Aggregating Diffuse Information with Subgroups

Kay-Yut Chen and Tad Hogg
HP Labs

November 20, 2001

Abstract

Decisions based on information spread among a group require mechanisms to efficiently uncover and aggregate this information. One approach, particularly effective with relatively small groups, is direct queries to the individuals combined with some reward for correct answers. Since such queries can be costly, we experimentally examine the tradeoff between the accuracy of the aggregation and the number of individuals queried. We compare the results to the behavior if all individuals were risk-neutral and rational with respect to expected returns. In terms of providing enough information to make correct decisions, we show queries to a few members of the group can be sufficient, even if selected randomly. We also show individual performance in information markets does not correlate well with ability to improve query results, and hence does not give a criterion to select particularly informative subgroups.

1 Introduction

Building on the observation that competition acts as a decentralized discovery process [10], information markets can aggregate information distributed among many people [9]. Existing instances, such as those predicting election outcomes [2], movie revenues[1], and technological progress[2], illustrate the potential for information markets [17]. Laboratory studies provide additional examples, under more controlled conditions, to identify their behaviors with fewer participants [8, 16, 18, 19].

Information markets could be useful in two situations. In the first, a crowd could contain a few experts. In this case, the market provides incentives for these experts to reveal their information through potentially rewarding trades. The second situation, which we focus on in this paper, consists of diffuse information in which each person has relatively little information but the group as a whole has a great deal. In this case, the market can provide not only an incentive to participate, but also a method to aggregate the information through determining the market equilibrium price.

[1] The Hollywood exchange at `http://www.hsx.com`.
[2] The Foresight exchange at `http://www.ideosphere.com`.

In the case of diffuse information, where participants have roughly the same information, they can nevertheless differ in their ability to exploit that information, risk preferences, trading skill, etc. This diversity can be especially important for small markets, as may appear for very specialized questions or when such markets are used entirely within an organization (e.g., to estimate future sales). In particular, experiments show markets do not work well as an information aggregation mechanism with small groups, with problems such as illiquidity [23], manipulation [7, 15] and lack of equilibrium [1].

For small groups, direct queries to its members provide better predictions than the market itself [3]. Moreover, such queries can address related questions, not directly appearing in the underlying information market, thereby avoiding the overhead of creating additional tradeable goods for these new questions. However, unlike the passive observation of market prices, direct queries could be expensive (e.g., if people require payments to participate), or introduce incentive issues in which participants are reluctant to reveal their information (e.g., due to privacy concerns). To minimize this expense, in this paper we experimentally examine the extent to which queries to a subset of the groups can suffice.

Experiments allow precisely controlling the information available to participants and hence specific evaluation of deviations in behavior from that predicted by utility theory [14, 24]. More generally, experiments allow evaluating the actual performance of various economic mechanisms [4, 20, 22].

In the remainder of this paper we first describe the experiments, and then measures to evaluate the performance of subgroups of the participants. We then present the experimental results and compare them to risk-neutral rational ("ideal") behavior. We conclude with a discussion of their implications for designing information aggregation mechanisms for relatively small groups.

2 Experimental Setup

A previous set of economic experiments compared an information market to direct queries to the participants [3]. This paper presents a further analysis of these experiments, namely the behavior of queries to subsets of the participants.

In these experiments, participants attempted to identify one item from among $T = 10$ possibilities, based on information given to them. The number of participants, N, in the experiments ranged from 8 to 13. Each experiment consisted of two stages, using distinct information aggregation mechanisms: markets and direct queries.

Each stage consisted of a number of periods. In each period, one of the T items was randomly selected as the correct one to be identified. The information available to each participant was generated from a list of the items with $C = 2$ additional copies of the correct one, for a total of $T + C$ items in the list. Each participant was given a certain number of values drawn uniformly at random from this list, with replacement. Participants were told the values of T and C, thus they knew each draw had probability $(C + 1)/(C + T)$ to be the correct item.

The first stage aggregated information through markets in which participants traded securities representing the items. At the end of each period, the security corresponding to the correct item paid a prespecified amount, and the others paid nothing. During each period, the securities were traded in a double sided call market. With a large number of participants, the equilibrium price for a security in such markets should reflect the likelihood its corresponding item is the correct one.

In the second stage, participants were queried about the likelihood that each possible item was in fact the correct one. Specifically, they were asked to divide 100 tickets into 10 buckets, one for each item. They were paid according to an increasing linear function of $\log(p)$ where p is the fraction of tickets they placed for the correct item. This payoff functional form induces a risk neutral expected utility maximizer to reveal his beliefs [3], which in this case is the probability distribution. Participants did not communicate with each other so their reports reflect only their own information.

