Endogenous Timing and the Clustering of Agents' Decisions

Faruk Gul

Northwestern University

Russell Lundholm

University of Michigan

This paper presents a model in which agents choose an action and a time at which to take the action. We show that when agents choose when to act, their decisions become clustered together, giving the appearance of an information cascade, even though information is actually being used efficiently. This occurs because the passage of time allows the first acting agent to anticipate something about the second agent's information, and for a large class of delay cost functions, the equilibrium orders agents in such a way that the most extreme information is revealed first.

I. Introduction

In many economic situations, agents base their decisions largely on the observed decisions of other agents. On observing an empty restaurant, one typically concludes that the food is bad. There is evidence that money managers tend to choose their portfolios on the basis of the observed choices of other money managers, currency traders tend to gather the same information as other currency trad-

This paper was formerly titled "On the Clustering of Agents' Decisions: Herd Behavior versus the Endogenous Timing of Actions." We thank Ayman Hindy, David Hirshleifer, 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.

[*Journal of Political Economy*, 1995, vol. 103, no. 5]
© 1995 by The University of Chicago. All rights reserved. 0022-3808/95/0305-0005$01.50

ers, industries are unusually slow to deviate from standard practices, and analysts tend to bias their forecasts toward the previously made forecasts of other analysts (see, respectively, Scharfstein and Stein 1990; Froot, Scharfstein, and Stein 1992; Zwiebel 1991; Bikhchandani, Hirshleifer, and Welch 1992; Stickel 1990, 1992). Deferring to conventional wisdom may not be irrational. Indeed a basic message of information economics is that one 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 result is known as an "information cascade," and its potential causes have been studied extensively.

The purpose of this paper is to analyze the observed similarity of agents' decisions. In this paper we shall distinguish between *clustering,* which is the observation that agents' decisions tend to be unexpectedly similar, and an *information cascade,* which is a situation in which agents ignore their own information entirely.[1] We provide a formal definition of clustering, illustrate a simple mechanism for clustering that does not involve an information cascade, and argue that clustering is likely to be a common phenomenon even without information cascades. In particular, we consider a setting in which 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 an information cascade are absent. Furthermore, the clustering of agents' choices that results from an information cascade 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 we describe our approach in more detail, however, a brief review of the information cascade literature is in order.

The existing literature offers two rational explanations for ignoring information, which we label as *statistical cascades* and *reputational cascades.* Examples of statistical cascades are given by Banerjee (1992), Bikhchandani et al. (1992), Welch (1992), and Chamley and Gale (1994). Consider the situation in Bikhchandani et al. (1992), where agents make binary decisions (accept or reject) in a predetermined sequence. Each agent has a conditionally independent signal about

[1] The observation that agents tend to make similar decisions has frequently been labeled "herding." In the contexts studied, herding has been inefficient; useful information has been ignored or suboptimal actions taken. Because the similarity of agents' actions is *not* inefficient in our model, we use the more neutral term "clustering."

the value of each choice and can observe the choices of all predecessor agents. Suppose that 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 renders the economy informationally inefficient in the sense that useful information is ignored. Chamley and Gale present a model similar to that of Bikhchandani et al., but one in which agents choose the timing of their actions strategically, as in our analysis. They note that statistical cascades also occur within their framework. Note, however, that these results hinge on the finiteness of the agents' choice set. In equilibrium, agents with slightly different information do not have the opportunity (or incentive) to take slightly different actions. In general, for any fixed number of players, the larger the choice set, the less likely statistical cascades are 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 (Lee 1993). Each agent in the sequence uses his 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 cascades are given in Scharfstein and Stein (1990) and Trueman (1994).[2] Consider the situation in the paper by Scharfstein and Stein, in which 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 whereas uninformative ones are independent, and agents do not know whether their signal is informative. An agent 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

[2] Other models in which agents imitate one another's actions include the model in Froot et al. (1992), who show in a market setting that agents might gather information about what other agents know rather than about the fundamental outcome, and the model in Zwiebel (1991), who shows that agents may choose the common action rather than a superior innovation because the common action provides a less risky benchmark for relative performance evaluation.

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; if the signal errors were uncorrelated, then common decisions would not indicate the presence of informed agents. Reputational cascades 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 cascade literature explains the similarity of agents' decisions by showing that, for either statistical or reputational reasons, agents rationally ignore their own information and mimic other agents' actions. However, when agents choose the timing of their actions strategically, there is another reason for their actions to be close together. 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 cascade 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 signal realizations, then in equilibrium they act first. When the improved decision of the first agent is held aside, the fact that the agent with the most extreme information acts first is itself a source of clustering. The reason is that differences in the agents' actions are caused by the first agent's inability to perfectly forecast the second agent's signal. Consequently, if the second agent always has the less extreme signal, then the first agent does not make the most extreme forecast errors; hence, the actions of the two agents are, on average, closer together. This source of clustering is labeled *ordering.* We isolate the effect of ordering with two examples from a setting in which 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 and ordering contributes to the clustering of agents' decisions. In another case in which the cost of delay is higher for agents with less extreme signals, the first agent makes the most extreme forecast errors, and agents' forecasts become dispersed rather than clustered together. In sum, clustering occurs when the agent who chooses to act second in equilibrium has a less extreme signal than a typical

agent (ordering) or when the agent who chooses to act first in equilibrium can infer some of the other agents' information (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 on the basis of 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 trade-off between more informed decisions and the urgency to make a decision is the main ingredient of our model.[3] 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 Sections II and III, we focus primarily on a game in which the cost of delay increases as the value of the unknown variable increases. While none of the previously studied causes of information cascades is present, we demonstrate that the agents' decisions are closer together than when the timing of actions is exogenous. We then show that clustering occurs for any delay cost that is either a strictly monotone or strictly convex function of the unknown variable. In Section IV we discuss how our results carry over to a model with many agents. We conclude the paper in Section V by comparing the general set of conditions necessary for clustering and information cascades.

