Information Markets vs. Opinion Pools: An Empirical Comparison

Yiling Chen
Chao-Hsien Chu
Tracy Mullen
School of Information Sciences & Technology
The Pennsylvania State University
University Park, PA 16802
{ychen|chu|tmullen}@ist.psu.edu

David M. Pennock
Yahoo! Research Labs
74 N. Pasadena Ave, 3rd Floor
Pasadena, CA 91103
pennockd@yahoo-inc.com

ABSTRACT

In this paper, we examine the relative forecast accuracy of information markets versus expert aggregation. We leverage a unique data source of almost 2000 people's subjective probability judgments on 2003 US National Football League games and compare with the "market probabilities" given by two different information markets on exactly the same events. We combine assessments of multiple experts via linear and logarithmic aggregation functions to form pooled predictions. Prices in information markets are used to derive market predictions. Our results show that, at the same time point ahead of the game, information markets provide as accurate predictions as pooled expert assessments. In screening pooled expert predictions, we find that arithmetic average is a robust and efficient pooling function; weighting expert assessments according to their past performance does not improve accuracy of pooled predictions; and logarithmic aggregation functions offer bolder predictions than linear aggregation functions. The results provide insights into the predictive performance of information markets, and the relative merits of selecting among various opinion pooling methods.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*economics*

General Terms

Economics, Performance

Keywords

Information markets, opinion pools, expert opinions, prediction accuracy, forecasting

EC'05, June 5–8, 2005, Vancouver, British Columbia, Canada.

1. INTRODUCTION

Forecasting is a ubiquitous endeavor in human societies. For decades, scientists have been developing and exploring various forecasting methods, which can be roughly divided into statistical and non-statistical approaches. Statistical approaches require not only the existence of enough historical data but also that past data contains valuable information about the future event. When these conditions can not be met, non-statistical approaches that rely on judgmental information about the future event could be better choices. One widely used non-statistical method is to elicit opinions from experts. Since experts are not generally in agreement, many belief aggregation methods have been proposed to combine expert opinions together and form a single prediction. These belief aggregation methods are called *opinion pools*, which have been extensively studied in statistics [20, 24, 38], and management sciences [8, 9, 30, 31], and applied in many domains such as group decision making [29] and risk analysis [12].

With the fast growth of the Internet, information markets have recently emerged as a promising non-statistical forecasting tool. Information markets (sometimes called prediction markets, idea markets, or event markets) are markets designed for aggregating information and making predictions about future events. To form the predictions, information markets tie payoffs of securities to outcomes of events. For example, in an information market to predict the result of a US professional National Football League (NFL) game, say New England vs Carolina, the security pays a certain amount of money per share to its holders if and only if New England wins the game. Otherwise, it pays off nothing. The security price before the game reflects the consensus expectation of market traders about the probability of New England winning the game. Such markets are becoming very popular. The Iowa Electronic Markets (IEM) [2] are real-money futures markets to predict economic and political events such as elections. The Hollywood Stock Exchange (HSX) [3] is a virtual (play-money) exchange for trading securities to forecast future box office proceeds of new movies, and the outcomes of entertainment awards, etc. TradeSports.com [7], a real-money betting exchange registered in Ireland, hosts markets for sports, political, entertainment, and financial events. The Foresight Exchange (FX) [4] allows traders to wager play money on unresolved scientific questions or other claims of public interest, and NewsFutures.com's World News Exchange [1] has

popular sports and financial betting markets, also grounded in a play-money currency.

Despite the popularity of information markets, one of the most important questions to ask is: how accurately can information markets predict? Previous research in general shows that information markets are remarkably accurate. The political election markets at IEM predict the election outcomes better than polls [16, 17, 18, 19]. Prices in HSX and FX have been found to give as accurate or more accurate predictions than judgment of individual experts [33, 34, 37]. However, information markets have not been calibrated against opinion pools, except for Servan-Schreiber et. al [36], in which the authors compare two information markets against arithmetic average of expert opinions. Since information markets, in nature, offer an adaptive and self-organized mechanism to aggregate opinions of market participants, it is interesting to compare them with existing opinion pooling methods, to evaluate the performance of information markets from another perspective. The comparison will provide beneficial guidance for practitioners to choose the most appropriate method for their needs.

This paper contributes to the literature in two ways: (1) As an initial attempt to compare information markets with opinion pools of multiple experts, it leads to a better understanding of information markets and their promise as an alternative institution for obtaining accurate forecasts; (2) In screening opinion pools to be used in the comparison, we cast insights into relative performances of different opinion pools. In terms of prediction accuracy, we compare two information markets with several linear and logarithmic opinion pools (LinOP and LogOP) at predicting the results of NFL games. Our results show that at the same time point ahead of the game, information markets provide as accurate predictions as our carefully selected opinion pools. In selecting the opinion pools to be used in our comparison, we find that arithmetic average is a robust and efficient pooling function; weighting expert assessments according to their past performances does not improve the prediction accuracy of opinion pools; and LogOP offers bolder predictions than LinOP. The remainder of the paper is organized as follows. Section 2 reviews popular opinion pooling methods. Section 3 introduces the basics of information markets. Data sets and our analysis methods are described in Section 4. We present results and analysis in Section 5, followed by conclusions in Section 6.

