An Experimental Test of Combinatorial Information Markets

John Ledyard
California Institute of Technology
Robin Hanson*
George Mason University†
Takashi Ishikida
Net Exchange

December 2007
First Draft, February 2005

Abstract

While a simple information market lets one trade on the probability of each value of a single variable, a full combinatorial information market lets one trade on any combination of values of a set of variables, including any conditional or joint probability. In laboratory experiments, we compare the accuracy of simple markets, two kinds of combinatorial markets, a call market and a market maker, isolated individuals who report to a scoring rule, and two ways to combine those individual reports into a group prediction. We consider two environments with asymmetric information on sparsely correlated binary variables, one with three subjects and three variables, and the other with six subjects and eight variables (and so 256 states).

*For their comments, we thank David Porter, Ryan Oprea, and participants of seminars at George Mason University, Brigham Young University, University of Michigan, James Madison University, University of California at Los Angeles, Overture Research, Microsoft Research, and the DIMACS Workshop on Markets as Predictive Devices. We thank DARPA for financial support.

† *rhanson@gmu.edu* http://hanson.gmu.edu 703-993-2326 FAX: 703-993-2323 MSN 1D3, Carow Hall, Fairfax VA 22030

1 Introduction

Economists have long noticed that speculative markets, though created for other purposes, also aggregate relevant information. In fact, it is hard to find information that such market prices do not embody, because those who find neglected information can profit by trading on it, and thereby correcting this neglect (Lo, 2000). Recently, new markets have been created specifically to take advantage of this effect. Called information markets, prediction markets (Wolfers & Zitzewitz, 2004), virtual stock markets (Spann & Skiera, 2003), artificial markets (Pennock, Giles, & Nielsen, 2001), or idea futures (Hanson, 1990, 1995), such markets now estimate product sales, project completion dates, and election outcomes.

In every known head-to-head field comparison, information markets have been no less accurate than other social institutions. Orange Juice futures improve on National Weather Service forecasts (Roll, 1984), horse race markets beat horse race experts (Figlewski, 1979), Oscar markets beat columnist forecasts (Pennock et al., 2001), gas demand markets beat gas demand experts (Spencer, 2004), stock markets beat the official NASA panel at fingering the guilty company in the Challenger accident (Maloney & Mulherin, 2003), election markets beat national opinion polls (Berg & Rietz, 2002), and corporate sales markets beat official corporate forecasts (Chen & Plott, 2002).

Of the many obstacles that limit the wider application of such markets, one of the most important is cost. While there are many applications where the benefits of such markets outweigh the current costs, lower costs would enlarge the set of cost-effective applications. One of the main cost drivers today is the fact that standard market mechanisms, such as the simple double auction, require many traders per asset. Because traders do not want to make offers that have little chance of being quickly accepted, a usual rule of thumb is that one needs several times more traders than the number of assets one wants to be traded. And prices should only change a small amount in the time between trades. When there are not enough people to trade often enough, markets are too thin, and most assets will not be traded.

Yet on most topics of interest there are usually hundreds, if not millions, of estimates that one might desire. For example, instead of estimating a company's total sales every quarter, one might want to estimate the sales of each product, in each region, every week. Instead of just estimating expected values, one might want a full probability distribution over each parameter. In addition to estimating the chances of simple events, one might also want the chances of conjunctions and disjunctions of those events. In the limit, one might want a full joint probability distribution over all relevant events.

Amid this vast combinatorial space of estimates of interest, *decision-conditional* forecasts stand out as being of special interest (Hanson, 1999; Berg & Rietz, 2003; Abramowicz, 2004; Hahn & Tetlock, 2004). For example, markets could estimate the sales of a particular product conditional on hiring various particular ad agencies to market that product. Such markets could directly advise particular decisions, by directly estimating the relevant consequences of those decisions.