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, with the accuracy of the prediction held 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, $i = 1$ or 2, observes the realization

[3] Other authors that demonstrate the cost of delayed decisions include Hendricks and Kovenock (1989), who study the trade-off 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, in the context of a public goods problem; and 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.

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 trade-off between the accuracy of the prediction and the time at which it is made, assume that agent i's utility is given by

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

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, without 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 in which the time cost is increasing in the squared deviation of w from its prior expectation, representing situations in which 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 preplay 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 in which 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. We focus primarily on the model without a relative performance term to highlight the source of clustering in the simplest possible setting.[4]

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

[4] Suppose that we add to agent i's utility (as given in [1]) a term $\beta(w - z_j)^2$, where z_j is the other agent's forecast. Here agent i no longer has an incentive to share information with agent j because she benefits from agent j's inaccuracy. Nevertheless, it can be shown that for β sufficiently small the results for this extended model are exactly as in our analysis.

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 s, and the elapsed time t. The forecast with this property is $s_i + E(S_j|\sigma, t)$, $j \neq i$. 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 $t_i: I \to \mathbf{R}^+$, where $t_i(s_i)$ specifies the *latest* possible time at which agent i with signal s_i will make her forecast (she will forecast earlier if the other agent forecasts before this time). We shall 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 that 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. In Gul and Lundholm (1993), we show 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

A. *The Symmetric Equilibrium*

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

Note that, because $t(\cdot)$ 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 $t(\cdot)$ by $s(\cdot)$. Because $t(\cdot)$ is downward sloping, if the game proceeds to time $\bar{\tau}$ without a forecast, each agent knows that the other agent's signal is not in the region $[s(\bar{\tau}), 1]$. Thus, if the first agent chooses to forecast at time $\bar{\tau}$, then her forecast is $s_i + [s(\bar{\tau})/2]$. Finding the time that maximizes agent 1's expected utility for the given strategy of agent 2 is equivalent to finding the signal $s \in I$, not necessarily s_1, that minimizes

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

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 $s_i + \frac{2}{3}z_j = s_1 + s_2$ immediately after her opponent's announcement (at time $t(\frac{2}{3}z_j)$). If agent i observes that $\tau_j \neq \frac{2}{3}z_j$, 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.[5] (The proof is given in the Appendix.)

The intuition for why $t(\cdot)$ is decreasing and continuous is straightforward. First, $t(\cdot)$ 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 in which $t(\cdot)$ is constant; if there were, an

[5] Note that off-equilibrium forecasts (i.e., forecasts at time $\tau_j \neq t(\frac{2}{3}z_j)$) play a very limited role in our model. The reason is that 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.

agent could wait an arbitrarily small amount of time and gain a strictly positive amount of additional information. Finally, $t(\cdot)$ 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.

B. *The Asymmetric Equilibria*

An asymmetric sequential equilibrium for our model is the following. Suppose that 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, 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 agents 1 and 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 uniformly 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.[6]

C. *Clustering*

In our economy neither reputational cascades nor statistical cascades 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 contin-

[6] As in the standard war of attrition problem, the following other equilibria exist: for all $s_i > \bar{s}$, agent i forecasts $s_i + 1/2$ at time 0, and for all $s_i \leq \bar{s}$, agents i and j play a suitably rescaled version of the equilibrium studied in subsection A.

uum, so the second agent always uses her own information to improve her decision. Nonetheless, agents' decisions are clustered together.

In the previous cascade 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 shall 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.[7] Because the forecasts that arise when the order is exogenous are the same as those 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. For instance, if the agent with s_1 is the first to forecast, then the difference between the forecasts is $z_2 - z_1 = s_2 - E(S_2|\sigma, t)$. More generally, 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 denote the second agent's signal by $Y = (1 - J)S_1 + JS_2$. With this, d_{en} can be written as $d_{en} = E(E\{[Y - E(Y|X)]^2|X\})$, 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 $\text{var}(Y|X)$, so $d_{en} = E[\text{var}(Y|X)]$. When the forecasting order is given exogenously, who forecasts first is uninformative. In this case the first agent's forecast of the second agent's signal is simply the prior mean, so $d_{ex} = E\{[S_i - E(S_i)]^2\} = \text{var}(S_i)$. The clustering measure can now be written as

$$d_{ex} - d_{en} = \text{var}(S_i) - E\{\text{var}(Y|X)\},$$

[7] 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.