2. REVIEW OF OPINION POOLS

Clemen and Winkler [12] classify opinion pooling methods into two broad categories: mathematical approaches and behavioral approaches. In mathematical approaches, the opinions of individual experts are expressed as subjective probability distributions over outcomes of an uncertain event. They are combined through various mathematical methods to form an aggregated probability distribution. Genest and Zidek [24] and French [20] provide comprehensive reviews of mathematical approaches. Mathematical approaches can be further distinguished into axiomatic approaches and Bayesian approaches. Axiomatic approaches apply prespecified functions that map expert opinions, expressed as a set of individual probability distributions, to a single aggregated probability distribution. These pooling functions are justified using axioms or certain desirable properties. Two of the most common pooling functions are the *linear opinion pool* (LinOP) and the *logarithmic opinion pool* (LogOP). Using LinOP, the aggregate probability distribution is a weighted arithmetic mean of individual probability distributions:

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta), \tag{1}$$

where $p_i(\theta)$ is expert i's probability distribution of uncertain event θ, $p(\theta)$ represents the aggregate probability distribution, w_i's are weights for experts, which are usually non-negative and sum to 1, and n is the number of experts. Using LogOP, the aggregate probability distribution is a weighted geometric mean of individual probability distributions:

$$p(\theta) = k \prod_{i=1}^{n} p_i(\theta)^{w_i}, \tag{2}$$

where k is a normalization constant to ensure that the pooled opinion is a probability distribution. Other axiomatic pooling methods often are extensions of LinOP [22], LogOP [23], or both [13]. Winkler [39] and Morris [29, 30] establish the early framework of Bayesian aggregation methods. Bayesian approaches assume as if there is a decision maker who has a prior probability distribution over event θ and a likelihood function over expert opinions given the event. This decision maker takes expert opinions as evidence and updates its priors over the event and opinions according to Bayes rule. The resulted posterior probability distribution of θ is the pooled opinion.

Behavioral approaches have been widely studied in the field of group decision making and organizational behavior. The important assumption of behavioral approaches is that, through exchanging opinions or information, experts can eventually reach an equilibrium where further interaction won't change their opinions. One of the best known behavioral approaches is the Delphi technique [28]. Typically, this method and its variants do not allow open discussion, but each expert has chance to judge opinions of other experts, and is given feedback. Experts then can reassess their opinions and repeat the process until a consensus or a smaller spread of opinions is achieved. Some other behavioral methods, such as the Nominal Group technique [14], promote open discussions in controlled environments.

Each approach has its pros and cons. Axiomatic approaches are easy to use. But they don't have a normative basis to choose weights. In addition, several impossibility results (e.g., Genest [21]) show that no aggregation function can satisfy all desired properties of an opinion pool, unless the pooled opinion degenerates to a single individual opinion, which effectively implies a dictator. Bayesian approaches are nicely based on the normative Bayesian framework. However, it is sometimes frustratingly difficult to apply because it requires either (1) constructing an obscenely complex joint prior over the event and opinions (often impractical even in terms of storage / space complexity, not to mention from an elicitation standpoint) or (2) making strong assumptions about the prior, like conditional independence of experts. Behavior approaches allow experts to dynamically improve their information and revise their opinions during interactions, but many of them are not fixed or completely specified, and can't guarantee convergence or repeatability.

3. HOW INFORMATION MARKETS WORK

Much of the enthusiasm for information markets stems from Hayek hypothesis [26] and efficient market hypothesis [15]. Hayek, in his classic critique of central planning in 1940's, claims that the price system in a competitive market is a very efficient mechanism to aggregate dispersed information among market participants. The efficient market hypothesis further states that, in an efficient market, the price of a security almost instantly incorporates all available information. The market price summarizes all relevant information across traders, hence is the market participants' consensus expectation about the future value of the security. Empirical evidence supports both hypotheses to a large extent [25, 27, 35]. Thus, when associating the value of a security with the outcome of an uncertain future event, market price, by revealing the consensus expectation of the security value, can indirectly predict the outcome of the event. This idea gives rise to information markets.

For example, if we want to predict which team will win the NFL game between New England and Carolina, an information market can trade a security "$100 if New England defeats Carolina", whose payoff per share at the end of the game is specified as follow:

$$\begin{cases} \$100 & \text{if New England wins the game;} \\ \$0 & \text{otherwise.} \end{cases}$$

The security price should roughly equal the expected payoff of the security in an efficient market. The time value of money usually can be ignored because durations of most information markets are short. Assuming exposure to risk is roughly equal for both outcomes, or that there are sufficient effectively risk-neutral speculators in the market, the price should not be biased by the risk attitudes of various players in the market. Thus,

$$p = \Pr(\textit{Patriots win}) \times 100 + [1 - \Pr(\textit{Patriots win})] \times 0,$$

where p is the price of the security "$100 if New England defeats Carolina" and $\Pr(\textit{Patriots win})$ is the probability that New England will win the game. Observing the security price p before the game, we can derive $\Pr(\textit{Patriots win})$, which is the market participants' collective prediction about how likely it is that New England will win the game.

The above security is a *winner-takes-all* contract. It is used when the event to be predicted is a discrete random variable with disjoint outcomes (in this case binary). Its price predicts the probability that a specific outcome will be realized. When the outcome of a prediction problem can be any value in a continuous interval, we can design a security that pays its holder proportional to the realized value. This kind of security is what Wolfers and Zitzewitz [40] called an *index* contract. It predicts the expected value of a future outcome. Many other aspects of a future event such as median value of outcome can also be predicted in information markets by designing and trading different securities. Wolfers and Zitzewitz [40] provide a summary of the main types of securities traded in information markets and what statistical properties they can predict. In practice, conceiving a security for a prediction problem is only one of the many decisions in designing an effective information market. Spann and Skiera [37] propose an initial framework for designing information markets.