Similar market thinness problems appear when one wants people to buy and sell combinations of items - there are far too many possible combinations of items to just create a separate market for each combination. These issues are relevant in spectrum auctions, for example, where buyers fundamentally have values for various packages of spectrum locations. To deal with these problems, increasingly powerful designs for combinatorial auctions and markets have been developed over the last two decades (Rassenti, Smith, & Bulfin, 1982; Banks, Ledyard, & Porter, 1989; DeMartini, Kwasnica, Ledyard, & Porter, 1999; Porter, Rassenti, & Smith, 2003). These successes raise our hopes that related solutions can be found to enable combinatorial information markets.

FutureMAP was a DARPA research program which in 2001 began funding two efforts to apply information markets in security contexts. One of those efforts was the Policy Analysis Market (PAM), which focused on estimating geopolitical stability in the Middle East, because this was an ambitious high-value application area, and on developing combinatorial markets for that application, because the PAM team had for many years pioneered combinatorial auction and market technologies (Polk, Hanson, Ledyard, & Ishikida, 2003).

As part of that PAM effort, the authors developed several concepts for combinatorial information markets, and then in 2002 and 2003 ran laboratory experiments comparing the prediction accuracy of two of those new mechanisms and three more traditional mechanisms. These comparisons were made regarding a laboratory information task intended to reflect some of the difficult information features of the intended geopolitical stability application.

This paper reports on the design and results of those experiments.[1] We describe the experimental environments, the tested mechanisms, the overall experimental design, and the experimental results.

2 Test Environments

There are many standard criteria used to choose laboratory environments. For example, laboratory environments are usually designed to require only a few minutes of time from a few people, to be easy to explain to random students, and be describable in a neutral unemotional manner, avoiding suggestions about how subjects should behave.

For experiments on information aggregation, another handy criteria is the ability to exactly calculate rational beliefs given individual information and given the sum of all individual information. This ability allows us to define a mechanism's accuracy as the distance between an ideal distribution and the actual probability distribution produced by the mechanism.

While several distance measures between probability distributions are available, the most commonly used measure is the relative entropy, or Kullback-Leibler (KL) information, distance[2] (Kullback & Leibler, 1951). This measures the added surprise or information (in the entropy sense) of one probability distribution, given that one had expected the other distri-

[1] PAM and all FutureMAP projects were canceled suddenly in July 2003, amid a media storm of criticism, just before PAM was about to begin trading (Hanson, 2006).

[2] The KL distance from p to q is $\sum_i q_i \log(p_i/q_i)$.

bution. For example, it gives the extra expected message length required to communicate a message using a code that was optimal for one distribution of messages, but actually following another message distribution. Unlike most distances between probability distributions, and unlike ordinary entropy, relative entropy is invariant under parameter transformations.

In virtually all laboratory experiments on information done so far, the basic information tasks that subjects have had to perform have been relatively simple; most of the complexity subjects faced was in inferring others' information from their behavior (Plott & Sunder, 1988; Forsythe & Lundholm, 1990; Sunder, 1995). In contrast, we wanted our laboratory testbed environments to model more of what we considered to be the essential information tasks posed by the real PAM environment.

2.1 A More Realistic Information Task

The initial PAM plan was to focus on forecasting military and political instability around the world, how US policies would effect such instability, and how such instability would impact US and global aggregates of interest. The plan later focused on the Mideast, because of the high cost of finding reputable sources to judge after the fact what instability had actually occurred in each nation.

The final PAM plan was to cover eight nations. For each nation in each quarter of a year, we would have traders predict numerical values for its military activity, political instability, economic growth, related US military activity, and US financial involvement. In addition, traders would predict US GDP, world trade, US military casualties, and western terrorist casualties, and a few to-be-determined miscellaneous items. This would require a hundred or so base markets. Most important, we wanted our traders to predict combinations of these events, such as how moving US troops out of Saudi Arabia would effect political stability there, how that would effect stability in Syria, and how all that might change oil prices.