Five experiments were conducted with varying concentrations of information. There were three treatments: all participants having the same amount (i.e., an equal number of draws), some having more than others and some having random amount of information (number of draws was random). This choice was the same in both stages of each experiment, i.e., a participant had the same process determining the number of draws for the market and the query states, whether that is random or fixed.

Participants can use their available information to estimate the probability distribution of the correct choice, i.e., for each item s the probability that item is the correct one. In the market, they use this estimate of payoffs for the items and make trades based on how their estimate compares with the current price. When queried directly, they report their estimated distribution, i.e., for each item s, they report a probability q_s such that $\sum_s q_s = 1$.

3 Evaluating Behavior

Previous work compared the results of the market to direct queries adjusted for risk preferences revealed by the market [3]. We now consider the behavior of queries directed at groups consisting of only some of the participants. Such subgroups will not have as much information as the full set of participants so will give lower performance.

To evaluate these mechanisms, suppose the observations in a group are available to all members and they estimate the probability distribution based on Bayes' theorem. Specifically, for a list of n_{obs} observations O, let n_s be the number of times item s is seen. Then the estimate for the probability item s is the correct one is

$$P(s|O) \propto \left(\frac{C+1}{C+T}\right)^{n_s} \left(\frac{1}{C+T}\right)^{n_{\text{obs}}-n_s} \tag{1}$$

with the overall constant of proportionality ensuring the probabilities sum to

one.

If the information market perfectly aggregates the information available to the participants and they are risk-neutral, the market prices should reflect this optimal distribution. Similarly, if a subgroup of the participants is queried, the best result would match this distribution for the observations available to that group.

A common criterion for the difference between two distributions $\{p_s | s = 1, \ldots, T\}$ and $\{q_s\}$ is the Kullback-Leibler divergence measure [12], or relative entropy, on distributions

$$KL(p, q) = \sum_s p_s \log \left(\frac{p_s}{q_s} \right) \tag{2}$$

This measure is not symmetric, so not a true distance, but is nonnegative, and zero only if the distributions are identical. Furthermore, the Kullback-Leibler measure based on two independent events is the sum of the measure based on each individual event. Thus, it provides a convenient way, namely by the use of averages, to summarize results of a multiple period experiment.

As a practical matter, information markets or queries are meant to provide information for some decision. Thus in addition to measuring how close to optimal the reported distributions are, we can also consider the likelihood they are sufficient for the correct decision. We specificly consider the case of selecting the correct item. That is, we suppose the groups inform a decision-maker, who receives a payoff for selecting the correct item. In addition, we suppose the decision-maker selects the item with the highest estimated probability. We also suppose group queries incur a cost proportional to the size of the group queried. With this tradeoff, the expected performance of the decision-maker may be largest with a smaller group, even though a larger group would provide somewhat more accurate estimates of the optimal distribution.

3.1 Rational, Risk-Neutral Behavior

For comparison with the experiments, we evaluate the distributions that would be reported by the average and best subgroup of various sizes, assuming rational and risk-neutral behavior.

Specifically, we exhaustively consider all $\binom{N}{k}$ groups $G \subset \{1, \ldots, N\}$ of k participants. Using the actual observations provided to the group members in each period of the experiment, we evaluate the rational belief of the distribution $p^{(i)}$ of the items for each individual i in the group by the use of Eq. (1). We then aggregate the individual distributions to produce an overall distribution for the group, assuming accurate reporting, i.e.,

$$p_s^{(G)} \propto \prod_{i \in G} p_s^{(i)} \tag{3}$$

This aggregation is equivalent to that from Eq. (1) using the combined observations of all members of the group. To quantify the performance of the group G,

we average $KL(p^{\text{optimal}}, p^{(G)})$ over all periods, where p^{optimal} is the distribution from Eq. (1) using all observations for all participants, not just those in the group G. With this procedure, we obtain a performance measure for the group. We then record the average and best performance over all the groups of size k.

An experiment with N participants has $2^N - 1$ nonempty groups. Thus this procedure is computationally intensive for experiments with many participants. On the other hand, by using the actual observations in the experiment, it provides a direct measure of the expected behavior for the groups under the assumptions of rationality and risk-neutrality. In particular, it allows comparing the best and typical group performances when differences are only due to the quality of information available to different groups.