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

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

so that

$$d_{ex} - d_{en} = \text{var}(S_i) - \text{var}(Y) + E\{[E(Y|X) - E(Y)]^2\}. \quad (3)$$

Label the first two terms together as *ordering* and the last term as *anticipation*. Anticipation is the expected improvement in the first agent's forecast that results from knowledge of her own signal (i.e., the conditioning on $X = s_i$). 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 second agent to act has the lower signal, so the ex ante forecast of the second agent's signal is $E(Y) = 1/3$.[8] However, when the first agent learns her own signal s_i, she can conclude that the second agent does not have a higher signal than she does (i.e., $y \leq s_i$), so her optimal forecast is $E(Y|X) = s_i/2$. 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 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 higher signals to forecast first, then the second agent's signal must be contained in a smaller interval than the support of the first agent's signal, so it will have a smaller variance. For instance, in the previous model, $\text{var}(S_i) = 1/12$ but $\text{var}(Y) = \text{var}(S_i|S_i < S_j) = 1/18$, so ordering contributes $1/36$ to clustering. Even when the equilibrium $t(\cdot)$ is not monotonic, if the equilibrium ordering causes the agent with the less extreme signal to move second, the first agent's forecast of this less extreme signal will be more accurate, on average, because the most extreme forecast errors will not occur. This contributes to clustering because it is the first agent's forecast error that causes the two agents to take different actions.

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

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 when the delay cost has the form $\alpha g(w)t$, where $g(\cdot)$ denotes the function g: $[0, 2] \to \mathbf{R}^+$ 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 $g(\cdot)$ is either strictly monotone or strictly convex, so clustering occurs even without any anticipation.

PROPOSITION 3. Suppose that 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 $g(\cdot)$ is either strictly monotone or strictly convex. In this case (i) there exists a symmetric equilibrium such that $t(\cdot)$ is strictly quasi-concave;[9] (ii) in all symmetric equilibria, $t(\cdot)$ is strictly quasi-concave; and (iii) in any symmetric equilibrium, ordering is positive, so clustering occurs. (The proof is in the Appendix.)

As seen in the proof, the existence of a strictly quasi-concave $t(\cdot)$, which is sufficient for clustering to occur, depends only on the strict quasi convexity of $g(\cdot)$. 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 costliest 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. This causes ordering to contribute to clustering.

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

D. *Isolating the Effect of Ordering*

In the next two examples, symmetric delay costs cause the strategy $t(\cdot)$ to be symmetric about $\frac{1}{2}$. Consequently, there is no anticipation; the first agent's forecast of the second agent's signal remains at the prior mean of $\frac{1}{2}$ as time passes. However, as the examples will illus-

[9] As in proposition 1, agent i forecasts $s_i + E(s_j | t(s_j) \geq \tau)$ 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.

trate, which agent forecasts first in equilibrium will determine whether agents' decisions become clustered or dispersed. To see how this can happen, recall that agents take different actions because the first agent cannot perfectly forecast the second agent's signal. If the equilibrium ordering is such that the second agent has a signal near the prior mean, then the first agent's point estimate will be relatively accurate and clustering will occur (the variance conditional on being the second agent will be less than the ex ante variance). However, if the equilibrium ordering is such that the second agent has an extreme signal, near either zero or one, then the first agent's point estimate of $\frac{1}{2}$ will be considerably less accurate and dispersion will occur (the variance conditional on being the second agent will be greater than the ex ante variance).

To illustrate how ordering can contribute to clustering, 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 a good example: 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: it is as valuable to predict extreme decreases in earnings as it is to predict extreme increases. This idea is the motivation for the following utility function:

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

We show that a symmetric equilibrium exists for this model, where $t(\cdot)$ is symmetric about $\frac{1}{2}$, increasing for $s_i \in [0, \frac{1}{2})$, and decreasing for $s_i \in (\frac{1}{2}, 1]$ 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 $t(\cdot)$ is symmetric about the prior mean of s_j, the optimal forecast does not change over time: $z_i = s_i + \frac{1}{2}$ for all t. Once the first agent makes her prediction, however, the second agent can use the observed forecast and the time at which it was made to recover the first agent's signal (as in the model with a 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

$$\begin{aligned} &\int_0^s \alpha(s_1 + s_2 - 1)^2 t(s_2)\, ds_2 + \int_{1-s}^1 \alpha(s_1 + s_2 - 1)^2 t(s_2) ds_2 \\ &\quad + \int_s^{1-s} \left(s_2 - \frac{1}{2}\right)^2 ds_2 + \int_s^{1-s} \alpha(s_1 + s_2 - 1)^2 t(s) ds_2. \end{aligned} \tag{5}$$

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 a two-sided delay cost. In this equilibrium, agent i predicts $s_i + 1/2$ at time $t(s_i) = -(3/4\alpha)\log(|2s_i - 1|)$ 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.)

The first agent's forecast does not change with the passage of time; hence there is no anticipation. However, clustering still occurs in this model because of ordering. In particular, $\text{var}(Y) = E\{\min[(S_1 - 1/2)^2, (S_2 - 1/2)^2]\} = 1/24$, so the measure of ordering equals $(1/12) - (1/24) = 1/24$.[10] 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 $\text{var}(Y)$ given above clearly demonstrates the source of clustering in this model. The equilibrium ordering causes the agent with the less extreme signal to act second, so the first agent's forecast error is the difference between the less extreme signal and the forecast of 1/2. The ordering of agents has not altered the first agent's point forecast. However, since the most extreme information is revealed first, the most extreme differences in agents' actions are avoided.

