Prediction Without Markets

Sharad Goel, Daniel M. Reeves, Duncan J. Watts, David M. Pennock
Yahoo! Research, 111 West 40th Street, New York, NY 10018
{goel, dreeves, djw, pennockd}@yahoo-inc.com

ABSTRACT

Citing recent successes in forecasting elections, movies, products, and other outcomes, prediction market advocates call for widespread use of market-based methods for government and corporate decision making. Though theoretical and empirical evidence suggests that markets do often outperform alternative mechanisms, less attention has been paid to the magnitude of improvement. Here we compare the performance of prediction markets to conventional methods of prediction, namely polls and statistical models. Examining thousands of sporting and movie events, we find that the relative advantage of prediction markets is surprisingly small, as measured by squared error, calibration, and discrimination. Moreover, these domains also exhibit remarkably steep diminishing returns to information, with nearly all the predictive power captured by only two or three parameters. As policy makers consider adoption of prediction markets, costs should be weighed against potentially modest benefits.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Algorithms, Measurement, Economics

Keywords

Forecasting, prediction markets, polls, statistical modeling

1. INTRODUCTION

Since at least Hayek [23], economists have recognized that market prices represent the aggregation of many different beliefs about the world. When the beliefs in question concern some future state of the world, be it about the impact of weather on crop yields or the possibility of armed conflict disrupting oil supplies, the corresponding prices can be interpreted as predictions about the relevant outcomes. Indeed, although designed to allocate resources or risk, traditional financial markets [23, 26, 39, 46] and sports betting markets [40, 43, 47, 53, 55] can be viewed as making implicit predictions.

More recently, researchers have begun to design markets—often called prediction or information markets—for which the generation of predictions is the explicit goal. In these markets, participants buy and sell securities that realize a value based on the occurrence of some future outcome, such as the result of an election, the box office revenue of an upcoming film, or the market share of a new product. For example, the day before the 2008 U.S. presidential election, you could have paid $0.92 for a contract in the Iowa Electronic Markets (`www.biz.uiowa.edu/iem`) that yielded $1 when Barack Obama won, implying a 92% market-estimated probability that Obama would win.

Considering the difficulty of outperforming index funds in equity markets, the notion that markets may be capable not only of making predictions, but of doing so optimally, is both plausible and appealing. Moreover, there are compelling theoretical reasons to expect that prediction markets should outperform other forecasting methods. First, they offer rewards for accuracy, incentivizing participants to gather and process information; and second, they weigh the opinions of confident agents more highly, where confidence is reflected in one's willingness to risk more money and overconfidence is penalized over time. Thus, prices in prediction markets can be set either by a small number of highly informed (and confident) participants, or by a large number of individuals each with one piece of the puzzle. Finally, the efficient market hypothesis [32, 11, 41] asserts that markets incorporate information attainable by any competing method. For example, if a poll of experts were to establish a track record of outperforming a prediction market, then at least one market participant would presumably exploit that advantage by arbitraging the difference between polls and market prices. As long as any performance difference remains, in fact, the participant could make money in the market; hence, prices should update to eliminate any performance disparity. In other words, prediction markets are designed to elicit information from whomever has it, and however it is distributed.

Inspired by such theoretical arguments, and also by a growing body of empirical findings that show markets beat alternatives, several authors have called for widespread application of prediction markets to real-world business strategy and policy development problems [1, 3, 19, 20, 37, 49,

EC'10, June 7–11, 2010, Cambridge, Massachusetts, USA.

56]. The theoretical and empirical analyses on which these claims are based, however, have focused primarily on the relative ranking of prediction methods. By contrast, the magnitude of the differences in question has received much less attention, and as such, it remains unclear whether the performance improvement associated with prediction markets is meaningful from a practical perspective. Here we compare the performance of prediction markets to polls and statistical models across several thousand sports and movie events. We find that all reasonable prediction methods perform roughly equally on three related, but distinct measures: squared error, calibration, and discrimination. For example, the Las Vegas market for professional football is only 3% more accurate in predicting final game scores than a simple, three parameter statistical model, and the market is only 1% better than a poll of football enthusiasts. That such elementary methods perform comparably to well designed and mature markets illustrates the surprisingly stark diminishing returns to information, and suggests, more generally, that there may be rather severe limits to prediction.

In the next section we review previous work on prediction markets. In Section 3 we describe the market data and detail our methodology, and in Section 4 we present our main results—analyses of football, baseball, and movie markets. We conclude in Section 5 by discussing the implications and limitations of our findings.

2. RELATED WORK

There is a substantial body of empirical evidence showing that prediction markets frequently make more accurate predictions than opinion polls and expert analysts [2, 3, 54, 56]. For example, a number of studies examine political election markets like the Iowa Electronic Markets (IEM) [2, 14, 15, 33, 34], while others examine markets on the Irish betting exchange TradeSports (now Intrade) [51, 50, 57]. In addition to field studies, laboratory experiments have been conducted to examine the performance of prediction markets [13, 35, 36, 48], and to identify various factors—like the number of traders [7], the market payment rules [28], and the design of the security to be traded [6]—that affect accuracy. A common concern about prediction markets is that wealthy traders with ulterior motives could manipulate prices. Rhode and Strumpf [38], however, analyze both controlled and uncontrolled manipulation attempts in real markets and find that the effects of manipulations are for the most part minimal and short lived. Hanson et al. [22] also find that markets appear robust to manipulation in a laboratory setting, while Hanson and Oprea [21] theorize that manipulators, like noise traders, can actually help market liquidity and accuracy.

Other evidence, however, suggests that the relative performance advantage of markets may be small, and that markets may not even be the best performers. In predicting the outcome of football games, pooled expert assessements are comparable in accuracy to information markets [5, 9]. Erikson and Wlezien [10], moreover, argue that previous studies showing that election markets outperform opinion polls make the wrong comparison. They point out that opinion polls reflect preferences on the day the poll is taken, and therefore overestimate the probability that the current poll leader will win—a bias that is particularly acute far in advance of the election. Correcting for this fact, Erikson and Wlezien generate predictions that are superior to those of the IEM. Graefe and Armstrong [18] have likewise found that a simple statistical model, based on single-issue voting preferences, outperformed the IEM with respect to election winners—although their model underperforms the IEM with respect to vote share. Furthermore, Healy et al. [24] show that iterative polls are more robust than markets with few people participating or many outcomes to predict. Finally, though financial incentives are often cited as a key reason for why markets should outperform alternatives, Servan-Schreiber et al. [44] find that play-money and real-money markets perform comparably.