4. DESIGN OF ANALYSIS

4.1 Data Sets

Our data sets cover 210 NFL games held between September 28th, 2003 and December 28th, 2003. NFL games are very suitable for our purposes because: (1) two online exchanges and one online prediction contest already exist that provide data on both information markets and the opinions of self-identified experts for the same set of games; (2) the popularity of NFL games in the United States provides natural incentives for people to participate in information markets and/or the contest, which increases liquidity of information markets and improves the quality and number of opinions in the contest; (3) intense media coverage and analysis of the profiles and strengths of teams and individual players provide the public with much information so that participants of information markets and the contest can be viewed as knowledgeable regarding to the forecasting goal.

Information market data was acquired, by using a specially designed crawler program, from TradeSports.com's Football-NFL markets [7] and NewsFutures.com's Sports Exchange [1]. For each NFL game, both TradeSports and NewsFutures have a winner-takes-all information market to predict the game outcome. We introduce the design of the two markets according to Spann and Skiera's three steps for designing an information market [37] as below.

- **Choice of forecasting goal:** Markets at both TradeSports and NewsFutures aim at predicting which one of the two teams will win a NFL football game. They trade similar winner-takes-all securities that pay off 100 if a team wins the game and 0 if it loses the game. Small differences exist in how they deal with ties. In the case of a tie, TradeSports will unwind all trades that occurred and refund all exchange fees, but the security is worth 50 in NewsFutures. Since the probability of a tie is usually very low (much less the 1%), prices at both markets effectively represent the market participants' consensus assessment of the probability that the team will win.

- **Incentive for participation and information revelation:** TradeSports and NewsFutures use different incentives for participation and information revelation. TradeSports is a real-money exchange. A trader needs to open and fund an account with a minimum of $100 to participate in the market. Both profits and losses can occur as a result of trading activity. On the contrary, a trader can register at NewsFutures for free and get 2000 units of Sport Exchange virtual money at the time of registration. Traders at NewsFutures will not incur any real financial loss. They can accumulate virtual money by trading securities. The virtual money can then be used to bid for a few real prizes at NewsFutures' online shop.

- **Financial market design:** Both markets at TradeSports and NewsFutures use the continuous double auction as their trading mechanism. TradeSports charges a small fee on each security transaction and expiry, while NewsFutures does not.

We can see that the main difference between two information markets is real money vs. virtual money. Servan-Schreiber

et. al [36] have compared the effect of money on the performance of the two information markets and concluded that the prediction accuracy of the two markets are at about the same level. Not intending to compare these two markets, we still use both markets in our analysis to ensure that our findings are not accidental.

We obtain the opinions of 1966 self-identified experts for NFL games from the ProbabilityFootball online contest [5], one of several ProbabilitySports contests [6]. The contest is free to enter. Participants of the contest are asked to enter their subjective probability that a team will win a game by noon on the day of the game. Importantly, the contest evaluates the participants' performance via the quadratic scoring rule:

$$s = 100 - 400 \times Prob\_Lose^2, \quad (3)$$

where s represents the score that a participant earns for the game, and $Prob\_Lose$ is the probability that the participant assigns to the actual losing team. The quadratic score is one of a family of so-called *proper* scoring rules that have the property that an expert's expected score is maximized when the expert reports probabilities truthfully. For example, for a game team A vs. team B, if a player assigns 0.5 to both team A and B, his/her score for the game is 0 no matter which team wins. If he/she assigns 0.8 to team A and 0.2 to team B, showing that he is confident in team A's winning, he/she will score 84 points for the game if team A wins, and lose 156 points if team B wins. This quadratic scoring rule rewards bold predictions that are right, but penalizes bold predictions that turn out to be wrong. The top players, measured by accumulated scores over all games, win the prizes of the contest. The suggested strategy at the contest website is "to make picks for each game that match, as closely as possible, the probabilities that each team will win". This strategy is correct if the participant seeks to maximize expected score. However, as prizes are awarded only to the top few winners, participants' goals are to maximize the probability of winning, not maximize expected score, resulting in a slightly different and more risk-seeking optimization.[1] Still, as far as we are aware, this data offer the closest thing available to true subjective probability judgments from so many people over so many public events that have corresponding information markets.

4.2 Methods of Analysis

In order to compare the prediction accuracy of information markets and that of opinion pools, we proceed to derive predictions from market data of TradeSports and NewsFutures, form pooled opinions using expert data from ProbabilityFootball contest, and specify the performance measures to be used.

4.2.1 Deriving Predictions

For information markets, deriving predictions is straightforward. We can take the security price and divide it by 100 to get the market's prediction of the probability that a team will win. To match the time when participants at the ProbabilityFootball contest are required to report their probability assessments, we derive predictions using the last trade price before noon on the day of the game. For more than half of the games, this time is only about an hour earlier than the game starting time, while it is several hours earlier for other games. Two sets of market predictions are derived:

- NF: Prediction equals NewsFutures' last trade price before noon of the game day divided by 100.
- TS: Prediction equals TradeSports' last trade price before noon of the game day divided by 100.