We judged the essence of the PAM information task to be using both theory and data to puzzle out the connections between many weakly related variables. For example, an expert on Syria might know enough to conclude that Syrian military activity tends to follow Syrian economic growth with a time lag, while an expert in regional political relations might conclude that Saudi Arabian political stability tends to follow Iraqi political stability with a time lag.

More abstractly, there should be many variables and different people should be experts in different overlapping sets of variables. While most pairs of variables should not be directly related, a few variable pairs should be more strongly, but imperfectly, related. The challenge for participants would be to combine general knowledge with empirical data on previous variable trends to infer the relations between particular variables.

A simple information environment design that roughly satisfies these design criteria is to give subjects exact knowledge of a specific probability distribution over some set of binary variables, but to then randomly relabel the names of all the variables. Each subject is shown, for certain renamed variables, the variable values in a set of random draws from the distribution. The subjects are then asked to combine their general knowledge of the distribution with their empirical data from random draws, to predict new random draws.

We created two environments with this general structure. In our "Challenging" environment six subjects predicted eight binary variables, and so estimated a probability distribution over $2^8 = 256$ possible states. Our "Training" environment had the same general structure, but was simpler. There three subjects predicted three binary variables, and so estimated a probability distribution over $2^3 = 8$ possible states. In both environments the subject's general knowledge was the fact that the true distribution was a certain chain of correlated variables, while their empirical data was partial descriptions of ten previous random draws.

2.2 Training Environment

In the training environment, the general knowledge given to subjects was that the cases were generated according to the following rule. First, the variable X had a 30% chance of being set to 0. Next, the variable Y had a 20% chance of being set equal to X. Finally, the variable Z was given a 50% chance of being zero, independently of X and Y. Thus only one of the three variable pairs has a strong relation.

Table 1 shows an example of the empirical information given to a subject in the training environment. This subject sees the values of variables A and B for ten random cases drawn from the true distribution. The other two subjects see the same ten random draws, except they see either the variables A and C or the variables B and C.

Initially, subjects do not know which variables in A,B,C correspond to which variables in X,Y,Z; all six possible permutations are equally likely. A subject who saw the data from Table 1, however, might reasonably infer that the variables A and B are unlikely to be strongly related, and that A was a good candidate to be the variable X.

Note that the sections "Sum" and "Same" of the table give redundant summaries of these ten cases; if forced to, subjects could compute these statistics for themselves. Note also that together all three subjects see all three variables and all three variable pairs, and that subjects can always just predict new cases based on the empirical statistics of the previous cases, ignoring their general knowledge of the true distribution.

2.3 Challenging Environment

In the challenging environment, the general knowledge given to subjects was that the cases were generated according to the following rule. First, the variable S had a 20% chance of being set to 0. Next, the variable T had a 20% chance of being set equal to S. Then the variable U had a 20% chance of being set equal to T. This pattern continued all the way down the chain of eight variables S,T,U,V,W,X,Y,Z. Thus out of 28 variable pairs, only 7 pairs, or 25%, were strongly and directly related.

Table 2 shows an example of the empirical information given to a subject in the challenging environment. This subject sees the values of the variables A,B,C,D for ten random draws from the true distribution. The other five subjects would see the values for E,F,G,H, for A,C,E,G, for B,D,F,H, for A,B,E,F, and for C,D,G,H. These variables can be arranged as

Case	A	B	C
1	0	1	–
2	1	0	–
3	0	0	–
4	1	1	–
5	1	1	–
6	1	1	–
7	1	0	–
8	0	1	–
9	1	0	–
10	1	0	–
Sum:	7	5	–
Same	A	B	C
A	–	4	–
B	–	–	–
C	–	–	–

