Combinatorial Prediction Markets: An Experimental Study

Walter A. Powell, Robin Hanson, Kathryn B. Laskey, and Charles Twardy

Volgenau School of Engineering,
George Mason University
4400 University Drive
Fairfax, VA 22030-4444 USA
{wpowell, klaskey, ctwardy}@gmu.edu

Abstract. Prediction markets produce crowdsourced probabilistic forecasts through a market mechanism in which forecasters buy and sell securities that pay off when events occur. Prices in a prediction market can be interpreted as consensus probabilities for the corresponding events. There is strong empirical evidence that aggregate forecasts tend to be more accurate than individual forecasts, and that prediction markets are among the most accurate aggregation methods. Combinatorial prediction markets allow forecasts not only on base events, but also on conditional events (e.g., "A if B") and/or Boolean combinations of events. Economic theory suggests that the greater expressivity of combinatorial prediction markets should improve accuracy by capturing dependencies among related questions. This paper describes the DAGGRE combinatorial prediction market and reports on an experimental study to compare combinatorial and traditional prediction markets. The experiment challenged participants to solve a "whodunit" murder mystery by using a prediction market to arrive at group consensus probabilities for characteristics of the murderer, and to update these consensus probabilities as clues were revealed. A Bayesian network was used to generate the "ground truth" scenario and to provide "gold standard" probabilistic predictions. The experiment compared predictions using an ordinary flat prediction market with predictions using a combinatorial market. Evaluation metrics include accuracy of participants' predictions and the magnitude of market updates. The murder mystery scenario provided a more concrete, realistic, intuitive, believable, and dynamic environment than previous empirical work on combinatorial prediction markets.

1. Crowdsourcing, predictions, and combinatorial markets

Forecasting future events is important to business, national security, and society in general. Despite decades of research and immense resources dedicated to developing forecasting methods, getting better forecasts has proven elusive. Traditionally, forecasting has relied on judgments of a few experts. Measured on predictive accuracy, experts repeatedly disappoint, even when compared to simple statistical models like "no change"

or linear models with equal or even random weights [1-6]. Furthermore, the data required for statistical models may be unavailable or inadequate. Recently, crowdsourcing has been shown to improve on the judgments of individual experts, or small groups of experts.

Common practice in crowdsourcing is to average the judgments of a large group of individuals who have some knowledge of the problem [7]. Theory suggests that giving more weight to better forecasters should outperform a simple average, but in practice simple averaging has been surprisingly hard to beat. Recently, however, prediction markets have been shown to improve accuracy not only over individual or small groups of experts, but also over simple averaging [8-9]. A prediction market allows forecasters to aggregate information into a consensus probability distribution by purchasing assets that pay off contingent on an event of interest. Since the resources available to make predictions are limited, forecasters self-select to make forecasts for which they have the most information. Over time, the market gives greater weight to more successful forecasters. More accurate forecasters acquire greater resources with which to make further predictions; less accurate forecasters will lack the resources to have much influence on the consensus probabilities.

We are especially interested in the problem of forecasting many interrelated variables. For such problems, graphical models such as Bayesian networks provide a principled approach to modeling dependencies among variables. Pennock and Wellman [10] suggested the use of graphical models for belief aggregation. A combinatorial prediction market [8-11] increases the expressivity of an ordinary prediction market by allowing conditional forecasts (e.g., the probability of B given A is p) and/or Boolean combinations of events (e.g., the probability of B and A is q). Theory suggests that this greater expressivity, if appropriately captured in market prices, should give rise to more accurate forecasts. This is almost trivially true on joint forecasts, but should also hold for marginal forecasts when knowledge is distributed among participants and communication is primarily through the market. It should be particularly apparent if knowledge of correlations and knowledge of facts is held by different participants who communicate primarily via the market.

Specifying dependencies among forecasts using a graphical probability model allows tractable computation of a joint probability distribution among a large number of interdependent questions. If asset prices are set using a logarithmic market scoring rule (LMSR), then the assets can be factorized in a similar manner to probabilities, giving rise to similarly efficient algorithms for managing forecasters' assets [12].