We apply LinOP and LogOP to ProbabilityFootball data to obtain aggregate expert predictions. The reason that we do not consider other aggregation methods include: (1) data from ProbabilityFootball is only suitable for mathematical pooling methods—we can rule out behavioral approaches, (2) Bayesian aggregation requires us to make assumptions about the prior probability distribution of game outcomes and the likelihood function of expert opinions: given the large number of games and participants, making reasonable assumptions is difficult, and (3) for axiomatic approaches, previous research has shown that simpler aggregation methods often perform better than more complex methods [12]. Because the output of LogOP is indeterminate if there are probability assessments of both 0 and 1 (and because assessments of 0 and 1 are dictatorial using LogOP), we add a small number 0.01 to an expert opinion if it is 0, and subtract 0.01 from it if it is 1.

In pooling opinions, we consider two influencing factors: weights of experts and number of expert opinions to be pooled. For weights of experts, we experiment with equal weights and performance-based weights. The performance-based weights are determined according to previous accumulated score in the contest. The score for each game is calculated according to equation 3, the scoring rule used in the ProbabilityFootball contest. For the first week, since no previous scores are available, we choose equal weights. For later weeks, we calculate accumulated past scores for each player. Because the cumulative scores can be negative, we shift everyone's score if needed to ensure the weights are non-negative. Thus,

$$w_i = \frac{cumulative\_score_i + shift}{\sum_{j=1}^{n}(cumulative\_score_j + shift)}. \quad (4)$$

where $shift$ equals 0 if the smallest $cumulative\_score_j$ is non-negative, and equals the absolute value of the smallest $cumulative\_score_j$ otherwise. For simplicity, we call performance-weighted opinion pool as weighted, and equally weighted opinion pool as unweighted. We will use them interchangeably in the remaining of the paper.

As for the number of opinions used in an opinion pool, we form different opinion pools with different number of experts. Only the best performing experts are selected. For example, to form an opinion pool with 20 expert opinions, we choose the top 20 participants. Since there is no performance record for the first week, we use opinions of all participants in the first week. For week 2, we select opinions of 20 individuals whose scores in the first week are among the top 20. For week 3, 20 individuals whose cumulative scores of week 1 and 2 are among the top 20s are selected. Experts are chosen in a similar way for later weeks. Thus, the top 20 participants can change from week to week.

The possible opinion pools, varied in pooling functions, weighting methods, and number of expert opinions, are shown

[1] Ideally, prizes would be awarded by lottery in proportion to accumulated score.

Table 1: Pooled Expert Predictions

#	Symbol	Description
1	Lin-All-u	Unweighted (equally weighted) LinOP of all experts.
2	Lin-All-w	Weighted (performance-weighted) LinOP of all experts.
3	Lin-n-u	Unweighted (equally weighted) LinOP with n experts.
4	Lin-n-w	Weighted (performance-weighted) LinOP with n experts.
5	Log-All-u	Unweighted (equally weighted) LogOP of all experts.
6	Log-All-w	Weighted (performance-weighted) LogOP of all experts.
7	Log-n-u	Unweighted (equally weighted) LogOP with n experts.
8	Log-n-w	Weighted (performance-weighted) LogOP with n experts.

in Table 1. "Lin" represents linear, and "Log" represents Logarithmic. "n" is the number of expert opinions that are pooled, and "All" indicates that all opinions are combined. We use "u" to symbolize unweighted (equally weighted) opinion pools. "w" is used for weighted (performance-weighted) opinion pools. Lin-All-u, the equally weighted LinOP with all participants, is basically the arithmetic mean of all participants' opinions. Log-All-u is simply the geometric mean of all opinions.

When a participant did not enter a prediction for a particular game, that participant was removed from the opinion pool for that game. This contrasts with the "ProbabilityFootball average" reported on the contest website and used by Servan-Schreiber et. al [36], where unreported predictions were converted to 0.5 probability predictions.

4.2.2 Performance Measures

We use three common metrics to assess prediction accuracy of information markets and opinion pools. These measures have been used by Servan-Schreiber et. al [36] in evaluating the prediction accuracy of information markets.

1. $Absolute\_Error = Prob\_Lose$,

   where $Prob\_Lose$ is the probability assigned to the eventual losing team. Absolute error simply measures the difference between a perfect prediction (1 for winning team) and the actual prediction. A prediction with lower absolute error is more accurate.

2. $Quadratic\_Score = 100 - 400 \times (Prob\_Lose^2)$.

   Quadratic score is the scoring function that is used in the ProbabilityFootball contest. It is a linear transformation of squared error, $Prob\_Lose^2$, which is one of the mostly used metrics in evaluating forecasting accuracy. Quadratic score can be negative. A prediction with higher quadratic score is more accurate.

3. $Logarithmic\_Score = log(Prob\_Win)$,

   where $Prob\_Win$ is the probability assigned to the eventual winning team. The logarithmic score, like the quadratic score, is a proper scoring rule. A prediction with higher (less negative) logarithmic score is more accurate.

5. EMPIRICAL RESULTS

5.1 Performance of Opinion Pools

Depending on how many opinions are used, there can be numerous different opinion pools. We first examine the effect of number of opinions on prediction accuracy by forming opinion pools with the number of expert opinions varying from 1 to 960. In the ProbabilityFootball Competition, not all 1966 registered participants provide their probability assessments for every game. 960 is the smallest number of participants for all games. For each game, we sort experts according to their accumulated quadratic score in previous weeks. Predictions of the best performing n participants are picked to form an opinion pool with n experts.

Figure 1 shows the prediction accuracy of LinOP and LogOP in terms of mean values of the three performance measures across all 210 games. We can see the following trends in the figure.

1. Unweighted opinion pools and performance-weighted opinion pools have similar levels of prediction accuracy, especially for LinOP.

