A Theory of Conformity

B. Douglas Bernheim

Princeton University and National Bureau of Economic Research

This paper analyzes a model of social interaction in which individuals care about status as well as "intrinsic" utility (which refers to utility derived directly from consumption). Status is assumed to depend on public perceptions about an individual's predispositions rather than on the individual's actions. However, since predispositions are unobservable, actions signal predispositions and therefore affect status. When status is sufficiently important relative to intrinsic utility, many individuals conform to a single, homogeneous standard of behavior, despite heterogeneous underlying preferences. They are willing to conform because they recognize that even small departures from the social norm will seriously impair their status. The fact that society harshly censures all nonconformists is not simply assumed (indeed, status varies smoothly with perceived type); rather, it is produced endogenously. Despite this penalty, agents with sufficiently extreme preferences refuse to conform. The model provides an explanation for the fact that standards of behavior govern some activities but do not govern others. It also suggests a theory of how standards of behavior might evolve in response to changes in the distribution of intrinsic preferences. In particular, for some values of the preference parameters, norms are both persistent and widely followed; for other values, norms are transitory and confined to small groups. Thus the model produces both customs and fads. Finally, an extension of the model suggests an explanation for the development of multiple subcultures, each with its own distinct norm.

I would like to thank Laurie Bagwell, Kyle Bagwell, and Avinash Dixit; seminar participants at Princeton, Yale, Chicago, and the Harvard–Massachusetts Institute of Technology joint theory seminar; and two anonymous referees for helpful comments. The work reported here was supported by the National Science Foundation through grant SES-9110211.

[*Journal of Political Economy,* 1994, vol. 102, no. 5]

I. Introduction

Most social scientists agree that individual behavior is motivated in large part by "social" factors, such as the desire for prestige, esteem, popularity, or acceptance. A large body of sociological, psychological, and anthropological research supports the view that these factors are widespread and that they tend to produce conformism. Social groups often penalize individuals who deviate from accepted norms, even when deviations are relatively minor.[1] Unfortunately, theories developed within other social science disciplines are not easily adapted to the problems and modes of analysis that are familiar to economists. Although many economists have acknowledged the potential importance of social and cultural influences in passing, few have examined these factors formally. We simply lack satisfactory formal models of custom and conformity.

Over the past few decades, there have been a number of attempts to develop such models. In most cases, economists have attempted to explain conformity without reference to the social factors mentioned above.[2] A variety of ingenious arguments have been used to demonstrate that traditional economic factors, under appropriate conditions, can generate conformism and the development of social norms. One line of research suggests that individuals obtain information by observing each others' actions and are therefore inclined to imitate those who are believed to be better informed (see, e.g., Conlisk 1980; Banerjee 1989; Bikhchandani, Hirshleifer, and Welch 1992). Another school of thought holds that agents act similarly because similar actions sometimes create mutual positive externalities (see, e.g., Katz and Shapiro 1986; Banerjee and Besley 1990). A refinement of this second view suggests that mutual interdependence may give rise to multiple equilibria and that social norms arise to coordinate the selection of some particular equilibrium (see, e.g., Schelling [1960] or, for a more modern treatment, Kandori, Mailath, and Rob [1993]).

These factors certainly help to explain behavior in a variety of situations. However, the available sociological research strongly suggests that they do not account for the full extent or degree of conformist behavior. Direct and indirect evidence confirms the importance of purely social influences (see, e.g., Ross, Bierbrauer, and Hoffman [1976] or, for a brief survey, Jones [1984]).

This paper adopts the alternative strategy of incorporating certain social factors directly into individual preferences. A key assumption

[1] A portion of this research is summarized by Akerlof (1980) and Jones (1984).

[2] Others explore the implications of an assumed preference for conformity without attempting to rationalize this preference. For example, Matsuyama (1991) examines dynamic behavior in a model with conformists and nonconformists.

is that individuals care directly about status (popularity, esteem, or respect).[3] There are at least three separate justifications for defining preferences directly over a social status variable. First, the assumption that individuals care about status is consistent with the psychological evidence cited above. Second, evolutionary pressures could well produce preferences of this form. On a purely biological level, individuals who are more highly regarded have greater opportunities to reproduce. Thus natural selection tends to favor those who are more concerned about esteem, popularity, or respect. Social evolution may also favor the development of preferences for esteem, since concern about the opinions of others fosters cooperative behavior. Social groups may also tend to protect individuals who are more highly regarded. Third, behavioral conditioning may foster the development of preferences for esteem. If esteemed individuals generally receive better treatment, then esteem-enhancing activities will be reinforced. Individuals may come to desire esteem, even when the enhancement of esteem serves no specific, concrete purpose.

Although the formulation of preferences adopted here is a departure from traditional modes of economic analysis, it does not require us to abandon the framework of consistent, self-interested optimization. Indeed, the motivational factors considered here are closely related to the more familiar and much-studied concept of altruism. An altruistic person cares about how someone else feels, whereas a person motivated by popularity or esteem cares about how someone else feels about *him* (or *her*).

In addition to esteem, individuals are also assumed to care about actions (e.g., consumption). The population is heterogeneous in the sense that, ceteris paribus, different individuals prefer different actions. I assume that status (esteem or popularity) depends on public perceptions about an individual's preferences over actions. Esteem does not depend on actions themselves, at least not directly. One possible justification of this assumption is that esteem is determined by expectations about future actions and that tastes and proclivities are the best predictors of future actions. I also assume that preferences over actions are not directly observable. Consequently, each individual must infer the preferences of others from their actions. This state of affairs creates a signaling problem. In the resulting

[3] The assumption that behavior is influenced by the desire for status has a long tradition in economics. Indeed, it was featured as the centerpiece of Veblen's (1899) seminal treatise on the "leisure class." More recent examples include Leibenstein (1950), Akerlof (1980), Jones (1984), Frank (1985), Besley and Coate (1990), Cole, Mailath, and Postlewaite (1992), Fershtman and Weiss (1992), Glazer and Konrad (1992), Ireland (1992), and Bagwell and Bernheim (1993).

equilibrium, status depends *indirectly* on actions (actions affect perceptions of preferences, which determine status).

A concrete illustration helps to fix concepts. "Generous" individuals are usually esteemed by society, whereas "selfish" individuals are disdained. One cannot, however, observe generosity directly. Rather, one must infer it from actions. Thus an individual who has taken generous actions is esteemed and thought of as generous. Yet most of us distinguish between people who merely *act* generous and people who truly *are* generous. Generosity is usually considered a personality trait, not simply a characterization of past actions. If, for example, it comes to light that some supposedly generous individual took apparently generous actions for selfish reasons, then the esteem accorded to that individual diminishes. Conversely, an apparently selfish act is typically discounted if it proves to be motivated by unselfish concerns. Thus status depends critically on motivations. The natural explanation for this is that motivations predict future actions: generous individuals are revered, in large part, because others expect them to act generously when the opportunity arises.

For a broad class of models, I demonstrate that equilibria have a number of striking properties. When popularity is sufficiently important relative to intrinsic utility (defined as the utility derived directly from consumption), many individuals conform to a single, homogeneous standard of behavior, despite heterogeneous underlying preferences. They are willing to suppress their individuality and conform to the social norm because they recognize that even small departures from the norm will seriously impair their popularity. The fact that society harshly censures all deviations from the accepted mode of behavior is not simply assumed (indeed, popularity varies smoothly with perceived type); rather, it is produced endogenously. In equilibrium, any departure from the norm is construed as evidence of extreme preferences (i.e., perceived type changes discontinuously when one deviates from customary behavior). Even so, agents with sufficiently extreme preferences refuse to conform. These "individualists" behave in ways that differ significantly from the social norm: there are no "trivial" nonconformists. Within the social fringe, heterogeneous preferences do result in heterogeneous behavior; these agents "express their individuality." Nevertheless, even individualists succumb somewhat to the desire for popularity and shade their choices toward the social norm.

The model provides an explanation for the fact that standards of behavior govern some activities but do not govern others. It also suggests a theory of how standards of behavior might evolve in response to a change in the distribution of intrinsic preferences. In particular, sufficiently small changes need not have any impact on the social norm, but large changes will upset the norm. Depending on the values of certain preference parameters, the resulting norm can be either very persistent or very transitory. Persistence is linked to the fraction of the population that adheres to the norm. When norms are widely obeyed, they are also persistent. On the other hand, when norms are obeyed by relatively small groups of individuals, they are transitory. Thus the model produces a theory that explains both customs and fads.

The paper is most closely related to previous work by Akerlof (1980) and Jones (1984). Akerlof assumes that deviations from social customs are punished by loss of social "reputation." In his model, this reputational effect gives rise to stable customs. There are at least two important differences between Akerlof's analysis and that presented here. First, Akerlof does not explain how customs come into being in the first place. For his model, there is always an equilibrium in which no one adheres to any custom. In the current paper, the existence or nonexistence of behavioral norms is explained by identifiable preference parameters. Second, Akerlof simply assumes that reputation changes discontinuously when one departs from the custom. In contrast, there are no structural discontinuities built into the current model.[4] Under some circumstances, popularity does vary discontinuously with actions, but this is derived as a consequence of equilibrium; it is not assumed.

