Research Support
School of Business Administration

September 1993

ON THE CLUSTERING OF AGENTS' DECISIONS: HERD BEHAVIOR VERSUS THE ENDOGENOUS TIMING OF ACTIONS

Working Paper #727

Faruk Gul
Stanford University
and
Russell Lundholm
University of Michigan

Research Paper #1195

On the Clustering of Agents' Decisions: Herd Behavior versus the Endogenous Timing of Actions

Faruk Gul* and Russell Lundholm**

latest version: September 1993

* Stanford University
** University of Michigan

We thank Ayman Hindy, Steve Huddart, Nahum Melumad, Ennio Stacchetti and Jeffrey Zwiebel for helpful suggestions. Gul gratefully acknowledges support from the Alfred P. Sloan Foundation and the National Science Foundation.

I. INTRODUCTION

In many economic situations agents base their decisions largely on the observed decisions of other agents. Upon observing an empty restaurant we typically conclude that the food is bad. There is evidence that money managers tend to choose their portfolios based on the observed choices of other money managers, currency traders tend to gather the same information as other currency traders, industries are unusually slow to deviate from standard practices, female grouse tend to enter male territories already populated by other female grouse, and analysts tend to bias their forecasts toward the previously-made forecasts of other analysts (see Scharfstein and Stein (1990); Froot, Scharfstein and Stein (1990), Zwiebel (1991), Bikhchandani, Hirshleifer and Welch (1991) and Stickel (1990, 1991), respectively). Deferring to conventional wisdom may not be irrational. Indeed a basic message of information economics is that we can often infer the information of other agents by observing their actions. Unfortunate outcomes can arise, however, when agents defer to others' decisions so much that they ignore their own information completely and simply take the same action that predecessor agents have taken. Such a strategy is known as "herding" and its potential causes have been studied extensively.

The purpose of this paper is to analyze the observed similarity of agents' decisions. In what follows we will distinguish between clustering, which is the observation that agents' decisions tend to be very similar, and herding, which is the statement that clustering occurs because some agents ignore their own information entirely. We provide a formal definition of clustering, illustrate a simple mechanism for clustering that does not involve herding and argue that clustering is likely to be a common phenomenon while herding is not. In particular, we consider a setting where agents choose both an action and the time at which to take the action. We show that allowing agents to choose when to act creates clustering, even when the previously studied motivations for herding are absent. Furthermore, the clustering of agents' choices that results from herding is informationally inefficient -- agents ignore useful information -- whereas the clustering of choices in our model is due to an informational efficiency; agents use their own information and can infer other agents' information as well. Before describing our approach in more detail, however, a brief review of the herding literature is in order.

The existing literature offers two rational explanations for herding, which we label as statistical herding and reputational herding. Examples of statistical herding are given by Bikhchandani, Hirshleifer and Welch (1991), and Banerjee (1990). Consider the situation in Bikhchandani, Hirshleifer and Welch (1991) where agents make binary decisions (accept or reject) in a pre-determined sequence. Each agent has a conditionally independent signal about the value of each choice and can observe the choices of all predecessor agents. Suppose the first two agents receive "high" signals and choose to accept. It is quite possible that the information implicit in the first two agents' actions overwhelms whatever information the third agent might have and hence she too will choose to accept. But now all subsequent agents are in exactly the same position as the third agent; they will each ignore their own information and choose to accept. This type of result, known as an "information cascade," renders the economy informationally inefficient in the sense that useful information is ignored. Note, however, that the result hinges on the binary nature of the agents' choice set. It can be shown that for any fixed number of players, the larger the choice set, the less likely statistical herding is to occur. In particular, if the choice variable is continuous and agents are rewarded according to the proximity of their choice to the full-information optimal choice, then no information goes unused. Each agent in the sequence uses their own information and any information recoverable from the predecessor agents' decisions. Although the agents' choices are closer together than if they were made simultaneously, the economy is informationally efficient.

Examples of reputational herding are given in Scharfstein and Stein (1990), Froot, Scharfstein and Stein (1990), Zwiebel (1991), and Trueman (1991)). Consider the situation in Scharfstein and Stein (1990) where each agent receives a signal about the value of alternative choices, but the signal may or may not be informative. Informative signals have correlated errors while uninformative ones are independent, and agents do not know whether their signal is informative. An agents does not attempt to make the most valuable decision; rather, she attempts to maximize the probability that an outsider will place on the possibility that she is an informed agent (i.e. her reputation). Because informed agents receive signals with correlated errors a subsequent agent maximizes her appearance as an informed agent by taking the same action as the predecessor agent, regardless of her information. Note that this result depends critically on the assumptions that agents' incentives are not aligned with the value of the actual outcome and that informed agents' signals have correlated errors. If an agent wanted only to make the most valuable decision then her own information would influence her decision and if the signal errors were uncorrelated then common decisions would not indicate the presence of informed agents. Reputational herding could be mitigated by a contract that would align the interests of the agent with the value of the outcome to the firm.

In short, the herding literature explains clustering by showing that, for either statistical or reputational reasons, agents rationally ignore their own information and mimic other agents' actions. Note, however, that in both statistical and reputational herding the order in which agents act is given exogenously. We offer a different and arguably more plausible explanation for the clustering of agents' decisions based on the idea that agents choose the timing of their actions strategically. If agents choose when to act then their timing choice may reveal some of their information. Furthermore, if the choice of when to act is informative then so is the choice of when not to act . Thus, the very first actor knows something about the other agents' information by the simple fact that they have not yet acted. The endogenous timing of actions creates an information leak that may enable the first actor to make a more informed decision. While it may appear that the second agent is biasing her action toward the first agent's choice (as in the herding models), we show that the first agent is actually altering her decision toward the forthcoming decision of the second agent. This source of clustering is labeled anticipation. In addition, if the cost of delaying an action is higher for agents with more extreme information, then in equilibrium they choose to act first. Holding aside the improved decision of the first agent, if the most extreme agent acts first and the second agent can recover the first agent's signal by observing the action, then the most extreme differences in the two agents' decisions cannot arise. This source of clustering is labeled ordering. In sum, the strategic timing of actions decreases the expected gap between the actions of the agents when those who have more extreme signals act first (i.e. ordering) and when the those who act first to infer some of the non-acting agents' information (i.e. anticipation).

Many economic situations present a trade-off between waiting for additional information to present itself and acting quickly on the basis of less information. A money manager may learn something about the optimal allocation between stocks and bonds by waiting to observe another manager's allocation choice, but the longer he waits the longer he holds a portfolio that is suboptimal based on his own information. A firm may wait to observe another firm's success with a new product before deciding how vigorously to enter the market, but the delay will cost the firm some market share if it subsequently chooses to enter. The simple discounting of future payoffs creates a delay cost. The tradeoff between more informed decisions and the urgency to make a decision is the main ingredient of our model.[1] In our setting, agents prefer to make decisions that are accurate (in the sense of being close to the full-information decision) and, for a given level of accuracy, they prefer to make decisions with as little delay as possible.

In the next two sections we focus primarily on a game where the cost of delay increases as the value of the unknown variable increases. While none of the previously-studied causes of "clustered" decisions are present, we demonstrate that the agents' decisions are closer together than when the timing of actions is exogenous. We then show

[1] Other papers that demonstrate the cost of delayed decisions include: Hendricks and Kovenock (1989), who study the tradeoff between waiting to see the results of another firm's oil exploration and the cost of delaying the profit if the results are favorable; Bulow and Klemperer (1991), who study how a buyer trades off waiting to get a lower price against the probability that the seller will run out of stock; and in the context of a public goods problem, Bliss and Nalebuff (1984), who show that the agent who suffers most by waiting for the public good is the first to supply it privately.

that clustering occurs for any delay cost that is either a strictly monotone or strictly convex function of the unknown variable. We illustrate this general theorem with two examples from a setting where the first agent's decision remains unchanged as time passes (so there is no anticipation). In one case the cost of delay is higher for agents with more extreme signals, so they forecast first. Clustering occurs because the most extreme news is available to the second agent and so she doesn't make the most extreme errors. However, in another case where the cost of delay is higher for agents with less extreme signals, the second agent makes the most extreme forecast errors and agents' forecasts become dispersed rather than clustered together.

In section IV we discuss two extensions to our model. The first demonstrates that our results hold in a setting where, in addition to time and accuracy, agents are also concerned about their relative performance. The second shows how our results carry over to a model with many agents. We conclude in section V by arguing that, in general, the set of conditions sufficient for clustering are quite mild and, in particular, they are considerably less stringent than the assumptions found in the herding literature.

II. THE MODEL

Consider a model with two agents. Each agent is interested in predicting the future value of a project, denoted by the realization of a random variable W and, holding the accuracy of the prediction constant, would prefer to make her prediction sooner rather than later. Each agent has information about the realization of W; in particular, $W = S_1 + S_2$ and agent i observes the realization s_i (uppercase characters denote random variables and lowercase variables denote their realizations). For simplicity we assume that the S_i's are independent and have a uniform distribution on the interval $I = [0,1]$. Denote agent i's prediction by z_i and the time of the prediction by t_i. Each agent makes only one prediction and the second agent observes the first agent's prediction.

To capture the tradeoff between the accuracy of the prediction and the time at which it is made, assume agent i's utility is given by

$$u(w,z_i,t_i) = - (w - z_i)^2 - \alpha w t_i, \qquad (1)$$

where $\alpha > 0$ is a constant. The utility function trades off the cost of an error in the agent's prediction (the first term) against the cost of delaying the prediction (the second term). Note that, absent some interaction with the other player, there is no reason to delay in making a prediction; the reason an agent may choose to wait is to observe the other agent's prediction. If the forecast of the agent who acts first depends in some way on her realized s_i then by observing this forecast the agent who acts second will be more informed about w. The delay cost is increasing in the realized w, capturing the idea that there is more urgency in forecasting more valuable projects. Later we present a model where the time cost is increasing in the squared deviation of w from its prior expectation, capturing situations where there is greater urgency in predicting extreme realizations in either direction. Finally, α parameterizes the utility function's relative sensitivity to accuracy versus delay.

For simplicity, the extensive form we consider precludes any pre-play communication between the agents. This is noteworthy because the utility of agent i given in (1) does not depend on agent j's actions, so both agents could achieve the highest utility by simply sharing their signals prior to the beginning of the game. Of course, the incentive to exchange information would be eliminated in a game where the players' utilities are decreasing in the accuracy of their opponent's decision. This would be the case in a model of relative performance evaluation, such as in Zwiebel (1991), or in any zero-sum game. Later we modify our model to include a term in agents' utilities that is decreasing in their opponent's accuracy and show that none of our conclusions are altered by the modification.

We focus primarily on the model without a relative performance term to highlight the source of clustering in the simplest possible setting.