3. METHODS

We examine predictions of over 7,000 U.S. National Football League (NFL) games, nearly 20,000 Major League Baseball (MLB) games, and box office revenue for approximately 100 feature films. Though political and policy markets are arguably of the greatest interest, we focus on sports and movies for two reasons: first, events in these domains happen with much higher frequency than presidential elections or product launches, greatly facilitating rigorous evaluation; and second, prediction markets for sports and entertainment are among the deepest and most mature. In the discussion, we consider whether and how our results generalize to other domains.

Market data are obtained from the Las Vegas sports betting markets, TradeSports (now Intrade), and Hollywood Stock Exchange (HSX). The Vegas and TradeSports markets are both real-money markets, and offer participants substantial financial incentives. In 2008, Nevada gamblers bet more than $1.1 billion dollars on football and more than $500 million on baseball [4]. TradeSports is much smaller but still relatively deep, with tens of thousands of members trading hundreds of thousands of contracts [45]. The play-money market Hollywood Stock Exchange is the world's leading online entertainment market, garnering about 25,000 unique visitors and 500,000 page views per month in the U.S.[1]

Performance Metrics. We assess the performance of prediction mechanisms along three dimensions: root mean squared error (RMSE), calibration, and discrimination.

RMSE quantifies an average difference between predicted and actual outcomes:

$$\texttt{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - X_i)^2}$$

where n is the number of events for which predictions are made, p_i is the predicted outcome for event i, and X_i is the actual outcome. In the case of football and baseball games, we mostly consider binary outcomes $X_i \in \{0, 1\}$, indicating whether the home team wins, where p_i is then the predicted probability of that occurring. For movies, we take X_i to be the logarithm of opening weekend box-office revenue.

Though RMSE is one of the most common measures of prediction accuracy, it is in some respects a crude test of performance. In particular, RMSE does not directly assess a prediction method's ability to distinguish between likely and unlikely events. Thus we additionally consider two other performance measures: calibration, which measures the agreement between predicted and observed prob-

[1] Web traffic data obtained from `quantcast.com`.

abilities; and discrimination, which captures the empirical variability of probabilities over outcomes.

To formally define calibration and discrimination, we first bin predictions into discrete categories. In predicting the probability the home team wins in football and baseball games, we round predictions to the nearest 5%, in which case predictions fall into 21 categories: $\{0, 0.05, \ldots, 0.95, 1\}$. For movies, where we predict the natural logarithm of box-office revenue, we round predictions to the nearest 0.5. Specifically, for each event $i = 1, \ldots, n$ define $\tilde{p}_i$ as the value of the prediction p_i rounded to the nearest category, and define $b_{\tilde{p}_i}$ to be the empirically observed average outcome in that category—for binary outcomes (e.g., indicating whether a team wins) this average is just the proportion of the events that occur. So, for example, if five events were predicted to occur with probability between 0.375 and 0.425, and three of those five events did ultimately occur, we would have $\tilde{p}_i = 0.4$ for all five events and $b_{0.4} = 3/5$. The calibration error is then the root mean squared error between predicted and empirically observed probabilities. Specifically:

$$\texttt{Calibration Error} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{p}_i - b_{\tilde{p}_i})^2}$$

Thus, when a mechanism with zero calibration error predicts an event to occur with probability 0.6, 60% of those events in fact happen.

On its own, low calibration error is not difficult to achieve. For example, knowing that New York City has approximately 121 days of precipitation annually, a perfectly calibrated, but minimally informative rule is to simply predict the chance of rain each day to be 0.33. We hence measure not only calibration, but also discrimination, or the variability of outcomes across prediction categories. Using the same notation as above:

$$\texttt{Discrimination} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_{\tilde{p}_i} - b)^2}$$

where $b = (\sum_i X_i)/n$ is the average outcome across all events. More informative mechanisms tend to have higher discrimination. In particular, though the extreme example of always predicting 33% chance of rain in New York City is perfectly calibrated, it has zero discrimination.

4. RESULTS

4.1 Football

In predicting outcomes for NFL games, we compare Vegas and TradeSports prediction markets against two poll variants and two simple statistical models. The first poll variant ("filtered polls") was run weekly on Amazon's Mechanical Turk (`mturk.com`), a web-based crowdsourcing [25, 27] service that permits requestors to post open solicitations for workers to perform tasks (called "human intelligence tasks," or HITs) along with a specified compensation. Workers elect to complete any number of these tasks for which they are then paid by the corresponding requestor. HITs range widely in size and nature, requiring from seconds to hours to complete, and compensation varies accordingly, but is typically on the order of \$0.01–\$0.10 per HIT. In our case, the HIT in question was to make a probabilistic prediction regarding the outcomes of football games. Specifically, in each of the 15 weeks of the 2008 NFL season, we asked 100 people to answer the question "What do you think is the likelihood A will beat B?" for each of the upcoming weekend's scheduled games. We also asked them to state whether they were "confident" or "not confident" in their predictions. We generated aggregate predictions by taking an unweighted average of predictions from confident respondents, where we emphasize that expressed confidence was purely self-reported. Participants were paid \$0.03 per prediction, regardless of their accuracy or confidence; thus poor performance was not subject to any penalties. Moreover, Mechanical Turk has no explicit sporting orientation, nor did we provide any incentives for experts to participate. Thus one would not expect respondents to have any particular expertise beyond what is typical in the general population.

Our second, incentivized poll uses data collected from Probability Sports (`probabilitysports.com`), an online contest in which participants compete for cash prizes by predicting the outcomes of sporting events. As with the filtered polls run on Mechanical Turk, participants made probabilistic predictions. However, participants on Probability Sports are scored according to a quadratic scoring rule [42], incentivizing and rewarding accuracy. Predictions are publicly visible, and we collected a total of 1.4 million such predictions for 2017 NFL games played over the course of eight years, from 2000 to 2007.[2] We generated an aggregate prediction for each game by taking the unweighted average of all individual predictions for that game. In this case we did not exclude any individual predictions when computing the average since those who decided to enter the contest had already presumably screened themselves.