Table 1: Training Environment, Sample Private Information

Case	A	B	C	D	E	F	G	H
1	0	1	0	1	–	–	–	–
2	1	0	0	1	–	–	–	–
3	0	0	1	1	–	–	–	–
4	1	0	1	1	–	–	–	–
5	0	1	1	1	–	–	–	–
6	1	0	0	1	–	–	–	–
7	0	1	1	1	–	–	–	–
8	1	0	0	1	–	–	–	–
9	1	0	0	1	–	–	–	–
10	1	0	0	1	–	–	–	–
Sum:	6	3	4	10	–	–	–	–
Same	A	B	C	D	E	F	G	H
A	–	1	2	6	–	–	–	–
B	–	–	7	3	–	–	–	–
C	–	–	–	4	–	–	–	–
D	–	–	–	–	–	–	–	–
. . .								

Table 2: Challenging Environment, Sample Private Information

the vertices of a cube, with the variable sets corresponding to the faces of that cube. Each subject sees six variable pairs, and of the 28 total pairs, twelve are seen by two subjects, twelve are seen by one subject, and four are seen by no subjects.

In table 2, the true name mapping is A=W, B=V, C=X, D=S, E=U, F=Z, G=Y, H=T. Initially, subjects only know that all 40,320 possible mappings are equally likely. A subject looking at table 2, however, might reasonably infer that A,B and A,C are strongly related, which in fact they are. She might also guess that B,D and C,D may be strongly related, which in fact they are not.

The challenging environment is clearly a much more difficult inference task than subjects are usually faced with in laboratory information experiments. It is arguably, however, still much simpler than found in most real world environments, such as PAM would have provided.

3 Mechanisms

What social institutions can induce subjects in the above two environments to combine their information into an accurate consensus forecast?

3.1 Simple Information Markets

The closest analogy to standard information markets for these environments is to have one asset per variable, and to trade each one in a simple double auction. That is, the asset "Pays $1 if A=1" is traded in the first market, the asset "Pays $1 if B=1" is traded in the second market, and so on. Subjects make offers to buy or sell a certain number of these assets for a given price, and when offers match a trade occurs. (Since short sales backed by cash deposits were allowed, there was no need for explicit markets in the matching assets "Pays $1 if A=0" and so on.)

We used *MarketScape* software provided by Charles Plott to run the simple information market sessions. We used the price of the last trade in each market as an estimate of P(A=1), P(B=1), and so on. We created an estimate of the full joint distribution over the variables by assuming variable independence.

Note that these markets are operating near the edge of feasibility in terms of traders per asset. In the training environment three subjects trade in three markets, while in the challenging environment six subjects trade in eight markets. Another possible way to use standard information markets in these environments would be to use one market per independent state. This is barely imaginable in the training case, where three subjects would have to trade in 7 markets. But this seems infeasible in the challenging environment, where six subjects would have to trade in 255 markets.

3.2 Combinatorial Call Markets

The second institution we compared was a combinatorial call market. It was similar to markets that have been successfully used in two sided combinatorial trading in other contexts, including in the closely related area of financial markets (Ledyard, Bossaerts, & Fine, 2002).

In a call market, there is an open time before the end of a round when subjects can submit offers and see some estimates of what would happen if no more offers were submitted. When the round ends, prices are chosen so as to allow as much as possible of as many offers as possible to be matched with each other, at least among offers that subjects can make good on. Some offers may be accepted for a reduced quantity, and may be given a more favorable per unit price than was requested. We allowed offers to be canceled during the open round, and passed unmet offers on to the next round.

In our combinatorial market, subjects could offer to buy and sell not only "Pays $1 if A=1", but also "Pays $1 if A=1 and B=0", "Pays $1 if A=1 given B=0", or "Pays $1 if A=1 and B=0 given C=1 and D=0." (A trade of an asset with a "given" condition is undone if the condition is false.) To make these offers as comparable as possible, they were internally translated into offers regarding packages of state assets. A state specifies the values of all the variables, and a state asset pays off only if that exact state is realized.