An agent's prediction z_i can be interpreted in a number of ways. It may literally be a forecast, as might be the case if the agents were financial analysts or macroeconomists. Alternatively, it may be a more tangible action choice. For example, z_i may be the size of an initial investment in a new project and w the optimal level of investment based on all available information. The agent prefers to make an investment as close to the full-information optimum as possible and, the larger the full-information level of investment, the more costly it is to delay the decision. In general, all that we require of z_i is that it be a one-to-one function of the agent's expectation of W.

Denote the strategy profile of the two agents by $\sigma = (\sigma_1,\sigma_2)$. This set is potentially quite large, but a few observations will greatly limit the set of possible best responses. First, agent i's prediction should minimize the mean squared error of the forecast (the first term in the utility expression) conditional on the agent's signal s_i, the equilibrium strategy profile σ and the elapsed time t. The forecast with this property is $s_i + E(S_j|\sigma,t)$. Second, note that once one agent has made a forecast there is no benefit to the other agent in delaying her forecast any longer. Thus, once one player makes her prediction the other player predicts immediately afterward. With these observations, a strategy for agent i is fully described by a function $t_i{:}I\rightarrow\Re^+$, where $t_i(s_i)$ specifies the latest possible time that agent i will make her forecast (she will forecast earlier if the other agent forecasts before this time). We will refer to the agent who chooses to forecast first as the first agent, although this may be either the agent with signal s_1 or the agent with signal s_2.

A final observation is in order. For games in continuous time, guaranteeing that strategies imply well-defined outcomes entails certain technical difficulties (see Stinchcombe (1988)). So, for example, the strategy that says one player will predict "immediately after" the other player is somewhat vague. To make this precise, our game should be considered as the limit of a series of discrete time games as the length of the time between periods goes to zero. We show in the appendix that there is a unique symmetric equilibrium outcome to the discrete time game, and that the equilibrium outcome converges to the outcome given for the continuous time game.

III. RESULTS

The Symmetric Equilibrium

Most of our attention will focus on the symmetric equilibrium where $t_1(s_i) = t_2(s_i) = t(s_i)$. In particular, we show that there exists a unique symmetric equilibrium. In this equilibrium $t'(s_i) < 0$ and $t(1) = 0$. Consider such a strategy profile.

Note that, because $t(s_i)$ is invertible, the second agent can infer the first agent's signal from the time the first agent made her forecast; denote the inverse of $t(s_i)$ by $s(\tau_i)$. Because $t(s_i)$ is downward sloping, if the game proceeds to time $\overline{\tau}$ without a forecast, each agent knows that the other agent's signal is not in the region $[s(\overline{\tau}),1]$. Thus, if the first agent chooses to forecast at time $\overline{\tau}$ then her forecast is $s_i + s(\overline{\tau})/2$. Finally, because $t(s_i)$ is invertible, finding the $t(s_i)$ that maximizes agent 1's expected utility for the given strategy of agent 2 is equivalent to finding the $s \in I$ that minimizes

$$\int_s^1 \alpha(s_1 + s_2)t(s_2)ds_2 + \int_0^s (s_2 - s/2)^2 ds_2 + \int_0^s \alpha(s_1 + s_2)t(s)ds_2 \qquad (2)$$

for each $s_1 \in I$.

Of course, agent 2 solves the analogous problem. To understand this expression note that for $s_2 \in [s,1]$ agent 2 will forecast first. Thus, for this region agent 1 will forecast immediately after agent 2, get the prediction exactly right, and incur time cost $\alpha(s_1 + s_2)t(s_2)$. This is the first term in (2). For $s_2 \in [0,s)$ agent 1 will forecast first. In this case she will forecast $s_1 + s/2$, her forecast accuracy cost will be $(s_2 - s/2)^2$, and her delay cost will be $\alpha(s_1 + s_2)t(s)$. These are the second and third terms in (2), respectively.

The following proposition gives the unique symmetric equilibrium of this game.

Proposition 1: There exists a unique symmetric Nash equilibrium outcome for the game described above. In this equilibrium agent i predicts $(3/2)s_i$ at time $t(s_i) = (1 - s_i)/6\alpha$ if her opponent has not made a prediction; otherwise she predicts $s_i + (2/3)z_j = s_1 + s_2$ immediately after her opponent's announcement (at time $t[(2/3)z_j]$). If agent i observes that $\tau_j \neq (2/3)z_j$ then agent i forms an arbitary conjecture about the distribution of s_j and forecasts s_i plus the mean of s_j given her new conjecture.[2] (The proof is given in the appendix.)

The intuition for why $t(s_i)$ is decreasing and continuous is straightforward. First, $t(s_i)$ cannot be increasing because it is more costly for agents with higher signal realizations to wait than it is for agents with lower signals, and the gain to waiting is not signal-dependent. Second, there can be no region where $t(s_i)$ is constant; if there were, an agent could wait an arbitrarily small amount of time and gain a strictly positive amount of additional information. Finally, $t(s_i)$ must be continuous because no agent would be willing to wait the strictly positive amount of time represented by the discontinuity to gain an infinitesimal amount of additional information.

The Asymmetric Equilibria

An asymmetric sequential equilibrium for our model is the following. Suppose agent 1's strategy is to make her prediction immediately and agent 2 is willing to wait indefinitely before being the first to forecast: $t_1(s_1) = 0$ and $t_2(s_2) = \infty$ for all s_1 and s_2. In this case agent 1 forecasts $s_1 + 1/2$ and agent 2, after observing agent 1's prediction,

[2] Note that off-equilibrium forecasts (i.e. forecasts at time $\tau_j \neq t[(2/3)z_j]$) play a very limited role in our model. This is because the utility of the first agent does not depend on the forecast of the second agent. Hence, the second agent's off-equilibrium beliefs do not affect the first agent's expected utility calculations and, consequently, there is a multiplicity of off-equilibrium conjectures associated with the unique symmetric equilibrium outcome.

forecasts $s_1 + s_2$. If agent 2 is willing to wait forever before making a forecast then agent 1's best response is to forecast immediately. Similarly, if agent 1 is going to forecast immediately then it is agent 2's best response to wait an instant, observe agent 1's prediction, and then forecast. Another asymmetric equilibrium reverses the roles of agent 1 and agent 2. These equilibria are given in the following proposition.

Proposition 2: Two asymmetric sequential equilibria in the game described above have the properties that agent i predicts $s_i + 1/2$ at time 0 and agent j predicts $s_j + (z_i - 1/2) = s_1 + s_2$ immediately afterward. If agent i fails to predict at time 0 agent j maintains her belief that s_i is distributed uniform on I and continues to wait.

Note that neither agent's strategy depends on her signal, so observing the time at which an agent acts is uninformative. Consequently, this equilibrium yields exactly the same forecasting and timing behavior that occurs if the order of the agent's actions is given exogenously. As such, it serves as a useful benchmark when evaluating the degree of clustering in the symmetric equilibrium created by the endogenous timing of forecasts.[3]

Herding and Clustering

In our economy neither reputational herding nor statistical herding arise. Agents' utilities do not depend on an outsider's perception of their ability, so the second agent has no reason to mimic the first agent in order to influence someone else's assessment of her ability. In addition, each agent's decision variable is chosen from a continuum, so the second agent always uses her own information to improve her decision. Nonetheless, agent's decisions are clustered together.

[3] As in the standard war of attrition problem, the following other equilibria exist: for all $s_i > \bar{s}$, agent i forecasts $s_i + 1/2$ with probability p at time zero and for all $s_i \leq \bar{s}$, agents i and j play a suitably rescaled version of the symmetric equilibrium.

In the previous herding studies, agents' decisions are clustered together in an extreme way: subsequent agents take the same action as the predecessor agent. In these studies there was no need to define a more general concept to capture the notion that "agents' decisions are too close together." For our purposes it is necessary to define a more sensitive metric of clustering. We will say that clustering has occurred when the squared difference between the two agents' predictions in the economy with endogenously-ordered forecasts is smaller in expectation than the squared difference between the two agents' predictions when the forecasting order is exogenously given.[4] Because the forecasts that arise when the order is exogenous are the same as in the asymmetric equilibria to our game, the appropriate benchmark is naturally defined from within the model.

Let $d_{en} = E\{[Z_1 - Z_2]^2\}$ be the expected squared difference between the two predictions when the forecasts are ordered endogenously and d_{ex} be the analogous measure when the forecasts are exogenously ordered, so that our clustering measure is $d_{ex} - d_{en}$. The second agent can always infer the first agent's signal from the first agent's forecast using the relation $z_i = s_i + E(S_j|\sigma,t)$, so the second agent always forecasts $z_j = s_1 + s_2$. Thus, the difference in forecasts is effectively the difference between the second agent's signal and the first agent's forecast of that signal. Denote the first agent's signal by $X = JS_1 + (1-J)S_2$, where $J=1$ if $t(s_1) < t(s_2)$ and $J=0$ otherwise, and the second agent's signal by $Y = (1-J)S_1 + JS_2$. With this, d_{en} can be written as $d_{en} = E\big(E\{[Y - E(Y|X)]^2|X\}\big)$; that is, the mean-squared error of the first agent's forecast of the second agent's signal, averaged over all possible realizations of the first agent's signal. Further, the inner expectation is $Var(Y|X)$, so $d_{en} = E\{Var(Y|X)\}$. When the forecasting order is given exogenously, who forecasts first is uninformative. In this case the first agent's forecast of

[4] More generally, for the models in propositions 1 and 4, the results that follow can be established for any measure of clustering that is increasing in the absolute difference between forecasts (i.e., the mean squared difference, the mean absolute difference etc.). For these models, the absolute difference between endogenously-ordered forecasts is first order stochastically dominated by the absolute difference between exogenously-ordered forecasts.

the second agent's signal is simply the prior mean, so $d_{ex} = E\{[S_i - E(S_i)]^2\} = Var(S_i)$. The clustering measure can now be written as

$$d_{ex} - d_{en} = Var(S_i) - E\{Var(Y|X)\};$$

that is, the difference between the exogenous variance of a signal and the expected variance conditional on the first agent's signal.

A different decomposition of d_{en} will illustrate two sources of clustering. Note that

$$d_{en} = E\{Var(Y|X)\} = Var(Y) - Var\{E(Y|X)\} = Var(Y) - E\{[E(Y|X) - E(Y)]^2\}, \text{ so that}$$

$$d_{ex} - d_{en} = \underbrace{Var(S_i) - Var(Y)}_{\text{ordering}} + \underbrace{E\{[E(Y|X) - E(Y)]^2\}}_{\text{anticipation}}. \qquad (3)$$

Label the first two terms together as ordering and the last term as anticipation. Anticipation represents the change in the first agent's forecast that results from the realization that she is indeed the first agent (i.e. the conditioning on X). This inference potentially informs her that the second agent's signal must not be in certain regions of I. For instance, in the previous model the ex ante forecast of the second agent's signal is $E(Y) = 1/3$.[5] However, when the first agent realizes that she is indeed the first agent she can conclude that the second agent does not have a higher signal than she does, so her optimal forecast is $s_i/2$. In expectation, then, anticipation contributes $\int_0^1 [s_i/2 - 1/3]^2 ds_i = 1/36$ to clustering in the previous model. Anticipation cannot be negative; the first agent's forecast cannot become less informed than the ex ante forecast E(Y). Consequently, unless the first agent's

[5] Note that the forecast is of the second agent's signal as opposed to the agent with signal S_2, which would be 1/2.

forecast is completely insensitive to the passage of time, the inference that the first agent can draw from the second agent's lack of action contributes to clustering.