For nearly two years the DAGGRE project [13-14] ran a public LMSR prediction market for geopolitical forecasting. The focus was on forecasting world events: usually questions with extended time horizons and significant irreducible uncertainty. These are the types of questions that have historically been the most vexing to intelligence analysts, economists and others. The DAGGRE market opened in October of 2011 as part of IARPA's Aggregate Contingent Estimation (ACE) program. The initial DAGGRE market was an ordinary ("flat") prediction market. In October of 2012, we launched a combinatorial prediction market. The market allowed users to forecast a question conditional on assumed values of another question. At any given time, there were on the order of 100

questions active on the market. Over time, some questions were removed as their outcomes became known, and new questions were added. Participants in the market were recruited from email solicitations, articles on blogs and newspapers, and personal recruiting at professional events. Participants received a small financial incentive for participating. Because of program restrictions, compensation did not depend on forecast accuracy, but the most accurate forecasters were recognized publicly on the DAGGRE site and listed on the leaderboard. Over the 20 months the market was open, more than 3000 participants contributed at least one forecast, with an average of about 150 forecasters per week. The market ran just over 400 total questions, about 200 of which were shared with four other teams in the IARPA-funded tournament.

Probability forecasts for the shared evaluation questions were reported daily to IARPA. When the outcome of an evaluation question became known, the question was scored by averaging the daily Brier score [15] over the period of time the target question was active. This approach has the benefit of rewarding forecasts that trend toward the correct outcome early during the period of time the question is being forecast. Forecasts were evaluated against a baseline system employing a uniformly weighted linear average of forecasts. Although early DAGGRE results were unreliable due to software issues, from February 2012 through May 2013, the DAGGRE market accuracy was about 38% greater than the baseline system. Accuracy was about the same before and after the launch of the combinatorial feature; however, usage of the combinatorial capability was low. About 10% of the users ever used the combinatorial feature, and only about 5% of the forecasts conditioned on another question. The DAGGRE prediction market closed in June of 2013 and will reopen in Fall 2013 with a change in focus to science and technology forecasting.

As a large-scale field study, the DAGGRE geopolitical market was not well suited to controlled experimentation. In this paper, we report a smaller-scale study that compares groups making the same predictions with the same information, one group using an ordinary flat market and the other using a combinatorial market.

2. Scope of the Experiment

The goal of the experiment was to investigate the effects on prediction market forecasts of allowing users to specify conditional probability links between questions. An experimental study [8] showed improved predictions with a combinatorial market, but on stylized forecasting problems with little face validity. The present experiment was designed to evaluate evidence and generate forecasts in a more concrete, realistic, intuitive, believable, and dynamic environment than previous work. Additionally, the experiment simulated the sequential nature of the flow of information in a prediction market.

3. Experimental design

Using the actual DAGGRE market and real-world questions would have provided the most concrete, realistic, and believable environment, but using the actual market would

cause experimental design challenges that could not easily be overcome. We therefore chose a "murder mystery" scenario to:

- Provide a concrete, realistic, intuitive, and believable, environment, in which relationships are based on statistical evidence familiar to the participants, e.g. men tend to be taller than women and people who wear bifocals tend to be older;
- Provide a common understanding of the basic relationships between the questions and clues upon which the belief structure (Bayes nets) could be built ;
- Use the same questions in a counterbalanced design;
- Control the delivery of information, providing clues sequentially (as in real prediction markets) and control the level of "expert" knowledge of the participants;
- Provide correct "gold standard" beliefs; and
- Control of the timing and order of the clues and outcomes of the questions, ensuring similar problems for different experimental runs.

The independent variables were:

- *Market* - the type of market, combinatorial or flat;
- *Market Order* - the order in which each type of market was used by the participants, combinatorial first or flat first.

The primary focus was on the effects of the *Market* variable. The *Market Order* variable allows analysis of interactions among the data due to learning effects. Each participant made predictions using both the combinatorial and flat markets.

In order to simulate the variation in levels of knowledge typical in prediction markets, different information concerning the relationships among the market questions and clues was distributed to the participants. In each session, each participant received conditional probability tables relating the market questions to each other and five of the ten conditional probability tables relating the clues (evidence) to the market questions. The conditional probability tables represented the expertise of each participant since the participants with a given table had the most accurate knowledge of the relationships between a specific clue and the market questions. Participants who didn't have access to specific tables had to rely on their general knowledge and the response of the market (including comments) to estimate the relationships between clues and questions.

