forecast accuracy [1].

The range of Brier scores is [0,2] where 0 indicates a 100% accurate prediction and 2 an inaccurate prediction. Applying the Brier scoring rule requires knowledge of the actual resolution for the forecasting problem. Consequently, Brier scoring can only be performed after the forecasting problem has closed and truth is known.

PILOT INVESTIGATIONS

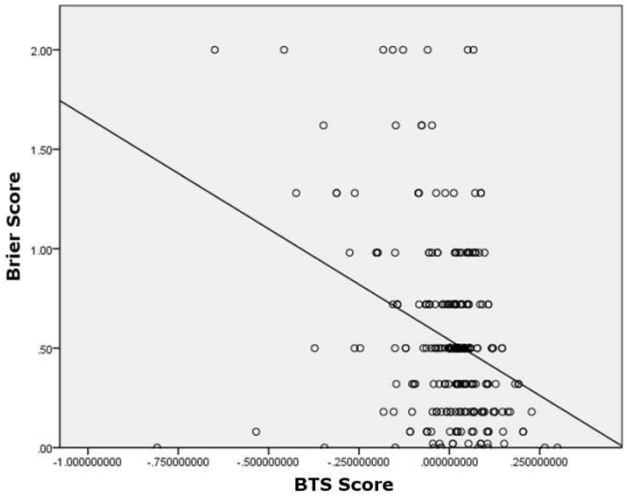
To begin investigating the usefulness of a knowledge network model for BTS-based forecasting, we designed a simple pilot experiment using students from [removed for blind review]. In this pilot study, 34 students answered 15 factual questions and 15 forecasting questions covering 4 topic areas: the Arab Spring, European Economics, North Korea, and Professional Sports.

In addition to providing a prediction (for forecasting questions) or a confidence (for factual questions), each student gave a BTS meta-prediction for the distribution of the groups' answers as well as predictions about the answers and confidences of other individual students. These other students included both students that they had previously indicated that they knew well as well as students that they did not know.

Modeling the Social Network of Student Forecasters

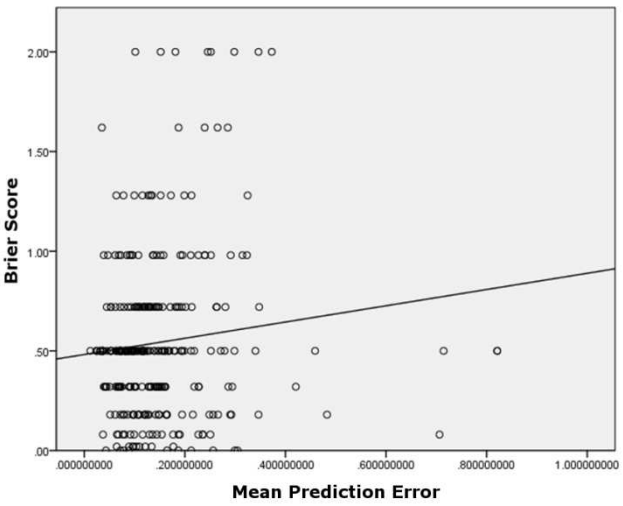
Before asking the factual and forecasting questions, we began by eliciting information from the 34 students that would help us build a simple knowledge network model. In this questionnaire, students were asked to rate their overall familiarity with each of the other students in the study on a 0-5 scale, with 0 indicating that they did not know the other student and 5 indicating that they knew the other student very well. With these connections in hand, we were able to build a social network model of this pilot study population. Figure 3 illustrates the model of the overall familiarity among students in the pilot study population.

Figure 3 clearly shows an important feature of our pilot population, namely that it is heterogeneous in terms of interconnectivity. After observing the two large cliques in