In addition to the two polls, we compared the markets' performance against that of two simple statistical models. The first uses only the historical probability of the home team winning in NFL match-ups. Based on 31 years of NFL data, we find this baseline probability is $b = 0.58$. Thus, our first model—the baseline model—predicts for each game that the home team will win with probability 0.58, regardless of which teams are playing. The second model—the win-loss model—incorporates both the home field advantage captured by the baseline, and the recent win-loss record of the two playing teams. Specifically, when teams A and B play each other on A's home field, the win-loss model estimates the probability A wins to be $b + (R_A - R_B)/2$, where $b = 0.58$ is again the baseline probability of the home team winning, R_A is the percentage of games team A has won out of its last 16 match-ups (the number of regular season games played annually by each team), and R_B is the corresponding percentage for team B.[3] This model, while more complicated than the baseline prediction, still ignores almost all the details of any particular game, incorporating only easily obtainable information.

The polls and models described above all generate predictions for the probability the home team wins. In contrast,

[2] Probability Sports was discontinued at the end of the 2007–2008 season.

[3] To motivate the win-loss model, we note that the approximate percentage of home games A wins is $b + (R_A - 1/2)$, and the approximate number of away games B loses is $1 - [(1 - b) + (R_B - 1/2)]$. Averaging these two quantities gives the model estimate. Alternatively, one could fit a logistic regression with R_A and R_B included as features; doing so yields similar results.

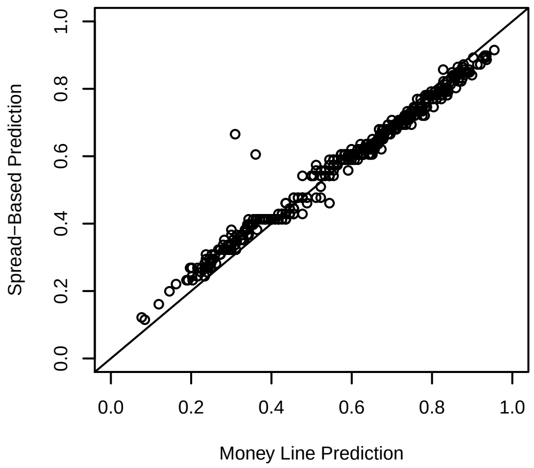

Figure 1: A comparison of money line predictions of the home team winning in NFL games to predictions generated via a model that converts point spreads to probabilities.

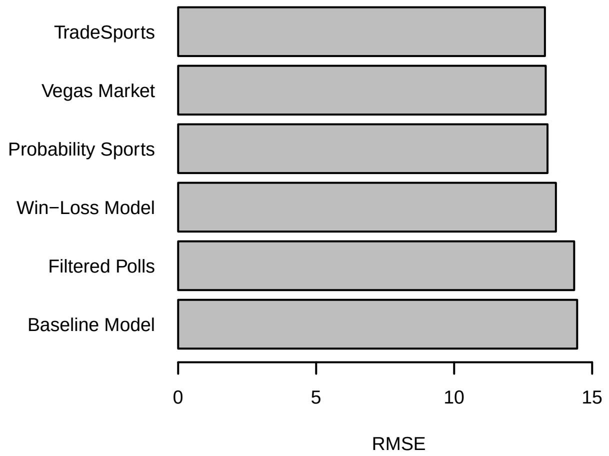

Figure 2: RMSE of six methods for predicting final point differences (i.e., home team score minus away team score) in NFL games.

football markets generally yield spread predictions on the final point difference between the playing teams (i.e., the home team score minus the away team score). Fortunately, spread and probabilistic predictions are statistically comparable [16]. To transform spread to probabilisitic predictions, on 7,152 NFL games from 1978 to 2008 we fit the logistic regression model

$$\Pr(\texttt{home team wins}) = \text{logit}^{-1}(\beta_0 + \beta_1 \times \texttt{spread})$$

where $\text{logit}^{-1}(x) = e^x/(1+e^x)$.[4] For a subset of 494 NFL games, we have both spread and probabilistic market predictions, from so-called money-line markets. On this subset we find that the spread-inferred and the probabilistic market predictions are in very good agreement, having a correlation of 0.99. This conversion is depicted in Figure 1, where each circle represents an NFL game. The probabilistic prediction is given on the x-axis, and the prediction inferred from the spread via the regression model is given on the y-axis. In light of this tight relationship, we convert between spread and probabilistic predictions as convenient.

Having described the six methods—two markets, two polls, and two statistical models—for predicting the probability the home team wins in NFL games, we consider the overall performance of each mechanism. Consistent with past empirical studies and theoretical arguments, the Vegas and the TradeSports markets are the best performers, both having an RMSE of 0.46. At the other extreme the baseline model is the worst performer, with an RMSE of 0.49. The performances of the remaining strategies lie in between that of the markets and the baseline model: Probability Sports, the win-loss model, and the filtered polls all have an RMSE of 0.47. To aid interpretation of these results, and also to ensure that the markets are not handicapped by our conversion of spread to probabilistic predictions, we consider the complementary problem of predicting the final point difference between the playing teams. Figure 2 shows that RMSE in this case ranges from 13.3 for the markets to 14.5 for the baseline model. On average, that is, the market predictions differ from the actual point difference by approximately 13.3 points, and predictions from the baseline model are off by 14.5 points. Overall, the ordering of these prediction methods is unsurprising: prediction markets beat models and polls, and all methods beat the baseline. What is surprising, however, is that the various mechanisms differ by so little: in predicting the final point difference, the win-loss model—which recall has only three parameters—is only 0.4 points (3%) worse than the markets, and Probability Sports is only 0.1 points (1%) worse than the markets. Figure 3 displays the difference between the Vegas market and the win-loss model from 1978 to 2008.