Jones presents a model in which utility depends on the extent to which an individual's action differs from those chosen by other members of his social group. Naturally, this generates some convergence of choices. However, utility changes smoothly as one deviates from some norm. As a result, convergence of choices is not complete for any subgroup; there is no true conformity, in the sense that heterogeneous agents behave identically. The difference between behavior with and without conformity in Jones's model is a matter of degree rather than of kind.

The current paper is organized as follows. Section II describes the basic signaling model. Section III discusses fully separating equilibria. The central result of that section describes a necessary and sufficient condition for the existence of fully separating equilibria. When this condition is violated, equilibrium must entail some pooling. Section IV analyzes equilibria in which separation is not complete. Typically,

[4] The absence of structural discontinuities distinguishes this paper not only from Akerlof (1980) but also from many other papers that generate equilibria with conformity. For example, in Banerjee (1989), agents *completely* ignore their own information and imitate the "herd" only because the choice set is discrete. With a continuous choice set, they would never ignore their own information entirely, and conformity (in the strict sense of *identical* behavior) would not be observed.

there are many equilibria of this type. However, a standard equilibrium refinement isolates equilibria with the properties described above. Section V elucidates various implications of the analysis, including a possible extension that suggests an explanation for the development of multiple subcultures, each with its own distinct norm. Section VI presents conclusions.

II. The Model

Consider a society consisting of many agents, each of whom selects some publicly observable variable x from the set X. For simplicity, I take X to be the normalized interval $[0, 2]$. Many interpretations of x are consistent with the model described below. For example, $x/2$ might represent the fraction of an agent's time spent on some activity or the fraction of his budget spent on some good; x could also measure qualities such as the brightness of clothing.

Each agent has *intrinsic* preferences over the set X. These preferences are summarized by a utility function, $g(x - t)$. The parameter t is the agent's *intrinsic bliss point* (IBP), in the sense that $g(x - t)$ is maximized at $x = t$. I shall also refer to t as an agent's *type.* I make the following assumption on the function $g(\cdot)$.

Assumption 1. The function $g(z)$ is twice continuously differentiable, strictly concave, and symmetric ($g(z) = g(-z)$) and achieves a maximum at $z = 0$.

Differentiability is assumed to ensure that conformism does not arise trivially from some structural feature of the model. Concavity and symmetry are assumed primarily to simplify the formal arguments.

I shall use T to denote the set of all possible types. Given the structure of preferences, it is natural to take $T = X$ (every point in X is the IBP for some potential type). I shall continue to use both symbols (T and X), despite obvious redundancy, since this helps to avoid confusion about whether types or actions are being discussed.

I assume that the actual population is a continuum. The distribution of types within the population is described by a cumulative density function $F(\cdot)$ defined on T and a corresponding density function $f(\cdot)$. I assume that all possible types are represented. Assumption 2 formally states this.

Assumption 2. $\text{supp}[f(\cdot)] = T$.

So far, the formulation of preferences is fairly standard. Now I shall suppose that, in addition to these intrinsic preferences, each individual also cares about esteem (alternatively, status or popularity). Esteem is, in turn, determined by public perceptions of an individual's type. I assume that all agents will, in equilibrium, form the same inferences (this is standard); consequently, it is possible to summarize an individual's perceived type by a single number, b. I use $h(b)$ to denote the esteem accorded to an individual who is perceived to be of type b.

Note that $h(\cdot)$ is assumed to depend only on b and not on type, t. In other words, all individuals assess and value esteem identically. This assumption is potentially controversial since, for example, each type may care about the opinions of different population subgroups (e.g., those more like themselves). At the cost of considerable analytic complexity, one can assume that $h(\cdot)$ depends on both b and t. I discuss this possibility as an extension of the model in Section V.

I make the following additional assumption on the function $h(\cdot)$.

ASSUMPTION 3. The function $h(\cdot)$ is twice continuously differentiable, strictly concave, and symmetric ($h(1 + z) = h(1 - z)$) and achieves a maximum at $b = 1$.

As before, differentiability is assumed to ensure that conformism does not arise trivially from the structure of the model, and concavity and symmetry simplify the formal analysis. The assumption that $h(b)$ reaches a maximum on the interior of X is central to my analysis and therefore requires further justification.

Intuitively, assumption 3 implies that extremists are esteemed less than centrists. In most contexts, this assumption seems natural. If, for example, one individual's regard for another is correlated with perceived similarity (people like "kindred spirits"), then centrists will generally be more popular in the aggregate than extremists. Alternatively, certain traits (e.g., generosity, bravery, studiousness, ambition, drive, or diligence) may be generally admired, but excessively virtuous people may be disdained as boring, stupid, or simply "different"; those who suffer by comparison may also resent the excessively virtuous. Likewise, although "fashion statements" may enhance status in some circles, extreme flamboyance may simply invite ridicule.

This treatment of esteem is, admittedly, somewhat stylized. It is, however, possible to derive an esteem function $h(\cdot)$ satisfying assumption 3 under more primitive assumptions about the structure of preferences and opinions. The interested reader is referred to Appendix A.

I assume that an individual's type is *not* directly observable. Hence, others must infer his type from his observable choices. Let $\phi(b, x)$ be the inference function. It is important to emphasize that this function will be determined endogenously, as part of an equilibrium. For any choice x, $\phi(\cdot)$ assigns a probability to each possible inference b about type t. The inference function must therefore satisfy

$$\int_T \phi(b, x)\, db = 1 \quad \text{for all } x \in X. \tag{1}$$

Each agent chooses an action x to solve

$$\max_{x \in X} U(x, t, \phi), \tag{2}$$

where

$$U(x, t, \phi) = g(x - t) + \lambda \int_T h(b)\phi(b, x)db. \tag{3}$$

The scalar λ summarizes the weight that each individual attaches to esteem, relative to intrinsic utility. Note that the inference function, though endogenous, is taken as parametric by each agent.

Although equation (3) might at first appear innocuous, it requires some discussion. Suppose that, for some action x, the inference function assigns positive probability to more than one type. How much esteem will an individual be accorded if he or she chooses x? According to equation (3), the answer is determined by taking the expected value of $h(\cdot)$ over types, using the inference function to assign probabilities. Although one can construe this as an expected utility calculation, that interpretation is somewhat forced.[5] It should be emphasized, however, that I use this particular formulation of utility for analytic convenience only. The central arguments in this paper would go through more generally under relatively mild regularity conditions, such as the requirement that those selecting x must be accorded a level of esteem between $\inf\{h(b) | b \in \text{supp}[\phi(\cdot, x)]\}$ and $\sup\{h(b) | b \in \text{supp}[\phi(\cdot, x)]\}$.

The functions $h(b)$ and $\phi(b, x)$ together provide the link between actions and esteem. This formulation is clearly related to Akerlof's notion that an individual sacrifices "reputation" by deviating from the social norm. However, in contrast to Akerlof, I have not built in preferences for conformity by assuming that esteem varies discontinuously with actions. Indeed, $h(b)$ is assumed to be a differentiable function. If a discontinuity occurs, it must appear in the inference function, $\phi(\cdot)$. It is therefore essential to bear in mind that $\phi(\cdot)$ is generated endogenously, as the result of informational equilibrium.

At various points in this paper, it will be helpful to consider a parametric example of this model. The following case will be used for illustrative purposes throughout.

Example.—Suppose that intrinsic preferences and esteem are both quadratic:

$$g(z) = -z^2 \tag{4}$$

[5] In particular, one would need to assume that each individual choosing x is randomly assigned some level of esteem, where the determination of this level is governed by the function $\phi(\cdot, x)$. It is more likely that each individual choosing x is assigned the same level of esteem.

and

$$h(b) = -(1 - b)^2. \tag{5}$$

Then indifference contours in the (x, b) plane for a type t agent are given by

$$(t - x)^2 + \lambda(1 - b)^2 = C, \tag{6}$$

where C is an arbitrary constant. Thus each contour is an ellipse around the point $(t, 1)$. I shall refer to this as the "spherical case" (with an obvious change of scale, indifference contours are spherical).

In general, under assumptions 1 and 3, indifference contours have the following characteristics: they are horizontal when $x = t$ and symmetric around the vertical line given by $x = t$, and they are vertical when $b = 1$ and symmetric around the horizontal line given by $b = 1$. I illustrate typical indifference contours in figure 1. It is immediately evident that the usual "single crossing property" (commonly assumed in signaling models) is *not* satisfied. This observation has a profound effect on the analysis.