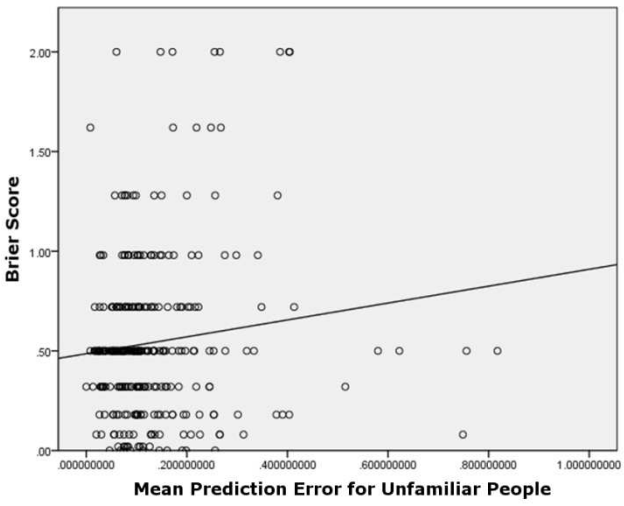
Pearson Correlation	-.326
Sig. (2-tailed)	<.001
N	309

Figure 4. BTS score is a good predictor of forecast performance. Lower Brier scores are better; higher BTS scores indicate expertise.



Pearson Correlation	.106
Sig. (2-tailed)	.062
N	309

Figure 5. Participants ability to predict other individuals' answers (X axis) and their own Brier score (Y axis).



Pearson Correlation	.117
Sig. (2-tailed)	.040
N	309

Figure 6. Ability to correctly predict the answers of other unfamiliar individuals (X axes) vs. one's own Brier score (Y axes). Lower error when predicting others' forecasts is correlated with better Brier scores for one's own predictions. Not shown, the nonsignificant correlation between ability to predict answers of other familiar individuals and Brier score.

this population, we discussed this finding with our graduate research assistant, who informed us that these cliques represented first and second-year students, respectively. The students who bridged these two groups are either taking courses ahead or behind the typical schedule, or acting as teaching assistants and student leaders.

Pilot Study Analysis and Results

First, we looked at the performance of BTS meta-predictions about the whole group. BTS scores were calculated as described in [19]. As expected, an individual's BTS score (which is immediately available when a forecast is made) was a good predictor of their performance as measured by the Brier score (which can only be calculated after truth is known). Figure 4 shows the relationship between BTS score and Brier score for this population.

Overall, we were happy to see that the BTS aggregation technique performed well with a population that had such a heterogeneous familiarity with one another. In addition to the group meta-predictions used to generate Figure 4, we asked participants to give meta-predictions about other individual students. These students included students that they knew well as well as those whom they did not know.

Overall, an individual's ability to predict other individuals' answers did not significantly correlate with Brier score. Figure 5 shows the relationship between participants' ability to predict other individuals' answers (X axis) and their own Brier score (Y axis). In this chart, the X axis shows the mean absolute error between an individual's prediction of others' forecasts and those people's actual forecasts. Lower (leftmost) positions on the X axes equate to better prediction of others' answers. While better predictions of others' answers did correspond to better (lower) Brier scores, this relationship was not significant.

While a participant's overall ability to predict others' answers was not predictive of their own ability to accurately answer questions and give forecasts (Figure 5), a different picture emerged when we looked separately at predictions of familiar and unfamiliar individuals. In general, one's ability to predict other individuals' answers, *whom one does not know*, correlates significantly with Brier score, while one's ability to predict other individuals' answers *whom one does know* did not. Figure 6 shows the significant relationship between participants' ability to correctly predict other unfamiliar individuals and their own Brier score.

Figure 6 suggests that the ability to correctly predict the answers of other unfamiliar individuals (X axes) is related to one's own Brier score (Y axes). Being able to predict others' answers indicates that you have a relatively high level of accuracy, but only for individuals whom you do not know well. This somewhat surprising relationship between familiarity and ability to predict others' answers and the accuracy of one's own forecasts indicates that there is value in modeling the social network of a group of forecasters. While interesting, the cause of this relationship is unclear. Perhaps the ability to predict the forecasts of a familiar group member demonstrates familiarity with that individual

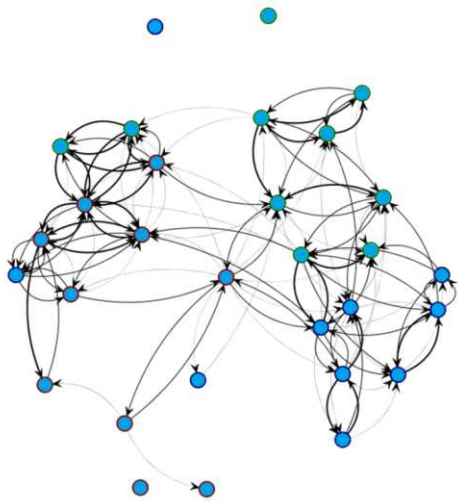


Figure 7. The social network of our study population. (Note that the weakest ties have not been drawn for clarity, which has the effect of isolating some nodes in this figure.)

and their opinions rather than knowledge about the forecast question itself. Perhaps other issues are at play. As this pilot experiment was not designed to answer this question definitively, one should be wary of putting too much confidence into the difference in correlation between familiar and unfamiliar individuals. What is clear is that additional research is needed to further explore this space.