The primary dependent variable was the accuracy of the predictions. For experiments on information aggregation, a common criterion is the ability to calculate ideal rational predictions given individual information and given the sum of all individual information. This ability allows us to define a mechanism's accuracy as the distance between an ideal distribution and the actual probability distribution produced by the mechanism. The measure of the accuracy of the predictions in each market was the Brier score. The Brier score is a proper scoring rule – that is, a forecaster minimizes his or her expected Brier score by accurately stating his/her true probability. The Brier score is often used as a measure to grade forecasts (Stevenson, et al. 2008). The Brier score is defined as

$$BrierScore = \frac{1}{N}\sum_{q=1}^{N}\sum_{s=1}^{R}(f_{qs} - o_{qs})^2 \ ,$$

where N is the number of questions, R is the number of possible outcomes for each question, f_{qs} is the probability forecast for outcome s of question q, and o_{qs} is an indicator (1 if yes; 0 if no) for whether the actual outcome for question q was s. Clearly, forecasts that predict the correct outcome with higher probabilities will result in lower Brier scores. The Brier score is a proper scoring rule, meaning that if outcomes are randomly generated according to a "gold standard" probability distribution, the Brier score is optimized by a forecaster who reports this "gold standard" distribution.

Also important in analyzing the participants' predictions in the expected Brier score, also known as the Brier prediction error. The expected Brier score is defined as

$$ExpectedBrierScore = \frac{1}{N}\sum_{q=1}^{N}\sum_{s=1}^{R} f_{qs}^{G}(f_{qs} - o_{qs})^2 \ ,$$

where f_{qs}^{BN} is the "gold standard" probability of a possible outcome. For our experiment, f_{qs}^{G} is the probability obtained from the Bayesian network used to generate the evidence, conditioned on the evidence that the forecaster has seen so far. The expected Brier score is measure of the inherent uncertainty in the problem – the best forecast that could be made given available evidence.

The experiment was conducted in sessions consisting of two paired trials designed so that each participant made predictions using both the combinatorial and flat markets. In each trial, the participants made predictions for five identical questions relating to characteristics of suspects in a "murder mystery." In the first trial of each session, half of the participants used an instance of the DAGGRE combinatorial market (Group A) and the remaining half used an instance of the DAGGRE flat market (Group B). In the second trial, each participant used the type of market he or she had not used on the first trial. In each trial, the instructions, training, market questions, and clue types were identical. Two scenarios – a series of clues and murderer characteristics, representing a specific murder scenario – were selected, one for each of the two trials. Each scenario used the same clues and murderer characteristics, but differed in the assigned values (e.g., female wearing heels vs. male wearing flats) and the order in which the clues were presented. The scenarios were constructed to be similar enough that effects due to scenario would be minimal.

Two sessions of the experiment have been conducted to date. The first session was conducted as part of the DAGGRE spring workshop in California for DAGGRE geopolitical market participants. Workshop attendees were all interested in geopolitical prediction markets, and ranged from novice to highly experienced. The second session was conducted at George Mason University using primarily third-year systems engineering students as participants. All GMU participants had passed a course in probability, and were consid-

ered to have the requisite critical thinking skills to understand quickly the functioning of the DAGGRE prediction markets. We wanted all participants to be familiar with the market, to reduce extraneous variation. All participants were given training such that they felt comfortable using both the flat and the combinatorial markets prior to beginning the experimental trials.

In order to provide the "gold standard" against which the participants' predictions could be compared, Bayesian networks representing the relationships between the market questions and clues were developed. The full Bayes net (Figure 1) was used as the basis for the relationships between market questions (in blue) and clues (in tan). The flat Bayes net (Figure 2) was obtained by removing links between market variables and setting their distributions to the marginal distributions obtained from the full Bayes net. The flat Bayes net was adopted as the "gold standard" for the non-combinatorial condition. Essentially, the combinatorial market and the flat market have identical market questions, clue types, and relationships between questions and clue types, but in the flat market the participants are unable to specify relationships between the market questions. The relationships among the market questions and the clue types were defined for the participants in conditional probability tables.