The similarity in performance of these prediction methods, moreover, is not due to any apparent anomaly in the markets. To test for obvious market inefficiencies, we predicted the final point difference in each game via a model that includes the market spread along with several other features. Specifically, we fit the linear regression model

$$\begin{aligned}\texttt{point difference} = \beta_0 + \beta_1 \times \texttt{spread} + \beta_2 \times X_{\texttt{spread}>0} \\ + \sum_i \beta_{\text{hometeam}[i]} \times H_i \\ + \sum_i \beta_{\text{awayteam}[i]} \times A_i + \epsilon\end{aligned}$$

where `spread` is the market predicted point spread, $X_{\texttt{spread}>0}$ is a dummy variable indicating whether the spread is greater than zero, and H_i and A_i are dummy variables indicating which teams are playing in the given game. In particular,

[4] To convert the probabilistic poll and model predictions to spread predictions, we analogously fit a linear regression:

$$\texttt{spread} = \beta_0 + \beta_1 \times \texttt{predicted probability} + \epsilon$$

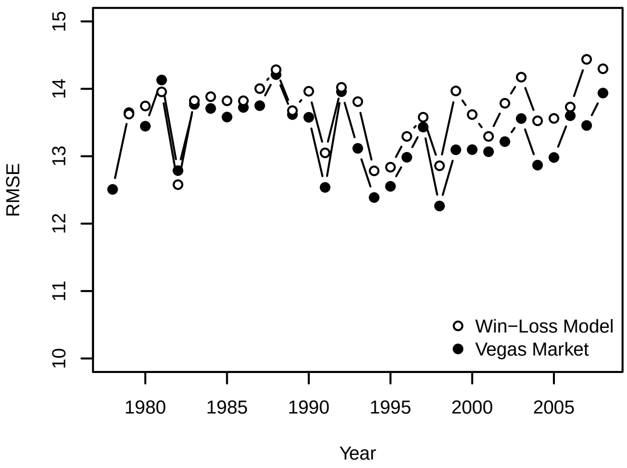

Figure 3: Yearly performance of the Vegas market and the win-loss model in predicting final point differences (i.e., home team score minus away team score) in NFL games.

	Calib. Err.	Discrim.	RMSE
Vegas Markets	0.02	0.17	0.46
TradeSports	0.05	0.19	0.46
Probability Sports	0.05	0.17	0.47
Win-Loss Model	0.02	0.14	0.47
Filtered Polls	0.10	0.18	0.47
Baseline Model	0.02	0.00	0.49

Table 1: Calibration error, discrimination, and RMSE for several methods in predicting the probability the home team wins in NFL games.

the model corrects for systematic bias that depends on which teams are playing. In an efficient market, a prediction based on such a model should perform on par with simply using the spread to predict the final point difference. We find this to be the case: with 5-fold cross-validation, the RMSE of 13.3 points for this model is identical to the RMSE of the spread alone.[5] Furthermore, despite differences in the fee structure of the Vegas and TradeSports markets, both perform identically.

We next move beyond RMSE to account separately for calibration and discrimination. Figure 4 shows the full distribution of predicted and empirically observed outcomes of the six forecasting methods, where predictions are binned into 5% intervals and the area of each circle represents the number of predictions in the corresponding probability range. As should be clear from the figure, all methods produce predictions that lie roughly on the diagonal. All methods are therefore reasonably well calibrated—predicted probabilities agree with observed probabilities within any given bin—however, they differ in their ability to discriminate. Most notably, whereas the baseline model includes all events in the same bin, thereby effectively treating high and low probability events as indistinguishable, other methods distinguish between empirically likely and unlikely events, as indicated graphically by the dispersion of bins along the diagonal. Table 1 confirms these visual impressions, quantifying the calibration and discrimination of each method, and also reveals two main findings. First, two out of the six methods are inferior to the others along one dimension or the other: filtered polls discriminate well but are not as calibrated as the other methods; whereas the baseline model is well calibrated but does not discriminate. And second, four of the six methods—the Vegas and Tradesports markets, Probability Sports, and the win-loss model—remain comparable both in terms of calibration and discrimination.

That the poor discrimination of the baseline model carries so small a penalty in terms of RMSE is due in part to specific features of the NFL (e.g., salary caps) that ensure that most games are played between closely matched teams, and hence are decided with probabilities close to 50%. In other words, although the baseline model does perform poorly for high and low probability events, the relative rarity of such events means that it is not penalized much for these failures. One might therefore suspect that in domains (such as policy analysis) where events are not designed to be coin tosses, and where possibly the predictions of greatest interest may be for extreme probability events, less discriminating methods would perform correspondingly worse than they do here. To test this idea, we recomputed the RMSE of the six prediction methods exclusively for lopsided pairings between "winning" teams (that have won at least 9 of their past 16 games) and "losing" teams (lost at least 9 of 16). This subset comprises 37% of the data. As expected, the baseline model performed worse on these games (RMSE increased from 14.5 to 15.1 points), but RMSE of the TradeSports and Vegas markets, Probability Sports, and the win-loss model were approximately unchanged, at 13.0, 13.2, 13.4, and 13.5, respectively. Thus, even for these more extreme events, we find prediction markets again have only a small advantage over conventional forecasting methods.

4.2 Baseball

Although we have considered a number of performance measures, it is possible that football remains a special case even in the domain of sports in that outcomes are dominated by hard to anticipant events—a hail Mary pass in the final minutes, for example, or an intercepted ball against the flow of play—for which there is relatively little real information on which to base sophisticated predictions. In addition to football, therefore, we consider Major League Baseball (MLB)—a sport for which very large amounts of data are collected, and where an entire field, sabermetrics, has been developed along with its own journal, the Baseball Research Journal, specifically for the purpose of analyzing performance statistics. In light of this considerable devotion to statistical models and prediction, one might assume that expert observers, and hence prediction markets, would outperform simplistic models by incorporating game-specific variables like pitching rotation, the recent batting performance of individual players, and so on. As described below,

[5] Cross-validation—also known as rotation estimation—protects against overfitting the model to the data. Events are first partitioned into $k = 5$ subsets of approximately equal size, and then predictions are made for events in each of the k subsets via a model trained on the remaining $k - 1$ subsets.