This procedure allows us to quantify the impact of increasing the amount of information by increasing group sizes. However, it does not take into account of participant behavior. Even with the same information, some subjects may be processing their information better than others. Thus, there may be a stronger reason to choose one subgroup over another.

We have also calculated the performance measure using a nonlinear aggregation [3] of the reports from the members of the subgroups. This technique is accurate if applied to the whole group [3]. This provides a comparison between the effects of information and the behavior of different group sizes.

4 Results

4.1 Group Performance

Fig. 1 shows the behavior of average and best groups as a function of size for an experiment with 9 subjects. For comparison, we also show the behavior one would obtain for rational, risk-neutral participants using the same observations. We see improvement with larger groups, as expected since they have more information. But they do not perform quite as well as possible based on the information they have.

For further insight into these results, Table 1 shows the individuals in the best group of each size. We examined all $\binom{n}{k}$ groups of size k. In most cases, the optimal group of size k is a subset of the optimal group of size $k+1$. Thus the best groups do not represent a diverse mix experts (as may be the case if participants were of several types, each type having similar information but distinct from others).

We thus see that queries to subgroups of the participants can give reasonable estimates of the distribution, even if those groups are selected at random. Using this behavior, a mechanism designer could determine whether the cost of larger groups is worth the gain in performance. Fig. 1 also shows gains are significantly larger if the high-performing groups can be identified a priori.

We see similar behaviors with other experiments, with different numbers of people and different numbers of observations given to the participants. Table 2 gives the numbers of participants and periods for the experiments.

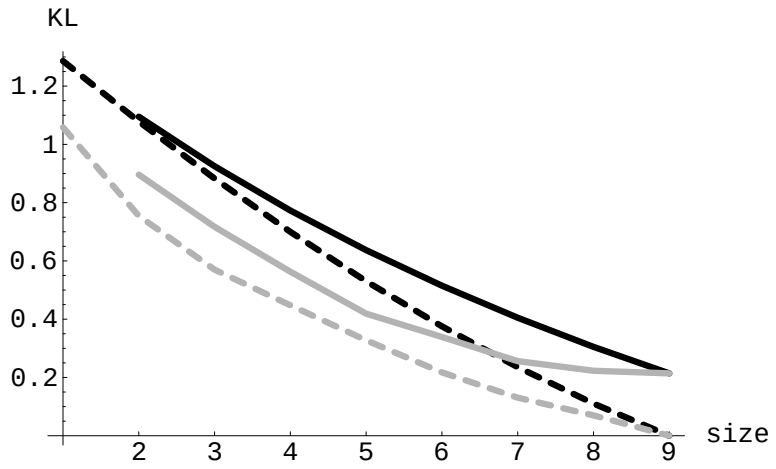

Figure 1: Performance of average (black) and best (gray) group as a function of group size for an experiment with 9 subjects and 18 periods in which each person received 3 draws. The solid curves are experimental data, dashed correspond to risk-neutral rational behavior.

size	experimental	theory
1		7
2	3 4	3 7
3	3 4 7	1 3 7
4	3 4 7 8	1 3 7 9
5	3 4 7 8 9	1 2 4 6 7
6	3 4 6 7 8 9	1 2 3 4 6 7
7	2 3 4 6 7 8 9	1 2 3 4 6 7 9
8	1 2 3 4 6 7 8 9	1 2 3 4 5 6 7 9

Table 1: Best groups for each size (experimentally observed and for the theory assuming rational, risk-neutral behavior) for the same situation as shown in Fig. 1.

	expt1	expt2	expt3	expt4	expt5
participants	13	9	11	8	10
periods	7	18	29	25	30

Table 2: Number of participants and periods in the experiments.

group	expt1	expt2	expt3	expt4	expt5
random	1.36	0.93	1.18	1.12	1.15
payoff	1.45	1.09	1.24	1.13	1.39
value	0.72	0.91	0.94	1.13	1.22
optimal	0.53	0.72	0.75	0.83	0.77

Table 3: KL measure of the distribution for groups of size 3 in five experiments. We compare four types of groups differing in how members are selected as described in the text. Behavior shown in Fig. 1 corresponds to experiment 2 in this table.