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 violates the conditions of proposition 3 and demonstrates how ordering can cause actions to be dispersed rather than clustered. Suppose that agents are more eager to act when the value of W is close to its prior mean than when it is extreme. Consider the problem facing uninformed market makers trading against both informed and uninformed traders. Sup-

[10] To compute this value, note that the cumulative 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).

pose that market makers make less money in periods of large price changes because they are trading with mostly informed traders, whereas they make more money in periods of constant prices because they are trading with primarily uninformed traders. Although delaying trade provides a market maker with information about the future price, this delay is costliest when the prices turn out to be constant. This type of delay cost is captured by the following utility function:

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

It can be shown that a symmetric equilibrium exists to this model in which $t(\cdot)$ is symmetric about $1/2$, decreasing for $s_i \in [0, 1/2)$, and 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. In contrast to the previous example, however, the agent with the less extreme signal is the first to forecast. Thus

$$\text{var}(Y) = E\left\{\max\left[\left(S_1 - \frac{1}{2}\right)^2, \left(S_2 - \frac{1}{2}\right)^2\right]\right\} = \frac{1}{8},$$

so the measure of ordering is $(1/12) - (1/8) = -(1/24)$.

By allowing agents to choose when to act, the equilibrium orders them in such a way that the agent with the least extreme information acts first. Consequently, the first agent knows that the second agent's signal is more extreme than hers but does not know in which direction. This causes more dispersion in agents' actions than if the order was exogenous; with an exogenous ordering, the agent with the least extreme information would act first only half the time.

E. Efficiency

In the information cascade literature the economy is informationally inefficient in the sense that subsequent agents ignore their own signals when the information is collectively still useful in making a superior decision. The observed clustering of behavior in many economic situations is seen as the undesirable outcome of information cascades. 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. 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 does not 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.

Chamley and Gale (1994) offer an interesting contrast to our results by studying a setting in which agents have a binary action choice (as in Bikhchandani et al. [1992]) and the timing of their action is endogenous (as in our model). In their model, the endogenous timing choice creates anticipation, but because there are only two action choices available, an information cascade can still occur. As in Bikhchandani et al., there is an insufficient set of action choices to completely communicate agents' information to one another.

Whenever the timing of actions is endogenous, there is a war of attrition effect that reduces welfare. In particular, 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 forecast. By delaying their forecasts, each agent imposes a negative externality on the other agent. In our model the cost of this externality is incurred through increased delay cost. A key result in Chamley and Gale is that, with binary action choices, in equilibrium agents will not incur significant delay costs in order to learn more about other agents' information. Consequently, the externality induced by endogenous timing in Chamley and Gale is less accurate forecasts.

The inefficiency that is created by the endogenous timing of actions is best illustrated by comparing the symmetric and asymmetric equilibria of our model. In contrast to the symmetric equilibrium, there is no delay cost in the asymmetric equilibria or in a setting in which 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 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.[11] Suppose that agent 1 forecasts first in the asymmetric equilibrium. Her ex ante expected utility (averaged over all realizations of s_1) is

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

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

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

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

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

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 at which her gain from increased accuracy equals her loss from additional delay.

IV. 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 $W = S_1 + S_2 + \ldots + S_n$. 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 $\{t^k(s, y)\colon k \in \{1, 2, \ldots, n\}\}$, 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.

When arguments similar to those given in the Appendix are used 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 nonincreasing 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 is defined formally and have shown that the trade-off between delayed decisions and more accurate decisions creates a tendency for clustered decisions even when there are no information cascades. Thus it would be wrong to conclude that agents are in an information cascade only on the basis of the observation that the agents are making similar decisions. Given the difficulty in observing agents' private information in most economic settings, it will be difficult to tell when clustered decisions are caused by an information cascade. One sign of a cascade is that the ex post accuracy of decisions does not improve with time in a cascade, whereas it will continue to improve in an efficient model

of clustering such as ours. Ex ante, whether an economic situation is likely to generate information cascades depends on the richness of the agent's action space and the observability of predecessor agents' actions.

Consider 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 who 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*B*.

When the "natural" assumption is used as a benchmark, note that in a model with exogenous timing of actions a cascade implies that, in equilibrium, the typical agent i *knows less* than under the natural assumption. In the statistical cascade models, the reason is that the discrete 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 a reputational cascade, 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 three-person model the third agent would not have complete information about the second agent's signal. Informational cascades 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 cascades, and aligning incentives with outcomes eliminates reputational cascades.

In contrast, 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 ith highest signal (or the ith most extreme signal in the case of a two-sided delay cost). Hence, information has leaked. In our model, this information leak is a result of the trade-off between accuracy and delay. Whenever the appropriate marginal calculations for this trade-off 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 cascade literature explains clustering by noting that agents may know less than one thought, whereas our model explains clustering by noting that they may know more than one thought.