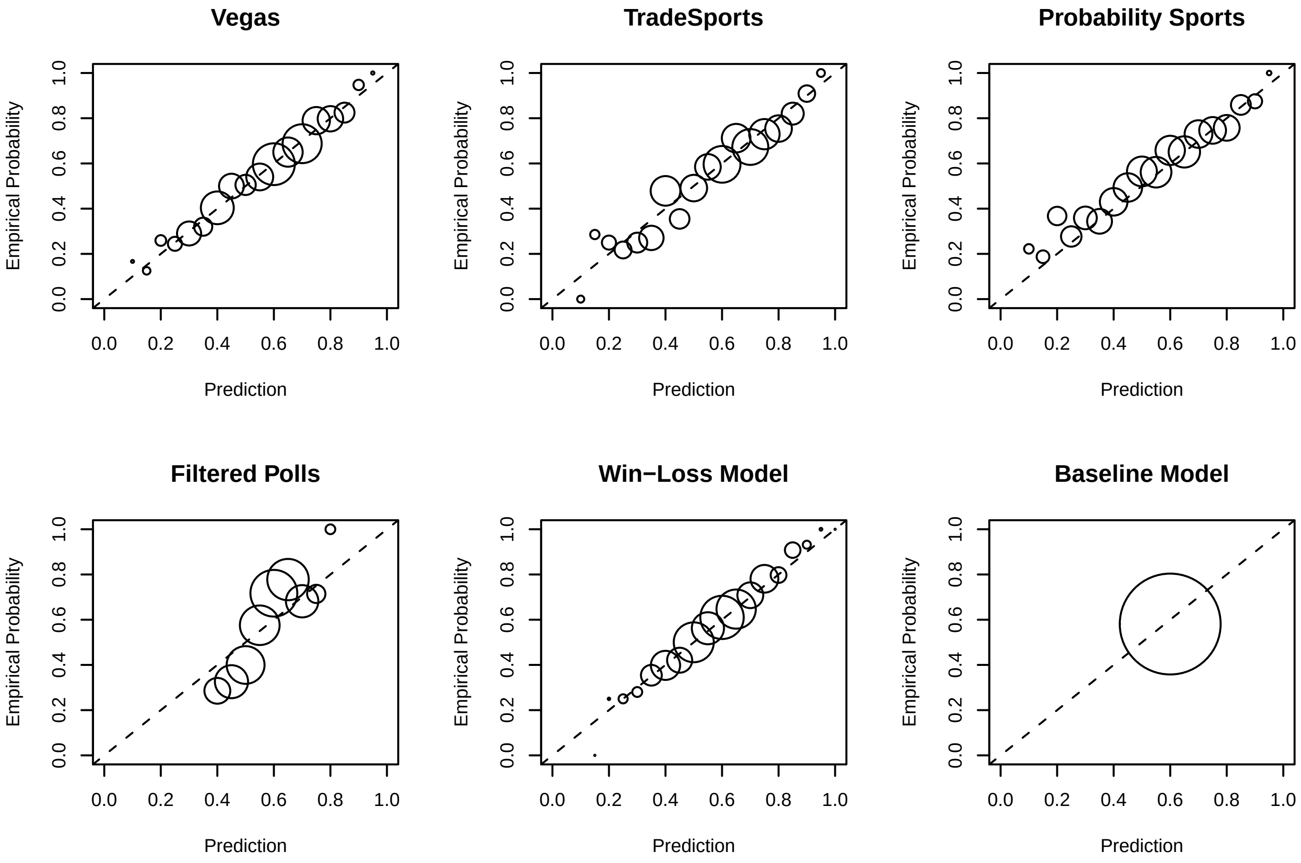

Figure 4: Distribution of predicted and empirical probability estimates for the home team winning in NFL games. The area of each circle represents the number of predictions in the corresponding probability range.

however, we find that baseball markets have only a small advantage over alternative forecasting tools.

We compare the performance of the Vegas market to the baseline and win-loss models for 19,633 Major League Baseball (MLB) games played over seven years, from 1999 to 2006, where the two models were constructed in the same manner as for football. Specifically, the baseline model ignores all game specific information, always predicting the home team wins with probability 0.54—the historical winning percentage of the home team in baseball. Correspondingly, the win-loss model for baseball was identical in form to that used for football predictions: when teams A and B play each other on A's home field, the probability A wins is estimated to be $b + (R_A - R_B)/2$, where $b = 0.54$ is the baseline probability of the home team winning, R_A is the percentage of games team A has won out of its last 162 match-ups (the number of regular season games each team plays annually), and R_B is the analogous percentage for team B.

In terms of the three performance measures introduced above—RMSE, calibration, and discrimination—we find once again that the win-loss model performs on par with the market (Figures 5 and 6; Table 2). In particular, the market and the win-loss model both have an RMSE of 0.49, slightly outperforming the baseline model, which has an RMSE of 0.50.[6] Furthermore, all three methods are well calibrated, with calibration errors of 0.02 for the market and the win-loss model, and 0.01 for the baseline model. Finally, although the shortcomings of the baseline model are apparent from its inability to discriminate between high and low probability events, the market and win-loss model remain comparable by this measure as well, with discrimination 0.09 and 0.07, respectively.

	Calib. Err.	Discrim.	RMSE
Vegas Markets	0.02	0.09	0.49
Win-Loss Model	0.02	0.07	0.49
Baseline Model	0.01	0.00	0.50

Table 2: Calibration error, discrimination, and RMSE for several methods in predicting the probability the home team wins in MLB games.

4.3 Movies

Given the amount of time, energy, and money dedicated to predicting the outcomes of baseball and football games, it is perhaps surprising that in both cases, a relatively simple statistical model can perform almost as well as the best avail-

[6] An RMSE of 0.5 is achievable for any probabilistic prediction by always predicting 1/2.

