The Marketcast Method for Aggregating Prediction Market Forecasts

Pavel Atanasov*, Phillip Rescober*, Eric Stone*, 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, aggregated using techniques associated with survey methods, rather than market matching algorithms. We discuss the process of conversion from market orders to probability estimates, and simple 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 model 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]. Prediction markets serve two separable functions: elicitation and aggregation of individual judgments. We show that separating these functions is possible and practical. Namely, 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, marketcasts can also be incorporated into more sophisticated statistical algorithms including unequal weighting of forecasters, temporal smoothing and transformation, which have been shown to improve accuracy of forecasts elicited through opinion polls [2].

* This research was supported by a research contract to the University of Pennsylvania and the University of California-Berkeley from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior/ National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding 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. Please address any correspondence to apav@sas.upenn.edu.

If the marketcast method demonstrates adequate performance, it could serve at least three valuable functions. First, the method could produce forecasts that are robust to market manipulation, a property that is especially beneficial in small, illiquid prediction markets where manipulation by a single individual could have long-lasting effects [3]. More generally, marketcasts could take advantage of all available data in the market rather than the latest matched orders. Examples of unused data include unmatched orders, heterogeneity of forecasting skills, expertise and risk preferences. To the extent such patterns persist over time, statistical methods could take advantage of them to produce better calibrated forecasts. Second, the marketcast method provides a bridge between elicitation platforms and could be applied in organizations that use multiple platforms and need to aggregate individual-level forecasts across these platforms. Third, marketcasts provide a measure of individual forecasting performance, in addition to market profits, and allow analysts to distinguish forecasters who gain advantage by placing accurate forecasts from those who simply exploit temporary market inefficiencies.

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 opinion polling 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 offer 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 the more and less knowledgeable forecasters [2]. Forecasts from prior low-performers could be removed or down-weighted 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. Other market-based mechanisms for scoring and aggregation of forecasts include pari-mutuel betting, dynamic pari-mutuel [5] and Robin Hanson's Market Scoring Rule [6]. A trade occurs if the bid price is higher than or equal to the ask price. Typically, the forecast is the last price at a certain time, although markets also provide related metrics such as typical (modal) and average (mean) price over the course of a day. In markets with few active participants, bid-ask spreads are often large and no trades occur for long periods. When trades do occur in illiquid markets, they can cause large market price fluctuations.

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 the possibilities of eliciting and aggregating information from prediction markets and opinion polls.

In opinion polls, forecasters are asked the question: "What is the probability that event X will occur by date Y?" In addition to stating their probabilistic beliefs, forecasters often state their perceived expertise in the question. Forecasts could be updated until the day the question is resolved. Individual forecasts could then be aggregated using techniques of varying sophistication. The simplest method takes the mean or median of the most recent forecast by each participant.

Such forecasts could be imported to prediction markets using trading agent algorithms that translate probabilistic beliefs into prediction market orders. A widely used class of algorithms for this purpose is known as Zero-Intelligence-Plus (ZIP), and could be applied to Continuous Double Auction markets [7]. The method results in stable prices that approach "true" values in prediction markets [8].

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 of 60%. If two other traders submit orders at $0.70 and $0.74, the unweighted mean probability from these forecasts would be 68%, and the median, 70%. In contrast, prediction market orders are matched in the market and statistical processing is not necessary for aggregation.

3 Aggregation Parameters

Because aggregation of survey forecasts is performed after the fact, researchers face some important choices in the process of converting individual orders to probability estimates and aggregating these into a single forecast. We discuss the influence of seven aggregation parameters below.

Order Size. In its simplest form, marketcasts ignore considerable information about the orders and interactions among forecasters. For example, each order is weighted equally, independently of its size. Such simplification would be optimal only if large orders are just as informative as small orders. If the order quantity 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 the 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, was given ten times the weight of a one-share order at the same price.

An alternative interpretation of order quantity is that it represents the forecaster's view that buying or selling shares at the order price would bring a large profit margin. This intuition is shared by Wolfers and Zitzewitz [9] who model the desired number of shares in a given market as a function of difference between the forecaster's personal probability estimate and the market price at this time.