Appendix

Proofs of Propositions 1, 3, and 4

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

Proof of Proposition 3

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

Denote by $h(\cdot)$ a function h: $[\underline{s}, s^*) \to (s^*, \bar{s}]$. 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 $g(\cdot)$. In one case, $\underline{s} = 0$, $h(0) = \bar{s}$, and $t(1) = 0$ is the initial condition. In the other, $\bar{s} = 1$, $h(\underline{s}) = 1$, 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. Hence, 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

$$t'(s)\alpha[G(s_1 + h(s)) - G(s_1 + s)] + \frac{[h(s) - s]^2}{4}[h'(s) - 1] = 0, \qquad \text{(A1)}$$

where $G(x) = \int_0^x g(w)dw$. Substituting s_1 for s in (A1) gives

$$t'(s_1)\alpha[G(s_1 + h(s_1)) - G(2s_1)] + \frac{[h(s_1) - s_1]^2}{4}[h'(s_1) - 1] = 0. \qquad \text{(A2)}$$

Equation (A2) 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 (A1) and then substitute s_1 for s to get

$$t'(s_1)\alpha[G(2h(s_1)) - G(h(s_1) + s_1)] + \frac{[h(s) - s_1]^2}{4}[h'(s_1) - 1] = 0. \qquad \text{(A3)}$$

Both (A2) and (A3) are satisfied if there exists a differentiable $h(\cdot)$ such that, for all $s_1 \in [0, s^*)$,

$$G(2h(s_1)) + G(2s_1) = 2G(h(s_1) + s_1). \tag{A4}$$

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 $G(\cdot)$ is continuous and because $G(2s^*) < 2G(s^*)$ since $g(\cdot)$ is strictly decreasing on $[0, 2s^*)$. Thus we are in the case in which $h(0) = \bar{s}$, and the appropriate initial condition is $t(1) = 0$. If the opposite inequality holds, then we are in the case, to be discussed later, in which there exists an $\underline{s}$ such that $h(\underline{s}) = 1$ and $t(0) = 0$. To verify the existence of a function h: $[0, s^*) \to (s^*, \bar{s}]$ that satisfies (A4) for all $s_1 \in [0, s^*)$, implicitly differentiate (A4) to get the first-order differential equation

$$h'(s_1) = \frac{g(h + s_1) - g(2s_1)}{g(2h) - g(h + s_1)}. \tag{A5}$$

Let $\Omega_\epsilon = [0, s^* - \epsilon) \times (s^*, \bar{s}]$. The right-hand side of (A5) is a continuously differentiable and bounded function from Ω_ϵ to R with a bounded derivative on Ω_ϵ for a fixed $\epsilon > 0$. Hence the function $h(\cdot)$ satisfies a Lipchitz condition sufficient to guarantee a unique solution to (A5) 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 $g(\cdot)$ and the right-hand side of (A5) can be extended as a differentiable function to $(-\epsilon, s^* - \epsilon) \times (s^*, \bar{s})$. To extend $h(\cdot)$ to the domain $[0, s^*)$, let $\epsilon \to 0$. This establishes that there exists a unique solution to (A4).

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

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

which yields the first-order condition

$$t'(s)\alpha[G(s_1 + s) - G(s_1)] + \frac{s^2}{4} = 0. \tag{A6}$$

Setting $s = s_1$ yields

$$t'(s_1)\alpha[G(2s_1) - G(s_1)] + \frac{(s_1)^2}{4} = 0. \tag{A7}$$

Since (A7) holds for $s_1 \in (\bar{s}, 1]$ and a solution can be obtained by integrating and using the initial condition $t(1) = 0$, $t(\cdot)$ is described by this solution on this region. Further, taking the calculated value of $t(\bar{s})$ from the solution of (A7) gives a boundary condition $t(0) = t(\bar{s})$ for (A2). When (A2) is integrated and this boundary condition is used, $t(\cdot)$ is described on $[0, s^*)$. On $(s^*, 1]$, $t(\cdot)$ is described by the relation $t(s_1) = t(h(s_1))$. Finally, define $t(s^*)$ as $\lim t(s_1)$ as $s_1 \to s^*$ 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 shall show that, for s and s_1 belonging to $[0, s^*)$, when $s_1 < s$, the left-hand side of (A1) is positive, and when $s_1 > s$, the left-hand side is negative.

The first-order condition states that, for $s_1 = s$, the left-hand side of (A1) equals zero. Note that the only term in (A1) involving s_1 is $G(s_1 + h(s)) - G(s_1 + s)$ and that $t'(s)$ is positive for all $s \in [0, s^*)$. If this term is greater or lesser at values of $s_1 \neq s$ than it is at $s_1 = s$, then the left-hand side of (A1) 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 $g(\cdot)$ 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 (A4) 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 $G(\cdot)$) 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 left-hand side of (A1) 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 $h(\cdot)$ and (A4). Thus $G(s_1 + h(s)) - G(s_1 + s) < G(s + h(s)) - G(s + s)$ for $s_1 > s$, so the left-hand side of (A1) 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 (A4) 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 (A2) and (A3) remain the same, and an equation analogous to (A7) is obtained for the segment $[0, \underline{s})$. Finally, the proof for the case in which $g(\cdot)$ is strictly monotone follows from the construction of the equilibrium $t(\cdot)$ on the segment $[\bar{s}, 1]$ when $g(\cdot)$ is increasing and from the construction of $t(\cdot)$ on the segment $[0, \underline{s})$ when $g(\cdot)$ is decreasing. Thus a symmetric equilibrium exists.