It is essential to realize that concern over popularity does not explain conformity by itself. To illustrate this point, suppose that esteem depends only on actions, rather than motivations, or equivalently that inferences are naive in the following sense:

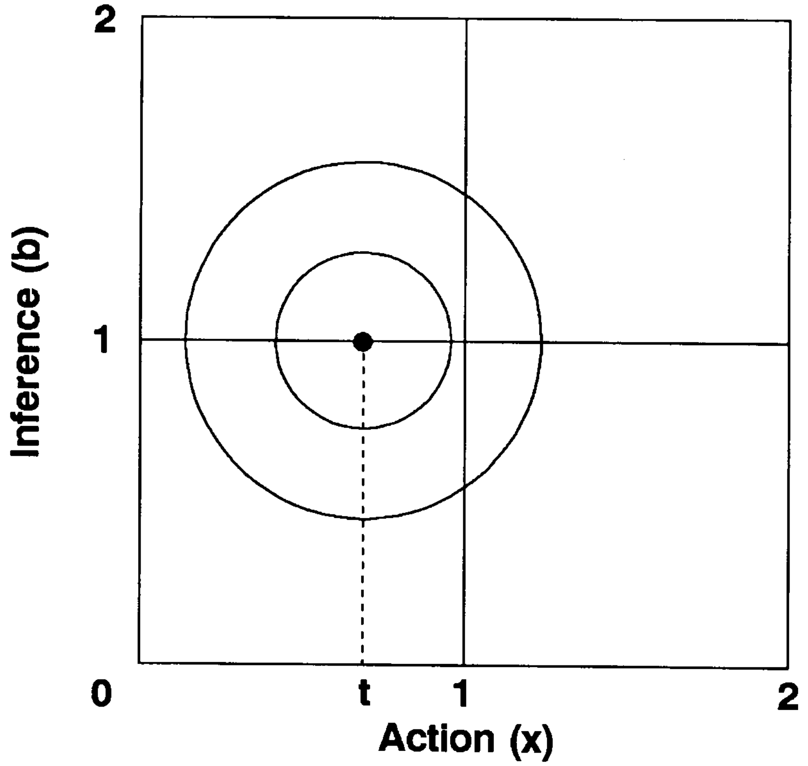

FIG. 1.—Indifference curves

$$\phi(b, x) = \begin{cases} 1 & \text{if } b = x \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

That is, on observing an agent choose x, one infers that x was the agent's IBP. Optimal choices would then be characterized by the following first-order condition:

$$g'(x - t) + \lambda h'(x) = 0. \tag{8}$$

Implicit differentiation yields

$$\frac{dx}{dt} = \frac{g''(x - t)}{g''(x - t) + \lambda h''(t)}. \tag{9}$$

Under the concavity assumptions, this term is *strictly* positive. Thus no two distinct types of agents make the same choice. The presence of the function $h(\cdot)$ makes the distribution of choices more concentrated than the distribution of IBPs (since $dx/dt < 1$), but there is no clustering at any point. Behavior in such a world would be observationally equivalent to that occurring in a society in which the distribution of IBPs was somewhat more concentrated and in which no one cared about popularity.

In addition to illustrating that the model does not build in conformity in some trivial sense, the preceding discussion also implies that naive inferences are not self-sustaining (each type would choose to imitate some type with an IBP closer to unity). Consequently, we have a signaling problem.

Formal definitions of signaling equilibria have been provided in other contexts (e.g., Kreps 1990; Fudenberg and Tirole 1991). Informally, a signaling equilibrium consists of an action function (mapping types to actions) and an inference function (mapping actions to inferences about type) such that actions are optimal given inferences, and inferences can be deduced from the action function using Bayes's law.[6] I shall focus on *pure strategy* equilibria, in which each individual makes a deterministic choice and all individuals of the same type make the same choice. The use of mixed strategies would add considerable notational complexity without altering any of the results.

[6] This definition is not quite standard. Usually, the inferring party also takes some action, and a description of this choice is included as part of the equilibrium. Here, agents care about inferences directly rather than about the inferring parties' actions. This difference has no formal significance, since one can also interpret my model as describing a situation in which agents care about the inferring parties' actions but each inferring party always strictly prefers to take an action that is the same as his or her inference. This interpretation is actually quite natural: individuals may value esteem because esteem affects the way others treat them (the model has, of course, abstracted from this by assuming that esteem affects utility directly).

Henceforth, I shall use $\mu: T \rightarrow X$ to denote the function that maps types to actions.

Before I proceed to the analysis of separating and pooling equilibria, it is useful to begin with a straightforward result that enormously simplifies the task of characterizing signaling equilibria. This result demonstrates that the action function μ is weakly monotonic. (Note: All proofs are contained in App. B.)

THEOREM 1. In any signaling equilibrium, if $t > t'$, then $\mu(t) \geq \mu(t')$.

In the next section, I analyze fully separating equilibria, where (by definition) no conformism occurs. The central result of that section isolates necessary and sufficient conditions for the existence of an equilibrium with complete separation. Sections IV and V characterize equilibria with incomplete separation and show that the characteristics of these equilibria are naturally interpreted as conformism.

III. Fully Separating Equilibria

A fully separating equilibrium is characterized by a function, $\phi_s(x)$, such that

$$\phi(b, x) = \begin{cases} 1 & \text{if } b = \phi_s(x) \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

For expositional purposes, it is useful to begin my analysis by ignoring agents with $t > 1$. In other words, I shall investigate the properties of $\phi_s(x)$ for $t \in [0, 1]$. Since the model is symmetric, the behavior of a type $2 - t$ agent will mirror that of a type t agent.

The slope of a type t indifference curve through any point (x, b) is given by

$$\frac{db}{dx} = -\frac{g'(x - t)}{\lambda h'(b)}. \tag{11}$$

Indifference curves must be tangent to $\phi_s(x)$ at the optimum choice for each agent type. Moreover, choices must be self-fulfilling, in the sense that $\phi_s(\mu(t)) = t$. Thus

$$\phi_s'(x) = -\frac{g'(x - \phi_s(x))}{\lambda h'(\phi_s(x))}. \tag{12}$$

This is easily recognized as a first-order differential equation for ϕ_s as a function of x. In the context of the current model, the appropriate initial condition is $\phi_s(0) = 0$: the least esteemed type is correctly identified in a fully separating equilibrium and therefore has no reason to depart from its IBP. This condition is guaranteed by standard equilibrium refinements (e.g., Bayesian perfection), and it generates the "Riley outcome" (see Riley 1979).

Let $\tilde{X} = \phi_s^{-1}([0, 1])$. The term $\tilde{X}$ represents the set of all choices made by individuals with $t \in [0, 1]$. Generally, $\tilde{X}$ is an interval, $[0, \tilde{x}]$. The existence and uniqueness of $\phi_s(\cdot)$ over $[0, \tilde{x}]$ are guaranteed by standard arguments (see, e.g., Courant and John 1974, 2:706). It is easy to verify that $\phi_s(x) < x$ for $x \in (0, \tilde{x})$.[7] This implies that the function $\phi_s(\cdot)$ remains below the 45-degree line. It follows immediately that $\tilde{x} \geq 1$. Thus in a fully separating equilibrium, individuals of type $t = 1$ would choose an action ($\tilde{x}$) greater than or equal to one.

Now I turn to the central question: Does a fully separating equilibrium exist? So far, we have considered only individuals with $t \in [0, 1]$. Recall that there is also a group of agents with $t \in [1, 2]$. Under our assumptions, the separating solution for this group is the mirror image of the solution just considered. Complete separation of $t \in [1, 2]$ therefore requires that inferences be governed by $2 - \phi_s(2 - x)$ over $[2 - \tilde{x}, 2]$.

One can now see that full separation is sustainable as a signaling equilibrium if and only if $\tilde{x} = 1$. When $\tilde{x} > 1$ (as with ϕ_s^1 in fig. 2), full separation of types $t \in [0, 1]$ is inconsistent with full separation of types $t \in [1, 2]$. If one attempted to piece together ϕ_s^1 with its mirror image, agents with $t = 1$ would be required to shade their choices both to the left and to the right to avoid imitation. As a formal matter, any attempt to combine $\phi_s(\cdot)$ and $2 - \phi_s(\cdot)$ would violate monotonicity (theorem 1). On the other hand, when $\tilde{x} = 1$ (as with ϕ_s^2 in fig. 2), there exists an equilibrium with full separation of all types, since one can simply combine $\phi_s(x)$ on [0, 1] with $2 - \phi_s(2 - x)$ on [0, 2].

This result anticipates the ultimate conclusion, that conformist behavior emerges precisely when $\tilde{x} > 1$. Of course, $\tilde{x}$ is endogenous, and I have not yet even ruled out the possibility that it is identically equal to unity in all cases (which would imply that conformity never arises in this model). In order to obtain some insight into the circumstances that give rise to conformity, it is necessary to describe the relations between $\tilde{x}$ and exogenous preference parameters.

Clearly, there are general conditions under which $\tilde{x} > 1$. Suppose, for example, that

$$g(0) + \lambda h(0) < g(1) + \lambda h(1). \tag{13}$$

[7] From (12) and the initial condition, it follows that $\phi_s'(0) = 0$, so $\phi_s(x) < x$ for small x. Suppose that, for $x' \in (0, \tilde{x})$, $\phi_s(x') > x'$. Then there exists $x'' \in (0, x')$ such that $\phi_s(x'') = x''$ and $\phi_s'(x'') \geq 1$. But eq. (12) implies that $\phi_s'(x'') = 0$ (since $\phi_s(x'') < 1$), which is a contradiction.