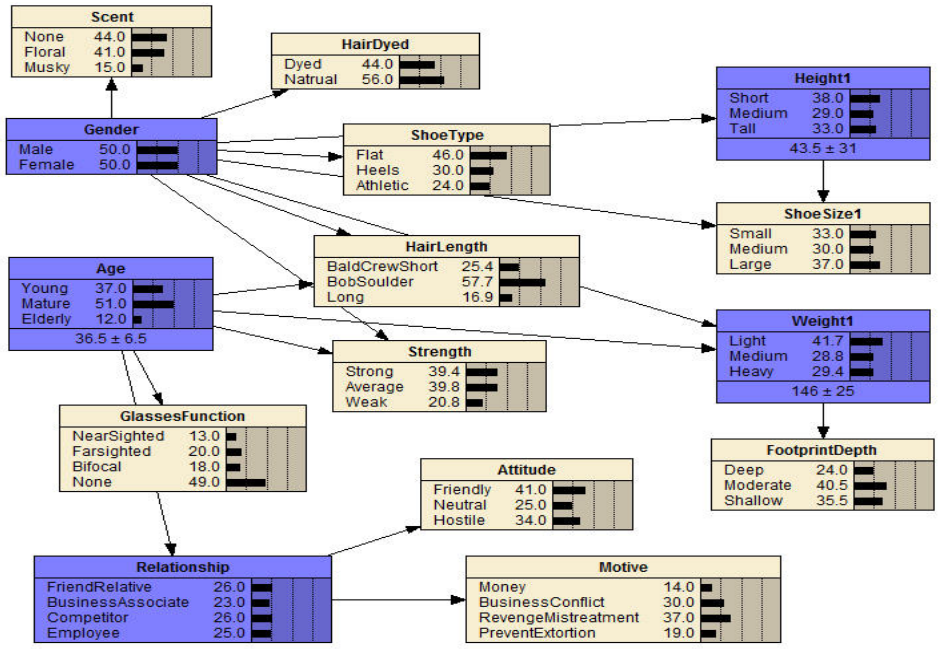

Figure 1: Combinatorial (Full) Bayes net

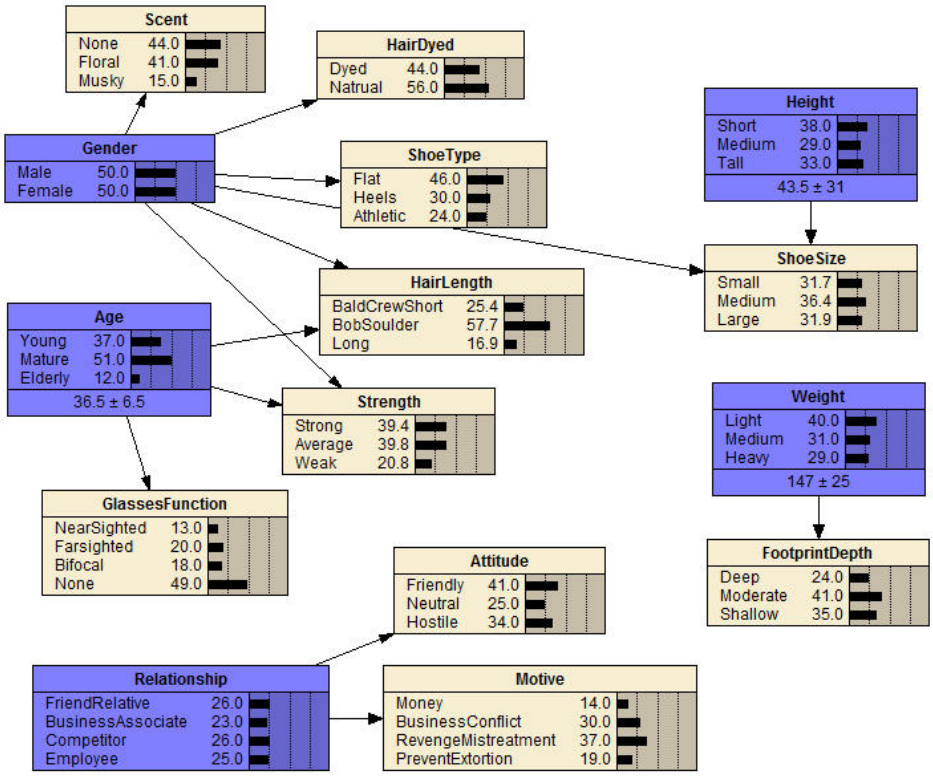

Figure 2: Flat Bayes net

Market questions and clues were chosen to be intuitive, plausibly related to a murder mystery, and related to each other (to provide a valid test of a combinatorial prediction market). The relationships between physical clues were based on statistical evidence in the population of the US. For example, men are on average taller and heavier than women; shoe size is correlated with height. The strength of the relationships was exaggerated over the actual correlations in the U.S. population. This reflects the stereotypical nature of typical murder mysteries, and helped to ensure strong correlations among clues and market questions.

Once the relationships among the questions and clue types were established in the full Bayes net, the Bayes net was sampled to generate 100 simulated individuals who, according to the scenarios, were attendees at the New Year's eve party where the murder occurred. Participants were told that the murderer was one of these 100 suspects. Table 1

contains examples of the characteristics for each clue type and market question associated with each case (attendee).

Because "gold standard" predictions were required for a baseline against which the participants' predictions could be evaluated, after the 100 individuals were simulated, the marginal and conditional probabilities both the full and flat Bayes nets were replaced with actual frequencies taken from the simulated guest list for the party. These frequencies are reflected in Figures 1 and 2, and were used to generate the "gold standard" predictions for the series of clues in each case.

