The Marketcast Method for Aggregating Prediction Market Forecasts

Pavel Atanasov*, Phillip Rescober*, Emile Servan-Schreiber**, Barbara Mellers*, Philip Tetlock*, Lyle Ungar*

* University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104
** Lumenogic, 48 rue du Cherche Midi, Paris 75006, France

Abstract. We describe a hybrid forecasting method called marketcast. Marketcasts are based on bid and ask orders from prediction markets, but the information is aggregated using techniques associated with survey methods. We discuss a variety of ways in which bids and asks could be used as input to survey aggregation methods. The performance of marketcasts is compared to a traditional prediction market and a traditional opinion poll. Overall, marketcasts perform approximately as well as prediction markets and opinion poll methods on most questions, and performance is stable across modeling specifications.

Keywords: Forecasting, Prediction Markets, Aggregation

1 Introduction

Prediction markets, also known as ideas futures, have been shown to produce accurate forecasts for political and sports events [1]. Most research demonstrating accuracy and stability of this method has focused on large markets with thousands of participants. Less attention has been paid to small markets, even though many practical applications feature limited numbers of forecasters. Small prediction markets, with 15 to 50 participants, tend produce well calibrated forecasts, but those with less than 15 active participants may be subject to manipulation [2].

Prediction markets serve two separable functions: they elicit individual opinions and aggregate those opinions. We show that separating these functions is possible and practical. Forecasts elicited through prediction markets can be aggregated using non-market mechanisms, producing what we call marketcasts. Marketcasts perform well even in their simplest forms. They can exploit information beyond the current price, for example using bids when no trades occur. As we demonstrate, marketcast can also be easily incorporated into statistical algorithms including unequal weighting of forecasters, temporal smoothing and transformation, which have been shown to improve accuracy of aggregate survey forecasts [3].

* Please address any correspondence to apav@sas.upenn.edu.

2 Prediction Markets vs. Survey Forecasts

Prediction markets offer one method for eliciting and aggregating crowd beliefs about uncertain events. An alternative method is to simply ask forecasters about their subjective probability of uncertain events, and average these values. Probably the most popular example of successful opinion pooling, albeit not of probability forecasts, was described by Galton [4], who showed that the median crowd estimate of an ox's weight was within 9 pounds (0.8%) of the correct answer. This method is known as an *opinion poll* and the resulting values are referred to as survey forecasts.

2.1 Elicitation

The prediction market interface, in its various forms, has several useful features for eliciting probability forecasts. First, markets offer incentives, financial or otherwise, that encourage forecasters to learn new information about specific questions and communicate it by placing orders on the markets. Second, order size is a measure of how confident participants are in their beliefs, as measured by the size of the bet they place. Third, the prediction market interface provides feedback about other participants' beliefs. Fourth, participation in prediction markets is a form of gambling and may lead to self-selection of participants who enjoy making such bets. Participants also face market selection, as consistent low-performers lose money and, unless they continue injecting funds, may lose liquidity and influence over market prices. In contrast, successful traders may gain influence if they choose to reinvest their winnings in future bets.

Opinion polls may share some of the useful elicitation features of prediction markets. They may provide feedback about crowd beliefs (e.g. the mean of outstanding forecasts), and provide performance feedback using metrics such as Brier scores. Expertise self-ratings could help distinguish between more and less knowledgeable forecasters [3]. Forecasts from prior low-performers could be removed or downweighted in the analysis stage.

2.2 Aggregation

While elicitation methods influence who expresses beliefs and how these beliefs are expressed, aggregation methods deal with the problem of merging crowd beliefs into a single forecast. Prediction markets usually solve this problem by matching bid and ask orders to produce a market price. In the continuous double auction (CDA) used in this study, buyers place bid and ask orders, specifying desired price and volume.[1] A trade occurs if a bid price is higher than or equal to ask price, that is when bid-ask spread is negative. Typically, the forecast is the last price at a certain time, although markets also provide related metrics

[1] Other market-based mechanisms for scoring and aggregation of forecasts include parimutuel betting, dynamic parimutuel [5] and Robin Hanson's Market Scoring Rule [6].

such as typical price, average price over the course of a day. In markets with few active participants, bid-ask spreads are often large and no trading occurs for long periods. Similarly, when the markets are thin and trades do occur, they can cause fluctuations in market price.

Market pricing is not the only way to aggregate beliefs among forecasts. An alternative method we propose and test in this study is to treat order prices as survey forecasts. Such marketcasts (as we call them) can potentially overcome many of the limitations of thin prediction markets. Table 1 shows possibilities of eliciting and aggregating information from prediction markets and opinion polls.