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 $0.10 (10%) would submit a bid order at $0.60 if she believes that the probability of an 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 these most closely approximates the link between participant beliefs and outcomes. After adding or subtracting the assumed profit margins, the imputed probability values were forced to the [0.01, 0.99] probability 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 consensus and is never matched. In practice, forecasters are discouraged from placing such orders because they limit

the funds available for trading on other questions. A sensitivity analysis focuses on the sub-sample of matched orders, ignoring all orders that remain unmatched at the time of aggregation. A lower Brier score for this sub-sample would imply that unmatched orders provide more noise than signal in 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 prices close to consensus 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 1 for each day since they were refreshed. For example, if the exponential decay constant is set to 0.5, today's forecasts receive a weight of 1, yesterday's forecasts are given a weight of 0.5, two-days-old forecasts receive a weight of 0.25, and so forth. An alternative method is to retain only the most recent forecasts, while tossing out older ones. Our core method retains 15% of the 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. [10] document, the logit aggregator is the maximum likelihood estimator of theoretically true probability and is theoretically and computationally simple to implement.

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 [11]. In practical terms, this means that aggregate predictions are less extreme than they should be. Extremizing forecasts improves accuracy in U.S. Presidential Elections prediction markets [12]. Extremizing aggregated survey forecasts has also been shown to improve survey forecasts [13]. In the sensitivity analyses below, forecasts were extremized in the manner described by Baron et al., with the constant set to 2. For example, a 40% forecast was transformed to 31%, while a 70% forecast was transformed to 84%.

Expertise Weights. When placing market orders, participants in the prediction market are asked to provide their self-assessments expertise in the domain of the question they bid on. More specifically, they are asked to provide an estimate of their relative expertise compared to other forecasters on a five-point scale. If participants hold accurate beliefs about their relative competence, this information could be used to improve aggregate performance by placing higher weight on more competent forecasters and lower weight on their less knowledgeable counterparts.

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. The main goal of this tournament is to develop innovative methods of assigning accurate probability estimates to events of national security interest. Each month, eight to ten new questions are added to the tournament, for a total of approximately 120 questions per year. While the teams are asked to suggest forecasting questions, an external party makes the final decision for inclusion in the tournament.

The current version focuses on forty-five binary (yes/no) questions that have resolved since the beginning of the 2012-2013 tournament year. Approximately seventy questions are expected to resolve by the end of March 2013, which may alter the current pattern of results.

Each question included in the tournament (e.g., "Will Victor Ponta resign or vacate the office of Prime Minister of Romania before 1 November 2012?") must satisfy the 10/90 rule: the moment a question is posed, a hypothetical knowledgeable observers should not place probability estimates outside the range between 10% and 90%. In other words, questions with seemingly obvious answers are not included in the tournament.

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. Dollar values represent play money so there no financial incentives for performance are provided. Participants, however, are given frequent feedback and face social incentives, including a leader-board for the top 20% participants in terms of total earnings. Financial incentives have been shown to exert minimal influence on prediction accuracy [14]. Forecasters are free to choose which questions to bid 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 possible questions. Two markets are run in parallel for all questions. Forecasters are randomly assigned to one of two parallel prediction market conditions. In the first one, they receive basic training on prediction markets. In the second condition, participants receive an additional one hour of training on forecasting and probability reasoning. Mellers et al. (2012) show that such training improves performance [2].

The Brier scoring rule is used to assess forecast accuracy [15]. According to this strictly proper rule, the penalty is the squared difference between the forecast value and the outcome (0 and 1), summed over the two answer options (e.g. yes/no). 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.

$$\begin{aligned} BS &= \frac{1}{N}\sum_{i=1}^{N} BS_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{k=1}^{D_i}\sum_{j=1}^{2}(f_{ijk}-x_{ij})^2}{D_i}\right) \end{aligned}$$

where i refers to a question (out of N total questions), $j(= 1, 2)$ refers to an outcome, k refers to a specific day, D_i is the number of days that question i is open, and x_{ij} equals 1 if outcome j for question i occurs and 0 otherwise.