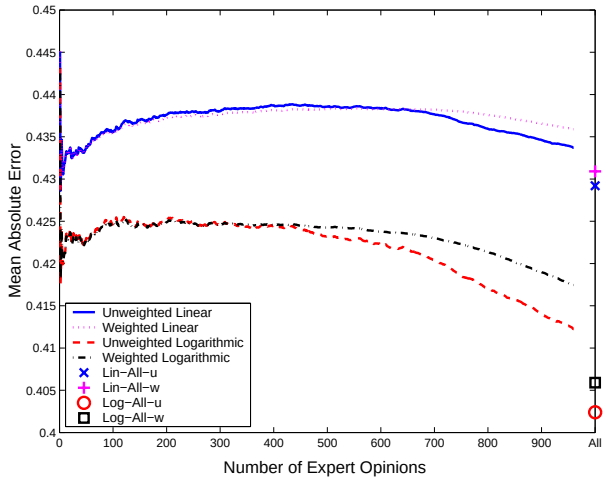
2. For LinOP, increasing the number of experts in general increases or keeps the same the level of prediction accuracy. When there are more than 200 experts, the prediction accuracy of LinOP is stable regarding the number of experts.

3. LogOP seems more accurate than LinOP in terms of mean absolute error. But, using all other performance measures, LinOP outperforms LogOP.

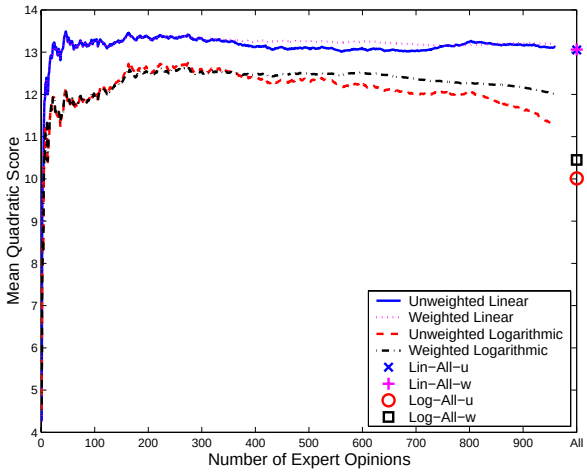
4. For LogOP, increasing the number of experts increases the prediction accuracy at the beginning. But the curves (including the points with all experts) for mean quadratic score, and mean logarithmic score have slight bell-shapes, which represent a decrease in prediction accuracy when the number of experts is very large. The curves for mean absolute error, on the other hand, show a consistent increase of accuracy.

The first and second trend above imply that when using LinOP, the simplest way, which has good prediction accuracy, is to average the opinions of all experts. Weighting does not seem to improve performance. Selecting experts according to past performance also does not help. It is a very interesting observation that even if many participants of the ProbabilityFootball contest do not provide accurate individual predictions (they have negative quadratic scores in the contest), including their opinions into the opinion pool still increases the prediction accuracy. One explanation of this phenomena could be that biases of individual judgment can offset with each other when opinions are diverse, which makes the pooled prediction more accurate.

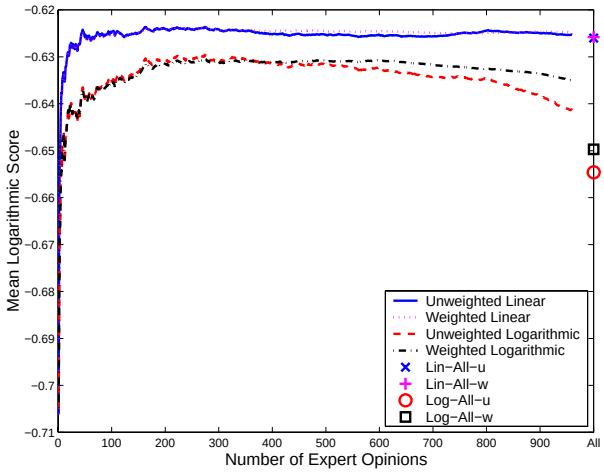
The third trend presents a controversy. The relative prediction accuracy of LogOP and LinOP flips when using different accuracy measures. To investigate this disagreement, we plot the absolute error of Log-All-u and Lin-All-u for each game in Figure 2. When the absolute error of an opinion



(a) Mean Absolute Error



(b) Mean Quadratic Score



(c) Mean Logarithmic Score

Figure 1: Prediction Accuracy of Opinion Pools

pool for a game is less than 0.5, it means that the team favored by the opinion pool wins the game. If it is greater than 0.5, the underdog wins. Compared with Lin-All-u, Log-All-u has lower absolute error when it is less than 0.5, and greater absoluter error when it is greater than 0.5, which indicates that predictions of Log-All-u are bolder, more close to 0 or 1, than those of Lin-All-u. This is due to the nature of linear and logarithmic aggregating functions. Because quadratic score and logarithmic score penalize bold predictions that are wrong, LogOP is less accurate when measured in these terms.

Similar reasoning accounts for the fourth trend. When there are more than 500 experts, increasing number of experts used in LogOP improves the prediction accuracy measured by absolute error, but worsens the accuracy measured by the other two metrics. Examining expert opinions, we find that participants who rank lower are more frequent in offering extreme predictions (0 or 1) than those ranking high in the list. When we increase the number of experts in an opinion pool, we are incorporating more extreme predictions into it. The resulting LogOP is bolder, and hence has lower mean quadratic score and mean logarithmic score.