MAIN EXPERIMENT

Given the evidence gathered in the pilot study indicating that there was value in looking at the social ties between individuals when considering their meta-predictions, we set out to design a follow-up study to investigate this relationship more closely. 30 students were recruited and asked to answer 10 forecasting questions, as well as provide meta-predictions and recommendations about fellow students.

To model our population, we began by asking each participant to rate their overall familiarity with every other student in the study on a 0-5 scale. This activity was repeated, with the variation that each student was asked to rate their familiarity with every other student's knowledge and opinions of four topic areas – Iran, North Korea, Europe, and Israel. The overall social network for this study population is shown in Figure 7. This study included 10 forecasting questions shown in Table 1. In addition to forecasting these 10 questions, participants gave BTS meta-predictions for the population as a whole, as well as predictions about other individuals in the study.

Additionally, knowing that peers are often good judges about their peers [2,6,13], we asked each student to pick the three students from the population that they felt would give the best answers to each of the 10 questions.

Will Israel officially announce that it recognizes the Armenian genocide before 1 April 2013? **(Mean Brier 0.24)**
Will the Republic of Macedonia be a NATO member before 1 April 2013? **(Mean Brier 0.19)**
Will any country officially announce its intention to withdraw from the Eurozone before 1 April 2013? **(Mean Brier 0.19)**
Will Iran sign an IAEA Structured Approach document before 1 April 2013? **(Mean Brier 0.16)**
Will Moody's issue a new downgrade of the long-term debt rating of the Government of Germany between 30 July 2012 and 31 March 2013? **(Mean Brier 0.36)**
Will Standard and Poor's downgrade the United Kingdom's foreign long-term credit rating at any point between 18 June 2012 and 1 April 2013? **(Mean Brier 0.49)**
Will the United Nations Security Council pass a new resolution directly concerning Iran between 17 December 2012 and 31 March 2013? **(Mean Brier 0.22)**
Will the sentence of any of the seven Italian experts convicted of manslaughter for failing to "adequately warn" about the L'Aquila earthquake be reduced, nullified, or suspended before 1 April 2013? **(Mean Brier 0.45)**
Before 1 April 2013, will substantial evidence emerge that Iran has enriched any uranium above 27% purity? **(Mean Brier 0.46)**
Will the United Kingdom's Liberal Democrats and Conservatives remain in a coalition through 1 April 2013? **(Mean Brier 0.23)**

Table 1. Forecasting questions from the main experiment and the mean Brier scores for study participants.

One-on-One Meta Predictions

Given the evidence from the pilot study, we designed a simple scoring mechanism that took advantage of all of the information gathered from the One-on-One predictions. This score would be used to weight individual forecasts when creating group aggregate forecasts. This score was calculated by looking at all possible pairs of forecasters in the group and transferring points among them in a zero-sum manner. Consider the following example, illustrated in Figure 8. The Red and Green participants make individual

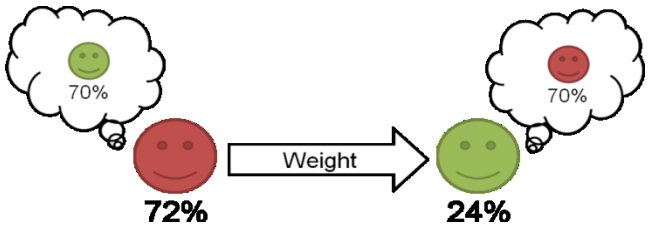


Figure 8. Weight transfer from One-on-One meta-predictions.

predictions about a question of 72% and 24%, respectively. They both predict that the other's answer was 70%. In this case, the Green participant did a much better job of predicting the Red participant than vice versa; as such, points are transferred from the Red participant to the Green participant. After all possible pairs of forecasters are examined in this manner, their individual forecasts are weighted by their final score. This has the combined effect of heavily weighting individuals who are good at predicting others while being difficult to predict themselves – both qualities of the BTS approach. Furthermore, this One-on-One method takes advantage of the nonuniform knowledge among participants as it samples all possible pairs rather than the group as a whole.