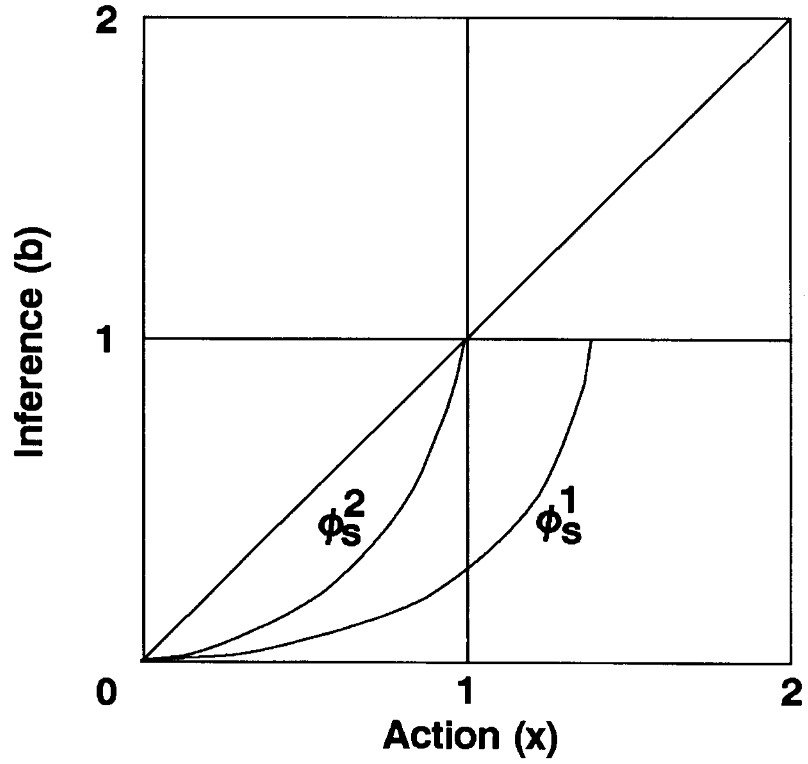

FIG. 2.—Illustration of separating functions

Then obviously one cannot have $\tilde{x} = 1$, since type 0 agents would imitate type 1 agents. Note that this condition holds whenever status is sufficiently important relative to intrinsic utility. It follows that fully separating equilibria do not exist when λ is sufficiently large. In fact, it is possible to prove a much stronger result.[8]

THEOREM 2. There exists $\lambda^* > 0$ such that a fully separating equilibrium exists if and only if $\lambda \leq \lambda^*$.

According to theorem 2, fully separating equilibria *do* exist when status is relatively unimportant ($\lambda \leq \lambda^*$) but *do not* exist when status is relatively important ($\lambda > \lambda^*$).

[8] This result should be contrasted with the analysis of Banks (1990). Banks studied an electoral model with the following features: (i) candidates announce platforms, which represent partial commitments to policies (deviations from platforms are costly); (ii) candidates have intrinsic preferences over policies; and (iii) candidates care about popularity, since it affects electoral outcomes. In this context, platforms can signal the candidate's intrinsic policy preferences (and, thus, his or her choices once elected). On a formal level, the structure of Banks's model is similar to that considered here. However, Banks finds that fully separating equilibria *never* exist, whereas I find that fully separating equilibria fail to exist only if the desire for status is sufficiently strong. The explanation for this discrepancy is that, in Banks's model, the function mapping the candidate's perceived preferences into the probability of electoral victory is necessarily kinked at the point at which perceived preferences coincide with those of the median voter (in contrast, I have assumed that esteem changes smoothly with perceived type). Because of the existence of this kink, "conformity" would arise in Banks's model even without signaling (i.e., even if platforms represented firm commitments).

One can get a more complete sense of the relation between $\tilde{x}$ and preference parameters by considering the spherical example. For this case, equation (12) becomes

$$\phi_s'(x) = \left(\frac{1}{\lambda}\right)\left[\frac{x - \phi_s(x)}{1 - \phi_s(x)}\right]. \tag{14}$$

Note that equation (14) is equivalent to the following linear dynamical system in (t, x):

$$\begin{bmatrix} dt/d\tau \\ dx/d\tau \end{bmatrix} = \begin{bmatrix} x - t \\ \lambda(1 - t) \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -\lambda & 0 \end{bmatrix}\begin{bmatrix} t - 1 \\ x - 1 \end{bmatrix} \equiv \mathbf{A}\begin{bmatrix} t - 1 \\ x - 1 \end{bmatrix}, \tag{15}$$

where τ is an index. It is easy to verify that the matrix $\mathbf{A}$ has real eigenvalues if and only if $\lambda \leq 1/4$. Thus when $\lambda > 1/4$, the separating function ϕ_s has the appearance of the curve labeled ϕ_s^1 in figure 2, and $\tilde{x} > 1$. It follows that no fully separating equilibrium exists if esteem is sufficiently important. As λ falls toward $1/4$, $\tilde{x}$ declines toward one. When $\lambda \leq 1/4$, the solution path starting at $(t, x) = (0, 0)$ converges monotonically to the unique steady state, $(t, x) = (1, 1)$. This implies that the separating function ϕ_s has the appearance of the curve labeled ϕ_s^2 in figure 2, and $\tilde{x} = 1$. Thus, when esteem is not sufficiently important, a fully separating equilibrium does exist. Once $\lambda < 1/4$, further reductions in λ do not alter $\tilde{x}$, but rather flatten ϕ_s^2 against the 45-degree line. For the spherical case, $\lambda \leq 1/4$ is therefore a necessary and sufficient condition for the existence of fully separating equilibria (i.e., $\lambda^* = 1/4$).[9]

It is natural to wonder whether λ^* is defined by expression (13). For the spherical case, this expression is equivalent to the condition that $\lambda > 1$. Thus, although (13) provides an upper bound on λ^*, in general this bound is not very tight. Fully separating equilibria fail to exist in a much wider range of circumstances than those described by (13).

IV. Equilibria with Incomplete Separation

When full separation is not sustainable, there always exist equilibria with incomplete separation. Indeed, there are so many equilibria with incomplete separation that it is difficult to obtain general results. Of course, some equilibria are implausible. This observation suggests that it may be possible to overcome the problem of multiplicity by invoking an equilibrium refinement.

[9] I am grateful to a referee for suggesting this method of analyzing eq. (14).

I shall refine the set of signaling equilibria through a device known as the D1 criterion (see Cho and Kreps 1987). In effect, this criterion insists that, on observing a deviation (defined as an action not taken with positive probability by any type of agent in the candidate equilibrium), an individual will infer that the deviating party belongs to the class of agents who had the greatest incentive to make the observed deviation. Aside from imposing a reasonably plausible restriction on inferences, the D1 criterion has proved hostile to pooling in a variety of contexts. Cho and Sobel (1990) have shown that when the single-crossing property is satisfied, pooling can occur only at a boundary point. Thus, in a standard model, the D1 criterion would rule out the existence of a central "conformist" pool. Moreover, in the current (nonstandard) model, the D1 criterion selects the fully separating equilibrium whenever it does exist. Thus the application of this criterion guarantees that conformity will not arise in this model simply because I have failed to rule out questionable instances of pooling.[10]

In discussing equilibria with incomplete separation, I shall make use of the following notation. For a given action function $\mu(x)$, define

$$T(x) = \{t \in X \mid \mu(t) = x\}, \tag{16}$$

$$t_l(x) = \inf T(x), \tag{17}$$

and

$$t_h(x) = \sup T(x). \tag{18}$$

Note that, by monotonicity (theorem 1), $T(x)$ is an interval. Since each type has measure zero, one can always treat any pool as including its endpoints:

$$T(x) = [t_l(x), t_h(x)]. \tag{19}$$

The following result establishes that, when a fully separating equilibrium fails to exist, any equilibrium satisfying the D1 test is necessarily characterized by the existence of a single, central pool.[11]

[10] It is also worth noting that, in conventional settings, the set of stable equilibria (in the sense of Kohlberg and Mertens [1986]) all pass the D1 test.

[11] It is worth mentioning that other authors have obtained equilibria featuring central pools in the context of other models. Lewis and Sappington (1989) demonstrate that countervailing incentives can produce a central pool in a principal agent model with hidden information. Green and Laffont (1990) consider a model in which one economic agent (the incumbent) operates in many environments at the same time and faces the potential for attack by another agent (the entrant) on each of these fronts. The incumbent is assumed to have private information about each environment. This information affects the incumbent's preferences over actions within each environment, as well as the entrant's preferences for launching an attack. Thus the entrant tries to infer the private information from the incumbent's action. The Bayesian perfect equilibria for this game yield central pools in specific parametric examples. Aside from the

THEOREM 3. If $\lambda > \lambda^*$, then for any signaling equilibrium that satisfies the D1 criterion, there exists at most one $x_p \in X$ such that $t_l(x_p) < t_h(x_p)$, and it satisfies $1 \in T(x_p)$.

Given this result, one can begin to visualize the characteristics of an equilibrium with incomplete separation. There is a single pool, and this pool necessarily contains the central portion of the population distribution (i.e., all types t within some neighborhood of unity). Specifically, for types $t \in [t_l(x_p), t_h(x_p)]$, an agent chooses x_p, and others draw inferences as follows:

$$\phi(b, x_p) = \begin{cases} f(b)[F(t_h) - F(t_l)]^{-1} & \text{if } t_l(x_p) \leq b \leq t_h(x_p) \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

The esteem accorded to individuals choosing x_p is given by $\xi(t_l(x_p), t_h(x_p))$, where

$$\xi(r, s) \equiv \int_r^s h(b)f(b)[F(s) - F(r)]^{-1}db. \tag{21}$$

Types $t \in [0, t_l(x_p))$ fully separate, choosing actions below x_p, with type $t = 0$ selecting $x = 0$ (provided that type 0 is not part of the pool). Separation within this lower tail is governed by the function $\phi_s(x)$; for $t < t_l(x_p)$, an agent chooses $\phi_s^{-1}(t)$, others infer that the agent is of type t, and he or she is accorded esteem of $h(t)$. Likewise, types $t \in (t_h(x_p), 2]$ fully separate, choosing actions above x_p, with type $t = 2$ choosing $x = 2$ (provided that type 2 is not part of the pool). Separation within this upper tail is governed by the function $2 - \phi_s(2 - x)$; for $t > t_h(x_p)$, an agent chooses $2 - \phi_s^{-1}(2 - t)$, others infer that the agent is of type t, and he or she is accorded esteem of $h(t)$.