Note that in this variation of the Brier score, numeric values are exactly twice as large as the values in another commonly used version in which scores vary between 0 and 1 and at-chance performance is 0.25. 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 and receive no crowd feedback. Moreover, no temporal smoothing, weighting or transformation is applied to individual or aggregate forecasts. Second, the prediction market condition features both the PM interface and the CDA order-matching algorithm. Finally, the marketcast uses values elicited through the prediction market but pooled using survey forecast aggregation methods.

5 Results

In total, 524 participants submitted at least one order for at least one of the 45 binary questions they faced. On average, 132 individuals submitted at least one order on any given question, resulting in 357 orders over the course of a typical question. Questions were open for an average of 100 days, with approximately 5.94 unique orders submitted per day per question. The first day after a question opened attracted the 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, 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 performed a series of paired t-tests to determine if the distributions of Brier Scores for marketcasts were significantly different from the core marketcast. All p-values are for the comparison between core marketcast specification and other methods. No Bonferroni adjustment for multiple comparisons was used, increasing the likelihood that some significant differences may have occurred by chance alone.

	Mean Brier Score			
ULinOp (Control)	0.368 (0.254)*			
Core Prediction Market	0.287 (0.374)			
Core Marketcast: Equal Weights, 10% Margin, All Orders, Most Recent 15%, Mean, Non-transformed	0.299 (0.368)			
	Marketcast Sensitivity Analyses			
1. Order Volume Weight	**Equal Weights**	**SqRt Weights**		
	0.299 (0.368)	0.301 (0.368)		
2. Profit margin (m)	**m=10%**	**m=0%**	**m=25%**	
	0.299 (0.368)	0.302 (0.360)	0.306 (0.359)	
3. Order matching	**All Orders**	**Matched Orders**		
	0.299 (0.368)	0.296 (0.368) *		
4. Temporal smoothing	**Most Recent 15%**	**c=0.10**	**c=0.50**	**c=0.85**
	0.299 (0.368)	0.308 (0.378)	0.300 (0.366)	0.315 (0.347)
5. Measure of central tendency	**Mean**	**Median**	**Logit**	
	0.299 (0.368)	0.302 (0.386)	0.303 (0.410)	
6. Transformation	**Non-Transformed**	**Transformed**		
	0.299 (0.368)	0.340 (0.508)*		
7. Expertise Weights	**Equal Weights**	**Expertise Weights**		
	0.299 (0.368)	0.295 (0.368)*		

Table 2. Mean Brier scores for 45 questions in the tournament.
* Denotes significant difference compared to core marketcast in two-tailed paired t-tests.

Overall, marketcast performance varied in a limited range. On the one hand, almost all marketcast methods yielded lower Brier scores than the ULinOp control condition. On the other hand, marketcasts produced higher Brier scores than the prediction market.

Sensitivity analyses, reported in the lower half of Table 2, 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 performed slightly better than those using all orders, which suggests that unmatched orders did not provide useful information.

Fourth, temporal smoothing parameters had a small impact on overall performance. A moderate level of exponential smoothing (c=0.50) yielded the best results but including only the last 15% of orders yielded the best results. Fifth, taking the mean marketcasts yielded slightly, but not significantly, lower Brier scores than either the median or the geometric mean of log odds (logit). However, the logit aggregator yielded lowest Brier score when the profit margin was set to zero. Finally, non-transformed marketcasts performed better than extremized ones, which suggested that marketcasts in this sample did not exhibit the long-shot bias.

In addition to the manual sensitivity analyses, we performed an optimization run using elastic net regularization (Zou & Hastie, 2005), in order to extract the optimal, Brier-score minimizing specification for aggregation parameters, while avoiding overfitting [16]. The optimal specification included only matched orders, used the logit aggregator with no transformation, made use of only the 15% most recent orders at a time, and gave higher weights to participants who provide higher self-rating of expertise and tend to submit more market orders. The mean Brier score for this combination was 0.277, slightly but not significantly better than the core prediction market, which yielded an average score of 0.287.