Equation 2 details the means by which the point-transfer is calculated. There are many such functions that should suffice, but this specific function was chosen as it is symmetric around zero (errors is either direction result in the same score transfer).

$$weightTransfer = \left| \ln \frac{metaPrediction}{actualPrediction} \right|$$

Equation 2. The transfer of points between pairs of participants is calculated using the ratio of meta-prediction to actual prediction.

Our study included two variations of the One-on-One approach in which the score transfer was scaled by the social ties between the pair. In one variation, familiarity increased the transfer of points, and in another variation, unfamiliarity did.

One limitation of the One-on-One approach is that to fully sample the pairs in a group, N*(N-1) meta-forecasts are needed. While any one individual only needs to make N-1 of these, this can still be burdensome for large groups. As such, we divided our population of 30 students into three groups of 10 (shown in Figure 9). These groups were built to include both strong and weak connections, and the use of multiple groups allowed for repeated measurements in our study, resulting in better statistical power.

Aggregation Methods and Design

Given our model of the social network and One-on-One aggregation design, we compared the following six aggregation techniques.

ULinOP. The mean forecast for the members of each group. This aggregate forecast is the traditional Wisdom of Crowds approach.

All-Stars. Because every participant picked the top students in the study to answer each question, we simply weighted each participant's answer by the inverse of their average rank when creating the aggregate forecast.

BTS Weighted. After calculating a BTS score using each participant's meta-prediction for the group as a whole, we weighted individual forecasts by the normalized BTS score to produce an aggregate forecast for the group.

One-on-One. This method used the One-on-One meta-predictions to calculate a weight for each forecaster using the method described in a previous section of this paper.

One-on-One Ties Make Stronger. In this variation, strong social ties (e.g., familiarity) increased the transfer of points between pairs of participants. This approach had the effect of giving participants who were good at meta-forecasting the answers of people they knew well more weight. As familiarity (F) was measured on a 0-5 scale, the point transfers were scaled by (F+1) in this variation.

One-on-One Ties Make Weaker. In this variation, strong social ties (e.g., familiarity) decreased the transfer of points between pairs of participants. This approach had the effect of giving participants who were good at meta-forecasting the answers of people they did not know well more weight. As familiarity (F) was measured on a 0-5 scale, the point transfers were scaled by 1/(F+1) in this variation.

Overall, our study design was:

10 binary forecasting questions x
3 groups x
6 aggregation methods =
180 measurements

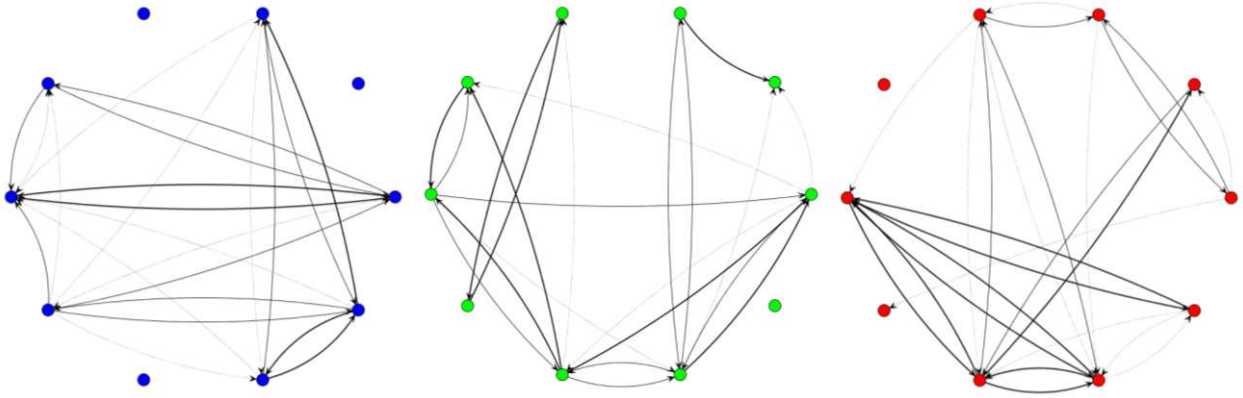


Figure 9. Each of our groups contained a mix of strong and weak social ties among participants.

