Accuracy of Simulated Flat, Combinatorial, and Penalized Prediction Markets

Kenneth C. Olson, Charles R. Twardy, and Kathryn B. Laskey, George Mason Univ.

There is preliminary empirical evidence that a combinatorial prediction market (CPM) can outperform a flat prediction market (FPM), but the conditions under which gains are realized are unclear. In a study with simulated agents, we found that CPMs fare better than FPMs when few agents have specialized knowledge and events are strongly related. By allowing conditional forecasts, a CPM is over 50% more accurate than an FPM in the FPM's worst-case scenario. We also consider a novel prediction market that directly penalizes probabilistic incoherence.

We consider accuracy in crowdsourced forecasts. Recent work has shown long-run accuracy well above a simple average [Mellers et al. 2014]. Two promising nondemocratic approaches addressed in this study are prediction markets and coherence weighting. In a prediction market, forecasters buy and sell contracts on verifiable future events. Market prices for contracts are interpreted as the probability of the event. When the event resolves, forecasters gain (lose) assets if they bought contracts representing the outcome that occurred (did not occur). Therefore accurate forecasters tend to increase their influence over time, while inaccurate forecasters tend to decrease their influence over time. More information can be found in [Arrow et al. 2008; Pennock et al. 2001].

Another nondemocratic approach weights by coherence under the assumption that more coherent forecasts are more accurate. Coherence is available at forecast time, and provides assessment even when estimating objective probabilities is difficult [Lindley et al. 1979]. Recent studies have shown that weighting forecasters by the degree to which their forecasts are probabilistically coherent can also substantially improve upon simple averages, at least in knowledge tasks [Karvetski et al. 2013; Olson and Karvetski 2013; Tsai and Kirlik 2012; Wang et al. 2011].

De Finetti showed that under proper scoring rules, any incoherent set of forecasts can be replaced by a coherent set that has a better score for every possible outcome [de Finetti 1937; 1981]. Therefore, if events in a market are related, we might expect that markets which disallow or penalize incoherence will outperform those which allow incoherence.

1. TYPES OF PREDICTION MARKETS

CPMs can express relationships among the events in the market; FPMs cannot. But it has been hard empirically to demonstrate the advantage of CPMs. Ledyard, Hanson, and Ishikida [2009] compared CPMs to FPMs on a small laboratory task involving either few variables or many interacting variables. Each of six participants only had information on a subset of variables. In one condition, information was relatively evenly distributed to participants, and in another condition, information was relatively unevenly distributed to participants. The one non-combinatorial mechanism, a simple double-auction prediction market, performed worst. The five combinatorial mechanisms had mixed results. Opinion pools only did well when information was evenly distributed to participants, and the market scoring rule for a combinatorial prediction market performed best in the complex environment. A follow-up study [Healy et al. 2011] yielded similar results, but with only three traders. A recent empirical comparison of CPM and FPM performance on a dynamic "whodunnit" was inconclusive [Powell et al. 2013].

These studies suggest that a CPM can provide more accurate aggregate forecasts than an FPM when there are many interrelated variables, few forecasters with expert knowledge, and strong relationships

Collective Intelligence 2014.

Table I. : Probabilities and conditional probabilities in strong- and weak-relations conditions.

Relations	True Event Probability												
	A	A|B	A|B^C	B	B|C	B|C^C	C	C|D	C|D^C	D	D|E	D|E^C	E
Strong	0.69	0.90	0.10	0.73	0.25	0.95	0.31	0.10	0.70	0.65	0.70	0.20	0.90
Weak	0.69	0.75	0.55	0.72	0.40	0.90	0.35	0.20	0.60	0.62	0.65	0.35	0.90

between variables. However, they only suggest. Neither do they isolate the reasons for any advantage. We use simulations to isolate reasons, and investigate the role of incoherence.

A forecaster may know that events A and B are exclusive yet state $p(A \cap B) = 0.0$, $p(A) = 0.2$, $p(B) = 0.5$, and $p(A \cup B) = 0.6$. His inaccuracy comes about by being incoherent. There is reason to expect such incoherence is common [Olson and Karvetski 2013; Mandel 2005], but a FPM on A and B can neither detect nor correct it. Karvetski et al. [2013] found ways to measure, discourage, and capitalize on probabilistic incoherence. Their best coherence weighting method obtained an almost perfectly coherent set of probability estimates in which more coherent judgments contributed more weight. The coherence weighting performed much better than coherentization (forcing coherence). It also operated better when incoherence was not discouraged, and thus happened more.