5.2 Comparison of Information Markets and Opinion Pools

Through the first screening of various opinion pools, we select Lin-All-u, Log-All-u, Log-All-w, and Log-200-u to compare with predictions from information markets. Lin-All-u as shown in Figure 1 can represent what LinOP can achieve. However, the performance of LogOP is not consistent when evaluated using different metrics. Log-All-u and Log-All-w offer either the best or the worst predictions. Log-200-u, the LogOP with the 200 top performing experts, provides more stable predictions. We use all of the three to stand for the performance of LogOP in our later comparison.

If a prediction of the probability that a team will win a game, either from an opinion pool or an information market, is higher than 0.5, we say that the team is the predicted favorite for the game. Table 2 presents the number and percentage of games that predicted favorites actually win, out of a total of 210 games. All four opinion pools correctly predict a similar number and percentage of games as NF and TS. Since NF, TS, and the four opinion pools form their predictions using information available at noon of the game

Table 2: Number and Percentage of Games that Predicted Favorites Win

	NF	TS	Lin-All-u	Log-All-u	Log-All-w	Log-200-u
Number	142	137	144	144	143	141
Percentage	67.62%	65.24%	68.57%	68.57%	68.10%	67.14%

Table 3: Mean of Prediction Accuracy Measures

	Absolute Error	Quadratic Score	Logarithmic Score
NF	0.4253	**15.4352**	-0.6136
	(0.0121)	(4.6072)	(0.0258)
TS	0.4275	15.2739	**-0.6121**
	(0.0118)	(4.3982)	(0.0241)
Lin-All-u	0.4292	13.0525	-0.6260
	(0.0126)	(4.8088)	(0.0268)
Log-All-u	**0.4024**	10.0099	-0.6546
	(0.0173)	(6.6594)	(0.0418)
Log-All-w	0.4059	10.4491	-0.6497
	(0.0168)	(6.4440)	(0.0398)
Log-200-u	0.4266	12.3868	-0.6319
	(0.0133)	(5.0764)	(0.0295)

*Numbers in parentheses are standard errors.
*Best value for each metric is shown in **bold.**

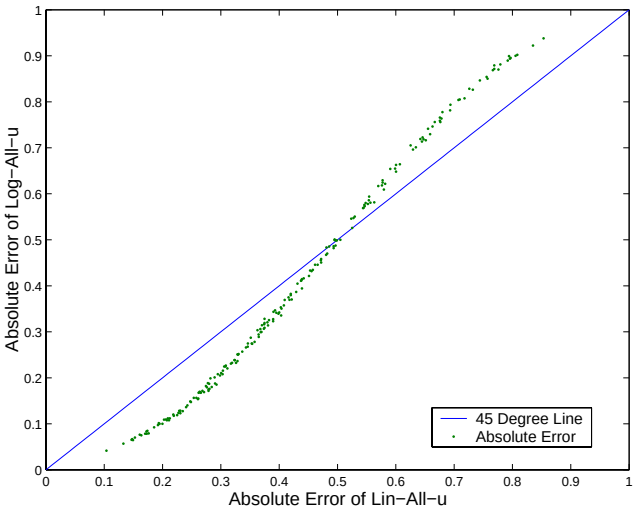


Figure 2: Absolute Error: Lin-All-u vs. Log-All-u

day, information markets and opinion pools have comparable potential at the same time point.

We then take a closer look at prediction accuracy of information markets and opinion pools using the three performance measures. Table 3 displays mean values of these measures over 210 games. Numbers in parentheses are standard errors, which estimate the standard deviation of the mean. To take into consideration of skewness of distributions, we also report median values of accuracy measures in Table 4. Judged by the mean values of accuracy measures in Table 3, all methods have similar accuracy levels, with NF and TS slightly better than the opinion pools. However, the median values of accuracy measures indicate that Log-All-u and Log-All-w opinion pools are more accurate than all other predictions.

We employ the *randomization test* [32] to study whether the differences in prediction accuracy presented in Table 3 and Table 4 are statistically significant. The basic idea of randomization test is that, by randomly swapping predictions of two methods numerous times, an empirical distribution for the difference of prediction accuracy can be constructed. Using this empirical distribution, we are then able to evaluate that at what confidence level the observed difference reflects a real difference. For example, the mean absolute error of NF is higher than that of Log-All-u by 0.0229, as shown in Table 3. To test whether this difference is statistically significant, we shuffle predictions from two methods, randomly label half of predictions as NF and the other half as Log-All-u, and compute the difference of mean absolute error of the newly formed NF and Log-All-u data. The above procedure is repeated 10,000 times. The 10,000 differences of mean absolute error results in an empirical distribution of the difference. Comparing our observed difference, 0.0229, with this distribution, we find that the observed difference is greater than 75.37% of the empirical differences. This leads us to conclude that the difference of mean absolute error between NF and Log-All-u is not statistically significant, if we choose the level of significance to be 0.05.