As a specific example of the performance with small groups, Table 3 shows the Kullback-Leibler measure for groups of size 3 in each of the experiments. The first row gives values for random groups (i.e., the average obtained from a sample of groups of size 3). The second and third rows are groups selected based on market performance, respectively either the total payoff or final holding value, including cash. The last row shows the performance of the optimal 3-person group. Note that for each type of selection, we report performance for a group facing the independent tasks of all periods rather than selecting a new group for each period. Thus as the number of periods increases, the variation in information quality among the groups decreases, leading to a smaller gap between random and optimal group performance when participants are rational and risk-neutral. This trend is also generally seen in the experimental data given in Table 3, although experiment 2 has somewhat less improvement for the best group than might be expected from comparison with the other experiments.

Interestingly, groups based on market payoff are worse than random according to this measure, i.e., their reported distributions have larger errors. Although one might expect participants with more accurate information to perform better in the market, actual payoff also reflects the ability to trade profitably on trends in the market, not just on the fundamental probabilities various securities will pay off at the end of the market. If we instead use participants' holding values as an indication of their information, the groups do better than using payoffs, although there is no definite improvement over random selection.

4.2 Providing Information for Decisions

Another way to access the group performance is the extent to which the reported distributions suffice to make correct decisions. In our case, the decision is to identify the correct item.

We measure the performance of a group by the fraction of "correct" decisions made by a group, based on the distribution aggregated from the group members using Eq. (3). In this context, we take the correct decision to be the most likely item according to the probability distribution resulting from combining draws available to all the participants. In most cases, this corresponds to the item actually selected prior to each period and is useful to show the extent to which

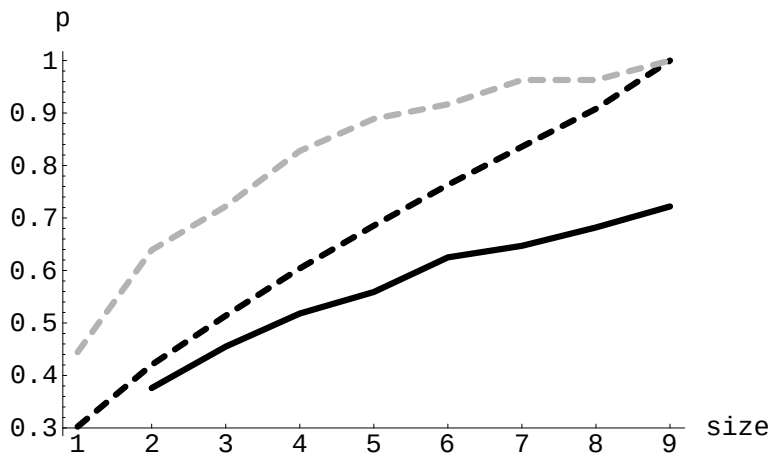

Figure 2: Decision performance of average (black) and best (gray) group as a function of size for the experiment shown in Fig. 1. The solid curve is experimental data, and dashed curves correspond to risk-neutral rational behavior.

group	expt1	expt2	expt3	expt4	expt5
random	0.47	0.46	0.47	0.52	0.43
payoff	0.57	0.28	0.59	0.48	0.33
value	1.00	0.44	0.66	0.48	0.57

Table 4: Fraction of correct decisions for groups of size 3 in five experiments. We compare three types of groups differing in how members are selected as described in the text discussing Table 3.

subgroups can match the best possible performance when all information is available to the decision-maker.

Fig. 2 shows group performance as a function of the group size. Groups are picked at random for a given size, so the performance corresponds to that of an average group. We see relatively little improvement in decision quality with the larger group sizes, especially compared to risk-neutral rational behavior. Thus, as with the difference in distributions shown in Fig. 1, we see the improvement in decision quality from asking larger groups may not be worthwhile when the cost to query groups increases with their size. Significantly, this is true even if the groups are selected at random.

Table 4 further illustrates this behavior, in five experiments with different concentrations of information. Selecting groups based on market behavior offers better performance than random for some of the experiments, but not always. Moreover, selection based on holding value (and cash) is better than selection based on payoff in the market. These observations are comparable to the con-

group	expt1	expt2	expt3	expt4	expt5
random	1.25	0.92	1.18	1.07	1.21
query	0.78	0.89	0.71	0.92	0.81
optimal	0.60	0.59	0.69	0.72	0.72

Table 5: KL measure of the distribution for groups of size 3 in five experiments. Here we consider performance only in the second half of the periods in each experiment. The random and optimal groups are selected as in Table 3. The query group is the best based on performance during the first half of the periods.

clusions drawn from measuring the difference in reported distributions, given in Table 3.