Next we show that when $g(\cdot)$ is either strictly monotone or strictly convex, any symmetric equilibrium $t(\cdot)$ 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 a 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 a time no later than τ by

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

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

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

which implies

$$l(\hat{\tau}) - l(\underline{\tau}) \leq c(\hat{s}, \underline{\tau}) - c(\hat{s}, \hat{\tau}) \tag{A8}$$

and

$$l(\hat{\tau}) - l(\bar{\tau}) \leq c(\hat{s}, \bar{\tau}) - c(\hat{s}, \hat{\tau}). \tag{A9}$$

Similarly, the optimality of $\underline{\tau}$ and $\bar{\tau}$ implies

$$l(\hat{\tau}) - l(\underline{\tau}) \geq c(\underline{s}, \underline{\tau}) - c(\underline{s}, \hat{\tau}) \tag{A10}$$

and

$$l(\hat{\tau}) - l(\bar{\tau}) \geq c(\bar{s}, \bar{\tau}) - c(\bar{s}, \hat{\tau}). \tag{A11}$$

We show that either (A10) contradicts (A8) or (A11) contradicts (A9). Consider $c(s_i, \bar{\tau} - c(s_i, \hat{\tau})$. Denote $A_1 = A_j(\hat{\tau})$, $A_2 = A_j(\hat{\tau})^C \cap A_j(\bar{\tau})$, and $A_3 = A_j(\bar{\tau})^C$, so that A_1, A_2, and A_3 are disjoint sets and their union equals I. Thus

$$\begin{aligned} c(s_i, \bar{\tau}) - c(s_i, \hat{\tau}) &= \int_{s_j \in A_1 \cup A_2} \alpha t(s_j) g(s_i + s_j)\, dF(s_j) + \int_{s_j \in A_3} \alpha \bar{\tau} g(s_i + s_j)\, dF(s_j) \\ &\quad - \int_{s_j \in A_1} \alpha t(s_j) g(s_i + s_j)\, dF(s_j) - \int_{s_j \in A_2 \cup A_3} \alpha \hat{\tau} g(s_i + s_j)\, dF(s_j) \\ &= \int_{s_j \in A_2} \alpha [t(s_j) - \hat{\tau}] g(s_i + s_j)\, dF(s_j) \\ &\quad + \int_{s_j \in A_3} \alpha (\bar{\tau} - \hat{\tau}) g(s_i + s_j)\, dF(s_j), \end{aligned} \tag{A12}$$

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 (A12) is positive; forecasting later yields higher expected delay costs.

From (A12), $[c(\bar{s}, \bar{\tau}) - c(\bar{s}, \hat{\tau})] - [c(\hat{s}, \bar{\tau}) - c(\hat{s}, \hat{\tau})]$ can be written as

$$\begin{aligned} &\int_{s_j \in A_2} \alpha [t(s_j) - \hat{\tau}][g(\bar{s} + s_j) - g(\hat{s} + s_j)]\, dF(s_j) \\ &+ \int_{s_j \in A_3} \alpha (\bar{\tau} - \hat{\tau})[g(\bar{s} + s_j) - g(\hat{s} + s_j)]\, dF(s_j). \end{aligned} \tag{A13}$$

For strictly monotone or strictly convex $g(\cdot)$, either $g(\bar{s} + s_j) > g(\hat{s} + s_j)$ or $g(\underline{s} + s_j) > g(\hat{s} + s_j)$. If $g(\bar{s} + s_j) > g(\hat{s} + s_j)$, (A13) 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 (A11) contradicts (A9). 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 (A13), $c(\underline{s}, \underline{\tau}) - c(\underline{s}, \hat{\tau}) > c(\hat{s}, \underline{\tau}) - c(\hat{s}, \hat{\tau})$, and (A10) contradicts (A8). This establishes that $t(\cdot)$ is quasi-concave.

To show that $t(\cdot)$ is *strictly* quasi-concave, suppose to the contrary that there exist $\underline{s}$ 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

$$l(\hat{\tau}) - l(\hat{\tau} + \Delta\tau) \leq c(s_i, \hat{\tau} + \Delta\tau) - c(s_i, \hat{\tau}) \tag{A14}$$

for $\Delta\tau > 0$. 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 $\epsilon > 0$ such that $l(\hat{\tau}) - l(\hat{\tau} + \Delta t) > \epsilon$ 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 (A14) is contradicted. Thus $t(\cdot)$ is strictly quasi-concave.

Finally, we show that the strict quasi concavity of $t(\cdot)$ implies that ordering is positive. Because anticipation is nonnegative by definition (see eq. [3] in the text), establishing that ordering is positive is sufficient to establish that clustering occurs. First, note that $\text{var}(S_i) = 1/12$, so ordering is positive if $\text{var}(Y) < 1/12$. By the strict quasi concavity of $t(\cdot)$, 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 cumulative distribution function of Γ (with realizations denoted by γ) is $P(\Gamma \leq \gamma) = P(s_1 \in [a(\gamma), a(\gamma) + \gamma]) \times P(s_2 \in [a(\gamma), a(\gamma) + \gamma]) = \gamma^2$, and so the density of Γ is 2γ on $\gamma \in [0, 1]$. With this,

$$\text{var}(Y) = \int_0^1 \int_{a(\gamma)}^{a(\gamma)+\gamma} (y - \mu)^2 \left(\frac{1}{\gamma}\right) dy 2\gamma d\gamma,$$