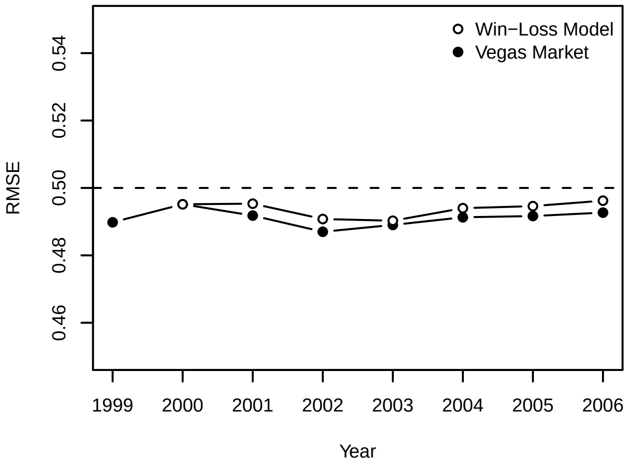

Figure 5: Yearly RMSE performance in predicting the probability of the home team winning in MLB games.

able prediction markets. Knowing this, however, one might still argue that our results merely illustrate that sporting events in general are designed to produce hard-to-predict outcomes, thereby providing the greatest amount of suspense, and hence enjoyment for fans. One might suspect, therefore, that sporting events are systematically different from other domains where events simply transpire in a way that, if planned at all, is certainly not designed to maximize uncertainty. To address this concern, we now consider a very different domain than sports, examining the relative performance of markets and statistical models in predicting the commercial success of movies. As different as movies are from sports, they do share two important features in common: first, they open regularly, and therefore provide a good source of data; and second, they are the subject of a very popular and well-developed prediction market, the Hollywood Stock Exchange (HSX), that has frequently been cited by advocates of prediction markets as evidence of their efficacy [49].

We compare the HSX prediction market to two simple statistical models in predicting opening weekend box-office revenues for 97 feature films released between September 2008 and September 2009.[7] Although our methods are largely similar to those used above, the nature of the phenomenon in question necessitates one modification. Revenue across movies varies over several orders of magnitude, from hundreds of thousands of dollars to hundreds of millions; therefore, all predictions are made and evaluated on the log scale. As with football and baseball, the baseline statistical model predicts each movie will earn the average amount among all recent movies, which in this case is \$8.1 million (15.9 on the log scale). The second, more informative model incorporates two additional features that have been shown to predict box office revenue [17]: the number of screens on which the movie opens, as reported by the Internet Movie Database (IMDB); and the total number of web searches for the movie in the week leading up to its opening, as recorded by Yahoo! Search.[8,9] We note that search counts are analogous to polling data, and thus this approach is similar in spirit to our analysis of football. Given screen and search data, predictions were generated with a linear model:

$$\log(\texttt{revenue}) = \beta_0 + \beta_1 \times \log(\texttt{screens}) \\ + \beta_2 \times \log(\texttt{search}) + \epsilon$$

To guard against overfitting, predictions were made via leave-one-out estimation. That is, a prediction for each of the 97 movies was generated by a model trained on the other 96 movies.

As with football and baseball, we find the market yields predictions that are better, but only slightly so, than those from a relatively simple statistical model (Table 3). Specifically, RMSE is 0.65 for HSX and 0.69 for the screens-search model—a difference of only 6%. Since we measure error for

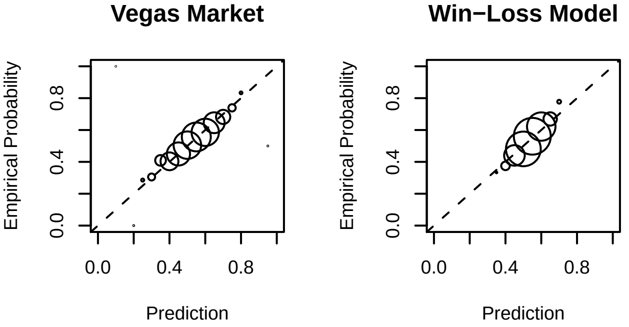

Figure 6: Distribution of predicted and empirical probability estimates for the home team winning in MLB games. The area of each circle represents the number of predictions in the corresponding probability range.

[7] Securities in the Hollywood Stock Exchange are initially tied to opening weekend box office revenue, but are later valued according to a movie's four-week domestic gross. To correct for the fact that an asset's price predicts two related, but distinct, outcomes, we infer the market prediction for opening weekend revenue via a linear model based on the stock price the day before a movie's release:

$$\log(\texttt{revenue}) = \beta_0 + \beta_1 \times \log(\texttt{hsx}) + \epsilon$$

In other words, by fitting the above model we convert raw market prices to box office predictions.

[8] To compute search query volume, a query was categorized as pertaining to a particular movie if an IMDB link to that movie appeared in the first page of search results. When multiple IMDB links appeared in the result set, the query was categorized according to the top-ranking result from IMDB. Though our analysis uses proprietary search data, query volume is also publicly available from Google Trends (`google.com/trends`).

[9] There are several other features that could potentially help predict box office revenue—including production and marketing budgets, genre, MPAA rating, and director and actor statistics—and more sophisticated models have in fact been developed that incorporate this additional information [12]. Opting for simplicity, we limit our analysis to screens and search volume.

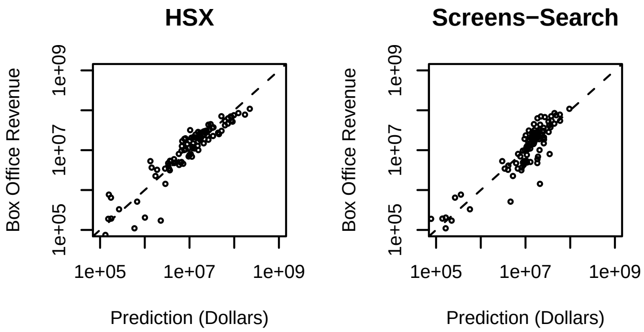

Figure 7: Actual opening weekend box office revenues compared to predictions from HSX and the screens-search model (log scale).