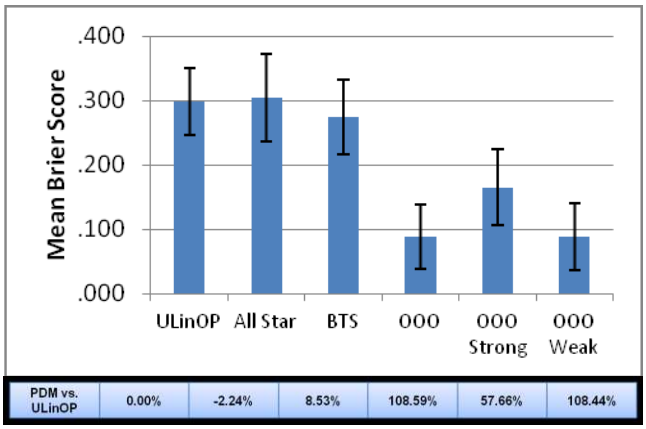


Figure 10. Mean Brier scores for each of the six aggregation methods. (Below) The PDM between each aggregation method and the ULinOP.

Results

Overall, we saw significant differences among the 6 aggregation methods for the 10 forecasting questions included in this study. A Repeated-Measures Analysis of Variance (ANOVA) using Brier Score as the dependent variable, Aggregation Method as a 6-level within-group contrast, and using Group ID as a 3-level between subjects factor showed a significant main effect for Aggregation Technique on Brier Score ($F_{5,130} = 4.871$, $p<0.001$, $\eta^2=0.158$). The mean Brier score (mean of each group for each forecast question) for each of the six aggregation techniques is shown in Figure 10. The Percent Difference in Means (PDM) between each aggregation technique and the ULinOP was -2.24%, 8.53%, 108.59%, 57.66%, and 108.44% for the All-Star, BTS, One-on-One, One-on-One Ties Make Stronger, and One-on-One Ties Make Weaker techniques, respectively.

Table 2 shows pairs of techniques that were significantly different from one another according to a post-hoc, pair-wise analysis of the results. After adjusting these multiple pair-wise comparisons using Bonferroni, our results suggest significant differences between the three One-on-One techniques and the ULinOP, All Star, and BTS techniques.

DISCUSSION

Overall, these results suggest that the One-on-One approaches do a significantly better job at identifying individual forecasters who should be weighed more heavily when creating aggregate forecasts.

Interestingly, the All-Star method performed poorly, indicating that these participants were not good at picking fellow students who would provide good forecasts. While the BTS approach did not produce a significantly improved forecast over the ULinOP in this study, we expect that this is due to low statistical power as we are seeing an absolute difference of around 9% PDM, which is in line with other BTS results. All variations of the OOO weighting approach work well. As found in the pilot study, one's ability to predict people you know well is not as good an indicator of expertise as one's ability to predict people you don't know.

The studies described in this paper were part of a two-year effort into forecast aggregation. As part of this larger effort, we did investigate models that looked toward an individual's previous performance. While this type of "leaderboard" weighing did help, it was not as valuable as one would hope. In the current study, it's not the case that a single amazing individual contributed to the good scores of the OOO techniques. It was the case that for any single question, there was often an individual whose weight was noticeably higher than the other individuals, but this person changed from question to question. An individual who could accurately answer such a diverse collection of questions would be a rarity indeed; however, the OOO technique seems to help identify the specific individual who is best able to answer a specific question.