The exact algorithm used to match offers, set prices, and estimate those results during the open round, is complex and proprietary to *Net Exchange* (the small business with the PAM contract). So subjects were only told that we would try to match as much of as many offers as possible. There were four trading rounds per period. For this and all combinatorial market mechanisms, the last combinatorial prices in a period, i.e., the prices for all states, were treated as the mechanism's probability distribution.

3.3 Proper Scoring Rules

As a reference for comparison, we ran sessions where individual subjects described their predictions to a proper scoring rule. A proper scoring rule is a rule for paying subjects based on both the true state that is realized and their stated probability distribution. A risk neutral Bayesian with state-independent utility should respond to a proper scoring rule by stating her true subjective probabilities (Brier, 1950; Savage, 1971).

We used a logarithmic scoring rule, which has many favorable properties (Good, 1952; Winkler, 1969). A less favorable property of this rule is that payments are unbounded. Since we imposed a budget constraint on subjects, subjects were in principle limited in how close to zero or one they could go in their predictions. We never actually had to impose this limit, however.

While the sessions with simple information markets used a different interface, for all other mechanisms the user interface for browsing and changing probabilities was made to be as uniform and standard as possible. This uniformity hopefully reduced a source of random variation in the relative mechanism performance.

The standard interface had a "get price" request that would return the current probability for any expression one cared to enter, such as "A=1 and B=0 given C=1 and D=0." One could also browse a display of probabilities. Initially one would see the unconditional probabilities for all the variables. One could then choose a particular variable and set the interface to assume that it had a particular value. After that one would see the probabilities for all variables conditional on that assumption. One could then make more assumptions.

In the proper scoring rule mechanism, subjects would browse the current probability values looking for a number they wanted to change. Upon asking to change a particular number, subjects would be told the consequences for their portfolio of making that change, and would then be asked if they wanted to continue.

The mechanisms also provided a uniform way to browse one's portfolio, showing how average payoffs would vary with different variable values. Mechanisms also had a uniform way of indicating whether one was already long or short on any particular asset one was thinking of changing.

3.4 Opinion Pools

The probability distributions given by individuals in response to a scoring rule can be combined in various ways to give a group forecast. It has long been known that if one has good models of how individual behavior responds to information, and a good model of the information task, one can use Bayesian inference to produce a good probability estimate from individual opinions (Genest & Zidek, 1986). This has even been demonstrated in some simple laboratory environments (Chen, Fine, & Huberman, 2003).

In the absence of such models, one can use some heuristic formulas to combine individual forecasts. Linear opinion pools, which give an arithmetic mean of individual opinions, and logarithmic opinion pools, which give a geometric mean of individual opinions, are popular.

A big problem with these approaches is the need to judge which participants are how informative on which topics. And there are some impossibility theorems that make all possibilities seem unsatisfactory in one way or another (Genest & Zidek, 1986). But this should not keep us from seeing how these rules perform in practice. We looked at simple linear and logarithmic opinion pools, where all subjects are weighted equally.

3.5 Market Scoring Rules

There is a way for many people to share, in sequence, a single proper scoring rule. The rule starts at an initial probability distribution, and then each person is invited to change that distribution whenever she is willing to be paid based on the *difference* between scoring rule payoffs for the previous distribution and the new distribution she reports (Hanson, 2003).

A risk-neutral Bayesian with state-independent utility should always expect to profit from changing the shared distribution in the direction of her current beliefs. And since every such change is equivalent to some bet, and vice versa, a market scoring rule is equivalent to a market maker that stands willing to make any combinatorial trades. It offers a fair price for infinitesimal quantities, and prices get worse for larger quantities.

While a scoring rule always suffers some expected loss, as the price of inducing participation, for a market scoring rule this loss only depends on the last report given to it, and is otherwise independent of all previous reports. For a combinatorial logarithmic rule the price only depends on the number of variables included, with no additional cost for allowing all combinations of these variables. Logarithmic versions are nicely modular (Hanson, 2007).