Ordering is the difference between the ex ante variance of an agent's signal and the variance of the second agent's signal. Thus, if the equilibrium strategy is for agents with more extreme signals to forecast first then the second agent's signal must be less extreme, so it will have a smaller variance. For instance, in the previous model $\mathrm{Var}(S_i) = 1/12$ but $\mathrm{Var}(Y) = \mathrm{Var}(S_i \mid S_i < S_j) = 1/18$, so ordering contributes 1/36 to clustering.

The next proposition shows that clustering is a general phenomenon. In particular, while the delay cost in the previous model is αwt, we show that clustering occurs for delay cost functions of the form $\alpha g(w)t$, where $g(w)$ denotes the function $g: [0, 2] \rightarrow \Re^+$ and is either strictly monotone or strictly convex; that is, when the expected cost of delay is higher for agents with relatively more extreme signals. Further, we show that ordering is positive when $g(w)$ is either strictly monotone or strictly convex, so clustering occurs even absent any anticipation.

Proposition 3 Suppose agent i's objective function is

$$u(w,z_i,t_i) = -(w - z_i)^2 - \alpha g(w)t_i,$$

where $\alpha > 0$ and $g(w)$ is either strictly monotone or strictly convex. In this case i) there exists a symmetric equilibrium such that $t(s_i)$ is strictly quasi-concave;[6] ii) in all symmetric equilibria, $t(s_i)$ is strictly quasi-concave; and iii) in any symmetric equilibrium, ordering is positive, so clustering occurs. (The proof is in the appendix.)

[6] As in proposition 1, agent i forecasts $s_i + E(s_j \mid t(s_j) \geq \tau)$ at time $t(s_i)$ if her opponent has not made a prediction; otherwise she predicts $s_1 + s_2$ immediately after her opponent's announcement.

As seen in the proof, the existence of a strictly quasi-concave $t(s_i)$ depends only on the strict quasi-convexity of g(w). Adding the requirement that g(w) is either strictly monotone or strictly convex allows us to prove that any $t(s_i)$ is strictly quasi-concave and clustering occurs for any symmetric equilibrium. With strictly convex or strictly monotone costs, it is most costly for agents with extreme signals to delay their forecast. Therefore,. the best response to any strategy profile in a symmetric equilibrium is for agents with more extreme signals to forecast first.

If the equilibrium results in agents with relatively more extreme signals forecasting second, then ordering can work against clustering. The next two examples illustrate how ordering can either contribute to or mitigate against clustering. To isolate the effect of ordering, each example uses symmetric delay costs, so that the equilibrium strategy allows no anticipation.

Symmetric Equilibrium with Two-Sided Cost of Delay

Suppose that the cost of delay increases as the squared deviation between W and its prior expectation increases. The idea here is that an agent is eager to act when the value of the unknown variable is extreme in either direction. A money manager's allocation between stocks and bonds is be a good example, where W is the optimal fraction to have invested in stocks given full information and E(W) is the existing fraction. Another example is an analyst's forecast, where it is as valuable to predict extreme decreases in earnings as it is to predict extreme increases. This idea is captured by the following utility function:

$$u(w,z_i,t_i) = -(w - z_i)^2 - \alpha(w - 1)^2 t_i. \qquad (4)$$

We show that a symmetric equilibrium exists for this model, where $t(s_i)$ is symmetric about 1/2, increasing for $s_i \in [0,1/2)$, decreasing for $s_i \in (1/2,1]$ and $t(1) = t(0) = 0$.

Note that as time passes without a prediction each agent learns that the other agent's signal is not in an extreme region of I. However, because $t(s_j)$ is symmetric about the prior mean of s_j, the optimal forecast does not change over time: $z_i = s_i + 1/2$ for all t. Once the first agent makes her prediction, however, the second agent can use the observed forecast and the time that it was made to recover the first agent's signal (as in the model with one-sided cost of delay). The $t(s_1)$ that maximizes agent 1's expected utility for the given strategy of agent 2 is determined by finding, for each $s_1 \in I$, the $s \in I$ that minimizes

$$\int_0^s \alpha(s_1+s_2-1)^2 t(s_2)ds_2 + \int_{1-s}^1 \alpha(s_1+s_2-1)^2 t(s_2)ds_2 + \int_s^{1-s} (s_2-1/2)^2 ds_2 + \int_s^{1-s} \alpha(s_1+s_2-1)^2 t(s)ds_2. \quad (5)$$

As before, agent 2 solves the analogous problem. To understand this expression, note that for $s_2 \in \{[0,s) \cup (1-s,1]\}$ agent 2 will be the first to forecast. In this case agent 1 will get the forecast exactly right and incur only the time cost. This is given by the first two terms in (5). For $s_2 \in [s,1-s]$ agent 1 will be the first to make a prediction. For this case the third term in (5) measures the cost of her forecast error and the fourth term measures her cost of delay. The following proposition gives the equilibrium.

Proposition 4: There exists a symmetric sequential equilibrium to the game with two-sided delay cost. In this equilibrium agent i predicts $s_i + 1/2$ at time

$t(s_i) = -\frac{3}{4\alpha}\log(|2s_i - 1|)$ if her opponent has not made a prediction; otherwise she

predicts $s_i + (z_j - 1/2) = s_1 + s_2$ immediately after her opponent's announcement of z_j. If her opponent forecasts at some time $\tau_j \neq t(z_j - 1/2)$ then agent i forms an arbitrary conjecture about the distribution of s_j and forecasts s_i plus the mean of s_j given her new conjecture. (The proof is in the appendix.)

While the first agent's forecast does not change with the passage of time, so there is no anticipation, clustering still occurs in this model due to ordering. In particular, Var(Y) = $E\{\min[(S_1 - 1/2)^2, (S_2 - 1/2)^2]\} = 1/24$ so the measure of ordering equals 1/12 - 1/24 = 1/12.[7] Thus, the agents' forecasts are closer together in the economy with endogenously-ordered forecasts than when the forecasting order is given exogenously; that is, clustering occurs.

The expression for Var(Y) given above clearly demonstrates the source of the clustering in this model. The equilibrium ordering reveals the most extreme signal first, leaving only the difference between the less extreme signal and the forecast of 1/2. While the ordering of agents has not altered their point forecasts, it does rule out the extreme regions of the joint distribution of S_1 and S_2.

Proposition 3 gave some sufficient conditions for clustering to occur. By eliminating anticipation and inverting the two-sided delay cost of the previous example, the next example demonstrates how ordering can cause forecasts to be dispersed rather than clustered. Suppose that agents were more eager to act when the value of W was closer to its prior mean. This idea is captured by the following utility function:

$$u(w,z_i,t_i) = -(w - z_i)^2 - \alpha[1 - (w - 1)^2]t_i. \qquad (6)$$

It can be shown that a symmetric equilibrium exists to this model where $t(s_i)$ is symmetric about 1/2, decreasing for $s_i \in [0,1/2)$, increasing for $s_i \in (1/2,1]$ and $t(1/2)=0$. As in the previous example, the optimal forecast remains $z_i = s_i + 1/2$ for all t. Unlike the previous example, however, the agent with the less extreme signal is the first to forecast. Thus, Var(Y) = $E\{\max[(S_1 - 1/2)^2, (S_2 - 1/2)^2]\} = 1/8$, so the measure of ordering is 1/12 - 1/8 = - 1/24. We offer this example to illustrate that ordering can be either positive or negative;

[7] To compute this value note that the cummulative distribution function of $v = (s_i - 1/2)^2$ is $2\sqrt{v}$ with support on (0,1/4) and so the probability density function of the minimum of two independent v_i's is $2(1 - 2\sqrt{u})/\sqrt{u}$ with support on (0,1/4).

not because we have a particular economic situation in mind that exhibits this type of delay cost.[8]

Efficiency

In the herding literature the economy is informationally inefficient in the sense that subsequent agents ignore their own signals when the information is still useful in making a superior decision. The observed clustering of behavior in many economic situations is seen as the undesirable outcome of herding. In contrast, our economy is informationally very efficient. Not only do both agents use their own information, but the second agent can recover the first agent's information by observing her forecast and, as long as the delay cost is not completely symmetric, the first agent can partially infer the second agent's signal from the passage of time. Even when the delay cost is two-sided, as long as the cost is higher for agents with more extreme signals, the first agent doesn't make the most extreme forecast errors. In all cases, both agents use their own information and all the information provided by the endogenous variables in the economy.

Another kind of inefficiency is present in the symmetric equilibrium of our model, however. In the symmetric equilibrium each agent trades off her own gain in accuracy with her own cost of delay without considering that the other player would also benefit from an earlier prediction. By delaying their predictions, each agent imposes a negative externality on the other agent.

In contrast to the symmetric equilibrium, there is no delay cost in the asymmetric equilibria, or in a setting where the order of forecasting is given exogenously; the first agent gains nothing from waiting, so she forecasts immediately. The asymmetric equilibria do not strictly dominate the symmetric equilibrium, however, because the first agent's

[8] Another way that anticipation may be zero is if the cost of delay does not depend on the agent's signal. For example, if the utility function is $u(x,z_i,t_i) = -(x-z_i)^2 - \alpha t_i$, then the passage of time does not reveal anything about an agent's signal and the optimal forecast remains $z_i = s_i + 1/2$. This is a standard war of attrition; its symmetric equilibrium strategy at time t is for each agent to mix between forecasting and waiting based on the probability density $f(t) = 12e^{-12\alpha t}$.

forecast is less accurate in the asymmetric equilibria. Nonetheless, the sum of agents' expected utility is higher in the asymmetric equilibria than in the symmetric equilibrium, so there is a sense in which the asymmetric equilibria are superior. In particular, if utility is transferable between agents then the asymmetric equilibria Pareto-dominate the symmetric equilibrium.[9] Suppose that agent 1 forecasts first in the asymmetric equilibrium. Her ex ante expected utility (averaged over all realizations of s_1) is

$$\int_0^1 \int_0^1 - (1/2 - s_2)^2 ds_2 ds_1 = - \frac{1}{12}.$$

Agent 2 forecasts in the next instant and so her expected utility is zero. In the symmetric equilibrium each player's expected utility is given by

$$\int_0^1 \left[- \int_{s_1}^{1} (s_1 + s_2)\frac{(1 - s_2)}{6} ds_2 - \int_0^{s_1} (s_1 + s_2)\frac{(1 - s_1)}{6} ds_2 - \int_0^{s_1} (\frac{s_1}{2} - s_2)^2 ds_2 \right] ds_1$$

$$= \int_0^1 \left[- \frac{(1 - s_1)^2(5s_1 + 1)}{36} - \frac{s_1^3}{12} - \frac{s_1^2(1 - s_1)}{4} \right] ds_1 = - \frac{1}{16}.$$

The sum of expected utilities in the asymmetric equilibria (- 1/12) is greater than the sum of expected utilities in the symmetric equilibrium (-1/8). Note also that the expected utilities do not depend on the players' sensitivity to the cost of delay, as parameterized by α. In particular, reducing the cost of delay does not reduce the public goods problem present in the symmetric equilibrium. Although it becomes less costly to wait as the delay cost diminishes, in equilibrium the second agent will wait longer before the first agent reaches the point where her gain from increased accuracy equals her loss from additional delay.