Figure 1 depicts performance of various methods by question, in increasing order of Brier scores for the core marketcast specification. In other words, questions on the left side (1, 2, 3) were correctly forecasted by the marketcast, while those on the right side resolved in unexpected ways: Brier scores above 0.5 mean that forecasts were, on average, on the wrong side of 50%. Marketcast performance tracked core prediction very closely, and both methods are visibly more accurate than the ULinOp control condition for the majority of questions.

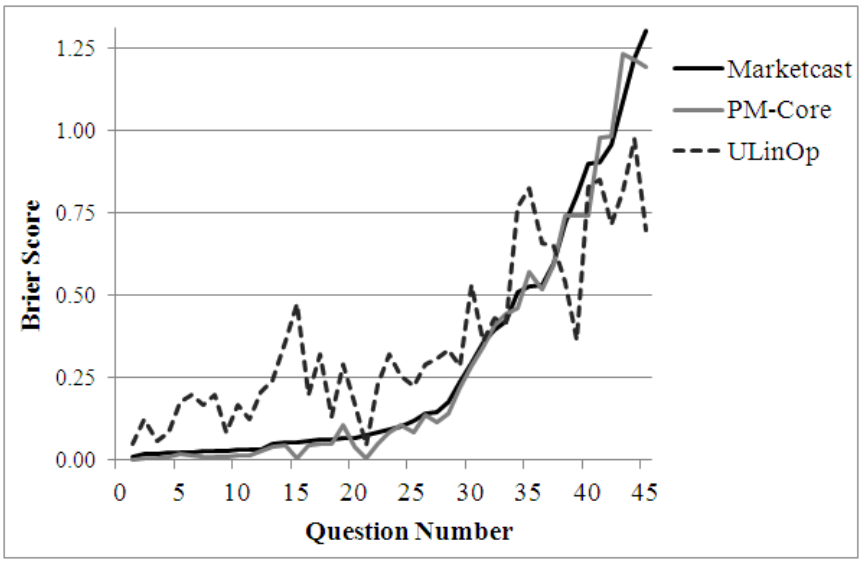

Figure 1. Brier scores per IFP for ULinOp, core prediction market and marketcast.

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. Some marketcast specifications slightly underperform traditional prediction markets, while the best specification produces forecasts that are 3% more accurate in terms of Brier score. Overall, the method produces stable and accurate forecasts

conditions and specifications. Future research should 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. 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.
3. Christiansen, J.D.: Prediction Markets: Practical Experiments in Small Markets and Behaviours Observed. J. of Prediction Markets. 1, 17-41 (2007)
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. Cliff, D., Bruten, J. Zero Not Enough: On The Lower Limit of Agent Intelligence For Continuous Double Auction Markets. HP Laboratories Technical Report HPL. (1997)
8. Othman, A. Zero-intelligence agents in prediction markets. In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2 (pp. 879-886). International Foundation for Autonomous Agents and Multiagent Systems. (2008)
9. Wolfers, J., Zitzewitz, E. Interpreting prediction market prices as probabilities (No. w12200). National Bureau of Economic Research. (2006)
10. 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.
11. Page, L., Clemen, R.T.: Do Prediction Markets Produce Well Calibrated Probability Forecasts? Econ. J. (2012)
12. Rothschild, D.: Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases. Public Opinion. Q. 73, 895–916 (2009)
13. Baron, J., Ungar, L.H., Mellers, B.A., Tetlock, P.E.: Two Reasons To Make Aggregated Probability Forecasts More Extreme. Manuscript under review.
14. Servan-Schreiber, E., Wolfers, J., Pennock, D., Galebach, B.: Prediction Markets: Does Money Matter? Electronic Markets. 243-251 (2004)
15. Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly weather review. 78, 1-3 (1950)
16. Zou, H., Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. (2005)