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

$$\mu = E\{E(Y|\Gamma)\} = \int_0^1 \left[a(\gamma) + \frac{\gamma}{2}\right] 2\gamma d\gamma = 2\int_0^1 a(\gamma)\gamma d\gamma + \frac{1}{3}. \tag{A15}$$

Now note that $\text{var}(Y)$ can be written as $E(Y - \frac{1}{2})^2 - (\frac{1}{2} - \mu)^2$, so

$$\text{var}(Y) = \int_0^1 \int_{a(\gamma)}^{a(\gamma)+\gamma} \left(y - \frac{1}{2}\right)^2 \left(\frac{1}{\gamma}\right) dy 2\gamma d\gamma - \left(\frac{1}{2} - \mu\right)^2. \tag{A16}$$

Integrating (A16) with respect to y and simplifying gives

$$\text{var}(Y) = 2\int_0^1 a(\gamma)\gamma[a(\gamma) + \gamma - 1]d\gamma + \frac{1}{12} - \left(\frac{1}{2} - \mu\right)^2. \tag{A17}$$

Note that the first term (involving an integral) and the last term in (A17) are nonpositive. The first term is zero only if $a(\gamma) = 0$ or $a(\gamma) + \gamma = 1$ for all

γ. By the strict quasi concavity of $t(\cdot)$, 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 (A15), when $a(\gamma) = 0$ for all γ, $\mu = 1/3$; 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, $\text{var}(Y) < 1/12$.

Proof of Proposition 1

This proposition is a special case of proposition 3. By substituting $g(w) = w$, so that $G(2s_1) - G(s_1) = 3(s_1)^2/2$ in (A7), we get the equilibrium condition

$$t'(s_1) = -\frac{1}{6\alpha}. \tag{A18}$$

Integrating (A18) and using the initial condition $t(1) = 0$ gives

$$t(s_1) = \frac{1 - s_1}{6\alpha}. \tag{A19}$$

The solution given in (A19) 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 (A19) describes the unique symmetric equilibrium. First, to show that $t(\cdot)$ is strictly decreasing, it is sufficient to note that $g(s_i + s_j)$ is strictly increasing in s_i for all $s_i \in I$. So from the proof of proposition 3, (A13) is positive and, if $t(\cdot)$ were nondecreasing, (A11) would contradict (A9).

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

$$\begin{aligned} L(s_1, s) = &\int_s^1 \alpha(s_1 + s_2)t(s_2)\,ds_2 \\ &+ \int_0^s \left(s_2 - \frac{s}{2}\right)^2 ds_2 + \int_0^s \alpha(s_1 + s_2)t(s)\,ds_2, \end{aligned} \tag{A20}$$

where the second argument of L is the agent's choice variable. If one relies on the invertibility of $t(\cdot)$, choosing s is equivalent to choosing t.

Suppose that $t(s)$ is discontinuous at s_1, and without loss of generality, suppose that the discontinuity takes the form $t(s_1) > \lim t(s_1 + \epsilon)$ as $\epsilon \downarrow 0$. 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

$$\int_{s_1}^1 \alpha(s_1 + s_2)t(s_2)\,ds_2 + \int_0^{s_1} \left(s_2 - \frac{s_1}{2}\right)^2 ds_2 + \int_0^{s_1} \alpha(s_1 + s_2)t(s_1)\,ds_2. \tag{A21}$$

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

$$\begin{aligned} &\int_{s_1+\epsilon}^1 \alpha(s_1 + s_2)t(s_2)\,ds_2 + \int_0^{s_1+\epsilon} \left(s_2 - \frac{s_1 + \epsilon}{2}\right)^2 ds_2 \\ &\quad + \int_0^{s_1+\epsilon} \alpha(s_1 + s_2)t(s_1 + \epsilon)\,ds_2. \end{aligned} \tag{A22}$$

As $\epsilon \downarrow 0$, the first two terms in (A22) converge to the first two terms in (A21) but the third term converges to

$$\int_0^{s_1} \alpha(s_1 + s_2)\lim[t(s_1 + \epsilon)]ds_2. \tag{A23}$$

Because $t(\cdot)$ is positive and decreasing, and by the discontinuity of $t(s)$ at s_1, (A23) is strictly less than the third term in (A21). This contradicts the optimality of $t(s)$ at s_1.

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

$$\frac{L(s_1, s_1 + \epsilon) - L(s_1, s_1)}{\epsilon} \geq 0. \tag{A24}$$

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

$$\frac{L(s_1 + \epsilon, s_1) - L(s_1 + \epsilon, s_1 + \epsilon)}{\epsilon} \geq 0. \tag{A25}$$

By substituting in the component terms of L and taking the difference term by term, we can express (A24) as

$$\begin{aligned} &\frac{1}{\epsilon}\left[\int_{s_1+\epsilon}^{1} \alpha(s_1 + s_2)t(s_2)ds_2 - \int_{s_1}^{1} \alpha(s_1 + s_2)t(s_2)ds_2\right] \\ &+ \frac{1}{\epsilon}\left[\int_0^{s_1+\epsilon} \left(s_2 - \frac{s_1 + \epsilon}{2}\right)^2 ds_2 - \int_0^{s_1} \left(s_2 - \frac{s_1}{2}\right)^2 ds_2\right] \\ &+ \frac{1}{\epsilon}\left[\int_0^{s_1+\epsilon} \alpha(s_1 + s_2)t(s_1 + \epsilon)ds_2 - \int_0^{s_1} \alpha(s_1 + s_2)t(s_1)ds_2\right] \geq 0. \end{aligned} \tag{A26}$$