4.3 Identifying Useful Groups

Comparing experiments with the behavior expected for rational, risk-neutral individuals we see the group performance is only partly due to the different information available to the participants. This raises the possibility that some characteristics of the individuals, possibly observable in their market behavior, may give useful indications of high-performing groups.

In Fig. 1, the performance gap between the average experimental and ideal performances (black solid and dashed lines, respectively) is entirely due to behavioral effects. Thus it should be possible to identify more effective groups than average to close this performance gap.

To identify high-performing groups, one might think performance in the market would be a good proxy for those with more information or greater capability to use it. However, when we select members of a group based on their market performance (i.e., payoff or holding values), the groups are generally no better than randomly selected groups, and often worse. This indicates that, at least in the small groups involved in our experiments, speculation and strategic aspects of the game outweigh the information differences among individuals. For example, it appears difficult to distinguish whether trading activity represents fundamental decisions based on available information or speculative trades based on expectations of what other traders will do in later periods. Thus although in principle it should be possible to reduce or eliminate the performance gap due to behavioral issues, it remains to be seen how to identify practical correlates of high performance in passively observed behaviors, such as market activity.

Another approach to identifying high-performing subgroups, shown in Table 5, uses the observed behavior in response to direct queries. Because the experiments consist of several independent periods, we can evaluate performance of various subgroups while all participants are queried. After a certain number of periods, we can use these observations to select the best subgroup. Subsequently, we can direct queries only to this subgroup. While this approach incurs

the cost of queries to the full group for some portion of the experiment, it nevertheless reduces costs compared to queries for all participants over all periods. Moreover, as shown in Table 5, by directly observing individual behavior in response to queries, it is able to identify better subgroups than selection based on market behaviors.

These observations indicate groups can be identified to reduce the performance gap between experimental and ideal behavior, on average. However, eliminating the gap between average and best ideal performance (the black and gray dashed lines, respectively, in Fig. 1) is probably not possible. This is because the gap is due to statistical effects: with the random draws, some participants receive better information than others by the luck of the draw. Because the draws in each period are independent, as the number of periods increases, the difference between average and best groups decreases. However, with the small number of periods used in these experiments (less than 30), this statistical effect is significant and might misleadingly suggest further mechanism improvements are possible beyond just removing biases from behavioral effects.

5 Discussion

When limited to small groups with diffuse information, markets do not aggregate information as well as direct queries to the participants [3]. In this paper, we extend these observations to the case where queries involve only some of the group members. While the smaller groups do not report distributions as accurately as asking all participants, they can be sufficient for decision-making, particularly if the queries involve costs increasing with group size.

We see this benefit even if the groups are selected randomly. However, potentially better performance is possible if the best groups can be identified. Unfortunately, group performance for queries does not correlate with observable market behavior, so it remains an open question to identify characteristics of the high-performing groups.

As one direction for future work, we could examine larger variation in available information among the participants. In particular, markets may be more useful as a way to identify "experts" when information is concentrated in a few's hand, unlike the setup in the experiments we discussed. This would establish market's performance as an "expert" identification mechanism as oppose to an aggregation mechanism.

As the groups get larger, we can expect market behavior to improve relative to direct queries, particularly since the cost involved in queries is likely to be larger. Thus an important experimental question is how information aggregation behavior tracks with the number of participants. In particular, how should size of subgroup scale with the number of participants to get, say, half the performance obtained with asking all participants? Obviously, this depends on how the information and skills are distributed amongst the group. It will be interesting to see if there are ways to determine information and skill distributions from market behavior.

Fig. 1 shows the performance difference in group size is not just a function of available information. Thus it would be interesting to determine how much of the variation is due to differences in skill, risk tolerance, or other properties of the participants. Moreover, to what extent do such differences appear in trading activity, e.g., to distinguish fundamental traders with more information from speculators?

The departures from rationality seen in these experiments provide an opportunity for automated agents with access to the same information available to the people, as for example with the Santa Fe token exchange [21] and various proposals for agent-based market mechanisms [5, 25, 27]. E.g., for equal performance we could manage with fewer agents than people. Moreover, if the agents cannot access the information directly, e.g., because it requires complicated judgements to recognize or evaluate, hybrid systems [6] could be effective. That is, the agents could use or combine the information more efficiently than people once it is available.

As a final observation, basing group selection (and hence any associated payments or gains in reputation or status [13, 26]) on visible behavior in markets raises interesting incentive issues if the selection could bias the market behavior. If the original market is meant to satisfy other purposes, it would be best to avoid such changes. One possible approach would be to rely on information-hiding techniques [11], such as zero-knowledge transactions, to minimize the amount of information revealed and hence reduce the impact on the market mechanisms.