The preceding paragraph implies that equilibria satisfying the D1 criterion can be characterized by three parameters: x_p, t_l, and t_h. Once these parameters are known, it is a simple matter to construct the equilibrium (provided that one has solved for ϕ_s). Of course, not all triplets (x_p, t_l, t_h) correspond to equilibria. For an equilibrium to obtain, several conditions must be satisfied.

Before I state these conditions, it is useful to define

$$\gamma_l(t_l, x_p) = g(\phi_s^{-1}(t_l) - t_l) + \lambda h(t_l) - g(x_p - t_l) \tag{22}$$

and

$$\gamma_h(t_h, x_p) = g(2 - \phi_s^{-1}(2 - t_h) - t_h) + \lambda h(t_h) - g(x_p - t_h). \tag{23}$$

existence of equilibria with central pools, these papers have relatively little in common with the current paper. For example, Green and Laffont build a structural discontinuity into their model by assuming that the entrant's action set is discrete. Banks (1990) also obtains central pooling equilibria in his model of electoral competition (see n. 8). The relation between my results and those of Banks is discussed again in n. 12.

The function $\gamma_l(t_l, x_p)$ denotes the level of utility from popularity that a type t_l would require in the central pool to be indifferent between the pool and separation. Similarly, the function $\gamma_h(t_h, x_p)$ denotes the level of utility from popularity that a type t_h would require in the central pool to be indifferent between the pool and separation.

The following three conditions are necessary and sufficient for the triplet (x_p, t_l, t_h) to constitute a *central pooling* equilibrium satisfying the D1 criterion:

$$\phi_s^{-1}(t_l) \leq x_p \leq 2 - \phi_s^{-1}(2 - t_h) \tag{24}$$

and

$$\gamma_l(t_l, x_p) \leq \lambda\xi(t_l, t_h) \tag{25}$$

with equality for $t_l > 0$, and

$$\gamma_h(t_h, x_p) \leq \lambda\xi(t_l, t_h) \tag{26}$$

with equality for $t_h < 2$. Expression (24) is required by monotonicity (theorem 1). With respect to expression (25), if the left-hand side ever exceeded the right-hand side, then type t_l agents would imitate a slightly lower type (for the case of $t_l = 0$, they would just choose $x = 0$). If the left-hand side was ever less than the right-hand side and $t_l > 0$, then some agent with t slightly less than t_l would choose x_p and join the central pool. A similar argument applies for expression (26). Thus (24)–(26) are clearly *necessary*. To establish *sufficiency*, one simply constructs the equilibrium, as described above. For the D1 criterion to be satisfied, out-of-equilibrium inferences must be made as follows: for $x \in [\phi_s^{-1}(t_l), x_p)$, $\phi(x, t_l) = 1$ (deviations below x_p are attributed to type t_l agents), and for $x \in (x_p, 2 - \phi_s^{-1}(2 - t_h)]$, $\phi(x, t_h) = 1$ (deviations above x_p are attributed to type t_h agents).

I have illustrated the equilibrium conditions (24)–(26) graphically in figure 3. The variable b_e is defined in the following way:

$$h(b_e) = \xi(t_l, t_h) \tag{27}$$

(and $b_e \leq 1$). Note that the type t_l indifference curve through the point $(\phi_s^{-1}(t_l), t_l)$, denoted I_l, passes through the point (x_p, b_e), thereby assuring that condition (25) is satisfied. A symmetric statement applies for $t_{h.}$

As a further step in the process of characterizing the equilibrium set, I establish next that, for each x_p, there is at *most* one equilibrium of the type described above.

THEOREM 4. For any given $x_p \in X$, there is at most one central pooling equilibrium, (x_p, t_l, t_h).

By itself, this result does not shed any light on the issue of whether or not an equilibrium actually exists for a given x_p. Define

$$X^* = \{x \in X \mid \exists \text{ some central pooling equilibrium } (x, t_l, t_h)\}. \tag{28}$$

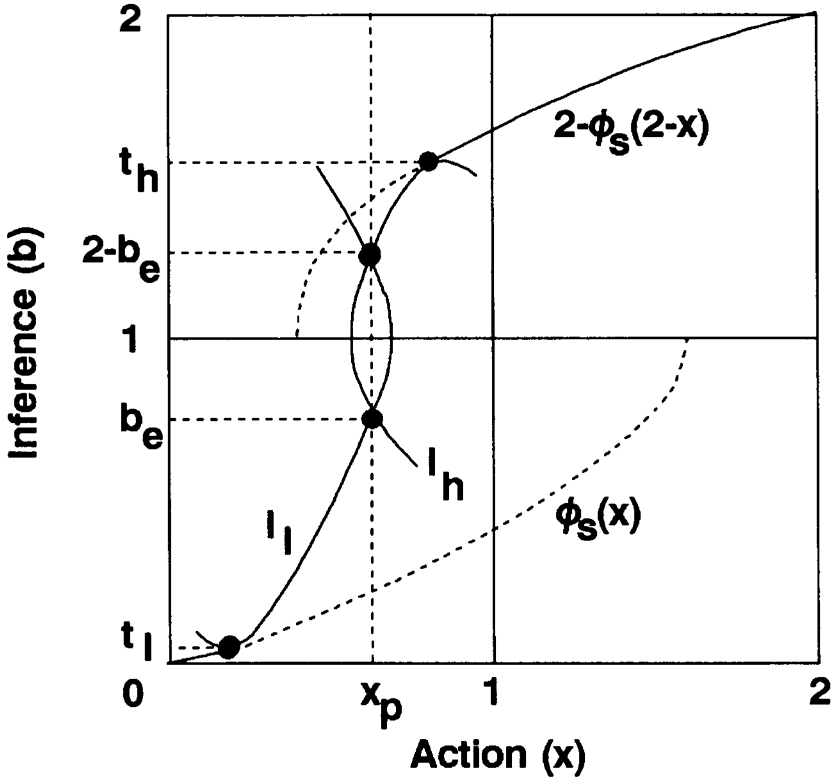

FIG. 3.—Illustration of equilibrium

The following result resolves questions about existence and multiplicity.

THEOREM 5. For $\lambda > \lambda^*$, there exists $\epsilon > 0$ such that $[1 - \epsilon, 1 + \epsilon] \subset X^* \subset [2 - \tilde{x}, \tilde{x}]$.

Thus, when a fully separating equilibrium fails to exist, there is always a central pooling equilibrium for values of x_p in a neighborhood of unity.[12] Moreover, the range of indeterminacy is strictly bounded: the central pool necessarily lies within $(2 - \tilde{x}, \tilde{x})$.[13]

V. Implications and Extensions

In elucidating the implications of this model, one should start by examining the properties of $\mu(t)$, the function that maps types into actions. Figure 4 illustrates a typical equilibrium action function. The

[12] In n. 8, I described the electoral model of Banks (1990). Banks also studied equilibria with incomplete separation. However, he restricted attention to equilibria that are symmetric in the preference space. The corresponding restriction in the current model would be that $x_p = 1$. This restriction would eliminate several important implications discussed in Sec. V.

[13] I suspect that there are natural conditions under which X^* is an interval, but I have not yet investigated this possibility.

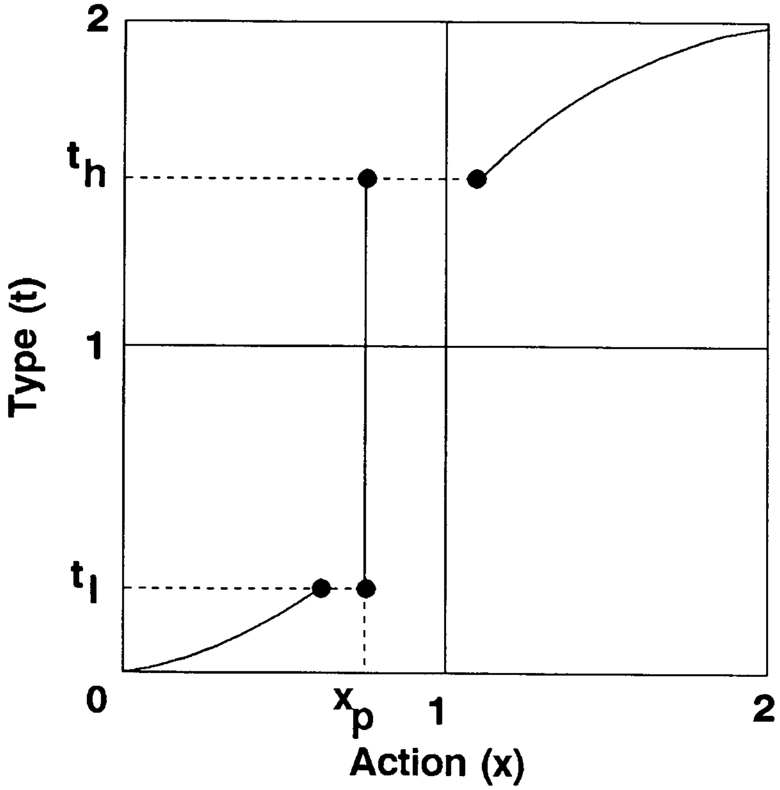

FIG. 4.—An equilibrium action function

important features of this function are (i) that it is constant over the region $[t_l, t_h]$ and (ii) that it is discontinuous at either t_l, t_h, or both. Formally, let

$$\mu_l = \lim_{t \uparrow t_l} \mu(t) \tag{29}$$

and

$$\mu_h = \lim_{t \downarrow t_h} \mu(t). \tag{30}$$

The following result obtains.[14]

THEOREM 6. Suppose that $\lambda > \lambda^*$. Consider a central pooling equilibrium, (x_p, t_l, t_h), such that $t_l > 0$ and $t_h < 2$. Then either $\mu_h > x_p$ or $\mu_l < x_p$. Moreover, there exists $\delta > 0$ such that if $x_p \in [1 - \delta, 1 + \delta]$, then $\mu_h > x_p > \mu_l$.