Table 5 and Table 6 are results of randomization test for mean and median differences respectively. Each cell of the table is for two different prediction methods, represented by name of the row and name of the column. The first lines of table cells are results for absolute error. The second and third lines are dedicated to quadratic score and logarithmic score respectively. We can see that, in terms of mean values of accuracy measures, the differences of all methods are not statistically significant to any reasonable degree. When it

Table 4: Median of Prediction Accuracy Measures

	Absolute Error	Quadratic Score	Logarithmic Score
NF	0.3800	42.2400	-0.4780
TS	0.4000	36.0000	-0.5108
Lin-All-u	0.3639	36.9755	-0.5057
Log-All-u	**0.3417**	**53.2894**	**-0.4181**
Log-All-w	0.3498	51.0486	-0.4305
Log-200-u	0.3996	36.1300	-0.5101

*Best value for each metric is shown in **bold.**

Table 5: Statistical Confidence of Mean Differences in Prediction Accuracy

	TS	Lin-All-u	Log-All-u	Log-All-w	Log-200-u
	8.92%	22.07%	75.37%	66.47%	7.76%
NF	2.38%	26.60%	50.74%	44.26%	32.24%
	2.99%	22.81%	59.35%	56.21%	33.26%
		10.13%	77.79%	68.15%	4.35%
TS		27.25%	53.65%	44.90%	28.30%
		32.35%	57.89%	60.69%	38.84%
			82.19%	68.86%	9.75%
Lin-All-u			28.91%	23.92%	6.81%
			44.17%	43.01%	17.36%
				11.14%	72.49%
Log-All-u				3.32%	18.89%
				5.25%	39.06%
					69.89%
Log-All-w					18.30%
					30.23%

*In each table cell, row 1 accounts for absolute error, row 2 for quadratic score, and row 3 for logarithmic score.

comes to median values of prediction accuracy, Log-All-u outperforms Lin-All-u at a high confidence level.

These results indicate that differences in prediction accuracy between information markets and opinion pools are not statistically significant. This may seem to contradict the result of Servan-Schreiber et. al [36], in which NewsFutures's information markets have been shown to provide statistically significantly more accurate predictions than the (unweighted) average of all ProbabilityFootball opinions. The discrepancy emerges in dealing with missing data. Not all 1966 registered ProbabilityFootball participants offer probability assessments for each game. When a participant does not provide a probability assessment for a game, the contest considers their prediction as 0.5.. This makes sense in the context of the contest, since 0.5 always yields 0 quadratic score. The ProbabilityFootball average reported on the contest website and used by Servan-Schreiber et. al includes these 0.5 estimates. Instead, we remove participants from games that they do not provide assessments, pooling only the available opinions together. Our treatment increases the prediction accuracy of Lin-All-u significantly.

6. CONCLUSIONS

With the fast growth of the Internet, information markets have recently emerged as an alternative tool for predicting future events. Previous research has shown that information markets give as accurate or more accurate predictions than individual experts and polls. However, information markets, as an adaptive mechanism to aggregate different opinions of market participants, have not been calibrated against many belief aggregation methods. In this paper, we compare prediction accuracy of information markets with linear and logarithmic opinion pools (LinOP and LogOP) using predictions from two markets and 1966 individuals regarding the outcomes of 210 American football games during the 2003 NFL season. In screening for representative opinion pools to compare with information markets, we investigate the effect of weights and number of experts on prediction accuracy. Our results on both the comparison of information markets and opinion pools and the relative performance of different opinion pools are summarized as below.

1. At the same time point ahead of the events, information markets offer as accurate predictions as our selected opinion pools.

   We have selected four opinion pools to represent the prediction accuracy level that LinOP and LogOP can achieve. With all four performance metrics, our two information markets obtain similar prediction accuracy as the four opinion pools.

Table 6: Statistical Confidence of Median Differences in Prediction Accuracy

	TS	Lin-All-u	Log-All-u	Log-All-w	Log-200-u
	48.85%	47.3%	84.8%	77.9%	65.36%
NF	45.26%	44.55%	85.27%	75.65%	66.75%
	44.89%	46.04%	84.43%	77.16%	64.78%
		5.18%	94.83%	94.31%	0%
TS		5.37%	92.08%	92.53%	0%
		7.41%	**95.62%**	91.09%	0%
			95.11%	91.37%	7.31%
Lin-All-u			**96.10%**	92.69%	9.84%
			95.45%	**95.12%**	7.79%
				23.47%	**95.89%**
Log-All-u				26.68%	93.85%
				22.47%	**96.42%**
					91.3%
Log-All-w					91.4%
					90.37%

*In each table cell, row 1 accounts for absolute error, row 2 for quadratic score, and row 3 for logarithmic score.
*Confidence above 95% is shown in **bold.**

2. The arithmetic average of all opinions (Lin-All-u) is a simple, robust, and efficient opinion pool.

   Simply averaging across all experts seems resulting in better predictions than individual opinions and opinion pools with a few experts. It is quite robust in the sense that even if the included individual predictions are less accurate, averaging over all opinions still gives better (or equally good) predictions.

3. Weighting expert opinions according to past performance does not seem to significantly improve prediction accuracy of either LinOP or LogOP.

   Comparing performance-weighted opinion pools with equally weighted opinion pools, we do not observe much difference in terms of prediction accuracy. Since we only use one performance-weighting method, calculating the weights according to past accumulated quadratic score that participants earned, this might due to the weighting method we chose.

4. LogOP yields bolder predictions than LinOP.

   LogOP yields predictions that are closer to the extremes, 0 or 1.

An information markets is a self-organizing mechanism for aggregating information and making predictions. Compared with opinion pools, it is less constrained by space and time, and can eliminate the efforts to identify experts and decide belief aggregation methods. But the advantages do not compromise their prediction accuracy to any extent. On the contrary, information markets can provide real-time predictions, which are hardly achievable through resorting to experts. In the future, we are interested in further exploring:

- Performance comparison of information markets with other opinion pools and mathematical aggregation procedures.

