

Automated Market-Making in the Large: The Gates Hillman Prediction Market

Abraham Othman

Computer Science Department
Carnegie Mellon University
aothman@cs.cmu.edu

Tuomas Sandholm

Computer Science Department
Carnegie Mellon University
sandholm@cs.cmu.edu

ABSTRACT

We designed and built the Gates Hillman Prediction Market (GHPM) to predict the opening day of the Gates and Hillman Centers, the new computer science buildings at Carnegie Mellon University. The market ran for almost a year and attracted 169 active traders who placed almost 40,000 bets with an automated market maker. Ranging over 365 possible opening days, the market's event partition size is the largest ever elicited in any prediction market by an order of magnitude. A market of this size required new advances, including a novel span-based elicitation interface. The results of the GHPM are important for two reasons. First, we uncovered two flaws of current automated market makers: spikiness and liquidity-insensitivity, and we develop the mathematical underpinnings of these flaws. Second, the market provides a valuable corpus of identity-linked trades. We use this data set to explore whether the market reacted to or anticipated official communications, how self-reported trader confidence had little relation to actual performance, and how trade frequencies suggest a power law distribution. Most significantly, the data enabled us to evaluate two competing hypotheses about how markets aggregate information, the Marginal Trader Hypothesis and the Hayek Hypothesis; the data strongly support the former.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Economics; I.2.11 [Distributed Artificial Intelligence]: Multiagent systems; H.1.2 [User/Machine Systems]: Human factors

General Terms

Economics, Experimentation, Design, Theory

Keywords

Prediction Markets, Automated Market Making, Elicitation, Experimental Studies, Data Analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

EC'10, June 7–11, 2010, Cambridge, Massachusetts, USA.
Copyright 2010 ACM 978-1-60558-822-3/10/06 ...\$10.00.

1. INTRODUCTION

Prediction markets are powerful tools for aggregating information. Most prediction markets in use today, however, only generate a single data point. For simple binary events, like the probability of a sports team winning its next match, this is entirely satisfactory. However, for more complex events, this can be inappropriate. Consider a prediction market to estimate the expected number of US casualties in Afghanistan over the next year. Conceivably, market participants could be split between a very low estimate and a very high estimate. The resulting consensus of a middle value could be an accurate estimate of the expectation, but would be misleading to design policy around.

Recent theoretical work has suggested that eliciting interesting distribution properties (like the element that has maximum probability) is as difficult as eliciting an entire distribution (Lambert et al., 2008). In this paper, we discuss the design of a market, the *Gates Hillman Prediction Market*, that generated a complete distribution over a fine-grained partition of possibilities, while retaining the interactivity and simplicity of a traditional market.

Fundamental to our design is an automated market maker (Hansson, 2003). It has three primary benefits. First, the market maker provides a rich form of liquidity: it guarantees that participants can make any self-selected trade at any time. Second, it allows instant feedback to traders, rather than delayed, uncertain, potential feedback. A trader can always get actionable prices both on any potential trade she is considering and on the current values of the bets she currently holds. Third, the automated market maker obviates the need to match combinations of user-generated buy and sell orders—a problem that can be combinatorially complex (Fortnow et al., 2003; Chen et al., 2008)—making a large event space computationally feasible.

Equally important to the success of the GHPM was the user interface of the trading platform. Traditional economic theory would hold that all suitably complex interfaces are equivalent, but this is not the case given well-documented shortcomings in human reasoning. In particular, a large event space implies that the average probability of an event is small, and people have great difficulty discriminating between small probabilities (e.g., Ali (1977)). To solve this problem, the GHPM used a span-based interface with ternary elicitation queries, which we discuss in Section 2.3.

As the first test of automated market making in a large prediction market, the GHPM allowed us to discover two flaws in current automated market makers, which will help focus future design of market makers. Section 3 discusses the two flaws, *spikiness* and *liquidity-insensitivity*, in detail and explores their theoretical roots.

The GHPM generated a rare large corpus of identity-linked trading data. In general one can obtain identity-linked data from laboratory experiments, but those data sets are generally small due to

practical constraints like subject payments, training effort, and the viable duration of an experiment. For example, Healy et al. (2008) study behavior and prices in laboratory prediction markets in detail, but their experiment only had three traders. Alternatively one could examine data from large markets (such as stock markets), but that data is not released with trader identities attached to events, severely restricting the kinds of questions that can be studied. The GHPM dataset is unique because it is large, detailed, and linked to traders. In Section 4 we use this data set to explore questions related to trader behavior and its impact on prices and information aggregation.

2. MARKET DESIGN

The GHPM used a raffle-ticket currency tied to real-world prizes, an automated market maker, and a novel span-based ternary elicitation interface. In the following sections we discuss each of these in turn.

2.1 Incentives and setup

Due to legal concerns, the GHPM used raffle tickets as the currency rather than real money. Thanks to generous grants from Yahoo! and other sources, we secured the equivalent of about \$2,500 in prizes to distribute. These prizes were distributed by raffle after the market's close, with the prizes allocated randomly proportionally to the number of tickets each user amassed. This gives (risk-neutral) participants the same incentives as if real money were used—unlike the approach where the best prize is given to the top trader, the second-best prize to the second-best trader, etc.

The GHPM was publicly accessible on the web at whenwillwemove.com, but trading accounts were only available to holders of Carnegie Mellon e-mail addresses. For fairness, we did not allow people with direct control over the building process (e.g., members of the building committee) to participate. Upon signup, each user received 20 tickets, and each week, if that user placed at least one new trade, she would receive an additional bonus of two tickets. In a market with real money, we would expect that traders more interested or knowledgeable would stake more of their personal funds in the market. However, in a fake-money setting, we do not have this option. For instance, a mechanism that asked users if they were “very interested” in the market, and promised to give them extra tickets if they answered affirmatively would obviously not be incentive compatible. The two ticket weekly bonus was intended to simulate the impact of more interested traders having more influence.

One of the most challenging parts of running a prediction market over real events is defining contracts so that it is clear which bets pay out. For example, *InTrade*, a major commercial prediction market, ran into controversy over a market it administered involving whether North Korea would test missiles by a certain date. When North Korea putatively tested missiles unsuccessfully, but the event was not officially confirmed, the market was reduced to a squabble over definitions. We set out to study when the Computer Science Department would move to its new home in the Gates and Hillman Centers (GHC), but *move* is a vague term. Does it indicate boxes being moved? Some people occupying new offices? The last person occupying a new office? The parking garage being open? From discussions with Prof. Guy Blelloch, the head of the building committee, we settled on using “the earliest date on which at least 50% of the occupiable space of the GHC receives a temporary occupancy permit”. Temporary occupancy permits are publicly issued and verifiable, must be granted before the building is occupied, and are normally issued immediately preceding occupancy (as was the case in the GHC).