Elicitation	Aggregation	
	PM	Survey
PM	Core Prediction Market	Marketcast
Survey	Trading Agent	Survey Forecasts

Table 1. Possible combinations between elicitation and aggregation methods.

Consider a binary prediction market in which a share pays $1 if an event occurs and $0 if it does not. If a participant submits a bid order at $0.60, a simple marketcast algorithm would impute the probability forecast to 60%. If two other traders submit orders at $0.70 and $0.74, respectively, the unweighted mean probability from these forecasts is would be 68%, and the median, 70%.

3 Aggregation Parameters

Because aggregation of survey forecasts is performed after the fact, researchers face some important choices during data analysis. We discuss the influence of six aggregation parameters below.

Order Size. In its simplest form, marketcasts ignore much information about the orders and interactions among forecasters. For example, each order is weighted equally, independently of its size. Such a simplification would be ideal only if large orders are just as informative as small orders. If number of shares ordered does provide useful information, larger orders should be given more weight in the aggregation phase. In a sensitivity analysis, we weight each order by the square root of number of shares ordered. This weighting scheme is consistent with the intuition that large orders have more information value than small ones, but the value does not increase linearly with order size. A buy order of 100 shares at a given price, for example, has ten times the weight of a one-share order at the same price.

Bid vs. Ask Orders. The naïve marketcast method is insensitive to the distinction between bid and ask orders: all orders are taken at face value. It is possible that market participants act with a profit margin in mind. For example, a trader with a desired profit margin of 10% would submit a bid order at $0.60 if she believes that the probability of event occurring is 70%, and pre-sell shares at $0.60 if she believes the event is 50% likely to occur. Sensitivity analyses with profit margins of 0%, 10% and 25% were performed to determine which of

those most closely approximates the ground truth. In the profit margin analyses, imputed probability values were forced to the [0,1] range.

Order Matching. The naïve marketcast method ignores the distinction between matched and unfilled orders. In other words, each order is treated as a signal of belief, even if it is far from the concensus and is never matched. In practice, forecasters are discouraged from placing such orders because they limit the funds they have available for trading on other questions. A sensitivity analysis scores focuses on the sub-sample of matched orders. A lower Brier score for this sub-sample would imply that unmatched orders provide more noise than signal on the aggregate.

Temporal Smoothing. Prediction markets and survey forecasts deal with "stale information" in different ways. In prediction markets, orders are retained on the order book until they are canceled or executed. Unmatched orders do not affect the most recent price directly, but may do so indirectly, by influencing trader behavior. On the other hand, orders at close to consensus prices are quickly matched by new or existing orders, and are unlikely to stay on the order book very long. Survey forecasts lack this feature, so temporal smoothing is often used to limit the influence of old forecasts, without ignoring them altogether. Exponential decay is a popular approach, in which forecasts are multiplied by a constant between zero and one for each day since they were refreshed. For example, if the exponential decay constant is set to 0.5, today's forecats receive a weight of 1, yesterday's forecasts are given a weight of 0.5, two-day forecasts receive a weight of 0.25 and so forth. An alternative method is to retain only the most recent forecasts, while tossing out olders ones. One of our methods retains the 15% most recent forecasts.

Central Tendency. We report the marketcast mean as our core measure of central tendency. Median, the measure advocated by Francis Galton, is influenced less by outliers and may perform better than the mean if forecasts far from the consensus are misinformed. Finally, we use the geometric mean of probability forecasts in the log-odds space. As Satopaa et al. [7] document, this measure has desirable statistical properties: it is well-behaved and could be used in tandem with unequal weighting of inidviduals and forecasts.

Transformation. In its current use, transformation, also known as signal amplification, addresses the problem of miscalibration. For example, political prediction markets have been shown to exhibit long-shot bias: low probability events are overvalued, while high probability events are undervalued [8]. In practical terms, this means that predictions are less extreme than they should be. Extremizing forecasts improves accuracy in U.S. Presidential election prediction markets [9]. Extremizing aggregated survey forecasts has also been shown to improve survey forecasts [10]. In the sensitivity analyses below, forecasts were extremized in the manner described by Baron et al., with a constant set to 2. For example, a 40% forecast was transformed to 31%, while a 70% forecast was transformed to 84%.

4 Methods & Data

The study is conducted as part of a large ongoing forecasting tournament sponsored by the Intelligence Advanced Research Projects Activity (IARPA). Five teams, including ours, participate in the tournament. A main goal of the tournament is to develop innovative methods of assigning accurate probability estimates for events of national security interest. Each month, eight to ten new questions are added to the tournament, for a total of approximately 120 per year. While teams are asked to suggest the forecasting questions, an external party makes the final decision for inclusion in the tournament.