From these observations, it follows that the population distribution of actual *choices* must exhibit a number of interesting properties. This

[14] It should be possible to strengthen the second half of theorem 6. In particular, it appears that, generically, $\mu(t)$ is discontinuous at *both* t_l and t_h as long as $x_p \in \text{int}(X^*)$. If this is the case, then equilibria will ordinarily exhibit a double discontinuity. This conjecture has not yet been proved.

distribution is obtained by applying the function $\mu(t)$ to the distribution of IBPs. This transformation creates an atom in the distribution of choices at x_p, so that individuals belonging to some heterogeneous subset of the population all behave identically. The discontinuities in $\mu(t)$ (property ii and theorem 6) create a region of zero density around x_p. Since $\mu(t) > t$ for $t < t_l$ and $\mu(t) < t$ for $t > t_h$, application of $\mu(t)$ to the population distribution of IBPs also thins out the tails of the distribution.

Further insight into the equilibrium can be gained by considering the relation between actions and esteem. The D1 criterion isolates equilibria with the property that agents deviating to some $x \in [\phi_s^{-1}(t_l), x_p)$ will be perceived as type t_l and accorded esteem of $h(t_l)$. Likewise, agents deviating to some $x \in (x_p, \phi_s^{-1}(t_h)]$ will be perceived as type t_h and accorded esteem of $h(t_h)$. Consequently, esteem is a discontinuous function of action. In particular, agents are penalized significantly for *any* deviation from the social norm, no matter how small. Thus the discontinuity that Akerlof assumes in his theory of social custom is produced *endogenously* in this model.[15]

It is useful to summarize the central results of this model verbally. When status is sufficiently important relative to intrinsic utility ($\lambda > \lambda^*$), many individuals will strive to conform to a single, homogeneous standard of behavior, despite heterogeneous underlying preferences. They are willing to suppress their individuality and conform to the social norm because they recognize that even small departures from the norm will seriously impair their popularity. The fact that society discontinuously censures all deviations from the accepted mode of behavior is not simply assumed (as in Akerlof's analysis), but rather is produced endogenously; in equilibrium, any departure from the norm is construed as evidence of extreme preferences. Even so, agents with sufficiently extreme preferences will refuse to conform. These "individualists" will behave in ways that differ significantly (rather than trivially) from the social norm. Within the "social fringe," heterogeneous preferences result in heterogeneous behavior; these agents "express their individuality." Nevertheless, even the individualists succumb somewhat to the desire for popularity and shade their choices toward the social norm.

Note that the model generates an explanation for the fact that

[15] Although the D1 criterion necessarily selects an equilibrium with a discontinuous inference function, it should be noted that the same outcome can be sustained as an equilibrium even if the inference function is continuous, as long as it is sufficiently kinked at x_p. The fact that the inference function is discontinuous, rather than simply nondifferentiable, is of no great consequence. Thus the central feature of the model is actually that nondifferentiability arises endogenously even when the underlying structure is assumed to be smooth.

standards of behavior govern some activities but do not govern others, and it allows us to identify the kinds of activities for which social norms arise. In particular, it suggests that a norm is more likely to develop for an activity when preferences over possible choices have a potentially large effect on esteem and when deviations from the most preferred choice do not involve the sacrifice of much intrinsic utility (i.e., $\lambda > \lambda^*$).[16] Larger values of λ generally imply that a larger set of individuals adhere to the norm.[17]

The model also suggests a theory of how standards of behavior might evolve in response to changing preferences. In particular, one could nest the support of the preference distribution within some larger space (such as the real numbers) and allow this distribution to shift through time. Under appropriate assumptions, the equilibria for this dynamic model would correspond to sequences of static equilibria of the type described in this and previous sections.[18] The following discussion briefly explores the properties of these dynamic equilibria.

What happens as the distribution of preferences shifts? Since equilibria are not locally unique, comparative statics are formally indeterminate. However, when a model exhibits a large number of equilibria that are roughly comparable from a formal standpoint, this indeterminacy may be resolved in favor of *focal* choices (see Schelling 1960). Certainly, if a social norm has prevailed in the immediate past, it is more focal than the alternatives. Selecting some other equilibrium from a continuum of alternatives would require a much greater degree of coordinated action. Thus it is natural to assume that the norm remains fixed unless this is incompatible with equilibrium. Formally, if the equilibrium in the previous period was given by (x_p, t_l, t_p) and if, for the current period, there exists a central pooling equilibrium (x_p, t_l', t_p'), then the assumption is that this equilibrium will prevail.

Now recall that X^* contains at least one interval. If x_p lies on the interior of X^*, then sufficiently small changes in the distribution of preferences will not upset the social norm. On the other hand, if the underlying distribution of preferences changes enough, then the preexisting social norm can no longer be sustained, and the society jumps discontinuously to some new standard of behavior.

[16] It is possible to justify this conclusion formally in a model with many distinct decisions, where utility is separable in both the actions and inferences associated with each decision (see Bernheim 1993).

[17] It is easy to verify that, for $x_p = 1$, the size of the central pool increases monotonically with λ. Although I have not established a similar monotonicity property for other values of x_p, one can show that the size of any central pool must shrink to zero as $\lambda \downarrow \lambda^*$ and that, for sufficiently large λ, there will be no deviants.

[18] For a more formal treatment, see Bernheim (1993).

In practice, the degree of conformity and persistence of norms vary greatly over activities. Indeed, the difference between a social custom and a fad is primarily one of degree. Customs have two distinguishing features: they are respected by a large fraction of the population, and they are very persistent. In contrast, a much smaller segment of the population follows fads and fashions, and these norms are much more transient.

In the dynamic model, the value of λ is a critical determinant of whether some specific aspect of behavior is governed by customs, fads, or neither. When λ is very high, a large fraction of the population adheres to the choice x_p. Moreover, since the set X^* is generally large for such an activity, any given standard of behavior will tend to persist, even when the underlying distribution of preferences shifts significantly.[19] This corresponds to the notion of a custom. On the other hand, when λ is near (but not less than) λ^*, a much smaller fraction of the population adheres to the choice x_p. Also, since the set X^* is small for such an activity, standards of behavior are transitory and easily disturbed by relatively small shifts in the underlying distribution of preferences. This corresponds to the notion of a fad.[20] According to this theory, customs develop whenever the impact of an activity on popularity is large relative to the activity's effect on intrinsic utility. It is easy to imagine that something like techniques of eating might fall into this category; hence society develops customs called table manners. In cases in which popularity is somewhat less important relative to intrinsic enjoyment, choices should be influenced by fads and fashions. The characteristics of clothing might fall into this category. Finally, when an activity has little effect on popularity, individualistic behavior should prevail. Accordingly, there may be little conformity with respect to methods of organizing items stored in one's attic.

The fact that the model produces conformity *only* at the center of the population distribution may be seen as a limitation. In practice, for any given society, one may observe many cohesive subgroups, each with its own distinct norm. Indeed, codes of behavior are often more rigid for extremists than for centrists. However, another possi-

[19] I have not established that the range of indeterminacy (the diameter of X^*) expands monotonically with λ. However, theorem 5 implies that the upper and lower bounds on X^* both converge to unity (and the range of indeterminacy shrinks to zero) as $\lambda \downarrow \lambda^*$. Moreover, it is easy to verify that, for sufficiently large λ, $X^* = X$.

[20] The notion of a fad does not necessarily imply cyclicity. Indeed, cyclicity is usually taken to be the distinguishing characteristic of fashions rather than fads. According to Blumer (1968, p. 344), "The most noticeable difference [from fashion] is that fads have no line of historical continuity; each springs up independent of a predecessor and gives rise to no successor."

ble extension of the model suggests an explanation for the development of multiple subcultures, each with its own distinct norm.

In Section II, I assumed that the function $h(\cdot)$ depended only on b and not on type, t. In other words, all individuals assess and value esteem identically. I also mentioned that this assumption is potentially controversial, since, for example, each type may care about the opinions of different population subgroups (e.g., they may place more weight on the opinions of those more like themselves). Suppose instead that a type t individual (one whose intrinsic bliss point is t) has a *perception bliss point* of $p(t)$. That is, a type t individual would most like to be perceived as a type $p(t)$ individual. If a type t individual is publicly perceived to be of type b, his or her utility is given by

$$U(x, t, b) = g(x - t) + \lambda h(b - p(t)), \tag{31}$$

where $h(z)$ is twice continuously differentiable, strictly concave, and symmetric ($h(z) = h(-z)$) and achieves a maximum at $z = 0$. The special case of $p(t) = 1$ for all t corresponds to the model described in Section II.