The market was active from September 4th, 2008 to August 7th, 2009. On this latter date, the GHC received the first occupancy permit, and it covered slightly over 50% of the space in the building. The price of a contract of August 7th, 2009 converged to 1 about five hours before the public announcement that the building had received its permit.

In total, 210 people registered to trade and 169 people placed at least one trade. 39,842 bets were placed. Following the conclusion of the market, we conducted recorded interviews with traders we deemed interesting about their strategies and impressions of the GHPM. Excerpts of some of these conversations appear later in this paper.

2.2 Automated market maker

In this section, we provide a brief discussion of the automated market maker concept originated by Hanson (2003), and how we applied it to the specific setting of the GHPM. This idea behind automated market making has been widely applied in practice, including at *Inkling Markets*, a prediction market startup company, and (before its recent demise) *Tradesports*, a sports betting prediction market company.

We began by partitioning the event space into $n = 365$ events, one for each day from April 2, 2009 to March 30, 2010 with the addition of “April 1, 2009 and everything before” and “March 31, 2010 and everything after”, to completely cover the space of opening days. To our knowledge, the GHPM is by far the largest market (by event partition size) ever conducted. The largest prior prediction markets are probably markets over candidates for political nominations, where as many as 20 candidates could have contracts (of course, only a handful of candidates in these markets are actually traded).

For the GHPM we applied the most widely-used automated market maker for prediction markets, the *logarithmic market scoring rule (LMSR)*, originally designed by Hanson (2003). The market maker operates according to a cost function $C : \mathbb{R}^n \mapsto \mathbb{R}$, which maps a vector of quantities \mathbf{q} to a scalar representing how much money has been paid into the system. Each entry q_i in the vector \mathbf{q} represents how much money is to be paid out if the i -th event is realized. The cost function for the LMSR is

$$C(\mathbf{q}) = b \log \left(\sum_i \exp(q_i/b) \right)$$

where $b > 0$ is a constant fixed *a priori* by the market administrator. As Pennock and Sami (2007) discuss, the b parameter can be thought of as a measure of market liquidity, where higher values represent markets less affected by small bets. In the GHPM we fixed $b = 32$, and since the LMSR has worst-case loss of $b \log n$, at most about 80 surplus tickets would be won from the market maker by participating traders. (This is indeed the amount actually transferred from the market maker to the participants because probability mass converged to the correct day before the market ended.)

Prices are defined by the gradient of the cost function, so that

$$p_i(\mathbf{q}) = \frac{\exp(q_i/b)}{\sum_j \exp(q_j/b)}$$

is the price of the i -th event. The prices can also be directly thought of as event probabilities. The generated prices define a probability distribution over the event space: they sum to unity, are non-negative, and exist for any set of events.

There has been a flurry of recent research involving automated market makers (Chen and Pennock, 2007; Agrawal et al., 2009). Regardless, of the particular market maker used, we can mandate

that all automated market makers should be strictly monotonic, so that the price of an event increases in its payout quantity.

2.3 Span-based elicitation with ternary queries

In this section, we present the novel elicitation mechanism used in the GHPM. A similar interface was developed independently and contemporaneously by Yahoo! Research for Yoopick, an application for wagering on point spreads in sporting events that runs on the social network Facebook.

The major problem in implementing fine-grained markets in practice is one of elicitation: they are too fine for people to make reliable point-wise estimates. Consider the GHPM, which is divided into 365 separate contracts, each representing a day of a year. Under a traditional interaction model, traders would act over individual contracts. But with 365 separate contracts, the average estimate of each event is less than .3%. People have great difficulty reliably distinguishing between such small probabilities (Ali, 1977), and problems estimating low-probability events have been observed in prediction markets (Wolfers and Zitzewitz, 2006).

We solve this problem by simple span-based elicitation, which makes estimation of probabilities easy for users. In our system, the user can select a related set of events and gauge the probability for the entire set. Spans are a natural way of thinking about large sets of discrete events: people group months into years, minutes into hours, and group numbers by thousands, millions, or billions. The key here is that spans use the concept of distance between events that is intrinsic to the setting.

For example, let the market be at state $\mathbf{q}^0 = \{q_1^0, \dots, q_n^0\}$. (These are the quantities that will be paid out if each of the respective states is realized.) A user's interaction begins with the selection of an interval, from indices s to t . This partitions the indices into (at most) three segments of the contract space: $[1, s]$, $[s, t]$, and $(t, n]$. The user then specifies an amount, r , to risk. Our market maker offers two alternative bets to the user:

- The “for” bet. The agent bets *for* the event to occur within the contracts $[s, t]$. The user’s payoff if he is correct, π_f , is calculated from

$$C(q_1^0, \dots, q_{s-1}^0, q_s^0 + \pi_f, \dots, q_t^0 + \pi_f, q_{t+1}^0, \dots, q_n^0) = C(\mathbf{q}^0) + r$$

- The “against” bet. The agent bets *against* the event occurring within the contracts $[s, t]$. The user’s payoff if he is correct, π_b , is calculated from

$$C(q_1^0 + \pi_b, \dots, q_{s-1}^0 + \pi_b, q_s^0, \dots, q_t^0, q_{t+1}^0 + \pi_b, \dots, q_n^0 + \pi_b)$$

$$= C(\mathbf{q}^0) + r$$

Solving for π_f and π_b is not generally possible in closed form. These equations can be solved numerically using, for example, Newton’s method. Depending on the specific cost function and numerical solution method, there might be issues with solution instability that should be addressed; for instance, the GHPM used Newton’s method with a bounded step size at each iteration to discourage divergence.

Given a selected set of events, the simplest way to represent a bet for that set is to have each event in the set pay out an identical amount if the event is realized, as we do in the two equations above. This simplicity means we can significantly condense the language we use when eliciting a wager from an agent. Instead of asking a user whether he would accept an n -dimensional payout vector, we need only present a single value to the user. A screenshot of the elicitation process in the GHPM appears as Figure 1.

Yahoo! Research’s Yoopick does not have “against” bets, but the GHPM does. From discussions with traders in the GHPM, against bets were used frequently to bet against specific (single) contracts they feel are overvalued. Several successful traders had a portfolio consisting solely of bets against a large number of single contracts. The success of these traders was likely a combination of the misjudging of small probabilities by other traders as well as the spiky price phenomenon discussed in the next section.