  In this paper, we only compare information markets with two simple opinion pools, linear and logarithmic. It will be meaningful to investigate their relative prediction accuracy with other belief aggregation methods such as Bayesian approaches. There are also a number of theoretical expert algorithms with proven worst-case performance bounds [10] whose average-case or practical performance would be instructive to investigate.

- Whether defining expertise more narrowly can improve predictions of opinion pools.

  In our analysis, we broadly treat participants of the ProbabilityFootball contest as experts in all games. If we define expertise more narrowly, selecting experts in certain football teams to predict games involving these teams, will the predictions of opinion pools be more accurate?

- The possibility of combining information markets with other forecasting methods to achieve better prediction accuracy.

  Chen, Fine, and Huberman [11] use an information market to determine the risk attitude of participants, and then perform a nonlinear aggregation of their predictions based on their risk attitudes. The nonlinear aggregation mechanism is shown to outperform both the market and the best individual participants. It is worthy of more attention whether information markets, as an alternative forecasting method, can be used together with other methods to improve our predictions.

7. ACKNOWLEDGMENTS

We thank Brian Galebach, the owner and operator of the ProbabilitySports and ProbabilityFootball websites, for providing us with such unique and valuable data. We thank Varsha Dani, Lance Fortnow, Omid Madani, Sumit Sang-

hai, and the anonymous reviewers for useful insights and pointers.

The authors acknowledge the support of The Penn State eBusiness Research Center.

8. REFERENCES

[1] http://us.newsfutures.com
[2] http://www.biz.uiowa.edu/iem/
[3] http://www.hsx.com/
[4] http://www.ideosphere.com/fx/
[5] http://www.probabilityfootball.com/
[6] http://www.probabilitysports.com/
[7] http://www.tradesports.com/
[8] A. H. Ashton and R. H. Ashton. Aggregating subjective forecasts: Some empirical results. *Management Science*, 31:1499–1508, 1985.
[9] R. P. Batchelor and P. Dua. Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41:68–75, 1995.
[10] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
[11] K. Chen, L. Fine, and B. Huberman. Predicting the future. *Information System Frontier*, 5(1):47–61, 2003.
[12] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203, 1999.
[13] R. M. Cook. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York, 1991.
[14] A. L. Delbecq, A. H. Van de Ven, and D. H. Gustafson. *Group Techniques for Program Planners: A Guide to Nominal Group and Delphi Processes*. Scott Foresman and Company, Glenview, IL, 1975.
[15] E. F. Fama. Efficient capital market: A review of theory and empirical work. *Journal of Finance*, 25:383–417, 1970.
[16] R. Forsythe and F. Lundholm. Information aggregation in an experimental market. *Econometrica*, 58:309–47, 1990.
[17] R. Forsythe, F. Nelson, G. R. Neumann, and J. Wright. Forecasting elections: A market alternative to polls. In T. R. Palfrey, editor, *Contemporary Laboratory Experiments in Political Economy*, pages 69–111. University of Michigan Press, Ann Arbor, MI, 1991.
[18] R. Forsythe, F. Nelson, G. R. Neumann, and J. Wright. Anatomy of an experimental political stock market. *American Economic Review*, 82(5):1142–1161, 1992.
[19] 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 and Organization*, 39:83–110, 1999.
[20] S. French. Group consensus probability distributions: a critical survey. *Bayesian Statistics*, 2:183–202, 1985.
[21] C. Genest. A conflict between two axioms for combining subjective distributions. *Journal of the Royal Statistical Society*, 46(3):403–405, 1984.
[22] C. Genest. Pooling operators with the marginalization property. *Canadian Journal of Statistics*, 12(2):153–163, 1984.
[23] C. Genest, K. J. McConway, and M. J. Schervish. Characterization of externally Bayesian pooling operators. *Annals of Statistics*, 14(2):487–501, 1986.
[24] C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.
[25] S. J. Grossman. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies*, 48(4):541–559, 1981.
[26] F. A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.
[27] J. C. Jackwerth and M. Rubinstein. Recovering probability distribution from options prices. *Journal of Finance*, 51(5):1611–1631, 1996.
[28] H. A. Linstone and M. Turoff. *The Delphi Method: Techniques and Applications*. Addison-Wesley, Reading, MA, 1975.
[29] P. A. Morris. Decision analysis expert use. *Management Science*, 20(9):1233–1241, 1974.
[30] P. A. Morris. Combining expert judgments: A bayesian approach. *Management Science*, 23(7):679–693, 1977.
[31] P. A. Morris. An axiomatic approach to expert resolution. *Management Science*, 29(1):24–32, 1983.
[32] E. W. Noreen. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley and Sons, Inc., New York, 1989.
[33] D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. The real power of artificial markets. *Science*, 291:987–988, February 2002.
[34] D. M. Pennock, S. Lawrence, F. A. Nielsen, and C. L. Giles. Extracting collective probabilistic forecasts from web games. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 174–183, San Francisco, CA, 2001.
[35] C. Plott and S. Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56:1085–118, 1988.
[36] E. Servan-Schreiber, J. Wolfers, D. M. Pennock, and B. Galebach. Prediction markets: Does money matter? *Electronic Markets*, 14(3):243–251, 2004.
[37] M. Spann and B. Skiera. Internet-based virtual stock markets for business forecasting. *Management Science*, 49(10):1310–1326, 2003.
[38] M. West. Bayesian aggregation. *Journal of the Royal Statistical Society. Series A. General*, 147(4):600–607, 1984.
[39] R. L. Winkler. The consensus of subjective probability distributions. *Management Science*, 15(2):B61–B75, 1968.
[40] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.