[9] While the sum of utilities as defined here is larger at the asymmetric equilibrium, there are behaviorally equivalent representations of preferences (for example, the cube of the utility given here) such that the symmetric equilibrium yields the higher sum.

IV. EXTENSIONS

A Model with Relative Performance Evaluation

For the models presented so far, both agents could increase their expected utility by simply sharing their information prior to the beginning of the game. However, if an agent derived some utility from the forecast error of the other agent -- as would be the case if the agent was subject to some type of relative performance evaluation -- then she would no longer find it in her best interest to truthfully share her information. The following utility function captures this idea:

$$u(w,z_i,z_j,t_i) = -(1-\beta)(w - z_i)^2 - \alpha w t_i + \beta(w - z_j)^2, \tag{7}$$

where β is positive and small enough to keep agent i's focus primarily on her own forecast error (it is shown in the appendix that $\beta < 1/18$ satisfies the second order condition). For this model, agent i's utility is increasing in the forecast error of agent j so an agent's offer to truthfully her reveal information prior to the beginning of the game is no longer credible.

Consider the symmetric equilibrium and, as before, consider the strategy profile $t'(s_i) < 0$ and $t(1) = 0$. For each $s_1 \in I$, agent 1's objective is to choose $s \in I$ to minimize

$$\int_s^1 \alpha(s_1 + s_2)t(s_2)ds_2 - \int_s^1 \beta\left(s_1 - \frac{s_2}{2}\right)^2 ds_2$$

$$+ \int_0^s (1-\beta)\left(s_2 - \frac{s}{2}\right)^2 ds_2 + \int_0^s \alpha(s_1 + s_2)t(s)ds_2 - \int_0^s \beta(s_1 - s)^2 ds_2 \tag{8}$$

As before, agent 2 solves the analogous problem. The first, third and fourth terms of (8) are the same as in (2). The second term in (8) captures the effect of agent 2's forecast error when agent 2 is the first to forecast. In this case agent 2 forecasts $(3/2)s_2$ and agent 1 forecasts immediately afterward. The last term in (8) captures the effect of agent 2's forecast error when agent 1 is the first to forecast. In this case agent 2 forecasts $s + s_2$ (which at this point need not be equal to $s_1 + s_2$). Note that agent 1 internalizes the effect of agent 2's off-equilibrium belief through the last term in (8); choosing an $s \neq s_1$ misleads agent 2 by exactly the difference between s and s_1.

Proposition 5: There exists a symmetric sequential equilibrium outcome for the game with relative performance evaluation described above. In this equilibrium agent i predicts $(3/2)s_i$ at time $t(s_i) = (1 - s_i)/6\alpha$ if her opponent has not made a prediction; otherwise she predicts $s_i + (2/3)z_j = s_1 + s_2$ immediately after her opponent's announcement (at time $t[(2/3)z_j]$). If agent i observes that $\tau_j \neq (2/3)z_j$ then agent i continues to believe that $E(S_j) = s(\tau_j)$ and forecasts $z_i = s_i + s(\tau_j)$ immediately. (The proof is in the appendix.)

The equilibrium strategy described in proposition 5 is the same as in the original model given in proposition 1. In this model, at the margin an agent trades off the cost of waiting an instant against the benefit of improving her own forecast and the benefit of harming the other agent's forecast by misleading her. By weighting the agent's own forecast error and the other agent's forecast error as a convex combination, the combined marginal benefit is exactly as in the original model. It is not incentive-compatible for agents to share their information prior to the beginning of the game in this model, yet it exhibits the same behavior as our original model. This justifies our original assumption of no pre-play communication.

An N-Person Game

The basic results of the two-person game will continue to hold in an n-person version of our model, where the future value of the project is now $W = S_1 + S_2 + ... + S_n$. A strategy in this game specifies a set of functions, each specifying the maximum amount of time an agent will wait after the beginning of the game, and then wait after each observed forecast, before making her forecast. Denote a strategy in a symmetric equilibrium by the set $\{t^k(s,y)$: $k \in \{1,2,...,n\}\}$, so that $t^k(s_i,y)$ gives the maximum amount of additional time agent i will wait before forecasting, given that k agents have not yet forecast and the sum of the forecasts made so far is y. As in the two-person game, $t^1(s_i,y) = 0$ for all s_i and y; once everyone else has acted there is no reason to delay.

Note that in the n-person game each subgame that ensues initially and after a forecast is essentially a rescaled version of the two-person game. Two features in particular remain the same. First, for one-sided delay costs, the cost of delaying any increment of time is higher for agents with higher signals. While some of the unknown components of W may be realized, it is still the case that, for agents who have not yet announced, the expected value of W is higher for agents with higher signals. Second, the expected forecast accuracy does not depend on an agent's own signal. An agent's own signal is forecast without error, as are the realized signals an agent observes from previous agents' forecasts.

Using arguments similar to those given in the appendix for the two-person game, these two facts can be used to prove that in a symmetric equilibrium, for all k and y, t^k is strictly decreasing in s. The t^k are non-increasing in s because the expected cost of delay is higher for agents with higher signals. To see why the t^k are strictly decreasing, given that they are not increasing, suppose that for some k, t^k is constant in some region of I. An agent with a signal in this region can gain a strictly positive increase in forecast accuracy by waiting an arbitrarily small amount of time. This contradicts the supposition that the t^k is constant in some region.

Two implications follow from the fact that t^k is decreasing in s for all k and y in the n-person game. First, there will be intervals of time in which no forecasts are made, just as in the beginning of the two-person game, but there will never be a frenzy of forecasting activity. Basically, agents' timing choices are strategic substitutes; a more aggressive choice of when to forecast is met by other agents with less aggressive choices. They benefit by waiting to observe the other agent's forecast. Second, the two sources of clustering in the two-person model, anticipation and ordering, are also present in the n-person model. Because each t^k is invertible, the passage of time is informative. Each agent's forecast will incorporate some knowledge of the subsequent agents' signals, thus moving all forecasts toward the full-information prediction. Furthermore, agents with more extreme signals forecast sooner, so subsequent agents do not make the most extreme forecast errors.

V. CONCLUSION

We have provided a framework in which clustering and herding can be defined formally and have shown that the tradeoff between delayed decisions and more accurate decisions creates clustering without the informational inefficiencies that accompany herding. In this final section we argue that clustering is likely to be a general phenomenon while herding is not.

Given a finite sequence of observed actions, $z_1, z_2, \ldots z_n$, the "natural" assumption is that each agent i knows her own information and the information of all the agents $j \leq i$ at the time she takes her action z_i. That is, she knows the signals (i.e. types) of all agents that preceded her and knows nothing about the signals of agents who have not yet acted, other than what she can deduce from her prior and the realizations of the earlier signals. Thus, at the end of the game, all the agents' information is revealed. This "natural" level of information at each stage of the game corresponds to the level of information that players have in the two asymmetric equilibria discussed in section III.

Using the "natural" assumption as a benchmark, note that both statistical and reputational herding generate clustering because, in equilibrium, the typical agent i knows less than under the natural assumption. In the statistical herding models this is because the binary choice sets provide an insufficient vocabulary to sustain a fully separating equilibrium, given the incentives of the players. At some point in the sequence of decisions, the information available from observing predecessor agents' decisions overwhelms agent i's information. Thus, she ignores her information and, consequently, her decision does not transmit her information to subsequent players. Similarly, in reputational herding the sequence of decisions fails to aggregate information at some point because the incentives of agents are such that they maximize their reputation by pooling with their predecessor agent. Hence, in a 3-person model the third agent would not have complete information about the second agent's signal. Herding can be eliminated by enriching the setting in a way that allows prior agents' information to be transmitted; a finer set of action choices eliminates statistical herding and more appropriately aligned incentives eliminates reputational herding. It is in this sense that we feel that herding is not a particularly robust phenomenon.

In contrast to the herding models, clustering occurs in our model because, in equilibrium, a typical agent i knows more than under the natural assumption. In particular, she knows the exact signals of all agents who have announced before her and she knows that she is the i'th highest signal (or the i'th most extreme signal in the case of two-sided delay cost). Hence, information has leaked. In our model this information leak is a result of the tradeoff between accuracy and delay. Whenever the appropriate marginal calculations for this tradeoff are not identical across agents' different possible signal realizations, the choice of when to act will cause an information leak and may result in clustering. In sum, the herding literature explains clustering by noting that agents may know less than you thought, while we explain clustering by noting that they may know more than you thought.

APPENDIX

Part I of this appendix derives the symmetric equilibrium to a discrete-time version of our game with one-sided time cost and Part II presents the proofs of propositions 1, 3, 4 and 5.

Part I: The Symmetric Equilibrium for a Discrete-Time Model

The first proposition in the text is somewhat vague regarding certain aspects of the equilibrium strategy as they relate to choosing t_i from a continuum (in particular, the idea that the second agent acts "immediately after" the first agent). Here we make these ideas precise by considering a model with the same features as the continuous time model given in section II, but where an agent can act only in discrete time periods. We show that a symmetric equilibrium can be constructed, it is unique, and that as the time between periods goes to zero the equilibrium strategy converges to the strategy in the continuous time model.

Let Δ denote the time between subsequent periods. Thus, an agent who forecasts in period k does so at time $t_i = (k - 1)\Delta$. Agent i's utility if she announces z_i in period k is now

$$- (w - z_i)^2 - \alpha w(k - 1)\Delta. \qquad \text{(a1)}$$

As before, the optimal forecast is $z_i = s_i + E(S_j|s_i,\sigma,t)$. Furthermore, once one agent has forecast there is no additional benefit to waiting, so the remaining agent will forecast in the next period (this is the analog to forecasting "immediately afterward" in the continuous time model). Thus, a symmetric equilibrium is described by a function $t_\Delta(s)$ that specifies the latest time that an agent with signal s will forecast. If an agent with signal s is willing to wait indefinitely for her opponent to forecast then we will write this as $t_\Delta(s) = \infty$.