In our sessions we started the market scoring rule off with a uniform probability distribution over all states, just as one usually does implicitly for a simple proper scoring rule.

A simple scoring rule is equivalent to a market scoring rule where there is only one trader. The interface for using the market scoring rule was very much like the interface for a proper scoring rule; the main difference is that the probabilities can change even when a subject does nothing. To deal with this, subjects could specify whether they wanted their order to be canceled if the prices had changed between the last price they saw and the moment they sent in their new order.

4 Experiment Design

The experiments were run at the California Institute of Technology in 2002 and 2003, and the subjects were students there. Subjects made an average of about $30 for a roughly two hour session. All subjects participated in a *training* environment session, and then most also participated later in a *challenging* environment session. Training sessions had twelve subjects, broken into four groups of three, while challenging sessions had eighteen subjects, broken into three groups of six. Subjects stayed in the same group for all periods of a session.

Each experimental session used a single environment and a single mechanism. For both environments, all sessions using that environment used exactly the same schedule of subject information and true distributions. This hopefully reduced random variation in relative mechanism performance. Each session was composed of a practice period, and then six real periods were planned, though sometimes software problems forced an early end. The first two real periods lasted for fifteen minutes each, and the rest last for twelve minutes each.

At the end of each period, subjects were paid based not on a single new sample from the true distribution, but based on a sample of one hundred cases from the true distribution. This reduced random variation in subject payoffs. Instructions were as similar as possible across the treatments. They were available online and were read by subjects before they came to the experiments.

Subjects were given 120 francs (i.e., experimental currency) at start of session, and another 40 francs before each real period, including the first. All mechanisms ensured that subjects could never bet what they did not have, and spending was also limited to 120 francs per period. In principle subjects who lost a lot in the first few rounds might have fewer francs to spend in later rounds than other subjects, but this never actually happened.

In the combinatorial call sessions, there were four open rounds and four closings per period. Open rounds lasted two to three minutes. In the proper scoring rule and market scoring rule sessions, subjects had a slightly higher average payoff due to the small net subsidy these mechanisms provide.

The performance measure we used to compare mechanisms was the relative entropy, or Kullback-Leibler (KL), distance from the mechanism's predictions to the rational beliefs given all the information available to the individuals in a group.

For all mechanisms, the official prediction was taken to be the last price or report. Since the combinatorial call mechanism sometimes returned zero prices for some states in the challenging environment, those prices were rounded up to be a minimum of 2×10^{-6}, to avoid giving the mechanism an infinite penalty under the relative entropy (KL) distance metric.

5 Results

In the training environment the KL distances were clearly not normally distributed. These distances cannot be less than zero, and many were relatively close to zero. The logarithms of the KL distances, however, were distributed roughly normally. We thus used -Log(KL Distance) as the basis for our statistical analysis of our results. As we had only weak expectations about relative performance, all statistical tests were two-tailed tests.

Table 3 collects the main results of our experiments, giving mean and 95% confidence interval values for each mechanism in each environment. The KL distances were much larger for the challenging environment; since there was more information to acquire there, there was more information that they failed to acquire there.

The number of cases listed is the number of distinct group periods for the market mechanisms, the number of individual reports for the proper scoring rule, and the number of distinct combinations of reports of subjects in their respective roles for the opinion pools.

Direct statistical tests were also made comparing the mechanisms. In the training environment, the simple double auction was significantly worse than the other mechanisms, the market scoring rule was significantly better than the other mechanisms, and the remaining mechanisms were not significantly different from each other. In the challenging environment, the market scoring rule and the two opinion pools were significantly better than the other mechanisms, but there were no significant differences within these categories. All significant differences were significant at the 0.1% level, while all insignificant differences were not significant at the 5% level.

An analysis of the market scoring rule price accuracy as a function of time within each period shows that prices were as accurate as they were ever going to be within three to five minutes, and did not on average get any more accurate in the ten minutes that followed.