There are several relevant pieces of information the market administrator could provide the users for each potential bet:

- The agent’s direct payout if he is correct, π_f (or π_b). Both Yoopick and the GHPM display this information.
- The *averaged* payout probability on the span, r/π_f or $1 - r/\pi_b$. Yoopick does not display this information. The GHPM displays this as a *ternary* (three-way) query, where agents can select whether their probability estimate lies in one of three partitions, as in Figure 1. So, the user selects whether his probability for the span is less than $1 - r/\pi_b$, greater than r/π_f , or in-between. (By monotonicity of prices, $r/\pi_f \geq 1 - r/\pi_b$, with equality only in the limit as $r \rightarrow 0$.) If an agent’s belief lies in the middle partition, presumably they could reduce their bet size or find another span on which to gamble.
- The *marginal* payout probability, which is the sum of the prices on the relevant span after the π_f or π_b of additional quantity. Since agents who are acting straightforwardly will not want to move marginal prices beyond their private valuation, marginal prices could be more informative to decision making. Neither the GHPM nor Yoopick displays this information. Early trials of the GHPM included marginal prices in the interaction interface, but testers found the information confusing when combined with the averaged payout probabilities and so we removed the marginal payout probabilities from later versions of the interface. Even though they were not explicit in the interface, sophisticated traders could still produce marginal prices either by explicitly knowing the pricing rule or by making small tweaks in the number of ticks risked and observing how prices changed. We feel that for a market populated by mathematically adept traders, explicit marginal prices would be a helpful tool.

Finally, the span-based elicitation scheme is arbitrarily expressive. If the users are sophisticated enough to make discriminating judgments over small probabilities, to the point that they can express their actionable beliefs over every contract, then they can still express this sophistication using spans—e.g., by trading spans that contain only one element (one day in the case of the GHPM).

3. PROBLEMS REVEALED

There are two key findings from our study. The first is a large and interesting corpus of trades, which we analyze in Section 4. The second is that we discovered two real-world flaws in the automated market-making concept. These were the *spikiness* inherent in prices and the *liquidity-insensitivity* that made prices in the later stages of the market change too much. We proceed to discuss these flaws in the next two sections, respectively.

3.1 Spikiness of prices across similar events

A phenomenon that quickly arose in the GHPM was how spiky the prices are across events at any snapshot in time. There was extraordinary local volatility between days that one expects should

You selected between October 4th and December 1st, and you're risking 2.76 tickets.

**BET AGAINST: IF THE GHC DOES NOT OPEN IN THIS SPAN, YOU MAKE 3.46 TICKETS.
TAKE THIS BET IF YOU THINK THE GHC HAS LESS THAN A 20.3% CHANCE OF OPENING IN THIS SPAN.**

BET FOR: IF THE GHC DOES OPEN IN THIS SPAN, YOU MAKE 11.33 TICKETS. TAKE THIS BET IF YOU THINK THE GHC HAS MORE THAN A 24.4% CHANCE OF OPENING IN THIS SPAN.

Figure 1: A screenshot of the elicitation query for a user-selected span in the GHPM. The query is ternary because it partitions the user's probability assessment into three parts. The GHC is the Gates and Hillman Centers, the new computer science buildings at Carnegie Mellon. Because of legal concerns, the market used raffle tickets rather than money.

have approximately the same probability. This volatility is far more than could be expected from a rational standpoint—e.g., betting against weekends—and it persisted even in the presence of profit-driven traders whose inefficiency-exploiting actions mitigate the most egregious disparities. Figure 2 is a screenshot of the live GHPM. Spiky prices are clearly evident.

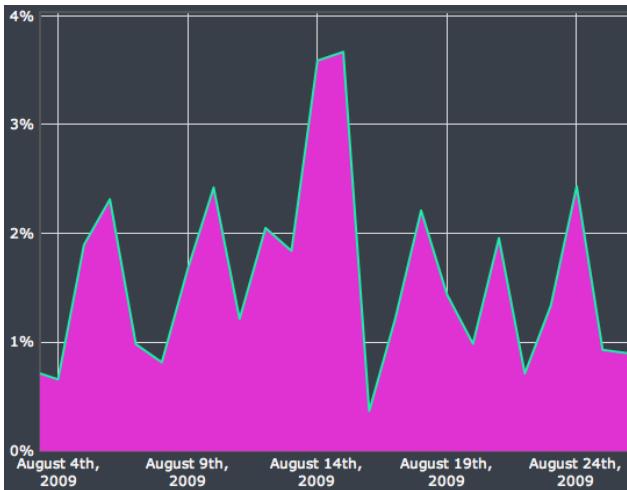


Figure 2: A screenshot of the GHPM that shows the spikiness of prices. The x-axis ranges over a set of potential opening days. The y-axis displays prices (as percentages; e.g., 1% means a price of 0.01).

Why did spikiness occur? Would there have been an automated market maker with a different cost function that would have resulted in a market where prices were not spiky? In this section, we show that on the one hand, spikiness was a consequence of using the LMSR. On the other hand, we show that any market maker that does not induce spikes causes a different problem. Specifically, we show that there is an inherent tension between a quality called *unlockability* and the ability of the market maker to not induce spiky prices. We prove that any market maker is either lockable or spike inducing (or both).

3.1.1 Why were prices spiky?

First, we must formalize what we mean by spikiness. From examining transactional data in the GHPM, we observed that spiky prices arose when agents' bets amplified small differences between individual days with low prices. These small differences arise nat-

urally by agents selecting slightly different intervals to wager on. Recalling that in our market, bets offer agents an identical payout for each day in a selected span, this kind of amplification is entailed if the market maker's price response is convex. This is because for (strictly) convex $f(\cdot)$, if $a < b$ and $x > 0$, then $f(b) - f(a) < f(b + x) - f(a + x)$. We provide a definition of what it means for a market maker to *not* induce spiky prices by that market maker providing a concave response to input in the limit. (Our definition avoids addressing obscure cases where the limit does not exist.) For simplicity, we assume that price functions are twice differentiable (the LMSR is differentiable an infinite number of times). We are interested in spikiness at low prices because with a large event space the price of any one event tends to be low.

Definition 1. A pricing rule is *not spike inducing* (at low prices) if for all i and all \mathbf{q}_{-i} , the price response is concave in the limit as prices approach 0. Formally,

$$\lim_{p_i \rightarrow 0} \frac{\partial^2}{\partial q_i^2} p_i(q_i, \mathbf{q}_{-i}) \leq 0$$

One advantage of using the LMSR in the GHPM was that it is *unlockable*. Recall that a cost function maps a vector of payout quantities to a scalar representing how much traders in aggregate have risked. The pre-image of some cost functions does not cover all of \mathbb{R}^n . If this is the case, then at the boundaries of the defined region the market maker has to force the market not to exit the defined region. The natural way to accomplish this is to “lock” prices at those boundaries, forcing the prices to stay between 0 and 1 and thus maintaining the correspondence between prices and probabilities. With a lockable market maker, additional code needs to be executed to check every interaction, both trades and price quotes, to ensure that prices never go below 0 or above 1. This adds substantial overhead to the market maker.

Definition 2. A market using a cost function with bounded pre-image is *lockable*. A market using a cost function with unbounded pre-image is *unlockable*.

PROPOSITION 1. Every market maker is either lockable or spike inducing (or both).