The current draft focuses on seventeen binary (yes/no) questions that have been resolved since the beginning of the 2012-2013 tournament year. Approximately sixty questions are expected to be resolved by the end of March 2013, which may alter the current pattern of results.

For each question (e.g., "Will Victor Ponta resign or vacate the office of Prime Minister of Romania before 1 November 2012?"), prediction market participants compete in a Continuous Double Auction market. Shares prices resolve to $0 if the event did not occur and $1 if the event occurred in the defined timeline. Forecasters are free to choose which questions to bet on, but are asked to submit at least one order on at least 30 questions over the course of the year, out of approximately 120 questions over the course of the year. Two markets are run in parallel for all questions. Forecasters are randomly assigned to one of two parallel prediction markets. In the first one, they receive basic training on prediction markets. In the second condition, participants receive training about basic principles in probability reasoning. Mellers et al. (2012) show that such training improves performance.

The Brier scoring rule is used to assess forecast accuracy [11]. According to this strictly proper rule, the penalty is the squared difference between the forecast value and the outcome (0 and 1). The best score is 0, the worst score is 2, and with binary questions, a probability forecast of 50% always results in a Brier score of 0.5. Daily Brier scores are averaged over the period for which a question is open. Each question is equally weighted in the determination of the aggregate score.

We report Brier scores for four conditions. First, unweighted linear opinion poll (ULinOp) is used as a baseline condition in the tournament. The method takes a simple mean of the latest survey forecast for each participant for each question. Participants in this condition undergo no special training, receive no crowd feedback, and no temporal smoothing, weighting or transformation are applied to individual or aggregate forecasts. Second, the individual forecast conditions features forecasts from all survey participants. Only the most recent 15% of forecasts are used in this condition. Third, the prediction market condition features both the PM interface and the CDA order-matching algorithm, and features play-dollar incentives. Financial incentives have been shown tohave minimal impact on accurcacy [12]. Finally, the marketcast uses values elicited through the prediction market but pooled using survey forecast aggregation methods.

5 Results

In total, 471 participants submitted at least one order for at least one of the 17 binary questions they faced. On average, 127 individuals submitted at least one order on any given question, resulting in 274 orders over the course of a typical question. Questions were open for 53 days on average, with 10 unique orders were submitted per day per question. The first day after a question opening attracted most activity, and the number of orders usually stabilized after the first three to five days of trading.

Table 2 shows the mean Brier scores for the four conditions of interest: ULinOp, individual surveys, core prediction markets and various marketcast specifications. For ease of presentation, we start with a core marketcast condition, and show the impact of varying settings, one change at a time. In the sensitivity analysis portion of the table, the core specification is repeated in the left-most column of every row. Standard deviations are shown in parentheses. We do not report p-values from inferential tests at this point, due to the high number of comparisons and small number of data points. Inferential tests will be conducted until the sample of questions reached thirty to forty questions. Most differences did not reach statistical significance.

	Mean Brier Score			
ULinOp (Control)	0.369 (0.258)			
Individual Survey Forecasts	0.307 (0.277)			
Core Prediction Market	0.293 (0.360)			
Core Marketcast: Equal Weights, 10% Margin, All Orders, Most recent 15%, Mean, Non-transformed.	0.337 (0.384)			
	Marketcast Sensitivity Analyses			
1. Order Volume Weight	**Equal Weights**	**SqRt Weights**		
	0.337 (0.384)	0.343 (0.385)		
2. Profit margin (m)	**m=10%**	**m=0%**	**m=25%**	
	0.337 (0.384)	0.347 (0.392)	0.330 (0.346)	
3. Order matching	**All Orders**	**Matched Orders**		
	0.337 (0.384)	0.332 (0.388)		
4. Temporal smoothing (c)	**Most Recent 15%**	**c=0.10**	**c=0.50**	**c=0.85**
	0.337 (0.384)	0.348 (0.380)	0.342 (0.379)	0.354 (0.371)
5. Measure of central tendency	**Mean**	**Median**	**Logit**	
	0.337 (0.384)	0.339 (0.411)	0.347 (0.431)	
6. Transformation	**Non-Transformed**	**Transformed**		
	0.337 (0.384)	0.395 (0.370)		

Table 2. Mean Brier scores for 17 questions in the tournament.

Overall, marketcast performance varied in a limited range. On the one hand, almost all Marketcast methods yielded lower Brier scores than the ULinOp con-

trol condition. On the other hand, marketcasts produced higher Brier scores than either pure survey forecasts or prediction markets.

Sensitivity analyses revealed several notable patterns. First, ignoring order size yielded slightly lower Brier scores than weighting orders by the square root of order size, which implies that order size did not provide useful information. Second, larger profit margins improved forecast accuracy, a result consistent with the intuition that bid and ask orders of the same price reflect different beliefs. Third, marketcasts based on matched orders did slightly better than those using all orders, which suggests unmatched orders did not provide useful information.