Replacing Δt with Δ in the proof that t is non-increasing (see part II of this appendix) establishes that t_Δ is non-increasing for all $\Delta > 0$. Now suppose that no type forecasts in period k. If this is the case then either all types have forecast by period k or there exists a first period $\bar{k} > k$ such that $t_\Delta(s) = (\bar{k} - 1)\Delta$. But if such a $\bar{k}$ existed then the agent with signal s would be strictly better off by forecasting in period k -- the time cost is lower and she learns nothing between periods k and $\bar{k}$. Thus, in every period k either some types forecast in that period or both players have announced prior to k (in which case the game is over). Together with the fact that t_Δ is non-increasing, this establishes the existence of a sequence of s^k such that $s^0 = 1$, $s^k < s^{k-1}$ and

$$t_\Delta(s) = \Delta(k - 1) \text{ for } s \in (s^k, s^{k-1}),\ k = 1, 2, ... \tag{a2}$$

The existence of the $t_\Delta(s)$ given in (a2) together with the continuity of utility in the agent's own type imply that the agent with signal s^k is indifferent between acting in period k and waiting to act in period k+1 (provided that the game does not end with probability one by period k). If the game has proceeded to period k without a forecast, then agent 1 with signal s^k knows that s_2 is below s^{k-1} and forecasts in period k. Her expected utility is

$$\left(\frac{1}{s^{k-1}}\right)\int_0^{s^{k-1}} \left(\frac{s^{k-1}}{2} - s_2\right)^2 ds_2 + \left(\frac{1}{s^{k-1}}\right)\int_0^{s^{k-1}} \alpha(s^k + s_2)(k - 1)\Delta ds_2. \tag{a3}$$

The term $\left(\frac{1}{s^{k-1}}\right)$ is the density of s_2, given that no forecast has been made prior to period k. Alternatively, if agent 1 with signal s^k waits to forecast in period k+1 then her expected utility is

$$\left(\frac{s^{k-1} - s^k}{s^{k-1}}\right)\int_{s^k}^{s^{k-1}} \alpha(s^k + s_2)k\Delta\left(\frac{1}{s^{k-1} - s^k}\right)ds_2 +$$

$$\left(\frac{s^k}{s^{k-1}}\right)\left[\frac{1}{s^k}\int_0^{s^k} \left(\frac{s^k}{2} - s_2\right)^2 ds_2 + \frac{1}{s^k}\int_0^{s^k} \alpha(s^k + s_2)k\Delta ds_2\right]. \tag{a4}$$

To understand the first term in (a4), note that with probability $\left(\frac{s^{k-1} - s^k}{s^{k-1}}\right)$ agent 2 forecasts in period k, so agent 1 incurs only the time cost, and in this case the density of s_2 is uniform on the interval $[s^k, s^{k-1}]$. With probability $\left(\frac{s^k}{s^{k-1}}\right)$ agent 2 does not forecast in period k. In this case agent 1 is in a position very similar to when she forecast in period k, except she has waited an additional Δ of time, and now knows that s_2 lies below s^k rather than below s^{k-1}. This is given in the second term in (a4).

Because agent 1 with signal s^k is indifferent between acting in period k or period k+1, the expected utility in (a3) must equal the expected utility in (a4). Equating these two expressions and evaluating the integrals gives

$$\frac{(s^{k-1})^2}{12} + \frac{\alpha(k - 1)\Delta(s^k s^{k-1} + \frac{(s^{k-1})^2}{2})}{s^{k-1}} = \frac{\alpha k\Delta(s^k s^{k-1} + \frac{(s^{k-1})^2}{2})}{s^{k-1}} + \frac{(s^k)^3}{12s^{k-1}},$$

which simplifies to

$$(s^k)^3 + 12\alpha\Delta s^{k-1}s^k + 6\alpha\Delta(s^{k-1})^2 - (s^{k-1})^3 = 0. \qquad \text{(a5)}$$

Thus, s^k is the root to a cubic that is parameterized by s^{k-1} and Δ. There is at most one solution to (a5) because the cubic's first derivative is strictly positive for $s^{k-1} > 0$. Furthermore, at $s^k = s^{k-1}$ the cubic is strictly positive, so the root is strictly less than s^{k-1}. Substituting in $s^k = s^{k-1}$ yields the results that the sequence $\{s^k\}$ is strictly decreasing. Also, for $s^{k-1} \geq 6\alpha\Delta$, the cubic is less than or equal to zero at $s^k = 0$ and greater than zero at $s^k = 1$, so there is exactly one root in the interval [0, 1] for $s^{k-1} \geq 6\alpha\Delta$. Thus, the sequence $\{s^k\}$ is uniquely determined for $s^{k-1} \geq 6\alpha\Delta$. Finally, for $s^{k-1} \leq 6\alpha\Delta$, $s^k \leq 0$, so the sequence$\{s^k\}$ decreases to zero (or lower).

To distinguish between the sequence of points s^{k-1} for k = 1, 2, ... and the generic parameter s^{k-1} to the cubic in (a5), denote the root of (a5) as a function $f(r, \Delta)$ where r corresponds to s^{k-1}. The cubic can now be re-expressed as

$$f(r, \Delta)^3 + 12\alpha\Delta r f(r, \Delta) + 6\alpha\Delta r^2 - r^3 = 0, \tag{a6}$$

where $s^k = f(r, \Delta)$ for $r = s^{k-1}$. Thus, the sequence $\{s^k\}$ defined by $s^0 = 1$ and $s^k = f(s^{k-1}, \Delta)$ for all $k \geq 1$ is uniquely determined. In any symmetric equilibrium this sequence determines the behavior of any $s \neq s^k$. For any $s = s^k$, arbitrarily specify $t_\Delta(s) = \Delta(k - 1)$. Thus, for any arbitrarily specified off-equilibrium path belief, the t_Δ as defined above is the unique equilibrium outcome (up to the behavior of type s^k agents) of the discrete time game.

We will now show that along any sequence $\Delta_n > 0$ such that $\lim \Delta_n = 0$ as $n \to \infty$, $\lim t_\Delta(s) = \dfrac{1 - s}{6\alpha}$ as $n \to \infty$, for all $s \in (0, 1]$. First, fix an $s \in (0, 1]$ and choose a $\bar{\Delta}$ such that $6\alpha\bar{\Delta} \leq s/2$. To see that the function $f: [s/2, 1] \times [0, \bar{\Delta}] \to [0, 1]$ is continuous (jointly in r and Δ), define a sequence (r_n, Δ_n) that converges to $(r, 0)$ as $n \to \infty$ and observe that by the definition of f, the ordered triplet $[f(r_n, \Delta_n), r_n, \Delta_n]$ solves (a6) for each $n = 1, 2, ...$ Further, if $\lim f(r_n, \Delta_n)$ exists as $n \to \infty$, then $[\lim f(r_n, \Delta_n), \lim r_n, \lim \Delta_n]$ solves (a6) as well. Thus, to verify the continuity of f, it is enough to show that $\lim f(r_n, \Delta_n)$ exists as $n \to \infty$. Because the range is compact, if the limit did not exist then two subsequences would exist such that each would converge to a different limit. But each of these limits would constitute a solution to (a6), which contradicts the fact that there is only one root to (a6) for $r \geq 6\alpha\Delta$.

Next we derive an expression that is analogous to the derivative of $t_\Delta(s)$. In particular, solve (a5) for Δ to get

$$\Delta = \frac{(s^{k-1} - s^k)((s^k)^2 + s^k s^{k-1} + (s^{k-1})^2)}{6\alpha(2s^k + s^{k-1})}, \tag{a7}$$

so that

$$\frac{\Delta}{s^k - s^{k-1}} = -\frac{1}{6\alpha}B(s^{k-1}, \Delta), \text{ where } B(s^{k-1}, \Delta) = \frac{(s^k)^2 + s^k s^{k-1} + (s^{k-1})^2}{s^{k-1}(2s^k + s^{k-1})}. \qquad \text{(a8)}$$

Note that (a8) loosely resembles $t'(s) = -1/6\alpha$ in the continuous time model. To establish the continuity of B we again use r to denote the generic value of the parameter s^{k-1} in (a5) in order to distinguish it from the particular point s^{k-1} in the sequence $\{s^k\}$. Recalling that $f(r, \Delta) = s^k$ for $r = s^{k-1}$, B can be expressed as

$$B(r, \Delta) = \frac{f(r,\Delta)^2 + f(r,\Delta)r + r^2}{r[2f(r,\Delta) + r]}. \qquad \text{(a9)}$$

Because f is continuous on $[s/2, 1] \times [0, \bar{\Delta}]$ and the denominator of the B is strictly positive for $r \geq s/2$, B is also continuous on $[s/2, 1] \times [0, \bar{\Delta}]$. Since this domain is compact, for a given Δ, B attains its minimum and maximum. Denote the minimum and maximum of B by m_Δ and M_Δ, respectively.

We now establish bounds for $t_\Delta(s)$ using m_Δ and M_Δ. First, note that by using (a7) we can express the Δ in $t_\Delta(s) = (k - 1)\Delta$ as

$$\Delta = \frac{(s^{k-1} - s^k)B(s^{k-1}, \Delta)}{6\alpha}, \qquad \text{(a10)}$$

which holds for all k = 1, 2, ..., N. Consider the lower bound

$$t^1_\Delta(s) = \frac{(s^1 - s)m_\Delta}{6\alpha}. \qquad \text{(a11)}$$

To verify that this is a lower bound to $t_\Delta(s)$ note that, for $s \in [s^k, s^{k-1})$, $t^1_\Delta(s)$ is at its highest point at s^k. Thus, consider

$$t^1_\Delta(s^k) = \frac{(s^1 - s^k)m_\Delta}{6\alpha} = \sum_{j=2}^{k} \frac{(s^{j-1} - s^j)m_\Delta}{6\alpha}. \qquad \text{(a12)}$$

Because (a10) holds for all k and m_Δ is the minimum of B, each of the k-1 terms in the summation of (a12) is less than or equal to Δ. Thus, $t^l_\Delta(s^k) \leq t_\Delta(s^k)$. Since this bound holds for each s^k, $t^l_\Delta(s) < t_\Delta(s)$ for all s.

As an upper bound consider

$$t^h_\Delta(s) = \frac{(1 - s)M_\Delta}{6\alpha}. \tag{a13}$$

For $s \in [s^k, s^{k-1})$, $t^h_\Delta(s)$ is at its lowest point at s^{k-1}. Thus, consider

$$t^h_\Delta(s^{k-1}) = \frac{(1 - s^{k-1})M_\Delta}{6\alpha} = \sum_{j=1}^{k-1} \frac{(s^{j-1} - s^j)M_\Delta}{6\alpha}, \tag{a14}$$

recalling that $s^0 = 1$. Since each of the k-1 terms in (a14) is greater than or equal to Δ, $t^h_\Delta(s^{k-1}) \geq t_\Delta(s^{k-1})$. Since this bound holds for each s^{k-1}, $t^h_\Delta(s) \geq t_\Delta(s)$ for all s.

The preceding expressions were derived for a given Δ. We now consider the behavior of the $t_\Delta(s)$ as the Δ goes to zero. First, define a sequence (r_n, Δ_n) that converges to (r, 0) as $n \rightarrow \infty$. Note that because f is continuous $\lim f(r_n, \Delta_n) = f(r, 0) = r$ as $n \rightarrow \infty$. This implies that the term $s^1(1)$, which is used in the lower bound $t^l_\Delta(s)$, converges to one as $n \rightarrow \infty$.