We extend their methods to compare a flat market (incoherence not discouraged), a fully combinatorial market (incoherence disallowed), and a coherence-penalized combinatorial market (incoherence penalized).

2. OBJECTIVES

We hypothesize that measured accuracy will be greater for an incoherence-penalized combinatorial prediction market (PPM) than for a fully combinatorial prediction market (CPM) and greater for a CPM than for an FPM (PPM > CPM > FPM). However, we expect market differences in accuracy to be greater when knowledge on the questions in the markets is unevenly distributed and when events are related, so that learning whether an event occurs provides information about the likelihood of another event.

3. EXPERIMENT DESIGN

In the FPM, by definition, the five main questions are not linked, in the CPM they are linked, and in the PPM they are linked and forecasters responses can incur a penalty to their assets. There are five binary questions common to all markets, and each simulated agent responds to all five questions. The questions are linked in a chain, $A \leftarrow B \leftarrow C \leftarrow D \leftarrow E$, creating two additional conditional probabilities to be judged in the combinatorial and penalized markets for each link in the chain. Four links make eight conditional probabilities for a total of 13 responses from each simulated agent in the CPM and PPM. The 13 true probabilities can be seen in Table I.

There are two factors, each with two levels, crossed to create four conditions. The two factors are strength of relations and spread of knowledge. In Table I, the conditional probabilities differ between the strong relations conditions and the weak relations conditions, but the marginal probabilities are nearly the same across conditions of the experiment.

Ten simulated agents forecast on the events and have beliefs about the probabilities. When knowledge about the events is widely distributed, four agents are very knowledgeable about each probability except $p(A)$. When knowledge about the events is narrowly distributed, three agents are very

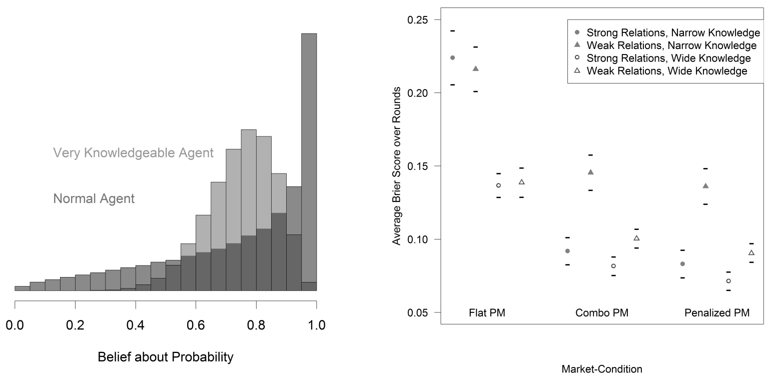

Fig. 1: (a) Example distributions for sampling beliefs of two different agents with mean 0.73.
(b) Mean Brier scores with standard errors for 3 markets in 4 conditions.

knowledgeable about each conditional probability, but only one agent is very knowledgeable about each marginal probability, except $p(A)$. Ignorant agents tend to defer to the market.

For every question, all agents have a belief distribution centered on the true probability. To be very knowledgeable about a probability means that the agent has a narrow distribution from which its belief is sampled, while to be less knowledgeable means that the agent has a wide distribution from which its belief is sampled. The random sampling of beliefs from the belief distributions ensures that no set of beliefs is perfectly coherent. A coherent set of forecasts is established for each agent by finding the set of probabilities closest to the agents set of beliefs while satisfying all probability constraints.

4. SIMULATIONS

For each agent and each market type, a user account is created on an instance of the SciCast LMSR combo prediction [Sun et al. 2012; Twardy et al. 2014]. Questions are created, and for CPM and PPM, links between them. Agents see the most recent market estimates and respond with their own updated beliefs. Trades are created from SciCast's "safe mode" which invests a maximum of 1% of the agent's assets to trade the estimate at most halfway towards their belief.