6 Interpretation

The only non-combinatorial mechanism we tested was the simple double auction. (We did not test a non-combinatorial call auction, for example.) This mechanism seems to be strongly punished for being non-combinatorial in environments where a lot of information is in the relations between variables. The combinatorial call market improved on this situation in the training environment, but not in the challenging environment, suggesting that there is a limit to the ability of such markets to add liquidity to thin markets.

	Training (3 Variables)			Challenging (8 variables)		
Mechanism	Cases	Mean	95%C.I.	Cases	Mean	95%C.I.
Simple Double Auction	24	1.52	(1.39,1.65)	18	-0.112	(-0.150,-0.073)
Combinatorial Call	24	2.60	(2.15,3.04)	18	-0.195	(-0.328,-0.062)
Proper Scoring Rule	72	2.44	(1.97,2.91)	144	-0.078	(-0.106,0.050)
Linear Opinion Pool	384	2.48	(2.33,2.63)	144	0.136	(0.116,0.156)
Log Opinion Pool	384	2.56	(2.40,2.72)	144	0.162	(0.142,0.183)
Market Scoring Rule	36	3.92	(3.18,4.66)	17	0.134	(0.024,0.244)

Table 3: Mechanism Performance, -Log(KL Distance)

The opinion pools did very well in the challenging environment, but less so in the training environment. This relative failing is understandable in that information was relatively equally distributed in the challenging environment, but unequally distributed in the training environment. In the training environment, only one subject saw the two variables that were strongly related, while in the challenging environment most subjects saw one to three strongly related variables. As many have noted, opinion pools can be sensitive to assumptions about the relative amount of information held by different participants.

The market scoring rule did the best of all, helping six people to combine their information to estimate 255 independent probabilities within a period of five minutes.

7 Conclusion

Information markets seem to hold great promise, but an important obstacle to their wider application is the fact that standard market mechanisms, such as the simple double auction, require many traders per asset. This typically makes it hard to obtain a combinatorial set of estimates from such markets.

We wanted a better mechanism to use in the DARPA-funded Policy Analysis Market. So we developed several concepts for combinatorial information markets, and then ran laboratory experiments comparing the prediction accuracy of two of those new mechanisms and some more traditional mechanisms. These comparisons were made regarding a laboratory information task intended to reflect some of the difficult information features of our intended geopolitical stability application.

In our combinatorial test environments, combinatorial mechanisms clearly dominated the one non-combinatorial market mechanism we tried. But even a combinatorial call market ran into trouble in our challenging environment. Opinion pools, which mechanically combine individual opinions, did well overall, but not quite as well in an environment where information is not evenly distributed to participants.

The market scoring rule had the best performance overall, clearly beating all other mechanisms in one environment, and doing as well as any other mechanism in the other environment. Based on these results, we intended to use a market scoring rule mechanism in the

Policy Analysis Market.

References

Abramowicz, M. (2004). Information Markets, Administrative Decisionmaking, and Predictive Cost-Benefit Analysis. *University of Chicago Law Review*, *71*(3).

Banks, J. S., Ledyard, J. O., & Porter, D. P. (1989). Allocating Uncertain and Unresponsive Resources: An Experimental Approach. *RAND Journal of Economics*, *20*, 1–25.

Berg, J., & Rietz, T. (2002). Accuracy and Forecast Standard Error of Prediction Markets. Tech. rep., University of Iowa, College of Business Administration.

Berg, J. E., & Rietz, T. A. (2003). Prediction Markets as Decision Support Systems. *Information Systems Frontiers*, *5*(1), 79–93.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*, 1–3.

Chen, K.-Y., Fine, L. R., & Huberman, B. A. (2003). Predicting the Future. *Information Systems Frontiers*, *5*(1), 4761.