PROOF. Suppose a market that is both unlockable and not spike inducing. Fix an arbitrary index i and \mathbf{q}_{-i} . Because the market is unlockable, q_i is unbounded in the negative direction, and because pricing rules are strictly monotonically increasing in their elements, there exists a direct correspondence between q_i and p_i . As a result,

we can rewrite the not spike-inducing limit condition as

$$\lim_{q_i \rightarrow -\infty} \frac{\partial^2}{\partial q_i^2} p_i(q_i, \mathbf{q}_{-i}) \leq 0$$

Now select arbitrary $\epsilon > 0$. Then by this limit condition there exists some Q such that

$$\frac{\partial^2}{\partial q_i^2} p_i(q'_i, \mathbf{q}_{-i}) < \epsilon \quad \text{for all } q'_i \leq Q$$

Let $z = \frac{\partial}{\partial q_i} p_i(Q, \mathbf{q}_{-i})$. We have $z > 0$ by monotonicity.

Now, consider the space of all pricing rules $p_i(q, \mathbf{q}_{-i}), q < Q', p_i \in \mathbb{C}^2$ with the property that $p'_i(Q, \mathbf{q}_{-i}) = z$ and $p''_i(q, \mathbf{q}_{-i}) < \epsilon$ (over all relevant q). On their domains, no pricing rules maps to higher values than the simple quadratic function $f(q)$ which is defined by $f''(Q) = \epsilon, f'(Q) = z$ and $f(Q) = p_i(Q, \mathbf{q}_{-i})$. But by the quadratic formula, this function is less than 0 for any q less than

$$\frac{-z + \epsilon Q - \sqrt{(z - \epsilon Q)^2 - 2\epsilon(p_i(Q, \mathbf{q}_{-i}) - zQ + \epsilon Q^2)}}{\epsilon}$$

Since this quadratic function maps to values at least as large as every pricing rule meeting the conditions on the first and second derivatives of the cost function, we know that p_i must be smaller than 0 for q_i less than the above value. Since prices go below 0, q_i is not unbounded. Thus the market is lockable, a contradiction. \square

Lockable market makers have been regarded as inferior because of the overhead involved in locking the market comes without any advantage (Pennock and Sami, 2007)). To our knowledge, this result is the first argument for choosing a lockable market maker over an unlockable market maker like the LMSR.

Spikiness represents a problem because it impacted trader behavior—not only were traders aware of spikiness, but this knowledge influenced their actions. In the next section, we discuss and analyze interviews with traders which suggest that spikiness played a large role in determining the way agents behaved in the GHPM.

3.1.2 Impact on trader behavior

Several successful traders based their strategies entirely around betting against spikes. Rob, a PhD candidate in the Computer Science Department ended with about 256 tickets, finishing in fourth place overall. In our interview with him, he described his strategy as follows:

I knew that the market was presumably figuring out the probabilities of events, and early on, those predictions were very uneven. I supposed some people were setting all their money down on a single day or small set of days, and that this was causing the probability graph to be very "spiky." I bet against the spikes.

Presuming (and I was correct) that as new people entered the market, the spikes would change radically and I'd cash out on the old spikes (making money) and bet against the new spikes.

Of course, on the other side of Rob's actions were traders like Jeff (a pseudonym). Jeff is another PhD student in the Computer Science Department with a background in finance; he worked as a quantitative analyst at a hedge fund before coming to graduate school. A frequent trader, Jeff finished with enough tickets to place himself in the top 15 traders. Of his experience, he said:

It seemed like every time I would make a trade the value [of the bet] would fall a little bit...it was frustrating, like everything I was doing was wrong.

Jeff's bets would fall in value because they would create spikes, which speculators like Rob would quickly sell.

Spiky prices are a problem because they create a disconnect between the user and the elicitation process. Users feel that the spiky prices they observe after interacting with the market maker do not reflect their actual beliefs. This is because users agree only to a specified potential payoff rather than to an explicit specification of prices after their interaction.

Moreover, because the difference between spiky prices and (putatively) efficient prices is so small, traders have little incentive to tie up their capital in making small bets to correct spikiness; there is almost certainly another interval where their actionable beliefs diverge more from posted prices. Our interview with Brian, a PhD student in the Machine Learning Department and the market's best-performing trader, was informative. He described a sophisticated strategy where he would check the future prospects of his current holdings against what he viewed as a risk-free rate of return—for instance, by betting against the building opening on a weekend. If the risk-free rate of return was higher, he would sell his in-the-money holdings and buy into the risk-free asset. So once a spike is small enough, damping it out can be less lucrative than other opportunities.

Finally, to bet against a spike, a trader accepts an equal payout on every other day. But moving the price function by an equal scalar is what caused spiky prices in the first place; that is, every bet against a spiky price has the tendency to amplify other spikes. In summary, in a spike-inducing market, spiky prices are easy to create and virtually impossible to eliminate.

3.2 Liquidity-insensitivity

Recall that the cost function used in the GHPM was

$$C(\mathbf{q}) = b \log \left(\sum_i \exp(q_i/b) \right)$$

and that prices are given by the gradient of this function

$$p_i(\mathbf{q}) = \frac{\exp q_i/b}{\sum_j \exp q_j/b}$$

Defining $\mathbf{1} \equiv (1, 1, \dots, 1)$, it is evident by inspection that

$$p_i(\mathbf{q}) = p_i(\mathbf{q} + \alpha \mathbf{1})$$

for scalar α . Hanson (2003) and Chen and Pennock (2007) present this relation as a property of any arbitrage-free market maker, because it ensures that

$$C(\mathbf{q} + \alpha \mathbf{1}) = C(\mathbf{q}) + \alpha$$

so that buying a guaranteed return of α regardless of the realized outcome should cost α .

A practical interpretation of this result is that the market maker is *liquidity insensitive*, so that quoted responses do not respond to the level of activity seen in the market. This implies that prices change exactly the same amount for a one dollar bet placed at the start of the market (say, at $\mathbf{q} = (0, 0, \dots, 0)$) as after the market maker has matched millions of dollars ($\mathbf{q} = (1000000, \dots, 1000000)$).

This is not the way we think of markets in the real world, operated by humans, as working. As markets grow larger with more frequent trading, they become deeper so that small bets have vanishingly small impact on prices. Liquidity-insensitivity is therefore a failure of current automated market makers.

3.2.1 Impact on trader behavior

Liquidity insensitivity had an impact on traders in the GHPM,

but unlike spikiness, which was publicly visible and a source of frequent consternation, it appears that only the most active and savvy traders were aware of liquidity insensitivity. Brian, the market's best trader, said this about the way he approached the market in its final weeks:

One thing I noticed was that at the end, these small bets would still make big jumps in the prices. So I would try to keep the amount that I bet really small...to try and minimize what would happen to the prices.

So, at least the savviest traders were aware of the disconnect between the automated market maker and the way a traditional market would function.

3.2.2 Relation to spikiness

Though spikiness and liquidity-insensitivity appear quite different, they are actually related. A market maker that is sensitive to liquidity would be able to temper spikiness, because in more liquid (deeper) markets, the market maker could move prices less per each dollar invested. Since spikes are the product of discrepancies in the amount that prices move, if prices move less, spikiness will be diminished.