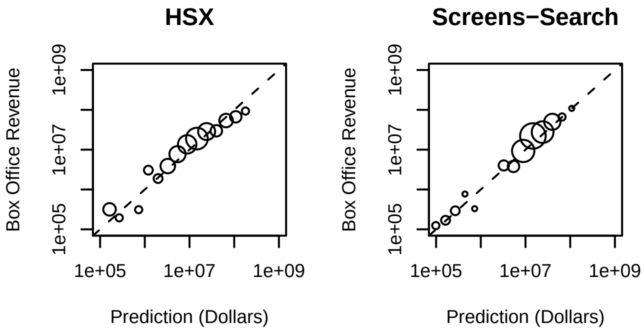

Figure 8: Binned distribution of predicted and actual opening weekend box office revenue (log scale).

movies on the log scale, one can interpret these results as indicating the approximate percent error of each method in predicting opening weekend box office revenue (i.e., HSX is off on average by about 65%, and the screens-search model is off by about 69%). Notably, and in contrast to sporting events, the baseline model does considerably worse than the market, with RMSE of 1.90. As shown in Figure 7, both the market and the screens-search model are reasonably well calibrated, where for ease of comparison with our previous results Figure 8 shows the same data binned. In particular, the calibration error for the model (0.27) is in fact lower than for the market (0.34). Finally, although the baseline model fails to discriminate at all, the market and the screens-search model are again comparable, having discrimination scores of 1.80 and 1.78, respectively.

5. DISCUSSION

Advocates of prediction markets tend to emphasize the fact that markets often perform better than alternative forecasting methods. Our results are consistent with this observation, but put them in a different light. Markets, we find, indeed outperform polls and statistical models in predicting outcomes of football and baseball games, as well as movie openings. However, regardless of which performance measure we use—squared error, calibration, or discrimination—simple forecasting techniques deliver results that are comparable to those of well designed and successful prediction markets. Given the amount of interest in predicting sports and entertainment events, and the plethora of available data, our results challenge the conclusion that markets are superior to alternative prediction mechanisms in substantively meaningful ways.

	Calib. Err.	Discrim.	RMSE
HSX	0.34	1.81	0.65
Screens-Search Model	0.27	1.78	0.69
Baseline Model	0.09	0.00	1.90

Table 3: Calibration error, discrimination, and RMSE for several methods in predicting the logarithm of opening weekend box office revenue for feature films.

A natural objection to this interpretation is that it is easier to make predictions for movies than for political outcomes, or that statistical models require a strict regularity and consistency to perform well, and hence question whether our results extend to other domains. Although reasonable, these doubts should be weighed against recent empirical evidence in political and policy analysis. As noted above, in predicting election winners, the Iowa Electronic Markets were outperformed both by statistically corrected polls [10] and by a model based on single-issue voting preferences [18]. Moreover, while not directly assessing markets, one study of expert political predictions found that statistical models outperformed not only individual experts, but also compared favorably with aggregate forecasts [52]. Presumably, properly designed election markets would in time adjust to incorporate predictions from these alternatives. Thus markets in the long run may still regain their performance advantages as suggested by theory. Nevertheless, these findings are consistent with our claim that market and non-market forecasting techniques are often comparable.

A second objection to our conclusion that small differences in performance are not of practical importance is that in some circumstances, such as the Vegas markets themselves or in applications like high-frequency quantitative trading, even small differences may translate into large cumulative advantages. In other words, our conclusion that simple forecasting methods perform on par with markets has useful implications only in domains where incremental improvements are not of practical value. Precisely what constitutes substantive improvement is a difficult question, and one which we do not address in detail; however, we would suggest that differences of the magnitude we have observed here—roughly a few percentage points—are unlikely to qualify in political, policy, and business applications, areas where markets are claimed to have the greatest potential. In part, this is because outcomes of interest in these domains occur relatively infrequently, and in part because any given prediction is likely to be just one component of a decision that may have many other sources of error. For example, it is not obvious how such small differences in, say, the predicted market share of a potential product line would influence a firm's decision about whether or not to invest in developing the product.

A final objection is that we have not analyzed the relative costs of markets, polls, and models, nor have we examined additional features of prediction mechanisms including real-time response. For example, the IEM contract for Colin Powell to win the 1996 Republican nomination fell precipitously within minutes of Powell's scheduling of a press conference, as traders inferred that he would announce his withdrawl [2]. Similarly, NFL markets on TradeSports update continuously as the games progress: as teams score points, commit turnovers, etc. It is possible, therefore, that markets are able to update their predictions in the face of new information or changing circumstances much faster than other methods, or that they could do so in a less costly manner, and that for this reason they could retain substantial practical advantages. We suspect, however, that most decision settings do not require instantaneous feedback, and that in many cases properly designed models and polls may be able to react almost as quickly as markets.

To conclude, we note that a body of related work suggests that the exercise of prediction in general is subject to strongly diminishing returns to sophistication, regardless of methodology and domain. For example, in reviewing the forecasting literature in psychology, statistics, and management science, Clemen [8] finds that simple methods of aggregating individual forecasts often work reasonably well relative to more complex combinations. And in a series of papers, Makridakis and colleagues [29, 30, 31] have compared the performance of various forecasting models for time series data, ranging from simple (e.g., exponentially weighted moving averages) to sophisticated (e.g., Box-Jenkins, neural networks, etc.). These studies were different from ours in some important respects: the objects of prediction were time series data, not discrete outcomes; the domain of application was largely economic and business, not sports or entertainment; they considered many statistical models, but no prediction markets or polls; and finally, they used different performance measures. Nevertheless, the high-level result—that simple methods perform almost indistinguishably from the most sophisticated methods—is essentially the same as what we find here. Although we remain enthusiastic about prediction markets, we hope that future research on prediction will place more emphasis on the magnitude of performance differences between alternative methods.

Acknowledgments

We thank Brian Galebach for providing the Probability Sports data, and Robin Hanson, Preston McAfee, and David Reiley for helpful conversations and comments.

6. REFERENCES

[1] 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. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

[2] J. E. Berg, R. Forsythe, F. D. Nelson, and T. A. Rietz. Results from a dozen years of election futures markets research. In C. R. Plott and V. Smith, editors, *Handbook of Experimental Economics Results, Volume 1*, pages 742–751. North Holland, 2008.

[3] J. E. Berg and T. A. Rietz. Prediction markets as decision support systems. *Information Systems Frontiers*, 5(1):79–93, 2003.

[4] Center for Gaming Research, University of Nevada, Las Vegas. 2008 Nevada gaming statewide revenue breakdown.

[5] Y. Chen, C. Chu, T. Mullen, and D. Pennock. Information markets vs. opinion pools: An empirical comparison. In *Proceedings of the 6th ACM conference on Electronic commerce*, page 67. ACM, 2005.