We now show that m_Δ and M_Δ converge to one as the sequence of Δ_n converges to zero. By definition, $M_{\Delta_n} = \max B(r_n, \Delta_n)$, for r_n chosen from [s/2, 1]. Because the domain of B is compact, there exists a value $\bar{r}_n$ where B reaches its maximum. Thus $\lim M_{\Delta_n} = \lim B(\bar{r}_n, \Delta_n)$ as $n \rightarrow \infty$. Suppose that $\lim B(\bar{r}_n, \Delta_n) \neq 1$. This would imply that there exists a neighborhood V of 1 and a subsequence $(\bar{r}_{nj}, \Delta_{nj})$ such that none of the elements of the subsequence $B(\bar{r}_{nj}, \Delta_{nj})$ belong to V. Now consider a subsequence of $(\bar{r}_{nj}, \Delta_{nj})$, denoted by $(\bar{r}_{nk}, \Delta_{nk})$, that converges to $(\bar{r}, 0)$. The compactness of the domain of B assures the existence of a convergent subsequence and, by definition, all subsequences of Δ_n converge to zero. Since $f(\bar{r}, 0) = \bar{r}$ and by the continuity of B, $\lim B(\bar{r}_{nk}, \Delta_{nk}) = B(\bar{r}, 0) = 1$. But this is a contradiction. If the subsequence

$B(\bar{r}_{nk}, \Delta_{nk})$ converges to 1 and this sequence is a subsequence of $B(\bar{r}_{nj}, \Delta_{nj})$, then some elements of $B(\bar{r}_{nj}, \Delta_{nj})$ must belong to V. An analogous argument shows that lim $m_{\Delta_n} = 1$ as $n \rightarrow \infty$.

With both the minimum and maximum bounds of B converging to 1 and $s^1(1)$ converging to 1, it follows that as $n \rightarrow \infty$,

$$\lim t^h_\Delta(s) = \lim t^l_\Delta(s) = \frac{(1 - s)}{6\alpha}, \qquad \text{(a15)}$$

which is t(s) of the continuous time model given in proposition 1. Since both the upper and lower bounds are converging to the same number, $t_\Delta(s)$ must also converge to the t(s) given by the continuous time model. This establishes the desired result.

Part II: Proof of Propositions 1, 3, 4, and 5

Parts of the proof of proposition 3 can be used to prove proposition 1, so we present it first.

Proposition 3

First we establish the existence of a symmetric equilibrium when g(w) is either strictly monotone or strictly convex. Consider the case where g(w) is strictly convex but not monotone. Denote by s^* the agent type with the lowest ex ante expected delay cost; that is, $s^* = 1/2 \text{ argmin } g(w)$, where $w \in [0, 2]$. We postulate a symmetric equilibrium t:[0,1] such that t is differentiable on $[0, s^*) \cup (s^*, 1]$, strictly increasing on $[0, s^*)$, strictly decreasing on $(s^*, 1]$ and satisfying either t(0) = 0 or t(1) = 0.

Denote by h(s) a function $h:[\underline{s}, s^*) \rightarrow (s^*, \bar{s}\,]$. This function will be used to identify the agents with signals s and h(s) who will both forecast at the same time. As the proof proceeds, two cases will present themselves, depending on the nature of g(w). In one case $\underline{s} = 0$, $h(0) = \bar{s}$ and t(1) = 0 is the initial condition. In the other, $\bar{s} = 1$, $h(1) = \underline{s}$

and t(0) = 0 is the initial condition. For clarity, we present the entire proof for the former case; the proof for the latter case is symmetric and is discussed briefly at the conclusion of the proof. With this, agent 1's optimization problem is to choose s from [0, s*) to minimize

$$\int_0^s \alpha t(s_2)g(s_1+s_2)ds_2 + \int_{h(s)}^1 \alpha t(s_2)g(s_1+s_2)ds_2 + \int_s^{h(s)} \alpha t(s)g(s_1+s_2)ds_2 + \int_s^{h(s)} \left[s_2 - \frac{h(s)+s}{2}\right]^2 ds_2$$

for all $s_1 \in [0, s^*)$. This yields the first order condition

$$t'(s)\alpha\{G(s_1+h(s)) - G(s_1+s)\} + \frac{[h(s) - s]^2}{4}[h'(s) - 1] = 0, \qquad \text{(a16)}$$

where $G(x) = \int_0^x g(w)dw$. Substituting s_1 for s in (a16) gives

$$t'(s_1)\alpha\{G(s_1+h(s_1)) - G(2s_1)\} + \frac{[h(s_1) - s_1]^2}{4}[h'(s_1) - 1] = 0. \qquad \text{(a17)}$$

Equation (a17) is the equilibrium condition for an agent with $s_1 \in [0, s^*)$. For an agent with $s_1 \in (s^*, 1]$, replace s_1 with $h(s_1)$ in (1) and then substitute s_1 for s to get

$$t'(s_1)\alpha\{G(2h(s_1)) - G(h(s_1)+s_1)\} + \frac{[h(s_1) - s_1]^2}{4}[h'(s_1) - 1] = 0. \qquad \text{(a18)}$$

Both (a17) and (a18) are satisfied if there exists a differentiable $h(s_1)$ such that, for all $s_1 \in [0, s^*)$,

$$G(2h(s_1)) + G(2s_1) = 2G(h(s_1)+s_1). \qquad \text{(a19)}$$

Consider two mutually exclusive cases; either $G(2) \geq 2G(1)$ or the opposite inequality holds. If the inequality is as given then there exists an $\bar{s} \in (s^*, 1]$ such that $G(2\bar{s}) = 2G(\bar{s})$. This follows because $G(x)$ is continuous and $G(2s^*) < 2G(s^*)$ since $g(w)$ is strictly decreasing on $[0, 2s^*)$. Thus, we are in the case where $h(0) = \bar{s}$ and the appropriate initial condition is $t(1) = 0$. If the opposite inequality holds then we are in the case, to be discussed later, where there exists a $\underline{s}$ such that $h(\underline{s}) = 1$ and $t(0) = 0$. To verify the existence of a function $h:[0, s^*) \rightarrow (s^*, \bar{s}]$ that satisfies (a19) for all $s_1 \in [0, s^*)$, implicitly differentiate (a19) to get the first order differential equation

$$h'(s_1) = \frac{g(h+s_1) - g(2s_1)}{g(2h) - g(h+s_1)}. \qquad \text{(a20)}$$

Let $\Omega_\varepsilon = [0, s^* - \varepsilon) \times (s^*, \bar{s}]$. The RHS of (a20) is a continuously differentiable and bounded function from Ω_ε to $\Re$ with a bounded derivative on Ω_ε for a fixed $\varepsilon > 0$. Hence the function h satisfies a Lipchitz condition sufficient to guarantee a unique solution to (a20) on Ω_ε (see Bartle 1976, p. 256). Bartle's theorem requires that the domain be open. This creates no problem in our case since both $g(w)$ and the RHS of (a20) can be extended differentially to $(-\varepsilon, s^* - \varepsilon) \times (s^*, \bar{s})$. To extend h to the domain $[0, s^*)$, let $\varepsilon \rightarrow 0$. This establishes that there exists a unique solution to (a19).

For $s_1 \in (\bar{s}, 1]$ agent 1's optimization problem is to choose s to minimize

$$\int_s^1 \alpha t(s_2) g(s_1+s_2) ds_2 + \int_0^s \alpha t(s) g(s_1+s_2) ds_2 + \int_0^s [\, s_2 - s/2]^2 ds_2,$$

which yields the first order condition

$$t'(s)\alpha\{G(s_1 + s) - G(s_1)\} + s^2/4 = 0. \qquad \text{(a21)}$$

Setting $s=s_1$ yields

$$t'(s_1)\alpha\{G(2s_1) - G(s_1)\} + (s_1)^2/4 = 0. \tag{a22}$$

Since (a22) holds for $s_1 \in (\bar{s}, 1]$ and a solution can be obtained by integrating and using the initial condition $t(1) = 0$, $t(s_1)$ is described by this solution on this region. Further, taking the calculated value of $t(\bar{s})$ from the solution of (a22) gives a boundary condition $t(0) = t(\bar{s})$ for (a17). By integrating (a17) and using this boundary condition, $t(s_1)$ is described on $[0, s^*)$. On $(s^*, 1]$, $t(s_1)$ is described by the relation $t(s_1) = t(h(s_1))$. Finally, define $t(s^*)$ as lim $t(s_1)$ as $s_1 \rightarrow s^*$ for $s_1 \in [0, s^*)\cup(s^*, 1]$. Note that this limit exists, but may be infinite.

We have shown that $t(s_1)$ satisfies the first order condition for all $s_1 \in [0, s^*)\cup(s^*, 1]$. To establish that $t(s_1)$ is a global minimum we will show that for s and s_1 belonging to $[0, s^*)$, when $s_1 < s$ the LHS of (a16) is positive and when $s_1 > s$ the LHS is negative.

The first order condition states that for $s_1 = s$ the LHS of (a16) equals zero. Note that the only term in (a16) involving s_1 is $\{G(s_1+h(s)) - G(s_1+s)\}$ and that $t'(s)$ is positive on $[0, s^*)$. If this term is greater or lesser at values of $s_1 \neq s$ than it is at $s_1 = s$ then the LHS of (a16) will be positive or negative, respectively, for these values. Consider $\frac{\partial}{\partial s_1}$ $\{G(s_1+h(s)) - G(s_1+s)\} = g(s_1+h(s)) - g(s_1+s)$. This derivative equals zero at most once because $g(w)$ is strictly convex. Further, $g(2s) > g(s + h(s))$. If this was not the case then the ordering $g(2s) \leq g(s+h(s)) < g(2h(s))$ would imply that $G(s+h(s)) - G(2s) < G(2h(s)) - G(s+h(s))$, which would contradict (a19) at $s_1 = s$. Thus, $\{G(s_1+h(s)) - G(s_1+s)\}$ is decreasing at $s_1 = s$. This, combined with the facts that $\{G(s_1+h(s)) - G(s_1+s)\}$ is also decreasing at $s_1 = 0$ (by the monotonicity of G) and changes direction at most once, establishes that for all $s_1 < s$, $\{G(s_1+h(s)) - G(s_1+s)\} > \{G(s+h(s)) - G(s+s)\}$. Thus, the LHS of (a16) is positive for all $s_1 < s$. For $s_1 > s$, it is possible that

$\{G(s_1+h(s)) - G(s_1+s)\}$ is increasing in s_1. However, even at $s_1 = s^*$, $\{G(s^*+h(s)) - G(s^*+s)\} < \{G(2h(s)) - G(s_1+h(s))\} = \{G(s+h(s)) - G(s+s)\}$ by the definition of h(s) and (a19). Thus, $\{G(s_1+h(s)) - G(s_1+s)\} < \{G(s+h(s)) - G(s+s)\}$ for $s_1 > s$, so the LHS of (a16) is negative for all $s_1 > s$. This establishes that choosing t(s) for $s < s_1$ or $s > s_1$ would yield a higher value, hence the derived $t(s_1)$ is indeed a global minimum.