Fourth, a moderate level of exponential smoothing (c=0.50) yielded best results, but temporal smoothing parameters had a small impact overall. Fifth, taking the mean marketcasts yielded lower Brier scores than the median or the logit transformed geometric mean. Finally, non-transformed marketcasts performed better than extremized ones, which suggested that marketcasts did not exhibit the long-shot bias in this sample.

Figure 1 depicts performance of various methods by question, in increasing order of Brier scores for the Core prediction market condition. In other words, the questions on the left side (A, B, C) were correctly predicted by the prediction market, while those on the right side resolved in unexpected ways (O, P, Q). A brier score above 0.5 means that forecasts were, on average, on the wrong side of p=50%.

Marketcasts perform approximately as well as core prediction markets on all but one question. The most notable disparity occurred for question K, which was open for only three days and resolved suddenly and unexpectedly. The use of temporal smoothing for marketcasts may have limited the method's potential to adjust to new information in this unusually short span of time.

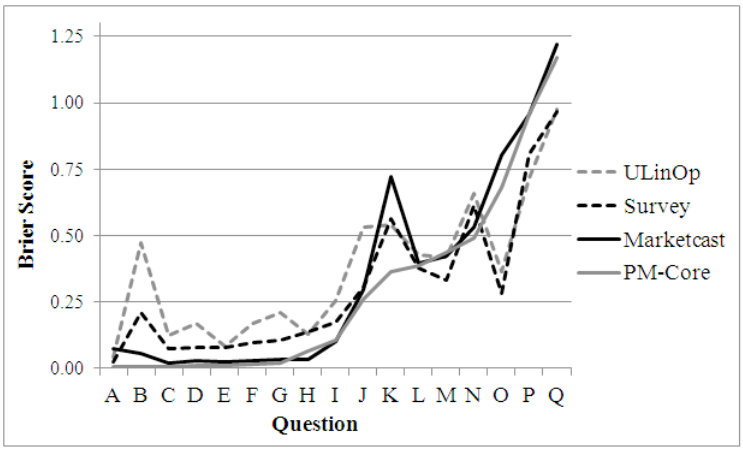

Figure 1. Brier scores per IFP for ULinOp, Survey Condition, Core prediction market and marketcast.

6 Conclusion

Marketcast analyses show that forecasts elicited by prediction markets perform well when used as inputs to non-PM aggregation algorithms. In other words, elicitation and aggregation elements are separable in principle and in practice. Although marketcasts slightly underperform traditional prediction market and pure survey forecasts, they produce stable and accurate forecasts under most conditions and specifications. Future research will examine the stability of the current results and illustrate novel applications of this promising method.

References

1. Wolfers, J., Zitzewitz, E.: Prediction Markets. J. Econ. Perspect. 18, 107-126 (2004)
2. Christiansen, J.D.: Prediction Markets: Practical Experiments in Small Markets and Behaviours Observed. J. of Prediction Markets. 1, 17-41 (2007)
3. Mellers, B.A., Ungar, L.H., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S., Moore, D., Atanasov, P., Swift, S., Tetlock, P.E.: Improving Geopolitical Forecasting with Teamwork, Training and Algorithms. Manuscript under review.
4. Galton, F. Vox Populi. Nature. 75:450-1 (1907).
5. Pennock, D.M.: A Dynamic Pari-Mutuel Market for Hedging, Wagering, and Information Aggregation. In: EC '04 Proceedings of the 5th ACM conference on Electronic commerce, 170-179. ACM New York, NY (2004)
6. Hanson, R.: Combinatorial Information Market Design. Inform. Syst. Front. 5, 107-119 (2003)
7. Satopaa, V.A., Baron, J., Foster, D.P., Mellers, B.A., Tetlock, P.E., Ungar, L.H.: Combining Multiple Probability Predictions Using a Simple Logit Model. Manuscript under Review.
8. Page, L., Clemen, R.T.: Do Prediction Markets Produce Well Calibrated Probability Forecasts? Econ. J. (2012)
9. Rothschild, D.: Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases. Public Opin. Q. 73, 895–916 (2009)
10. Baron, J., Ungar, L.H., Mellers, B.A., Tetlock, P.E.: Two Reasons To Make Aggregated Probability Forecasts More Extreme. Manuscript under review.
11. Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly weather review. 78, 1-3 (1950).
12. Servan-Schreiber, E., Wolfers, J., Pennock, D., Galebach, B.: Prediction Markets: Does Money Matter? Electronic Markets. 243-251 (2004)