Guest	Scent	Hair Dyed	Hair Length	Shoe Type	Strength	Glasses Function	Attitude	Motive	Footprint Depth	Shoe Size	Gender	Weight	Height	Age	Relation-ship
1	MSK	NAT	BCS	FLT	STR	NON	NUT	EXT	MOD	MED	M	MED	TAL	ELD	ASO
2	FLR	DYD	SHD	ATH	STR	BIF	NUT	EXT	DEP	SML	F	HVY	SHT	MAT	FOR
3	NON	NAT	BCS	FLT	WEK	FRS	FRN	EXT	MOD	LRG	M	MED	TAL	MAT	FOR
4	MSK	NAT	SHD	ATH	AVG	NON	HST	MNY	SHL	MED	M	HVY	TAL	MAT	CMP
5	NON	NAT	SHD	FLT	STR	NON	HST	EXT	MOD	LRG	M	MED	TAL	YNG	FOR

Table 1: Partial Attendee Characteristic Table

Of the 100 cases, two were selected as murderers, one for each of the two experimental trials. The differences between the Brier scores for the flat and full BNs, the differences expected Brier scores for the flat and full BNs, and the relative probability of each occurrence of each cases were used to select ten candidate cases. The candidate cases were those that:

- Had relatively large differences in the Brier scores;
- Had relatively large differences in the expected Brier scores;
- Had a variety of characteristics for clue types and outcomes for market questions; and
- Were above average in their probability of occurrence.

Larger differences in the Brier and Expected Brier scores indicated that there should be differences between the "gold standard" predictions and would provide opportunities for the participants to generate differences predictions in the combinatorial and flat markets. Cases with higher than average probability of occurrence were selected as representative of typical cases.

Once the candidate cases were selected, simulations using permutations of the ordering of clues were generated to evaluate the difference between the full and flat Bayes net "gold standard" predictions over time. The two cases used in the experiment and the ordering of the clues were chosen from those with larger average differences in Brier scores over time. Based on data gathered during a pilot test, the timing of the clues was established such that individuals would have sufficient time to analyze the impact of the clues and enter any updates to the market predictions.

At the beginning of the experiment, each group (combinatorial and flat) was subdivided into two subgroups, each of which was seated at a separate table. Each subgroup received marginal and conditional probability tables for the market variables, as well as conditional probability tables for a subset of the clues. Participants were given time to

enter information from the probability tables into the market. Figure 3 shows a screenshot of the interface used by participants to enter probabilities. The screen shows an assessment of the probability that the murder was a friend or relative, business associate, competitor, or employee of the victim. Participants see the current probability and a chart showing the history of probability values since the start of the experiment. Participants can use the "+" and "-" buttons to raise or lower the probability values. The interface also shows their expected score if each of the outcomes occurs. On the left-hand side of the screen, we see the current question highlighted in blue. The top part of the screen shows assumptions. In the combinatorial condition, participants can drag other questions up into the assumption area and select an assumed value. In this case, the participant is assuming that the murderer is young; thus the probabilities shown to the right are conditional probabilities of relationship of the murder to the victim, given that the murderer was young.

Only those in the combinatorial condition could enter information about relationships among market questions. Dragging questions into the assumptions area was disabled in the flat condition. All participants could enter marginal probabilities. After about ten minutes, clues were handed out at intervals of a few minutes. Near the end of the experiment, as a way to keep up interest in the game, the guest list was handed out and subjects were challenged to identify the murderer. At this point, participants had enough to identify the murderer with certainty.