Acknowledgements

We thank B. Huberman and L. Fine for helpful discussions.

References

[1] L. Anderson and C. Holt. Information cascades in the laboratory. *American Economic Review*, 87:847–862, 1997.

[2] Stanley W. Angrist. Iowa market takes stock of presidential candidates. *Wall Street Journal*, August 28 1995. See also www.biz.uiowa.edu/iem.

[3] Kay-Yut Chen, Leslie R. Fine, and Bernardo A. Huberman. Forcasting uncertain events with small groups. In *Proc. of the ACM Conference on E-commerce*, October 2001.

[4] Kay-Yut Chen and Charles R. Plott. Nonlinear behavior in sealed bid first price auctions. *Games and Economic Behavior*, 25:34–78, 1998.

[5] Scott H. Clearwater, editor. *Market-Based Control: A Paradigm for Distributed Resource Allocation.* World Scientific, Singapore, 1996.

[6] Rajarshi Das, James E. Hanson, Jeffery O. Kephart, and Gerald Tesauro. Agent-human interactions in the continuous double auction. In *Proc. of the 17th Intl. Joint Conf. on Artificial Intelligence (IJCAI-2001)*, pages 1169–1176, San Francisco, 2001. Morgan Kaufmann.

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

[8] R. Forsythe, T. Palfrey, and C. Plott. Asset valuation in an experimental market. *Econometrica*, 50:537–567, 1982.

[9] Robin Hanson. Decision markets. *IEEE Intelligent Systems*, pages 16–19, May/June 1999.

[10] Friedrich A. Hayek. Competition as a discovery procedure. In *New Studies in Philosophy, Politics, Economics and the History of Ideas*, pages 179–190. University of Chicago Press, Chicago, 1978.

[11] Bernardo A. Huberman, Matt Franklin, and Tad Hogg. Enhancing privacy and trust in electronic communities. In *Proc. of the ACM Conference on Electronic Commerce (EC99)*, pages 78–86, NY, 1999. ACM Press.

[12] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1952.

[13] C. H. Loch, B. A. Huberman, and S. Stout. Status competition and performance in work groups. Technical Report 99/49/TM/OB, INSEAD, Sept. 1999.

[14] Mark J. Machina. Decision-making in the presence of risk. *Science*, 236:537–542, 1987.

[15] M. Noeth and M. Weber. Information aggregation with random ordering: Cascades and overconfidence. Technical report, Univ. of Mannheim, 1998. Presented at the Summer 1998 ESA Meetings.

[16] J. O'Brien and S. Srivastava. Dynamic stock markets with multiple assets. *J. of Finance*, 46:1811–1838, 1991.

[17] David M. Pennock, Steve Lawrence, C. Lee Giles, and Finn Arup Nielsen. The real power of artificial markets. *Science*, 291:987–988, 2001.

[18] C. Plott and S. Sunder. Efficiency of experimental security markets with insider information: An application of rational expectations models. *J. of Political Economy*, 90:663–698, 1982.

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

[20] C. Plott, J. Wit, and W. Yang. Pari-mutuel betting markets as information aggregation devices: Experimental results. Social Science Working Paper 986, Caltech, 1997.

[21] John Rust, Richard Palmer, and John H. Miller. Behavior of trading automata in a computerized double auction market. In D. Friedman, J. Geanakoplos, D. Lane, and J. Rust, editors, *The Double Auction Market: Institutions, Theories and Evidence*. Addison Wesley, 1992.

[22] Vernon L. Smith. Experimental methods in the political economy of exchange. *Science*, 234:167–234, 1986.

[23] S. Sunder. Markets for information: Experimental evidence. *Econometrica*, 60:667–695, 1992.

[24] Richard H. Thaler. Anomalies: The ultimatum game. *J. of Economic Perspectives*, 2(4):195–206, 1988.

[25] Carl A. Waldspurger, Tad Hogg, Bernardo A. Huberman, Jeffery O. Kephart, and W. Scott Stornetta. Spawn: A distributed computational economy. *IEEE Trans. on Software Engineering*, 18(2):103–117, February 1992.

[26] Claus Wedekind and Manfred Milinski. Cooperation through image scoring in humans. *Science*, 288:850–852, 2000.

[27] Michael P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. *J. of Artificial Intelligence Research*, 1:1–23, 1993.