4. TESTING HYPOTHESES

The large corpus of trades linked to user accounts is a valuable product of the GHPM. In this section, we use this data set to explore questions related to trader behavior and performance, and its impact on prices and information aggregation.

4.1 The GHPM both predicted and reacted to official communications (and lack thereof)

A key question in prediction markets—and one of our driving motivations for developing the GHPM—is whether the market would have predictive power beyond official public communications. In this section, we discuss how the market both reacted to, but also anticipated, the official public communications. We also argue that the market responded to the lack of official communications.

Figure 3 shows how the distribution of prices changed over time and Table 1 shows the officially communicated moving dates. As we explained earlier, the moving day provides an upper bound on the issuance of the occupancy permit because people are not allowed to move into a building without a permit.

Date of Communication	Moving Day	Medium
October 15, 2008	July	Blog
February 14, 2009	August 3rd	E-mail
July 23, 2009	August 3rd	E-mail
July 28, 2009	Approx. August 10th	E-mail

Table 1: Officially communicated moving dates.

We can provide a rough narrative of the market from these two sources. Following some initial skepticism, market prices moved towards the correct prediction, becoming very prescient by the end of November. By then, the exterior framing of the buildings was complete. Over the next several months, the outside appearance of the buildings did not improve measurably, and there were no official communications during this period. Prices reflected this apparent lack of progress.

The weather may have further reinforced traders' beliefs in a delay; the winter of 2008-09 was particularly cold in Pittsburgh and featured the lowest temperatures in fifteen years. Pittsburgh's

average temperature in January 2009 was 22 degrees, six degrees colder than the historic average of 28 degrees. As Figure 4 shows, the market's probabilities for the building opening in early August peaked in late November and steadily fell throughout December and January.

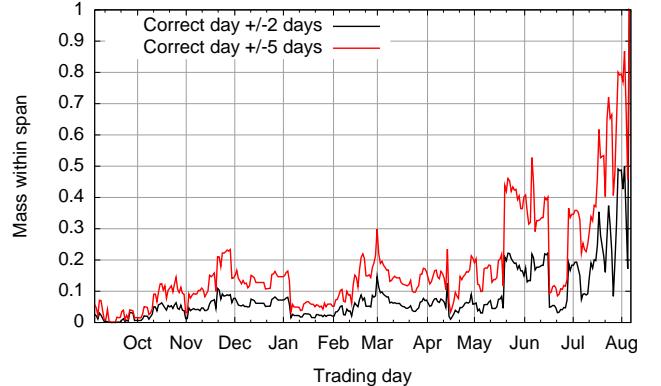


Figure 4: The amount of probability mass around the correct opening day. The x-axis ranges over days the market was open. The lines indicate the mass of spans around the opening day on each trading day; upper line is for August 2 to 12, the lower line for August 5 to 9.

The e-mail of February 14th dramatically shifted market prices, lowering their volatility and bringing them close to the correct opening date. It is clear from examining Figure 3 that the market did not anticipate this e-mail.

The GHPM did, however, predict the e-mail of July 28th that announced a delay in the structure. This was despite official assurances, including an e-mailed moving schedule, that affirmed the August 3rd date. The day before the e-mail announcing the delay arrived, more than 90% of the mass of prices was after August 3rd. So, in July, anyone interested in finding out the opening date would have been better off relying on the predictions from the GHPM than relying on the official communications.

4.2 Self-declared savviness

When traders signed up, they were asked "How savvy do you think you are relative to the average market participant?". They were given five choices, "Much less savvy", "Less savvy", "About the same" (the default selection), "More savvy", and "Much more savvy". Participants were informed that their answer to this question would not impact their payouts or the way they interacted with the market.

Because people are usually over-confident in various settings—and in prediction markets in particular (Forsythe et al., 1999; Graefe and Armstrong, 2008)—it was our expectation that traders would be over-confident in their own abilities relative to others. Instead, we found the opposite.

4.2.1 Background

Prior studies have suggested that overconfidence causes trade in prediction markets. In the standard setting (without a market maker) and with perfectly rational agents, theory provides no-trade results (Milgrom and Stokey, 1982), that are problematic because in real prediction markets trade does occur. Briefly, the no-trade argument is as follows: Speculative trade is zero-sum. A perfectly rational agent knows that no other perfectly rational agent would offer a trade with positive expected value, and so no trade occurs.