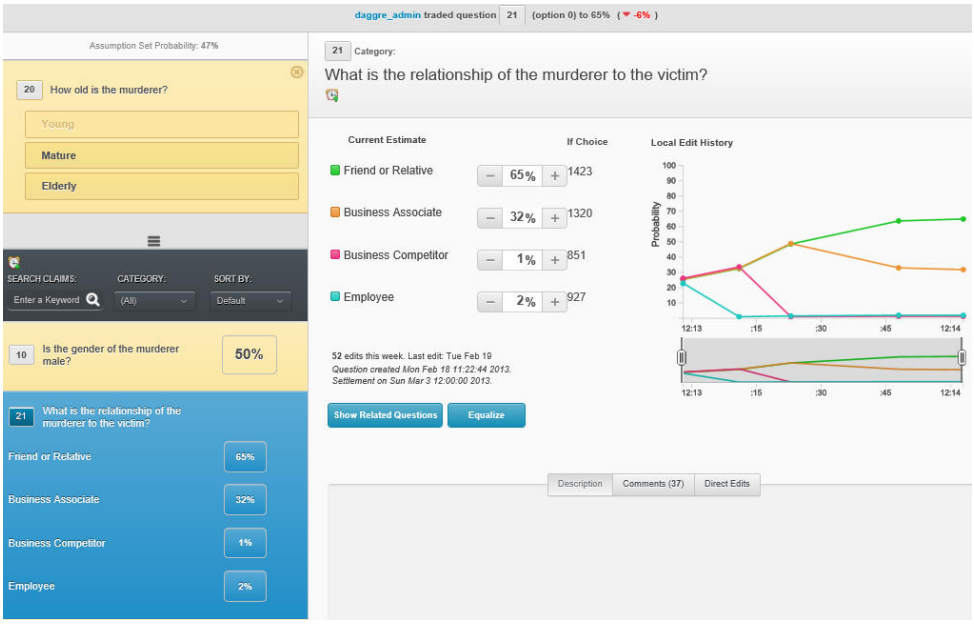

Figure 3: DAGGRE Prediction Market User Interface

4. Results and observations

The first clue was introduced about ten minutes after the probability tables were distributed. During those ten minutes, participants could use the market to establish the initial marginal and conditional probabilities for the market questions. Figure 4 compares the time series of Brier scores for combinatorial and flat prediction markets starting from the time the first clue was distributed. The figure also shows the Brier scores for the full and flat Bayesian networks. Those in the combo condition made more edits, reflecting extra effort to correlate the market questions. Those in the flat market could not express those correlations in the market. Since the full Bayesian network represents all the available information, theoretically, on average it should have the lowest Brier score for a representative case drawn from the Bayes net. The flat Bayes net, since it does not capture the relationships among the market questions, should not as accurately predict the outcomes of the market questions. Indeed, in Figure 4, the Brier scores for predictions from the flat Bayes net are always greater than those from the full Bayes net.

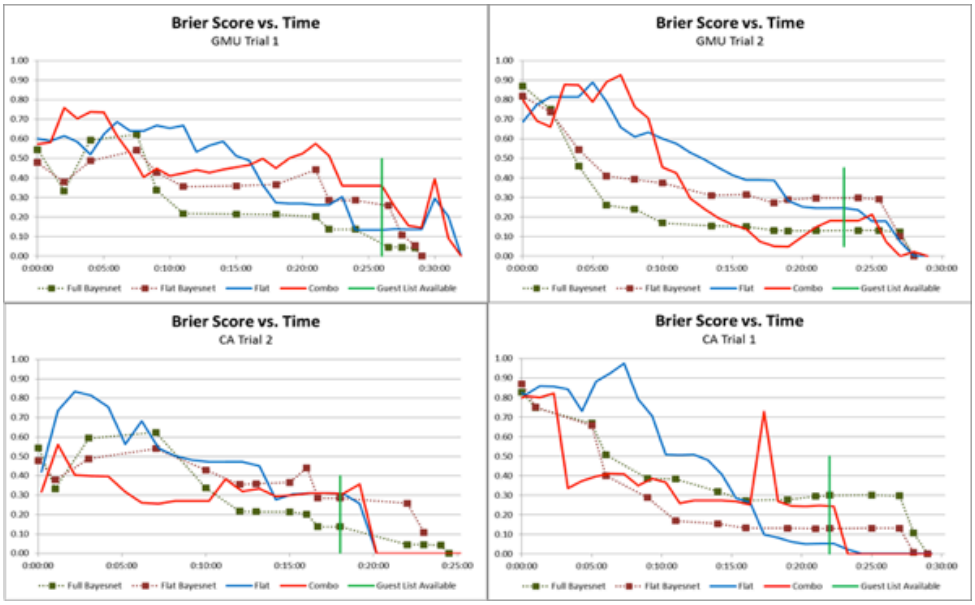

Figure 4: Comparison of Brier Scores

Figure 5 is similar to Figure 4, but compares expected Brier scores. These figures can be divided into regions in which the amount of information available to the participants differed. Prior to the lists of attendee characteristics being distributed (before the green vertical lines in the figures) the participants had available only the information contained in the clues and in the conditional probability tables. This information was reflected in the

Bayes nets used to construct the scenarios and clues. Near the end of the experiment, to keep up interest, a list of party guests and their characteristics was distributed. At this point, the participants had access to information not captured in the Bayes nets, and this information was sufficient to identify the murderer with certainty before the clues resolving the questions were distributed. Therefore, our analysis stops when the guest list was distributed