Recall that following (a19) we assumed that $G(2) \geq 2G(1)$. The proof when $G(2) < 2G(1)$ is symmetric. In this case $\bar{s} = 1$, so h is defined as $h:[\underline{s}\ , s^*) \rightarrow (s^*, 1]$, $h(1) = \underline{s}$ and the initial condition is t(0) = 0. The equilibrium conditions given in (a17) and (a18) remain the same and an equation analogous to (a22) is obtained for the segment $[0, \underline{s})$. Finally, the proof for the case where g(w) is strictly monotone follows from the construction of the equilibrium $t(s_1)$ on the segment $[\bar{s}\ ,1]$ when g(w) is increasing and from the construction of $t(s_1)$ on the segment $[0, \underline{s})$ when g(w) is decreasing. Thus a symmetric equilibrium exists.

Next we show that when g(w) is either strictly monotone or strictly convex, any symmetric equilibrium t(s) is strictly quasi-concave. Recalling that S_i and S_j are independent, denote the expected squared error of agent i's forecast made at time no later than τ as

$$l(\tau) = \int_{s_j \notin A_j(\tau)} [\, s_j - E(s_j|\tau)]^2 dF(s_j),$$

where $A_j(\tau)$ is the subset of I for which agent j forecasts before τ (recall that when agent j forecasts first, agent i's mean squared error is zero.) Also, denote the expected cost of delay to agent i who forecasts at time no later than τ as

$$c(s_i,\tau) = \int_{s_j \in A_j(\tau)} \alpha t(s_j) g(s_i+s_j) dF(s_j) \ + \int_{s_j \notin A_j(\tau)} \alpha\tau g(s_i+s_j) dF(s_j).$$

Note that the first term represents the delay cost when agent j forecasts first at time $t(s_j)$, so that $t(s_j) \leq \tau$ for $s_j \in A_j(\tau)$. With this, agent i's objective function is to choose τ to minimize $l(\tau) + c(s_i,\tau)$.

To show that $t(s_i)$ is quasi-concave, suppose the contrary (strict quasi-concavity will be shown later). If $t(s_i)$ is not quasi-concave then there exists a $\lambda \in (0,1)$, $\underline{s}$ and $\overline{s}$ such that $t(\lambda\underline{s} + (1-\lambda)\overline{s}) < \min[t(\underline{s}),t(\overline{s})]$. Let $\hat{s} = \lambda\underline{s} + (1-\lambda)\overline{s}$, $\hat{\tau} = t(\hat{s})$, $\underline{\tau} = t(\underline{s})$ and $\overline{\tau} = t(\overline{s})$. By the optimality of $\hat{\tau}$ we have

$$l(\hat{\tau}) + c(\hat{s},\hat{\tau}) \leq \min\{l(\underline{\tau}) + c(\hat{s},\underline{\tau}), l(\overline{\tau}) + c(\hat{s},\overline{\tau})\},$$

which implies

$$l(\hat{\tau}) - l(\underline{\tau}) \leq c(\hat{s},\underline{\tau}) - c(\hat{s}, \hat{\tau}) \qquad \text{(a23)}$$

and

$$l(\hat{\tau}) - l(\overline{\tau}) \leq c(\hat{s}, \overline{\tau}) - c(\hat{s}, \hat{\tau}). \qquad \text{(a24)}$$

Similarly, the optimality of $\underline{\tau}$ and $\overline{\tau}$ imply

$$l(\hat{\tau}) - l(\underline{\tau}) \geq c(\underline{s}, \underline{\tau}) - c(\underline{s}, \hat{\tau}) \qquad \text{(a25)}$$

and

$$l(\hat{\tau}) - l(\overline{\tau}) \geq c(\overline{s}, \overline{\tau}) - c(\overline{s}, \hat{\tau}). \qquad \text{(a26)}$$

We show that either (a25) contradicts (a23) or (a26) contradicts (a24). Consider $c(s_i, \overline{\tau}) - c(s_i, \hat{\tau})$. Denote $A_1 = A_j(\hat{\tau})$, $A_2 = A_j(\hat{\tau})^C \cap A_j(\overline{\tau})$ and $A_3 = A_j(\overline{\tau})^C$, so that A_1, A_2 and A_3 are disjoint sets and their union equals I. With this,

$$c(s_i, \bar{\tau}) - c(s_i, \hat{\tau}) = \int\limits_{s_j \in A_1 \cup A_2} \alpha t(s_j) g(s_i+s_j) dF(s_j) + \int\limits_{s_j \in A_3} \alpha \bar{\tau} g(s_i+s_j) dF(s_j)$$

$$- \int\limits_{s_j \in A_1} \alpha t(s_j) g(s_i+s_j) dF(s_j) - \int\limits_{s_j \in A_2 \cup A_3} \alpha \hat{\tau} g(s_i+s_j) dF(s_j).$$

$$= \int\limits_{s_j \in A_2} \alpha [t(s_j) - \hat{\tau}] g(s_i+s_j) dF(s_j) + \int\limits_{s_j \in A_3} \alpha [\bar{\tau} - \hat{\tau}] g(s_i+s_j) dF(s_j) \qquad \text{(a27)}$$

where the final simplification follows because A_1, A_2 and A_3 are disjoint sets. Recall that by supposition $\hat{\tau} < \min\{\bar{\tau}, \underline{\tau})$ and note that, for $s_j \in A_2$, $t(s_j) \geq \hat{\tau}$. Thus, (a27) is positive; forecasting later yields higher expected delay costs.

Using (a27), $[c(\bar{s}, \bar{\tau}) - c(\bar{s}, \hat{\tau})] - [c(\hat{s}, \bar{\tau}) - c(\hat{s}, \hat{\tau})]$ can be written as

$$\int\limits_{s_j \in A_2} \alpha [t(s_j) - \hat{\tau}][g(\bar{s}+s_j) - g(\hat{s}+s_j)] dF(s_j) + \int\limits_{s_j \in A_3} \alpha [\bar{\tau} - \hat{\tau}][g(\bar{s}+s_j) - g(\hat{s}+s_j)] dF(s_j). \qquad \text{(a28)}$$

For strictly monotone or strictly convex g, either $g(\bar{s}+s_j) > g(\hat{s}+s_j)$ or $g(\underline{s}+s_j) > g(\hat{s}+s_j)$. If $g(\bar{s}+s_j) > g(\hat{s}+s_j)$, (a28) is positive, implying that $[c(\bar{s}, \bar{\tau}) - c(\bar{s}, \hat{\tau})] > [c(\hat{s}, \bar{\tau}) - c(\hat{s}, \hat{\tau})]$ and (A26) contradicts (A24). Alternatively, if $g(\underline{s}+s_j) > g(\hat{s}+s_j)$ then, substituting $\underline{s}$ for $\bar{s}$ and $\underline{\tau}$ for $\bar{\tau}$ yields another positive (a28), and $[c(\underline{s}, \underline{\tau}) - c(\underline{s}, \hat{\tau})] > [c(\hat{s}, \underline{\tau}) - c(\hat{s}, \hat{\tau})]$ and (A25) contradicts (A23). This establishes that $t(s_i)$ is quasi-concave.

To show that $t(s_i)$ is <u>strictly</u> quasi-concave, suppose to the contrary that there exists $\underline{s}$ and $\bar{s}$ such that for $s_i \in (\underline{s}, \bar{s})$, $t(s_i) = \hat{\tau}$. The optimality of $s_i \in (\underline{s}, \bar{s})$ implies that

$$l(\hat{\tau}) - l(\hat{\tau} + \Delta\tau) \leq c(s_i, \hat{\tau} + \Delta\tau) - c(s_i, \hat{\tau}) \qquad \text{(a29)}$$

for at $\Delta\tau > 0$. However, $l(\hat{\tau})$ integrates over $A_j(\hat{\tau})^C$ -- it includes the region $(\underline{s}, \bar{s})$ -- while $l(\hat{\tau}+\Delta\tau)$ integrates only over $A_j(\hat{\tau}+\Delta\tau)^C$ -- it excludes the $(\underline{s}, \bar{s})$ region. Hence, since the integrand is always positive, there exists $\varepsilon > 0$ sukch that $l(\hat{\tau}) - l(\hat{\tau}+\Delta t) > \varepsilon$ for all $\Delta\tau > 0$.

Further, $c(s_i, \hat{\tau}+\Delta\tau)$ is continuous and increasing in $\Delta\tau$, so there exists a $\Delta\tau$ sufficiently small to make $c(s_i,\hat{\tau}+\Delta\tau)$ - $c(s_i, \hat{\tau})$ sufficiently close to zero such that (a29) is contradicted. Thus, $t(s_i)$ is strictly quasi-concave.

Finally, we show that the strict quasi-concavity of $t(s_i)$ implies that ordering is positive. Because anticipation is non-negative by definition (see (3) in the text), establishing that ordering is positive is sufficient to establish that clustering occurs. First, note that $Var(S_i) = 1/12$, so ordering is positive if $Var(Y) < 1/12$. By the strict quasi-concavity of $t(s_i)$, when the first agent conditions her belief on X she knows that the second agent's signal, Y, is uniformly distributed on an interval $[a(\gamma), a(\gamma) + \gamma]$ and, as the notation implies, the interval is uniquely determined by its length γ. The cummulative distribution function of Γ (with realizations denoted by γ) is $P(\Gamma \leq \gamma) = P(s_1 \in [a(\gamma), a(\gamma) + \gamma]) * P(s_2 \in [a(\gamma), a(\gamma) + \gamma]) = \gamma^2$, and so the density of Γ is 2γ on $\gamma \in [0, 1]$. With this,

$$Var(Y) = \int_0^1 \int_{a(\gamma)}^{a(\gamma)+\gamma} (y - \mu)^2 (1/\gamma) dy 2\gamma d\gamma,$$

where $\mu = E(Y)$. To find μ, note that $E(Y|\gamma) = a(\gamma) + \gamma/2$, so

$$\mu = E\{E(Y|\Gamma)\} = \int_0^1 [a(\gamma)+\gamma/2] 2\gamma d\gamma = 2 \int_0^1 a(\gamma)\gamma d\gamma + 1/3. \tag{a30}$$

Now note that $Var(Y)$ can be written as $E(Y - 1/2)^2 - (1/2 - \mu)^2$, so

$$Var(Y) = \int_0^1 \int_{a(\gamma)}^{a(\gamma)+\gamma} (y - 1/2)^2 (1/\gamma) dy 2\gamma d\gamma \ - (1/2 - \mu)^2. \tag{a31}$$

Integrating (a31) with respect to y and simplifying gives

$$\mathrm{Var}(Y) = 2\int_0^1 a(\gamma)\gamma[a(\gamma) + \gamma - 1]d\gamma + 1/12 - (1/2 - \mu)^2. \tag{a32}$$

Note that the first term (involving an integral) and the last term in (a32) are non-positive. The first term is zero only if $a(\gamma) = 0$ or $a(\gamma) + \gamma = 1$ for all γ. By the strict quasi-concavity of $t(s_i)$ it cannot be the case that $a(\gamma) = 0$ for some of the domain and $a(\gamma) + \gamma = 1$ for the remaining domain. But, from (a30), when $a(\gamma) = 0$ for all γ, $\mu = 1/3$ and when $a(\gamma) + \gamma = 1$ for all γ, $\mu = 2/3$. Thus, when the first term is zero, the last term is strictly negative. Hence, $\mathrm{Var}(Y) < 1/12$.