[6] Y. Chen and A. M. Kwasnica. Security design and information aggregation in markets, 2006.

[7] J. D. Christiansen. Prediction markets: Practical experiments in small markets and behaviours observed. *Journal of Prediction Markets*, 1(1), 2006.

[8] R. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989.

[9] V. Dani, O. Madani, D. Pennock, S. Sanghai, and B. Galebach. An empirical comparison of algorithms for aggregating expert predictions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. Citeseer, 2006.

[10] R. S. Erikson and C. Wlezien. Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, 72(2):190–215, 2008.

[11] E. Fama. The behavior of stock-market prices. *Journal of business*, 38(1):34, 1965.

[12] M. Ferrari and A. Rudd. Investing in movies. *Journal of Asset Management*, 9(1):22–40, 2008.

[13] R. Forsythe and R. Lundholm. Information aggregation in an experimental market. *Econometrica*, 58(2):309–347, 1990.

[14] R. Forsythe, F. D. Nelson, and G. R. Neumann. Anatomy of an experimental political stock market. *American Economic Review*, 82(5):1142–1161, 1992.

[15] R. Forsythe, T. A. Rietz, and T. W. Ross. Wishes, expectations, and actions: A survey on price formation in election stock markets. *Journal of Economic Behavior & Organization*, 39:83–110, 1999.

[16] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall, 2003.

[17] S. Goel, J. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. What can search predict? Technical Report.

[18] A. Graefe and J. S. Armstrong. Predicting elections from the most important issue facing the country, 2009.

[19] R. W. Hahn and P. C. Tetlock, editors. *Information Markets: A New Way of Making Decisions*. AEI-Brookings Press, 2006.

[20] R. Hanson. Decision markets. *IEEE Intelligent Systems*, 14(3):16–19, 1999.

[21] R. Hanson and R. Oprea. Manipulators increase information market accuracy, 2004.

[22] R. Hanson, R. Oprea, and D. Porter. Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior & Organization*, 60(4):449–459, 2006.

[23] F. A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.

[24] P. Healy, J. Ledyard, S. Linardi, and R.J.Lowery. Prediction market alternatives for complex environments. In *Conference on Auctions, Market Mechanisms and Their Applications*, 2009.

[25] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business, New York, 2008.

[26] J. C. Jackwerth and M. Rubenstein. Recovering probability distributions from options prices. *Journal of Finance*, 51(5):1611–1631, 1996.

[27] F. Kleeman, G. G. Voss, and K. Rieder. Un(der)paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, Technology & Innovation Studies*, 4(1):5–26, 2008.

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

[29] S. Makridakis and M. Hibon. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16:451–476, 2000.

[30] S. Makridakis, M. Hibon, and C. Moser. Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society. Series A*, 142(2):97–145, 1979.

[31] S. Makridakis, R. M. Hogarth, and A. Gaba. Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, In press, 2009.

[32] J. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29(3):315–335, 1961.

[33] K. Oliven and T. A. Rietz. Suckers are born, but markets are made: Individual rationality, arbitrage and market efficiency on an electronic futures market. *Management Science*, 50(3):336–351, 2004.

[34] D. M. Pennock, S. Debnath, E. J. Glover, and C. L. Giles. Modeling information incorporation in markets, with application to detecting and explaining events. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 405–413, Edmonton, CA, 2002. Association for Uncertainty in Artificial Intelligence.

[35] C. R. Plott and S. Sunder. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *Journal of Political Economy*, 90(4):663–698, 1982.

[36] C. R. Plott and S. Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56(5):1085–1118, 1988.

[37] C. Polk, R. Hanson, J. Ledyard, and T. Ishikida. Policy analysis market: An electronic commerce application of a combinatorial information market., 2003.

[38] P. W. Rhode and K. S. Strumpf. Manipulating political stock markets: A field experiment and a century of observational data, 2006.

[39] R. Roll. Orange juice and weather. *American Economic Review*, 74(5):861–880, 1984.

[40] R. N. Rosett. Gambling and rationality. *Journal of Political Economy*, 73(6):595–607, 1965.

[41] P. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Management Review*, 6(2), 1965.

[42] L. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[43] C. Schmidt and A. Werwatz. How accurately do markets predict the outcome of an event? The Euro 2000 soccer championships experiment, 2002.

[44] E. Servan-Schreiber, J. Wolfers, D. Pennock, and B. Galebach. Prediction markets: does money matter? *Electronic Markets*, 14(3):243–251, 2004.

[45] A. Serwer. Making a market in (almost) anything. Fortune, Monday, July 25, 2005.

[46] B. J. Sherrick, P. Garcia, and V. Tirupattur. Recovering probabilistic information from options markets: Tests of distributional assumptions. *Journal of Futures Markets*, 16(5):545–560, 1996.

[47] W. W. Snyder. Horse racing: Testing the efficient markets model. *Journal of Finance*, 33(4):1109–1118, 1978.

[48] S. Sunder. Experimental asset markets. In J. H. Kagel and A. E. Roth, editors, *The Handbook of Experimental Economics*, pages 445–500. Princeton University Press, Princeton, NJ, 1995.

[49] C. R. Sunstein. Group judgements: Statistical means, deliberation, and information markets. *New York Law Review*, 80(3):962–1049, 2005.

[50] P. C. Tetlock. Does liquidity affect securities market efficiency?, 2006.

[51] P. C. Tetlock. How efficient are information markets? evidence from an online exchange, 2006.

[52] P. E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, NJ, 2005.

[53] R. H. Thaler and W. T. Ziemba. Anomalies: Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, 2(2):161–174, 1988.

[54] G. Tziralis and I. Tatsiopoulos. Prediction markets: An extended literature review. *Journal of Prediction Markets*, 1(1), 2006.

[55] M. Weitzman. Utility analysis and group behavior: An empirical study. *Journal of Political Economy*, 73(1):18–26, 1965.

[56] J. Wolfers and E. Zitzewitz. Prediction markets. *The Journal of Economic Perspectives*, 18(2):107–126, 2004.

[57] J. Wolfers and E. Zitzewitz. Using markets to inform policy: The case of the Iraq war, 2006.