In the PPM, a penalty is imposed for deviation of an agent's response from the coherent probability. The penalty is equivalent to the cost of moving the market estimate from the coherent probability to the belief probability. This penalty is invoked before the trade in the market, reducing the assets available for the trade, and therefore reducing the influence of the agent on the market.

5. DATA ANALYSIS

Figure 1 shows the statistics obtained from 50 simulations yielding 12 distributions of 50 market-wide Brier scores. The FPM performed worse than either the CPM or PPM on matched conditions; all markets performed worse when agents had narrow knowledge. The PPM performed slightly but not significantly better than the CPM across conditions ($t = 1.2, df = 198, p = 0.12$). The average Brier scores were 0.18 for the FPM, 0.09 for the CPM, and 0.08 for the PPM. Strong dependencies between events benefit the CPM significantly more than they do the FPM ($t = 2.2, df = 167, p = 0.01$). The greatest difference between the CPM and the FPM was when events were strongly related but few agents had knowledge of the events or their relations. Under these circumstances, the CPM improved the Brier score on marginal probabilities from 0.22 to 0.09, almost 59% better than the FPM.

ACKNOWLEDGMENTS

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Dept. 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.

REFERENCES

K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz. 2008. The promise of prediction markets. *Science* 320 (2008), 877–878.

R. T. Clemen. 2008. Comment on Cooke's classical method. *Reliability Engineering & System Safety* 93 (2008), 760–765.

R. M. Cooke and L. L. H. J. Goossens. 2008. TU Delft expert judgment data base. *Reliability Engineering & System Safety* 93(5) (2008), 657–674.

B. de Finetti. 1937. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7 (1937), 1–68. English translation in Studies in Subjective Probability, H. Kyburg and H. Smokler, eds., Krieger Publishing, Huntington, 1980.

B. de Finetti. 1981. The role of Dutch books and proper scoring rules. *British Journal for the Philosophy of Science* 32 (1981), 55–56.

P. J. Healy, S. Linardi, R. Lowery, and J. O. Ledyard. 2011. Prediction markets: Alternative mechanisms for complex environments with few traders. *Management Science* 56(11) (2011), 1977–1996.

C. W. Karvetski, K. C. Olson, D. R. Mandel, and C. R. Twardy. 2013. Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis* 10(4) (2013), 305–326.

J. Ledyard, R. Hanson, and T. Ishikida. 2009. An experimental test of combinatorial information markets. *Journal of Economic Behavior and Organization* 69(2) (2009), 182–189.

D. V. Lindley, A. Tversky, and R. V. Brown. 1979. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society* 142(2) (1979), 146–180.

D. R. Mandel. 2005. Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied* 11(4) (2005), 277–288.

B. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. Swift, T. Murray, E. Stone, and P. Tetlock. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25(5) (2014), 1106–1115.

K. C. Olson and C. W. Karvetski. 2013. Improving expert judgment by coherence weighting. *Proceedings of IEEE International Conference on Intelligence and Security Informatics* (2013), 197–199. Piscataway, NJ: IEEE, Inc.

D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. 2001. The real power of artificial markets. *Science* 291(5506) (2001), 987–988.

W. Powell, R. Hanson, K. Laskey, and C. Twardy. 2013. Combinatorial prediction markets: An experimental study. *Proceedings of the Seventh International Conference on Scalable Uncertainty Management* (2013), Washington, DC: Scalable Uncertainty Management.

Wei Sun, Robin Hanson, Kathryn Blackmond Laskey, and Charles Twardy. 2012. Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*. AUAI Press, Catalina, (CA). http://mason.gmu.edu/~wsun/publications/uai2012.htm

J. Tsai and A. Kirlik. 2012. Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting* (2012), 313–317. Thousand Oaks, CA: Sage Publications.

Charles Twardy, Robin Hanson, Kathryn Laskey, Tod Levitt, Brandon Goldfedder, Adam Siegel, Bruce D'Ambrosio, and Daniel Maxwell. 2014. SciCast: Collective Forecasting of Innovation. *Collective Intelligence* (2014).

G. Wang, S. R. Kulkarni, H. V. Poor, and D. N.. Osherson. 2011. Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis* 8(2) (2011), 128–144.