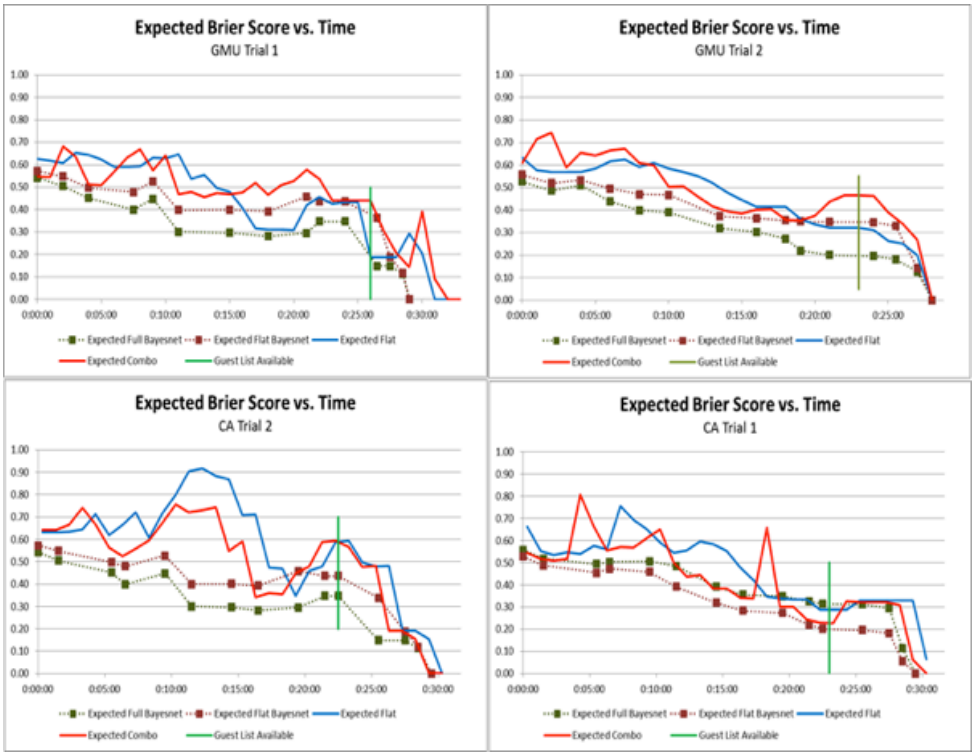

Figure 5: Comparison of Expected Brier Scores

As can be seen in the figures, there was a lot of noise in the markets -- compared to the Bayes nets, and no clear tendency. Indeed, it is hard to tell from inspection which curve had the lower time-averaged Brier score. Table 2 summarizes the the average Brier scores for the flat and combinatorial markets in each trial and compares them to the corresponding difference in the "gold standard" Brier scores obtained from the flat and full Bayes nets. In three of the trials, the average Brier scores from the combinatorial markets were lower than those from the flat markets indicating that, on average, the predictions made in the combinatorial market were more accurate than those made in the flat markets. The exception to this was GMU Trial 1 in which the average flat market predictions were

more accurate. As expected, in three of the four trials, the average difference between the flat and combo participants' scores were less than those from the corresponding Bayes nets, indicating that on average the participants had not integrated all the available knowledge into their predictions. There was an exception to this trend also; in CA trial 2 the difference between the participants' Brier scores was greater than that between the Bayes net scores. Inspecting Figure 4, it appears that the CA trial 2 combo participants were overconfident: their combinatorial Brier scores were below those of the full Bayes net, indicating that their predictions were stronger than they "should" have been with the information available; however, this overconfidence may have been warranted by the knowledge that they were participating in an experiment.

	Bayes Net Average Difference	Participants' Average Difference
CA Trial 1	0.1085	0.0993
CA Trial 2	0.0281	0.1537
GMU Trial 1	0.0622	-0.0542
GMU Trial 2	0.1252	0.0951

Table 2: Average Brier Scores

Overall, the Brier score analysis suggests that the use of combinatorial markets had an effect on the forecasts made by users, but although the results are suggestive, the effects seen in this experiment do not conclusively demonstrate an effect of combinatorial markets on accuracy.

Like the Brier scores, the expected Brier scores (Figure 5 and Table 3) for the combinatorial market are not consistently lower than those for the flat market indicating that that the certainty in the participant's predictions was not consistently less for the combinatorial market than for the flat market. As can be seen in Table 3, at various time in the trials, the expected Brier scores in the combinatorial and flat markets approached the theoretical minimum generated by the full Bayes net expected Brier scores, though neither the combinatorial nor the flat market expected Brier score did so consistently. This lack of consistency is also evident in the average difference between the participants' expected Brier scores. Though the average expected Brier scores for the combinatorial market were less than those for the flat market in the California trials, they were slightly greater in the GMU trials.