As $\epsilon \downarrow 0$, the limit of the first term in (A26) is $-2s_1\alpha t(s_1)$, the limit of the second term is $s_1^2/4$, and the limit of the third term is

$$2s_1\alpha t(s_1) + \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1 + \epsilon) - t(s_1)}{\epsilon}\right].$$

Thus the limit of (A24) as $\epsilon \downarrow 0$ is

$$\frac{s_1^2}{4} + \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1 + \epsilon) - t(s_1)}{\epsilon}\right] \geq 0. \tag{A27}$$

A similar exercise yields the limit of (A25) to be

$$-\frac{s_1^2}{4} - \frac{3}{2}\alpha s_1^2 \lim\left[\frac{t(s_1 + \epsilon) - t(s_1)}{\epsilon}\right] \geq 0. \tag{A28}$$

Because the left-hand side of (A28) has a sign opposite to that of the left-hand side of (A27), the only way to satisfy both inequalities is for each to equal zero. This yields

$$\lim\left[\frac{t(s_1 + \epsilon) - t(s_1)}{\epsilon}\right] = -\frac{1}{6\alpha}.$$

The conditions in (A24) and (A25) are given for $\epsilon > 0$. If $\epsilon < 0$, the inequalities in both conditions are reversed (as is the limit direction), and hence the inequalities in (A27) and (A28) are reversed. As before, the only way both conditions could be satisfied is for each to equal zero. Thus

$$\lim\left[\frac{t(s_1 + \epsilon) - t(s_1)}{\epsilon}\right] = -\frac{1}{6\alpha}$$

regardless of the direction in which the limit is taken; consequently, $t(\cdot)$ is differentiable.

Thus the unique symmetric equilibrium strategy for the model with a 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 (A5) is $h(s) = 1 - s$ (which we guessed on the basis of the symmetry of $g(\cdot)$), and $s^* = \frac{1}{2}$. Substituting $1 - s_1$ for $h(s_1)$ and $(w - 1)^2$ for $g(w)$ gives the equilibrium condition

$$t'(s_1) = -\frac{3}{2\alpha(2s_1 - 1)}. \tag{A29}$$

Integrating (A29) and using either initial condition $t(1) = 0$ or $t(0) = 0$ gives

$$t(s_1) = -\frac{3}{4\alpha}\log(|2s_1 - 1|). \tag{A30}$$

References

Banerjee, Abhijit V. "A Simple Model of Herd Behavior." *Q.J.E.* 107 (August 1992): 797–817.

Bartle, Robert G. *The Elements of Real Analysis.* 2d ed. New York: Wiley, 1976.

Bikhchandani, Sushil; Hirshleifer, David; and Welch, Ivo. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *J.P.E.* 100 (October 1992): 992–1026.

Bliss, Christopher, and Nalebuff, Barry. "Dragon-Slaying and Ballroom Dancing: The Private Supply of a Public Good." *J. Public Econ.* 25 (November 1984): 1–12.

Bulow, Jeremy I., and Klemperer, Paul D. "Rational Frenzies and Crashes." Technical Paper no. 112. Cambridge, Mass.: NBER, September 1991.

Chamley, Christophe, and Gale, Douglas. "Information Revelation and Strategic Delay in a Model of Investment." *Econometrica* 62 (September 1994): 1065–85.

Froot, Kenneth A.; Scharfstein, David S.; and Stein, Jeremy C. "Herd on the Street: Informational Inefficiencies in a Market with Short-Term Speculation." *J. Finance* 47 (September 1992): 1461–84.

Gul, Faruk, and Lundholm, Russell. "On the Clustering of Agents' Decisions: Herd Behavior versus the Endogenous Timing of Actions." Working paper. Ann Arbor: Univ. Michigan, 1993.

Hendricks, Kenneth, and Kovenock, Dan. "Asymmetric Information, Infor-

mation Externalities, and Efficiency: The Case of Oil Exploration." *Rand J. Econ.* 20 (Summer 1989): 164–82.
Lee, In Ho. "On the Convergence of Informational Cascades." *J. Econ. Theory* 61 (December 1993): 395–411.
Scharfstein, David S., and Stein, Jeremy C. "Herd Behavior and Investment." *A.E.R.* 80 (June 1990): 465–79.
Stickel, Scott E. "Predicting Individual Analyst Earnings Forecasts." *J. Accounting Res.* 28 (Autumn 1990): 409–17.
———. "Reputation and Performance among Security Analysts." *J. Finance* 47 (December 1992): 1811–36.
Stinchcombe, M. "Maximal Strategy Sets for Continuous-Time Game Theory." Working paper. Berkeley: Univ. California, 1988.
Trueman, Brett. "Analyst Forecasts and Herding Behavior." *Rev. Financial Studies* 7 (Spring 1994): 97–124.
Welch, Ivo. "Sequential Sales, Learning, and Cascades." *J. Finance* 47 (June 1992): 695–732.
Zwiebel, Jeffrey. "Corporate Conservatism, Herd Behavior and Relative Compensation." Working paper. Stanford, Calif.: Stanford Univ., 1991.