	ULinOP	All Star	BTS	One-on-One	One-on-One Strong	One-on-One Weak
ULinOP	-	p=1.0	p=1.0	**p<0.001**	**p<0.001**	**p<0.001**
All Star	p=1.0	-	p=1.0	**p<0.001**	**p<0.001**	**p<0.001**
BTS	p=1.0	p=1.0	-	**p<0.001**	**p<0.05**	**p<0.001**
One-on-One	**p<0.001**	**p<0.001**	**p<0.001**	-	**p<0.005**	p=1.0
One-on-One Strong	**p<0.001**	**p<0.001**	**p<0.05**	**p<0.005**	-	**p<0.01**
One-on-One Weak	**p<0.001**	**p<0.001**	**p<0.001**	p=1.0	**p<0.01**	-

Table 2. Post-hoc, pair-wise comparisons between aggregation techniques. Significantly different pairs are shown in bold. Adjusted for multiple comparisons using Bonferroni.

As with any experiment, there are considerations that the reader should make when interpreting the results and their applicability to real-world scenarios. The following paragraphs attempt to outline some of the threats to the external validity of this study and our results.

The major limitation of applying this technique to many real-world scenarios centers on the difficulty and expense of sampling the social ties among the members of a group. Our studies used a relatively stable and reasonably-sized population of students. Through questionnaires, we were able to model the social network of these populations once and apply this model to our analysis. In other areas, it may not be possible or practical to gather the necessary relationship data to apply the One-on-One technique. Many populations are too large to measure in their entirety (e.g., movie reviewers on Netflix) or are too unstable to build a static model of the network (e.g., product reviewers on Amazon.com). Other populations, such as relatively static organizational teams, are more amenable to modeling. While not explored in the current research, existing social networks (such as Facebook and LinkedIn) may be good environments to explore these techniques. The reader should consider their population of interest and the potential difficulties in building their social network model when thinking about the application of this paper's technique to their aggregation problem.

This work was primarily motivated by the needs of the Intelligence Community and aims to help improve their ability to make accurate group forecasts about future events. While we have discussed the application of these techniques to other important fields (medicine, education, economics, etc.), it is hard to argue that in any field, group forecasts would not improve by aggregating the individual forecasts of experts. While it is widely understood that experts themselves can often fail when making predictions [24], better input into an aggregation engine should produce better output. Our study participants consisted of students enrolled in intelligence programs. While these students are impressive and quite knowledgeable, as students they should not yet be considered experts in their field. As such, future work should include studying the aggregation of expert forecasters so that the results of the study better match the intended use cases.

In this set of experiments, we investigated a small number of binary forecasting questions concerning world events in four topic areas. While binary questions have the benefits of being relatively easy to score and are relatively easy to determine the outcome of, these considerations should not limit the types of forecasting questions that new aggregation techniques are subjected to. Many real-world forecasting questions have multiple possible outcomes, or are better phrased as point predictions rather than discrete outcomes (e.g., the future price of oil). Future work should include extending the One-on-One techniques to these other question types, and should include extending the forecast topic areas to other important domains, such as medicine, politics, education, and economics. Similarly, forecasting is only one of a number of worthwhile judgments to explore in this line of research. Medical diagnostics, pathology, and forensics are all types of judgments that might benefit from the techniques described in this paper, although they were not studied per se. Future work should include investigations into other types of collective judgments, the aggregation of which is made easy by our increasingly networked culture.

Looking ahead, many of the open questions in the field of forecast aggregation center on the accurate elicitation of individual forecasts and (in our case) the elicitation of the meta-forecast and a population's social network. The CHI community is in an excellent position to contribute to both of these challenges, and we are pursuing collaborative efforts to extend the art in these areas.

CONCLUSION

Overall, we were encouraged to see the value in One-on-One meta-predictions and the value of incorporating familiarity (e.g., social ties) into meta-prediction methods. Looking ahead, we are interested in further improving our ability to apply these techniques to larger groups (with significant N*(N-1) problems) and would like to investigate subsampling pairs from the population. Similarly, our approach used a simple score transfer function, and the space of such functions should be looked into. Finally, we recognized that this study used three small groups with a mixture of levels of connectivity. It would be good to investigate the value of social network influenced meta-predictions in populations with strong and weak social ties.

The results show that there are significant benefits to employing a One-on-One meta-prediction weighting scheme when aggregating individual forecasts into group forecasts. The base One-on-One technique produced a 108.59 PDM score when compared with ULinOP aggregation. It is our hope that the strong results presented in this initial study, mediated by the reservations discussed in this paper, motivate others to join us as we continue to investigate improvements to aggregate judgments.

ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

REFERENCES

1. Brier, G.W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1950, pp. 1-3.
2. Chi, E.H., "Information Seeking Can Be Social," *IEEE Computer*, Vol. 42, No. 3, IEEE Press, pp. 42-46.
3. Clemen, R.R., "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 1989, pp. 559-583.
4. Dawid, A., DeGroot, M., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M., Lindley, D., McConway, K., and Winkler, R., "Coherent Combination of Experts' Opinions," *TEST*, Vol. 4, Issue 2, 1995, pp. 263-313.
5. Du, J., "Hybrid Ensemble Prediction System: a New Ensembling Approach," preprints, Symposium on the 50th Anniversary of Operational Numerical Weather Prediction, University of Maryland, College Park, Maryland, June 14-17, 2004 (5 pp.).
6. Ehrlich, K., Shami, N.S. "Searching for Expertise," *Proceedings of the ACM Conference on Human Factors in Computing Systems*, (CHI 2008), ACM Press, 2008.
7. Ertekin, S., Hirsh, H., and Rudin, C., "Learning to Predict the Wisdom of Crowds," Presented at Collective Intelligence Conference (arXiv: 1204.3611), 2012.
8. Galton, F., "Vox Populi," *Nature*, Vol. 75, 1907, pp. 450-451.
9. Genre, Kenny, Meyler, and Timmermann, "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting*, Vol. 29, Issue 1, 2013, pp. 108–121.
10. Graefe, A., Armstrong, J.S., Jones, R., and Cuzan, A., "Combining Forecasts: An Application to Election Forecasts," APSA Annual Meeting, 2011.
11. Hegselmann, R. and Krause, U., "Opinion Dynamics Driven by Various Ways of Averaging," *Computational Economics*, 25, 2005, pp. 381-405.
12. IARPA Aggregative Contingent Estimation (ACE). http://www.iarpa.gov/Programs/ia/ACE/ace.html
13. Jeong, J.W., Morris, M.R., Teevan, J., and Liebling, D., "A Crowd-Powered Socially Embedded Search Engine," *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, Boston, MA, July 2013.
14. Law, E. and von Ahn, L., "Input-Agreement: a New Mechanism for Collecting Data Using Human Computation Games," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '09)*, ACM, New York, 2009, pp. 1197-1206.
15. Le, J., Edmonds, A., Hester, V., Biewald, L., "Ensuring Quality in Crowd-Sourced Search Relevance Evaluation: The effects of Training Question Distribution," *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, 2010.
16. Miller, S., Forlines, C., Regan, J., "Exploring the Relationship Between Topic Area Knowledge and Forecasting Performance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56, No. 1, 2012, pp. 318-322.
17. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., Biewald, L., "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowd Sourcing," *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '11)*, ACM, New York, NY, USA, 2011.
18. Ipeirotis, P.G., Provost, F., and Wang, J., "Quality Management on Amazon Mechanical Turk," *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*, ACM, New York, NY, USA, 2010, pp. 64-67.
19. Prelec, D., "A Bayesian Truth Serum for Subjective Data, *Science,* 306, 2004, pp. 462-466.
20. Ranjan, R. and Gneiting, T., Combining Probability Forecasts," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, pp. 71–91.
21. Sun, Y. and Dance, C., "When Majority Voting Fails: Comparing Quality Assurance Methods for Noisy Human Computation Environment," Presented at Collective Intelligence Conference, 2012.
22. Sun, Y., Roy, S., Little, G., "Beyond Independent Agreement: A Tournament Selection Approach for Quality Assurance of Human Computation Tasks," *Proc. of: Human Computation,* San Francisco, California, USA, 2011.
23. Surowiecki, J., *The Wisdom of Crowds*, New York: Doubleday, 2004.
24. Tetlock, P., *Expert Political Judgment*, Princeton University Press, Princeton, NJ, 2006.
25. Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J., "The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions," *AAAI Fall Symposium Series,* 2012.
26. Wolfers, J. and Zitzewitz, E., "Prediction Markets, *Journal of Economic Perspectives, American Economic Association*, V 18, No. 2, '04, pp. 107-126.
27. Yaniv, I., "Weighting and Trimming: Heuristics for Aggregating Judgments under Uncertainty," *Organizational Behavior and Human Decision Processes*, 69, 1997, pp. 237-249.