Proof of Proposition 1

This proposition is a special case of proposition 3. By substituting w for g(w) in (a22), so that $G(2s_1) - G(s_1) = \frac{3(s_1)^2}{2}$, we get the equilibrium condition

$$t'(s_1) = -\frac{1}{6\alpha}. \tag{a33}$$

Integrating (a33) and using the initial condition $t(1) = 0$ gives

$$t(s_1) = \frac{1 - s_1}{6\alpha}. \tag{a34}$$

The solution given in (a34) is unique within the class of differentiable strategies. Next we show that any symmetric equilibrium to the game with $g(w) = w$ must be strictly decreasing and differentiable, so (a34) describes the unique symmetric equilibrium. First, to show that $t(s_i)$ is strictly decreasing it is sufficient to note that $g(s_i+s_j)$ is strictly increasing in s_i so, from the proof of proposition 3, (a28) is positive and, if $t(s_i)$ were non-decreasing, (a26) would contradict (a24).

Next we show that $t(s_i)$ is continuous. With it established that $t(s_i)$ is strictly decreasing, we can express agent 1's optimization problem as minimizing

$$L(s_1,s) = \int_s^1 \alpha(s_1 + s_2)t(s_2)ds_2 + \int_0^s (s_2 - s/2)^2 ds_2 + \int_0^s \alpha(s_1 + s_2)t(s)ds_2 \qquad \text{(a35)}$$

where the second argument of L is agent's choice variable. By relying on the invertibility of t(s), choosing s is equivalent to choosing t.

Suppose t(s) is discontinuous at s_1 and, without loss of generality, suppose that the discontinuity takes the form $t(s_1) > \lim t(s_1 + \varepsilon)$ as $\varepsilon \downarrow 0$. We show that if this is the case then by forecasting a little sooner agent 1 can reduce her cost of delay by a positive amount without reducing her forecast accuracy. In particular, for agent 1 with signal s_1, $L(s_1,s_1)$ equals

$$\int_{s_1}^1 \alpha(s_1 + s_2)t(s_2)ds_2 + \int_0^{s_1} (s_2 - s_1/2)^2 ds_2 + \int_0^{s_1} \alpha(s_1 + s_2)t(s_1)ds_2. \qquad \text{(a36)}$$

However, if agent 1 chooses to predict at time $t(s_1+\varepsilon)$ then $L(s_1,s_1+\varepsilon))$ equals

$$\int_{s_1+\varepsilon}^1 \alpha(s_1 + s_2)t(s_2)ds_2 + \int_0^{s_1+\varepsilon} (s_2 - (s_1+\varepsilon)/2)^2 ds_2 + \int_0^{s_1+\varepsilon} \alpha(s_1 + s_2)t(s_1 + \varepsilon)ds_2. \qquad \text{(a37)}$$

As $\varepsilon \downarrow 0$ the first two terms in (a37) converge to the first two terms in (a36) but the third term converges to

$$\int_0^{s_1} \alpha(s_1 + s_2)\lim[t(s_1 + \varepsilon)]ds_2. \qquad \text{(a38)}$$

Because t(s) is positive and decreasing, and by the discontinuity of t(s) at s_1, (a38) is strictly less than the third term in (a36). This contradicts the optimality of t(s) at s_1.

Finally, we show that $t(s_i)$ is differentiable. If agent 1 with signal s_1 chooses $s=s_1$, it must be the case that for all $\varepsilon > 0$

$$\frac{L(s_1,s_1+\varepsilon) - L(s_1,s_1)}{\varepsilon} \geq 0. \qquad \text{(a39)}$$

Similarly, if agent 1 with signal $s_1 + \varepsilon$ chooses $s{=}s_1 + \varepsilon$, it must be the case that for all $\varepsilon > 0$

$$\frac{L(s_1{+}\varepsilon,s_1) - L(s_1{+}\varepsilon,s_1{+}\varepsilon)}{\varepsilon} \geq 0. \qquad \text{(a40)}$$

By substituting in the component terms of L and taking the difference term by term, (a39) is expressed as

$$\frac{1}{\varepsilon}\left[\int_{s_1+\varepsilon}^{1} \alpha(s_1{+}s_2)t(s_2)ds_2 - \int_{s_1}^{1} \alpha(s_1{+}s_2)t(s_2)ds_2\right] + \frac{1}{\varepsilon}\left[\int_0^{s_1+\varepsilon} (s_2 - \frac{s_1{+}\varepsilon}{2})^2 ds_2 - \int_0^{s_1} (s_2{-}\frac{s_1}{2})^2 ds_2\right]$$

$$+\frac{1}{\varepsilon}\left[\int_0^{s_1+\varepsilon} \alpha(s_1 + s_2)t(s_1{+}\varepsilon)ds_2 - \int_0^{s_1} \alpha(s_1 + s_2)t(s_1)ds_2\right] \geq 0. \qquad \text{(a41)}$$

As $\varepsilon \downarrow 0$, the limit of the first term in (a41) is $-2s_1\alpha t(s_1)$, the limit of the second term is $\frac{s_1^2}{4}$ and the limit of the third term is $2s_1\alpha t(s_1) + \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1{+}\varepsilon) - t(s_1)}{\varepsilon}\right]$. Thus, the limit of (a39) as $\varepsilon \downarrow 0$ is

$$\frac{s_1^2}{4} + \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1{+}\varepsilon) - t(s_1)}{\varepsilon}\right] \geq 0. \qquad \text{(a42)}$$

A similar exercise yields the limit of (a40) to be

$$-\frac{s_1^2}{4} - \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1{+}\varepsilon) - t(s_1)}{\varepsilon}\right] \geq 0. \qquad \text{(a43)}$$

Because the LHS of (a43) is of opposite sign of the LHS of (a42), the only way to satisfy both inequalities is for each to equal zero. This yields $\lim\left[\frac{t(s_1{+}\varepsilon) - t(s_1)}{\varepsilon}\right] = -\frac{1}{6\alpha}$.

The conditions in (a39) and (a40) are given for $\varepsilon > 0$. If $\varepsilon < 0$ the inequalities in both conditions are reversed (as is the limit direction) and hence the inequalities in (a42) and (a43) are reversed. As before, the only way both conditions could be satisfied is for each to equal 0. Thus, $\lim\left[\frac{t(s_1+\varepsilon) - t(s_1)}{\varepsilon}\right] = -\frac{1}{6\alpha}$ regardless of the direction that the limit is taken; consequently, t(s) is differentiable.

Thus, the unique symmetric equilibrium strategy for the model with one-sided delay cost is strictly decreasing and differentiable. Hence, the solution given in proposition 1 is the unique symmetric equilibrium.

Proof of Proposition 4

Proposition 4 is a special case of proposition 3. The solution to (a20) is h(s) = 1-s (which we guessed based on the symmetry of g(w)), and $s^* = 1/2$. Substituting $1 - s_1$ for $h(s_1)$ and $(w - 1)^2$ for g(w) gives the equilibrium condition

$$t'(s_1) = -\frac{3}{2\alpha(2s_1 - 1)}. \tag{a45}$$

Integrating (a45) and using either initial condition t(1) = 0 or t(0) = 0 gives

$$t(s_1) = -\frac{3}{4\alpha}\log(|2s_1 - 1|). \tag{a46}$$

Proof of Proposition 5

Differentiating equation (8) in the text with respect to s gives the first order condition

$$0 = \beta(s_1 - \tfrac{s}{2})^2 + \frac{(1-\beta)s^2}{4} + \alpha(s_1 + \tfrac{s}{2})st'(s) + 2s\beta(s_1 - s) - \beta(s_1 - s)^2.$$

Solving for t'(s) and evaluating at $s = s_1$ gives

$$t'(s_1) = -\frac{1}{6\alpha};$$

integrating and using the initial condition gives

$$t(s_1) = \frac{(1-s_1)}{6\alpha}.$$

To verify that this is a global minimum, substitute the derived t(s) into the objective function and note that the first derivative is $\left(\frac{1 - 18\beta}{6}\right)\left(s^2 - s_1 s\right)$. This quadratic has a root at $s=s_1$ and at $s=0$. For $\beta<1/18$, the first derivative is negative between the two roots and positive elsewhere. This guarantees that $s=s_1$ is the unique minimum when $\beta<1/18$.

REFERENCES

Banerjee Abhijit, 1992, "A Simple Model of Herd Behavior," The Quarterly Journal of Economics, 107:797-817.

Bartle, R. The Elements of Real Analysis, Wiley: New York, 1976.

Bikhchandani, Sushil, David Hirshleifer and Ivo Welch, 1992, "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades," Journal of Political Economy, 100:992-1026.

Bliss, Christopher and Barry Nalebuff, 1984, "Dragon-Slaying and Ballroom Dancing: The Private Supply of a Public Good," Journal of Public Economics, 25:1-12.

Bulow, Jeremy and Paul Klemperer, 1991, "Rational Frenzies and Crashes," NBER Technical Paper No. 112.

Froot, Kenneth, David Scharfstein and Jeremy Stein, 1990, "Herd on the Street: Informational Inefficiencies in a Market With Short-Term Memory," NBER Working Paper No. 3250.

Hendricks, Kenneth and Dan Kovenock, 1989, "Asymmetric Information, Information Externalities, and Efficiency: The Case of Oil Exploration," RAND Journal of Economics, 20:2, 164-82.

Scharfstein, David and Jeremy Stein, 1990, "Herd Behavior and Investment," American Economic Review 80:3, 465-79.

Stickel, Scott, 1990, "Predicting Individual Analyst Earnings Forecasts," Journal of Accounting Research, 28, 409-17.

Stickel, Scott, 1992, "Reputation and Performance Among Security Analysts," The Journal of Finance, 48:1811-36.

Stinchcombe, M., 1988, "Maximal Strategy Sets for Continuous-Time Game Theory," University of California-Berkeley Working Paper.

Trueman, Brett, 1991, "Analyst Forecasts and Herding Behavior," University of California-Berkeley Working Paper.

Zwiebel, Jeffrey, 1991, "Corporate Conservatism, Herd Behavior and Relative Compensation," Stanford University Working Paper.