5. Conclusion

The four trials in this experiment do not show a clear advantage for combo markets over flat markets on this "murder mystery" scenario. These results set limits on the conditions and range where a clear advantage may be seen. First, given the noise in the market estimates, the scenarios used in these trials provided insufficient theoretical difference (~10%) between the flat and combo Brier scores (as generated by the full and flat Bayes

nets). Although each of our trials involved ~10 people working for several hours, the effective sample size is simply the number of trials, four (4). A scenario with a larger theoretical difference might show a consistent difference between the groups.

	Bayes Net Average Expected Difference	Participants Average Expected Difference
CA Trial 1	0.0569	0.0381
CA Trial 2	0.0764	0.0814
GMU Trial 1	0.0786	-0.0133
GMU Trial 2	0.0718	-0.0232

Table 3: Average Brier Scores

Second, to level the playing field, we provided direct evidence for all the market questions. But the most likely benefit to using a combinatorial market lies in the ability to propagate the effect of evidence through the market and influence predictions for market questions that are not directly related to the evidence. Examination of the data shows that changes in the predictions due to direct evidence seemed to overwhelm the changes in predictions due to evidence that was only indirectly related to each question. A clearer advantage for the combinatorial market might be seen if some questions could only be predicted from evidence relating to other correlated questions.

In designing future trials based on the experimental trials reported here, several modifications may increase the effects on the dependent variables (Brier scores and expected Brier scores). Possible modifications to the experimental design include making the correlations among the questions and between evidence and the question stronger; simulating more specialize knowledge i.e. lower percentage of the participants receive each conditional probability table; making it more difficult for participates to retain knowledge of specific relationships among the markets question and types of evidence; and adding questions for which no direct evidence is provided, but which are correlated with other questions for which there is evidence. Additionally, designing trials that take less time could result in more trials being run with the same number of participants and thus provide an overall increase in the statistical power of the experiment. Also, the experiment could be instrumented to provide data that would support the analysis of other metrics, e.g. joint probability distributions and conditionals. The basic design of these experimental trials seems sound, and improvements to the experimental design have been identified that should increase the ability of the Combinatorial Market experiment to determine the effects of using probabilistically linked questions on prediction markets.

Acknowledgements

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number

D11PC20062. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained 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. The authors are grateful to Shou Matsumoto, Brandon Goldfedder and Jamie Ostheimer for software support.

References

1. Meehl, P. E.: Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. University of Minnesota Press (1954)
2. Dawes, R., Faust, D. and Meehl, P. E.: Clinical Versus Actuarial Judgment. Science 243:4899; 1668–74 (1989)
3. Marchese, M.C.: Clinical Versus Actuarial Prediction: a Review of the Literature. Perceptual and Motor Skills 75:2, 583–94 (1992)
4. Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. and Nelson, C.: Clinical Versus Mechanical Prediction: a Meta-Analysis. Psychological Assessment 12:1, 19–30 (2000)
5. Tetlock, P.: Expert Political Judgment: How Good Is It? How Can We Know? Princeton University Press, (2005)
6. Silver, N.: The Signal and the Noise: Why So Many Predictions Fail — but Some Don't (1st ed.). Penguin Press HC (2012)
7. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
8. Hanson, R. Combinatorial information market design. Information Systems Frontiers 5:1, 107-119 (2003)
9. Hanson, R. Logarithmic market scoring rules for modular combinatorial information aggregation. The Journal of Prediction Markets 1:1 3-15 (2007)
10. Pennock, D. M. & Wellman, M. P.: Graphical models for groups: Belief aggregation and risk sharing. Decision Analysis, 3, 148-164 (2005).
11. Chen, Y, and Pennock, D.M.: Designing markets for prediction. AI Magazine 31:4 42-52. (2010)
12. Sun, W., Hanson, R., Laskey, K. B. and Twardy, C.: Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets. Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA (2012)
13. Berea, A., Maxwell, D. and Twardy, C.: Improving Forecasting Accuracy Using Bayesian Network Decomposition in Prediction Markets. Proceedings of the AAAI Fall Symposium Series (2012)
14. Berea, A., and Twardy, C.: Automated Trading in Prediction Markets. Social Computing, Behavioral-Cultural Modeling and Prediction, pp. 111-122. Springer Berlin Heidelberg, (2013)
15. Brier, G. W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review. 75:1-3 (1950)