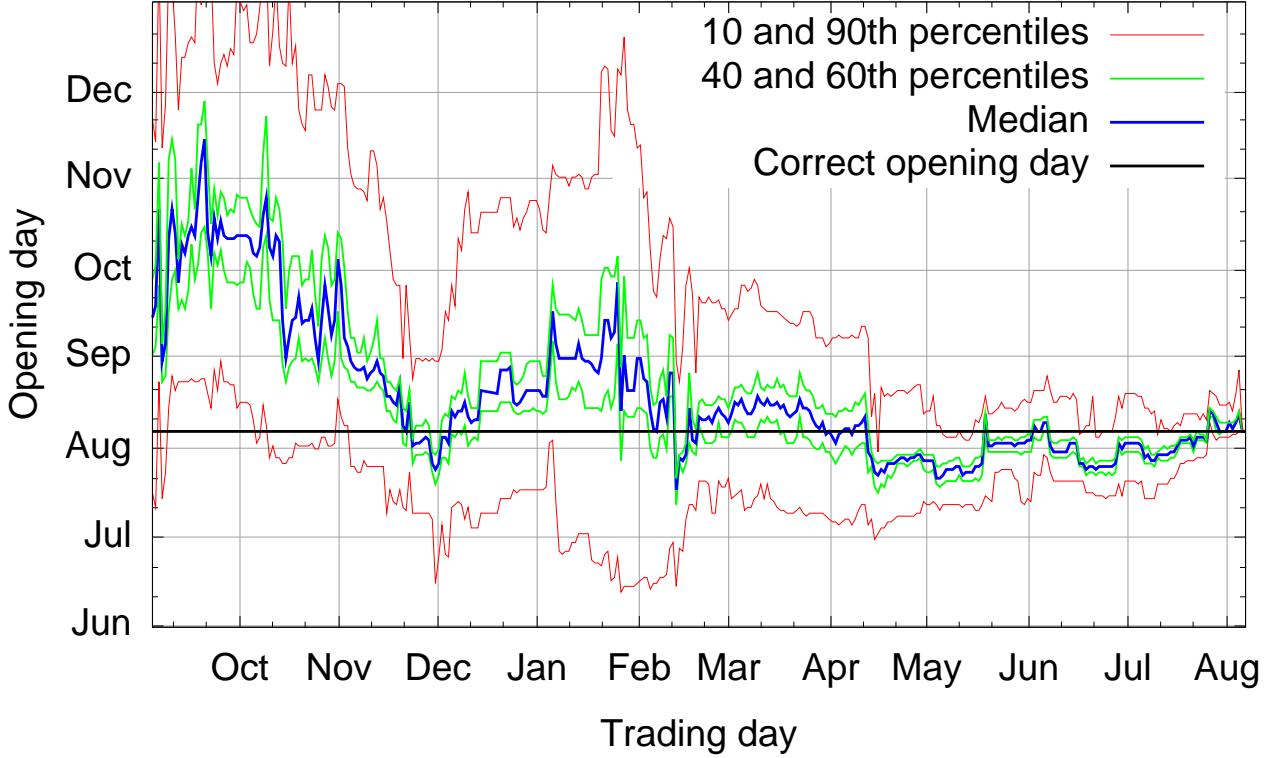


Figure 3: The set of prices offered by the market maker corresponds to a probability distribution. This figure shows the level percentile curves of the probability distribution. The x-axis ranges over the trading days of the market, while the y-axis ranges over the possible opening days (the contracts in the market). The market actually spanned opening dates from April 1, 2009 to March 31, 2010; the y-axis is truncated here for clarity.

One way around the no-trade results is to assert that trade is not speculative but rather a risk-hedging tool. However, stakes in current prediction markets are too small to provide serious hedging opportunities (Wolfers and Zitzewitz, 2006). No traders we spoke to mentioned hedging as a motivation for participating.

A perhaps more plausible explanation of trade in prediction markets is for agents to not be perfectly rational. In particular, if an agent is overconfident in her beliefs or abilities, she may see market prices as errors that she can profitably correct. This was the explanation offered by Forsythe et al. (1999) for traders' actions in the *Iowa Electronic Market*, the longest-running prediction market. It is further backed up by Graefe and Armstrong (2008), who found that traders (incorrectly) believed they could achieve higher payoffs by adjusting consensus market-price estimates in laboratory experiments.

4.2.2 Reported under-confidence

Based on previous studies of over-confidence in markets, we would expect to see most traders rate themselves as at least comparable to the average trader in the market. Table 2 shows our survey results. 77 traders described themselves as less or much less savvy than average, while only 13 traders described themselves as more savvy than average. Most surprisingly, not a single trader listed themselves as much more savvy than the average trader.

Why did we find traders under-confident, instead of over-confident, in their own abilities? Recent research by Moore and Healy (2008) on confidence sheds some light on this issue. They find that

Self-Declared Savviness	Number of Traders
Much Less than Average	30 (17.8%)
Less than Average	47 (27.8%)
Average	79 (46.7%)
More than Average	13 (7.7%)
Much More than Average	0

Table 2: Self-assessment of savviness.

On difficult tasks, people...mistakenly believe that they are worse than others; on easy tasks, people...mistakenly believe they are better than others.

A novel market setting, such as the web-based automated market maker with span-based elicitation we used in the GHPM, is unfamiliar enough to a new trader as to seem potentially difficult. Prior market studies, because they have used traditional market interfaces that even the most casual participant is familiar with, would seem potentially less difficult and therefore would be susceptible to overconfidence.

4.2.3 Traders poorly predicted their own performance

We found that traders' self-reported savviness relative to other traders had little bearing on their relative performance. Table 3 groups traders by self-reported savviness and displays the group medians. The median over all traders was 17.46 tickets, identical to the least-savvy group and within a ticket of the two next-savvy

groups. Ironically, traders identifying themselves as more savvy than the average trader performed more than 10 tickets worse than any other group.

Self-Declared Savviness	Median Tickets
Much Less than Average	17.46
Less than Average	16.78
Average	18.36
More than Average	6.05
Much More than Average	N/A

Table 3: Traders who self-identified as “more savvy than the average participant” in the market had dramatically lower median performance than other traders, while those traders identifying as “much less savvy than the average participant” had the same overall median performance as the general population.

4.3 Trade frequencies suggest a power law

The numbers of bets made by traders appear to closely fit a power law distribution. Figure 5 shows the relationship in terms of the probability of a trader having more than a certain number of trades from our data set, and the best-fitting power law distribution. (We also tried a log-normal distribution and the fit was poor.)

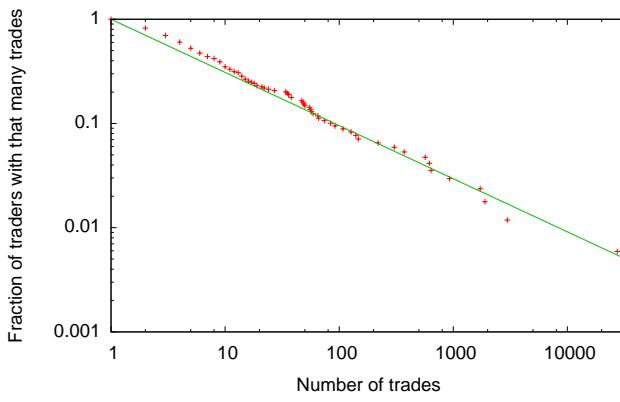


Figure 5: A log – log plot showing the relationship between traders and the number of trades they placed. The straight line shows a power law fit for $\alpha = .51$.

Unfortunately, with only 169 traders we can not assert an appropriate level of statistical significance, so we cannot rule out the data being generated by other distributions. However on a log-log plot, the data does appear to snugly fit the canonical straight line of a power law distribution.

Why might one expect a power law distribution of trade frequency? It seems reasonable to suggest that a trader both makes new bets and sells old bets in proportion to the number of bets she has currently outstanding, with the constraint that she never go under one bet outstanding (in order to collect her two free tickets each week). As Mitzenmacher (2004) discusses, this type of generative model yields a power law distribution.

4.4 Trading by a bot

Conventionally, when we think about prediction markets, we think about a collection of individuals making probability judgments. This is a quality distinct from traditional exchanges, in which automated trading is common and frequent. But as Berg et al. (2001) discuss, trading bots make up a large fraction of the

observed volume in the Iowa Electronic Markets, and must be considered in any sort of qualitative summary of the properties of prediction markets. We found that trade in the GHPM was also dominated by a bot.

This was surprising because we did not make automated trading easy. The GHPM did not use an API, so any trading bot would have to come up with a way to parse the web page and simulate its user’s actions on the page. Jim, a graduate student in the Computer Science Department, took two days to write a trading bot which operated on fitting the market data to a mixture of Gaussian distributions and identified trading opportunities based on deviations from the fit.

The bot made 68.5% of the trades in the market (27,311 of 39,842). The median number of trades placed for all traders was five.

Jim’s bot did well in the market; at its peak it was the second-highest-valued trader. Jim turned his bot off after the e-mail of February 14th and began trading manually. He ended up losing the bulk of his tickets by betting on the building opening earlier than it actually did, finishing 158th of the 210 registered users and 117th out of the 169 traders.