Chen, K.-Y., & Plott, C. R. (2002). Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem. Tech. rep. 1131, California Institute of Technology.

DeMartini, C., Kwasnica, A. M., Ledyard, J. O., & Porter, D. (1999). A New and Improved Design For Multi-Object Iterative Auctions. Tech. rep. 1054, California Institute of Technology.

Figlewski, S. (1979). Subjective Information and Market Efficiency in a Betting Market. *Journal of Political Economy*, *87*(1), 75–88.

Forsythe, R., & Lundholm, R. (1990). Information Aggregation in an Experimental Market. *Econometrica*, *58*(2), 309–347.

Genest, C., & Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, *1*(1), 114–135.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114.

Hahn, R. W., & Tetlock, P. C. (2004). How Information Markets Could Change the Policy World. Tech. rep., AEI-Brookings Joint Center for Regulatory Studies.

Hanson, R. (1990). Market-Based Foresight - A Proposal. *Foresight Update*, pp. 1,3,4.

Hanson, R. (1995). Could Gambling Save Science? Encouraging an Honest Consensus. *Social Epistemology*, *9*(1), 3–33.

Hanson, R. (1999). Decision Markets. *IEEE Intelligent Systems*, *14*(3), 16–19.

Hanson, R. (2003). Combinatorial Information Market Design. *Information Systems Frontiers*, *5*(1), 105–119.

Hanson, R. (2006). Decision Markets for Policy Advice. In Patashnik, E., & Gerber, A. (Eds.), *Promoting the General Welfare: American Democracy and the Political Economy of Government Performance*, pp. 151–173. Brookings Institution Press.

Hanson, R. (2007). Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation. *Journal of Prediction Markets*, *1*(1), 1.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Ledyard, J., Bossaerts, P., & Fine, L. (2002). Inducing Liquidity In Thin Financial Markets Through Combined-Value Trading Mechanisms. *European Economics Review*, *46*(9), 1671–1695.

Lo, A. W. (2000). Finance: A Selective Survey. *Journal of the American Statistical Association*, *95*(45), 629–635.

Maloney, M. T., & Mulherin, J. H. (2003). The complexity of price discovery in an efficient market: the stock market reaction to the Challenger crash. *Journal of Corporate Finance*, *9*(4), 453–479.

Pennock, D. M., Giles, C. L., & Nielsen, F. A. (2001). The Real Power of Artificial Markets. *Science*, *291*, 987–988.

Plott, C., & Sunder, S. (1988). Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets. *Econometrica*, *56*(5), 1085–1118.

Polk, C., Hanson, R., Ledyard, J. O., & Ishikida, T. (2003). The policy analysis market: an electronic commerce application of a combinatorial information market. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pp. 272–273 New York. Association for Computing Machinery.

Porter, D., Rassenti, S., & Smith, V. (2003). Combinatorial Auction Design. Tech. rep., Interdisciplinary Center for Economic Science.

Rassenti, S., Smith, V., & Bulfin, R. (1982). A Combinatorial Auction Mechanism for Airport Time Slot Allocation. *Bell Journal of Economics*, *13*(2), 402–417.

Roll, R. (1984). Orange Juice and Weather. *American Economic Review*, *74*(5), 861–880.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Spann, M., & Skiera, B. (2003). Internet-Based Virtual Stock Markets for Business Forecasting. *Management Science*, *49*(10), 1310–1326.

Spencer, J. (2004). New ICAP-Nymex Derivatives Have U.S. Gas Market's Number. *Wall Street Journal*.

Sunder, S. (1995). Experimental Asset Markets. In Kagel, J. H., & Roth, A. E. (Eds.), *The Handbook of Experimental Economics*, pp. 445–500. Princeton University Press, Princeton New Jersey.

Winkler, R. L. (1969). Scoring Rules and the Evaluation of Probability Assessors. *Journal of the American Statistical Association*, *64*(327), 1073–1078.

Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, *18*(2), 107–126.