The shape of the function $p(t)$ will dictate the nature of the resulting equilibria. When $p(t) > t$, individuals of type t are inclined to "lean to the right," in the sense that they most wish to be perceived as being of a type greater than t. On the other hand, when $p(t) < t$, individuals of type t instead "lean to the left," in the sense that they most wish to be perceived as being of a type less than t. In the preceding sections, we have assumed that $p(t) = 1$ for all t, so that $p(t) > t$ if and only if $t < 1$. This implies that all individuals "lean toward the center." It is also possible, however, that some individuals could "lean toward the extremes" (e.g., $p(t) < t$ for $t < 1$) if, for example, they actually dislike being esteemed by those who are sufficiently different from themselves.

Figure 5 depicts an arbitrary perception bliss point function, $p(t)$. In this figure, the function $p(t)$ crosses the 45-degree line four times (at t_1, t_2, t_3, and t_4), dividing T into five regions (labeled I, II, III, IV, and V). As indicated by the arrows, individuals lean to the right in regions I, III, and V; they lean to the left in regions II and IV. Thus one can construct an equilibrium as follows. Divide T into three segments, the first (A) consisting of regions I and II, the second (B) consisting of regions III and IV, and the third (C) consisting of region V. Note that each segment can be treated as a completely separate signaling problem. Depending on the strength of preferences for esteem, for reasons analogous to those explored in the preceding sections, one might obtain central pools within segments A and B, as well as a pool at the upper boundary of segment C. In such an equilibrium, different groups would adhere to different norms. Thus this

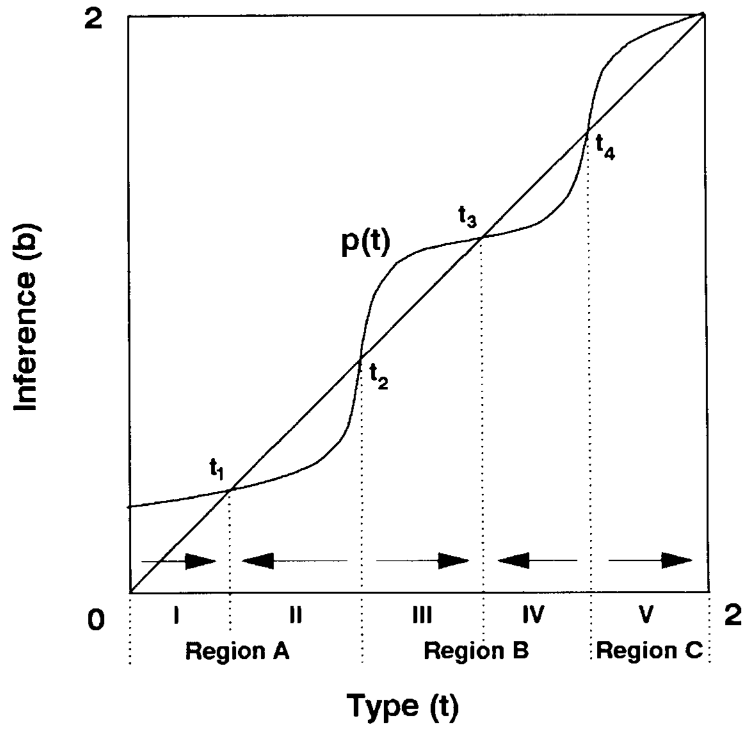

FIG. 5.—Formation of subcultures

extension suggests a possible explanation for the development of cliques and subcultures.[21]

VI. Conclusions

In this paper, I have analyzed a model of social interaction in which individuals care about status as well as "intrinsic" utility (which refers to utility derived directly from consumption). Status is assumed to depend on public perceptions about an individual's predispositions rather than on the individual's actions. However, since predispositions are unobservable, actions signal predispositions and therefore affect status. When status is sufficiently important relative to intrinsic utility, many individuals conform to a single, homogeneous standard of behavior, despite heterogeneous underlying preferences. They are willing to suppress their individuality and conform to the social norm because they recognize that even small departures from the norm will seriously impair their popularity. The fact that society harshly censures all deviations from the accepted mode of behavior is not

[21] Unfortunately, formal analysis of the extended model raises technical issues that do not arise in the simple case in which $p(t) = 1$ for all t. For example, it is easy to construct examples with two types in which there exist nonmonotonic equilibria. A comprehensive treatment of the extended model is left for future work.

simply assumed (indeed, popularity varies smoothly with perceived type); rather, it is produced endogenously. Despite this penalty, agents with sufficiently extreme preferences refuse to conform. The model provides an explanation for the fact that standards of behavior govern some activities but do not govern others. It also suggests a theory of how standards of behavior might evolve in response to a change in the distribution of intrinsic preferences. In particular, for some values of the preference parameters, norms are both persistent and widely followed; for other values, norms are transitory and confined to small groups. Thus the model produces both customs and fads. Finally, a possible extension of the model suggests an explanation for the development of multiple subcultures, each with its own distinct norm.

The theory detailed in this paper is potentially applicable to a variety of economic and social situations. This section offers some preliminary comments on a few of them.

Casual empiricism suggests that concerns over social status and esteem are more pronounced in some cultures than in others. Given this premise, the theory predicts greater conformity, as well as more stable norms, in those cultures that attach more importance to esteem. Casual empiricism supports these predictions. The theory also implies that the range of indeterminacy for x_p should be larger in cultures that attach more importance to esteem. This creates a potential policy role for government. If we assume that the government can, through publicity, make one value of x_p more focal than another, then policy makers may be able to affect real decisions without traditional forms of economic intervention.[22] (This conclusion would be strengthened if, in addition, individuals also attached weight to the "opinion" of the government.) Under this assumption, governmental announcements and publicity campaigns are more potent when the range of indeterminacy for x_p is larger.

A second interesting application concerns the economic performance of worker-owned firms. There is some evidence that workers are more productive when they share in profits. For large firms this phenomenon is puzzling. Although a worker can boost the firm's profits by working harder or by monitoring his co-workers more closely, this has a negligible effect on his own compensation (in other words, each worker has an incentive to free-ride).

There is, however, evidence indicating that work effort is heavily influenced by social pressures that produce conformity to some norm

[22] The analysis therefore provides a formal theory linking public exhortations to behavior. For a discussion of such policies in the context of saving (in both the United States and Japan), see Bernheim (1991).

(Mayo 1945). Even if worker ownership leaves pure economic incentives essentially unchanged, it creates a situation in which a norm of high effort Pareto dominates (for the workers) a norm of low effort.[23] Consequently, the highest values of effort within the region of indeterminacy—rather than the lowest values—may become focal. In addition, worker ownership may alter certain economic parameters. It is natural to think that $p(t)$ would rise, since workers might well have more respect for hardworking co-workers, and that λ would rise, since workers would care more about what their co-workers were doing. The first effect would directly increase equilibrium effort. The second would produce a larger region of indeterminacy, which could then be resolved in favor of a high-effort, Pareto-dominating alternative.

The theory may also help to explain norms of economic equality. For example, there appear to be instances in which employers pay the same wage to all workers belonging to some well-defined group (e.g., "secretary, class I," or first-year assistant professors), despite the fact that some members of the group are noticeably more productive than others. Unequal pay invites the inference that one worker is more valuable than another. This inference may offend the less valued worker, causing him or her to become even less productive. An employer may refrain from tying compensation to relatively small differences in productivity, since the modest gain from improved incentives may be outweighed by the discontinuous loss of morale. Instead, the employer may choose to reward only outstanding employees (and to punish only very poor ones).

This list of applications is by no means exhaustive. Central pooling may account for other phenomena as well, such as particular instances of product standardization (due to signaling on either the supply side or the demand side) or the fact that arbitrators do not appear to exhibit symmetric biases. These and other applications are left for future work.

Appendix A

In this Appendix, I provide a more detailed model of the process through which esteem enters preferences. I use this model to demonstrate that the properties of the function $h(\cdot)$ can be derived from more primitive assumptions.

Let $L(b, t)$ denote the esteem of a type t agent for someone whom he

[23] Without worker ownership, high effort might eventually induce the employer to pay workers more, but this is more distant and speculative. Further, the benefits of higher wages in the future will not be enjoyed by workers who leave the firm in the interim.

believes to be of type b, and let $M(s, t)$ denote the weight assigned by a type t individual to the opinion of a type s individual (these weights are assumed to integrate to unity). Then if a type t individual is perceived as a type b individual, his or her total utility will be given by

$$U(x, t, b) = g(x - t) + \lambda h(b, t), \tag{A1}$$

where

$$h(b, t) = \int_T M(s, t) L(b, s) f(s) ds. \tag{A2}$$

Note that $h(\cdot)$ will be independent of t as long as either $M(s, t)$ is independent of t (all individuals attach the same relative weights to the opinions of others) or $L(b, t)$ is independent of t.