4.5 The Marginal Trader Hypothesis versus the Hayek Hypothesis

How do markets incorporate information and generate good prices? Two competing hypotheses about how markets work are the *Marginal Trader Hypothesis* (MTH) and the *Hayek Hypothesis* (HH). In this section, we discuss how the results of our market support the MTH over the HH.

4.5.1 Background

The MTH holds that a small fraction of market traders are responsible for setting good prices. The other traders have negative expectation and essentially subsidize the experts. The MTH was formulated by the team of researchers responsible for the Iowa Electronic Markets to explain why their 1988 and 1992 presidential election markets worked well (Forsythe et al., 1999; Berg et al., 2001; Oliven and Rietz, 2004). They argued that marginal traders were able to ignore any personal biases they had towards the candidates and act objectively within the markets, setting good prices.

The trade-level data from the IEM is closely held by that research team, so independent verification or analysis of the data is not possible. However, an argument against the MTH being a factor in other markets is that the IEM markets involve political events that people feel strongly about. It seems reasonable that people feel much less strongly about a building opening in August or October than about a presidential election. As a result, there may not be any biases to cloud the judgment of participants, suggesting that the MTH may not be relevant to the GHPM. Furthermore, the 1996 IEM vote-share market failed miserably, with prices diverging sharply from accurate values over the last few weeks of the market (Berg et al., 2001). Where were the marginal traders in 1996?

The Hayek Hypothesis (HH) derives from the work of Smith (1982) analyzing the ideas of Hayek (1945) that markets could function informatively despite the general ignorance of participants in trading environments. Later computational studies have taken this idea literally, populating both continuous double auctions (Gode and Sunder, 1993) and prediction markets (Othman, 2008) with zero-intelligence agents that do not learn or optimize. Surprisingly, these two studies showed that markets composed of zero-intelligence trading bots produced results qualitatively similar to markets composed of human traders. Of course, simply suggesting that markets *could* function without marginal traders is much dif-

ferent than showing that they *do* function without marginal traders. The HH has been supported only by computer simulations, not large-scale market experiments with real people.

4.5.2 What would each hypothesis entail?

We can use the GHPM data to test these two hypotheses. In this section we will discuss what kind of data each of the hypotheses would generate, and in the next section we show what we actually found in the data.

The two hypotheses involve the informational content of markets (how prices become good), rather than their speculative content. Both information provisioning and speculation can affect the final distribution of wealth among the participants. Therefore, “total tickets” is not necessarily a good measure of a trader’s provisioning of information. Tickets could be accumulated by adding valuable information to the market, like from buying the correct span or selling against an incorrect peak. But they could also be accumulated simply by placing a bet each week, or through purely speculative activities that do not increase the amount of (good) information in the market. Consider the following sequence of events:

1. The speculator buys an incorrect span.
2. Another trader buys the same span. This increases the value of the speculator’s bet.
3. The speculator sells their original bet.

The speculator ends without a net position but with an increase of tickets.

Since looking at accumulated tickets alone might not be indicative of the quality of information injected into the market, we came up with another measure, which we dub *Information Addition Ratio (IAR)*. This measure attempts to separate a trader’s return from speculative activities from a trader’s return from information-adding activities. It answers the question “If we see a trader making a one-ticket bet, what is her expected return if she were to hold that bet until the market closes?”. A return of one ticket on each ticket invested is always available to a trader by betting on the entire range of exhaustive contracts. Traders who inject valuable information into the market will have an IAR greater than one, while traders who have a deleterious impact on information will have an IAR of less than one. Essentially, IAR measures how much each trader pushed the price of August 7th, the correct opening day, higher. IAR is an attempt to compress a complex concept into a scalar, and such an enormous dimensionality reduction is inherently lossy. IAR places a focus exclusively on rewarding traders for making bets that raised the price of the correct opening day.

We can gauge how skewed a distribution is at a glance by measuring its Gini coefficient, a standard measure of inequality. Assuming we have the data points $x_1 \leq x_2 \leq \dots \leq x_n$, the Gini coefficient, G , of the sample is given by

$$G = \frac{2 \sum_i i x_i}{n \sum_i x_i} - \frac{n+1}{n}$$

The Gini coefficient ranges from zero to one and can be thought of as a measure of how unequal drawn samples are, with particular sensitivity to large outliers.

The MTH entails a very skewed distribution of IARs (a high Gini coefficient) and for the median trader to have an IAR of much less than one. This is because, according to the MTH, the mass of traders subsidizes a small, elite cadre of knowledgeable traders.

The HH entails a more balanced distribution of IARs (a lower Gini coefficient), and for the median trader to have an IAR of about one. If every trader is essentially the same, some traders should

inject correct information and other traders incorrect information, and the median trader should be close to breaking even.

4.5.3 Data supports the MTH

Figure 6 displays the distribution of IARs, and Table 4 displays the Gini coefficients for the GHPM in context with other distributions. Both the distribution of tickets and the distribution of IARs was heavily skewed and unequal. The median trader had a return of .16 tickets per ticket bet, and 79 traders (47%) did not place a single bet on a span including the correct date, August 7th. This is surprising for several reasons. First, a return of one ticket per ticket bet was *always* available to traders by betting on the entire span. Second, in the ternary elicitation interface, one of the bets offered will always include August 7th, and the other will not, so there was no inherent bias against traders making correct bets. Finally, the median number of bets per trader was five, meaning that the mass of traders made poor judgements several times, not just once. Taken as a whole, these results indicate that the majority of market participants consistently made judgements that hurt the accuracy of the GHPM. On the other side, there was clearly a small and select group of traders responsible for actually making the GHPM produce meaningful prices. Only 37 traders (22%) had an IAR of more than one, and only 13 traders (8%) had an IAR of more than two. Our results therefore strongly support the Marginal Trader Hypothesis over the Hayek Hypothesis.

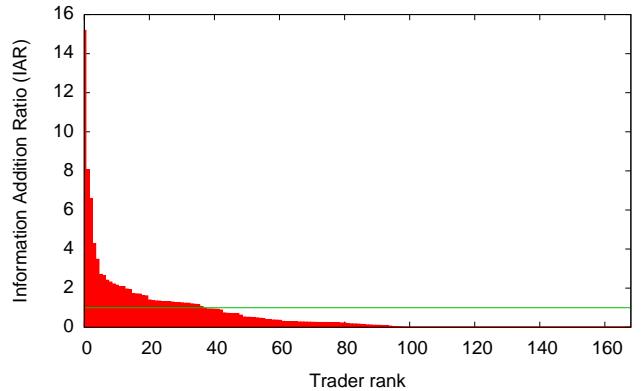


Figure 6: The distribution of IAR of traders ordered by rank. The straight line shows an IAR of 1 (one ticket expected per ticket wagered).

Data Set	Gini Coefficient
Normal Distribution $\mu = 5, \sigma = 1$.113
Denmark Income	.247
Uniform Distribution	.333
United States Income	.408
Log-normal Distribution $\mu = 5, \sigma = 1$.521
GHPM Tickets	.700
Namibia Income	.743
GHPM IARs	.762

Table 4: Gini coefficients are a standard measure of the degree of inequality of a distribution. As this table shows, the distribution of both information (IARs) as well as overall performance (tickets) were extremely unequal. For reference, we include country income inequality coefficients from the United Nations (2008); Denmark had the lowest coefficient and Namibia the highest.

5. DISCUSSION

The Gates Hillman Prediction Market (GHPM) represents, to our knowledge, the largest test faced by automated market making in prediction markets. It was a long-lived market with hundreds of participants, hundreds of events, and tens of thousands of trades. By testing the boundaries of automated market making through actual implementation, we can get a better perspective on where new research should be directed. The GHPM uncovered in practice two significant shortcomings in current market maker designs:

- Spikiness of prices of similar events at any snapshot in time. We proved that this is an unavoidable consequence of using an unlockable market maker. Future research should pay more attention to lockable market makers which may be able to avoid spiky prices. Also, new interaction interfaces could be developed that lead traders to place bets that might better reflect their beliefs while still being simple enough for unsophisticated users.
- Liquidity insensitivity that led to price volatility even in mature stages of the market. As we discussed, future market makers should be liquidity sensitive: they should make prices “stiffer” (i.e., the price changes less as a function of the amount that is bet) in markets where lots of trade volume has been (and will be) placed. One such market maker is discussed in Othman et al. (2010).

From an experimental perspective, our real-world study complements (and is arguably more valuable than) laboratory studies, especially in light of a recent strain of experimental economics which has called into question the reliability of laboratory experiments. Researchers have found that often the behavior seen in the lab, under scrutiny, does not mirror the way people behave outside of the lab (Levitt and List, 2008). We believe that the GHPM was able to avoid these issues because it was long running, and had large numbers of unsupervised participants.

With this unique data set, we were able to provide in-depth study of the market’s microstructure. We showed that the market both predicted and reacted to official communications—and lack thereof. While over-confidence has been suggested as the reason that no-trade theorems get circumvented in prediction markets, we actually found that traders were under-confident. Furthermore, self-confidence did not predict performance; greater-than-average confidence was actually negatively correlated with performance. The data also suggests that the trade frequency across traders follows a power law distribution, and that a trading bot was active and successful. Most significantly, we were able to take two competing hypotheses about how markets work to aggregate information, the Marginal Trader Hypothesis and the Hayek Hypothesis, discuss what one should expect to observe under each hypothesis, and then examine our data to determine that it strongly supports the Marginal Trader Hypothesis.

The GHPM yielded a valuable data set, which we plan to release openly to the community. There are interesting questions about trader behaviors in the market that we only scratched the surface of in this paper. We are especially interested in the application of machine learning techniques towards the data set, studying whether we can reliably cluster and categorize traders from seeing snapshots of their behaviors.

Acknowledgements

We thank Don McGillen and Yahoo! for their generous lead sponsorship of the GHPM prizes. We also recognize Peter Lee, erstwhile CMU Computer Science Department head, as well as Chuck

Schneider and Rick Zuchelli of the CMU Bookstore for their donations. We are also very thankful to David Pennock and Daniel Reeves of Yahoo! Research for discussions in the GHPM’s formative stages. This work was supported by NSF grant IIS-0905390.

References

- S. Agrawal, E. Delage, M. Peters, Z. Wang, and Y. Ye. A unified framework for dynamic pari-mutuel information market design. In *Proceedings of the tenth ACM conference on Electronic Commerce (EC)*, pages 255–264, 2009.
- M. Ali. Probability and Utility Estimates for Racetrack Bettors. *The Journal of Political Economy*, 85(4):803, 1977.
- J. Berg, R. Forsythe, F. Nelson, and T. Rietz. Results from a Dozen Years of Election Futures Markets Research. *Handbook of Experimental Economics Results*, 2001.
- Y. Chen and D. Pennock. A utility framework for bounded-loss market makers. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 49–56, 2007.
- Y. Chen, L. Fortnow, N. Lambert, D. M. Pennock, and J. Wortman. Complexity of combinatorial market makers. In *Proceedings of the 9th ACM conference on Electronic Commerce (EC)*, 2008.
- R. Forsythe, T. Rietz, and T. Ross. Wishes, expectations and actions: a survey on price formation in election stock markets. *J. of Econ. Behavior and Organization*, 39(1):83–110, 1999.
- L. Fortnow, J. Kilian, D. M. Pennock, and M. P. Wellman. Betting boolean-style: a framework for trading in securities based on logical formulas. In *Proceedings of the 4th ACM conference on Electronic commerce (EC)*, pages 144–155, 2003.
- D. Gode and S. Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *J. Pol. Econ.*, pages 119–137, 1993.
- A. Graefe and J. S. Armstrong. Can you beat the market? Accuracy of Individual and Group Post-Prediction Market Judgments. In *Third workshop on prediction markets, part of the ACM conference on Electronic Commerce (EC)*, 2008.
- R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003.
- F. Hayek. The use of knowledge in society. *American Economic Review*, 35:519–530, 1945.
- P. Healy, J. Ledyard, S. Linardi, and J. Lowery. Prediction Market Alternatives for Complex Environments. In *AMMA*, 2008.
- N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on Electronic Commerce (EC)*, 2008.
- S. Levitt and J. List. Homo economicus evolves. *Science*, 319 (5865):909–910, 2008.
- P. Milgrom and N. Stokey. Information, Trade and Common Knowledge. *Journal of Economic Theory*, 26(1):17–27, 1982.
- M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2): 226–251, 2004.
- D. Moore and P. Healy. The trouble with overconfidence. *Psychological review*, 115(2):502–517, 2008.
- K. Oliven and T. Rietz. Suckers are born but markets are made: Individual rationality, arbitrage, and market efficiency on an electronic futures market. *Man. Sci.*, 50(3):336–351, 2004.
- A. Othman. Zero-intelligence agents in prediction markets. In *Proc. of the 7th Int’l joint conf. on Autonomous agents and multiagent systems (AAMAS)*, pages 879–886, 2008.
- A. Othman, D.M. Pennock, D.M. Reeves, T. Sandholm. A Practical Liquidity-Sensitive Automated Market Maker. In *Proceedings of the 11th ACM conference on Electronic Commerce (EC)*, 2010.
- D. Pennock and R. Sami. Computational Aspects of Prediction Markets. In *Algorithmic Game Theory*, chapter 26, pages 651–674. Cambridge University Press, 2007.
- V. Smith. Markets as Economizers of Information: Experimental Examination of the “Hayek Hypothesis”. *Economic Inquiry*, 20 (2):165–79, 1982.
- United Nations. Human Development Report 2007/08, 2008.
- J. Wolfers and E. Zitzewitz. Five Open Questions About Prediction Markets. Technical Report 1975, Institute for the Study of Labor (IZA), Feb. 2006.