It can be shown that assumption 3 holds under the following conditions: (i) $L(b, t)$ can be written as $L(b - e(t))$, and $L(z)$ achieves a maximum at $z = 0$ (so that $e(t)$ is the type most esteemed by type t); (ii) $L(\cdot)$ is twice continuously differentiable, strictly concave in z, and symmetric; (iii) $e(t)$ is differentiable and symmetric ($e(t) = 2 - e(2 - t)$); and (iv) $M(s, t)$ is independent of t, nonnegative, differentiable in s, and symmetric ($M(1 + s) = M(1 - s)$). It should be noted that assumption iii subsumes the case in which individuals esteem "kindred spirits" ($e(t) = t$), as well as cases in which opposites attract ($e(t) = 2 - t$) and all individuals esteem the same characteristics ($e(t) = 1$). In this last instance ($e(t) = t$), assumption iv is unnecessary.

Appendix B

This Appendix contains proofs of the theorems that are stated in the text. In some instances, I sketch proofs in order to conserve space.

Proof of Theorem 1

Let r be the status level associated with choosing $x = \mu(t)$, and let r' denote the status level associated with choosing $x' = \mu(t')$. Assume $x' > x$. In order to have an equilibrium, it must be the case that

$$g(x - t) + r \geq g(x' - t) + r' \tag{B1}$$

and

$$g(x' - t') + r' \geq g(x - t') + r, \tag{B2}$$

which implies

$$g(x' - t') - g(x - t') \geq g(x' - t) - g(x - t). \tag{B3}$$

But

$$\begin{aligned} &[g(x' - t') - g(x - t')] - [g(x' - t) - g(x - t)] \\ &= \int_x^{x'} [g'(w - t') - g'(w - t)dw = \int_x^{x'} \int_{t'}^{t} g''(w - s) ds dw. \end{aligned} \tag{B4}$$

Since $g(\cdot)$ is strictly concave, this term is strictly negative, which contradicts (B3). Q.E.D.

Proof of Theorem 2

First I show that, if λ is sufficiently large, no fully separating equilibrium exists. Choose $\lambda > [g(0) - g(1)]/[h(1) - h(0)]$ and suppose that there is a fully separating equilibrium. Then if $\mu(1) \leq 1$, type 0 agents would have an incentive to imitate type 1 agents, which contradicts the supposition. On the other hand, if $\mu(1) > 1$, type 2 agents would have an incentive to imitate type 1 agents, which again contradicts the supposition.

Second, I argue that, if $\lambda > 0$ is sufficiently small, a fully separating equilibrium does exist. To establish this property, I prove that, for small λ, $\tilde{x} = 1$. Choose some $\theta > 1$, and define the line segment $B(x) = (1 - \theta) + \theta x$ over the interval $[(\theta - 1)/\theta, 1]$. Note that $B((\theta - 1)/\theta) = 0$ and $B(1) = 1$.

I claim that there is some $K > 0$ such that, for $x \in [(\theta - 1)/\theta, 1)$,

$$G(x) \equiv \frac{g'(x - B(x))}{h'(B(x))} > K. \tag{B5}$$

Since $x > B(x)$ and $B(x) < 1$ for $x < 1$, the term in the middle of (B5) is strictly positive. Thus the claim can be false only if there is some sequence $\langle x_k \rangle_{k=0}^{\infty}$ such that $\lim_{k \to \infty} G(x_k) = 0$. Without loss of generality, suppose that this sequence converges to a single limit point, x^*. Suppose $x^* < 1$. Then, by continuity of $G(\cdot)$, we must have $G(x^*) = 0$. But this cannot be the case under assumption 1, since $x - B(x) > 0$. Now suppose that $x^* = 1$. Then the limit of $G(x_k)$ can be computed through application of L'Hospital's rule:

$$\lim_{x \to 1} G(x) = \frac{(1 - \theta) g''(0)}{\theta h''(1)} > 0, \tag{B6}$$

which again is a contradiction. So the claim is established.

Now choose λ such that $\lambda\theta < K$. I shall argue that $\phi_s(x) > B(x)$ for all $x \in [(\theta - 1)/\theta, 1)$. Since $(\theta - 1)/\theta > 0$, $\phi_s((\theta - 1)/\theta) > B((\theta - 1)/\theta) = 0$. Suppose that there exists $x' \in ((\theta - 1)/\theta, 1)$ such that $\phi_s(x) < B(x)$. Then there must be some $x'' \in ((\theta - 1)/\theta, x')$ such that $\phi_s(x) = B(x)$ and $\phi_s'(x) \leq B'(x)$. But $\phi_s'(x) > K/\lambda > \theta = B'(x)$, which is a contradiction. This argument implies that $\phi_s(x)$ must remain above $B(x)$ when $x < 1$. But that rules out the possibility that $\tilde{x} > 1$.

Third, I argue that $\tilde{x}$ is (weakly) monotonically increasing in λ. So far, for convenience, I have suppressed λ in much of the notation. Here, it will be useful to use $\phi_s(x|\lambda)$ to denote the separating function corresponding to the particular value λ and, similarly, to use $\tilde{x}(\lambda)$ to denote the value of $\tilde{x}$ corresponding to a particular value of λ. Consider λ', λ'' with $\lambda' > \lambda''$. Using equation (12) along with the fact that $\phi_s'(0) = 0$, one can show that

$$\phi_s''(0) = -\left(\frac{1}{\lambda}\right)\left[\frac{g''(0)}{h'(0)}\right] > 0. \tag{B7}$$

Note that this expression is decreasing in λ. It then follows that, for small x,

$\phi_s(x|\lambda'') > \phi_s(x|\lambda')$. Now suppose that $\tilde{x}(\lambda'') > \tilde{x}(\lambda')$. Then there exists some $x < \tilde{x}(\lambda')$ such that $\phi_s(x|\lambda') = \phi_s(x|\lambda'')$ and $\phi_s'(x|\lambda'') \leq \phi_s'(x|\lambda')$ (i.e., the two curves must cross). But in that case,

$$\begin{aligned}
\phi_s'(x|\lambda') &= -\left(\frac{1}{\lambda'}\right)\left[\frac{g'(x - \phi_s(x|\lambda'))}{h'(\phi_s(x|\lambda'))}\right] \\
&= -\left(\frac{\lambda''}{\lambda'}\right)\left(\frac{1}{\lambda''}\right)\left[\frac{g'(x - \phi_s(x|\lambda''))}{h'(\phi_s(x|\lambda''))}\right] \qquad \text{(B8)} \\
&= \left(\frac{\lambda''}{\lambda'}\right)\phi_s'(x|\lambda'') < \phi_s'(x|\lambda''),
\end{aligned}$$

which is a contradiction.

From the monotonicity of $\tilde{x}$, it follows immediately that, if a fully separating equilibrium exists for λ, then one also exists for $\lambda' < \lambda$; likewise, if a fully separating equilibrium does not exist for λ, then no such equilibrium exists for $\lambda' > \lambda$. Thus there exists λ^* such that fully separating equilibria exist for $\lambda < \lambda^*$ and do not exist for $\lambda > \lambda^*$. It can be shown that $\tilde{x}(\lambda)$ is continuous, from which it follows that a fully separating equilibrium also exists for $\lambda = \lambda^*$. Q.E.D.

Proof of Theorem 3

Before I prove this theorem, it will be helpful to introduce some additional notation. Let $U^*(t)$ denote the equilibrium payoff received by a type t agent, and let $H(t)$ denote the equilibrium status of a type t agent:

$$H(t) \equiv \lambda \int_T h(b)\phi(b, \mu(t))\,db. \qquad \text{(B9)}$$

Finally, let

$$I(x, t) = U^*(t) - g(x - t), \qquad \text{(B10)}$$

where $I(x, t)$ denotes the status that would make type t indifferent between choosing x and his or her equilibrium choice. I now establish two claims.

CLAIM 1. Consider any t', t'' with $t' < t''$ and any x, b such that $U(x, t'', b) \geq U^*(t'')$ and $U(x, t', b) \leq U^*(t')$. Then for any $z > x$, $I(z, t') > I(z, t'')$.

To prove this claim, note that

$$\begin{aligned}
I(z, t') - I(z, t'') &= [U^*(t') - g(z - t')] - [U^*(t'') - g(z - t'')] \\
&\geq [U(x, t', b) - g(z - t')] - [U(x, t'', b) - g(z - t'')] \qquad \text{(B11)} \\
&= -\int_x^z \int_{t'}^{t''} g''(q - w)\,dq\,dw > 0.
\end{aligned}$$

CLAIM 2. Consider any t', t'' with $t' < t''$ and any x, b such that $U(x, t', b) \geq U^*(t')$ and $U(x, t'', b) \leq U^*(t'')$. Then for any $z < x$, $I(z, t'') > I(z, t')$.

The proof is completely symmetric to the proof of claim 1.

Now I prove the theorem. Suppose, contrary to the theorem, that there exists (in equilibrium) some pool at x_p, with $t < 1$ for all $t \in T(x_p)$ (since $T(x_p)$ is an interval, we must have either $t < 1$ or $t > 1$; for the case of $t > 1$, the argument is symmetric). Let H_p be the status conferred on members of this pool. For the remainder of this proof, for notational simplicity I shall use t_h to denote $t_h(x_p)$ (likewise for t_l).

In equilibrium, type t_h agents must receive utility

$$U^*(t_h) = g(x_p - t_h) + H_p. \tag{B12}$$

This is obvious if the pool includes its endpoints. Even if the pool does not include its endpoints, equation (B12) must still hold: if the right-hand side exceeded the left-hand side, then type t_h agents would join the pool; if the left-hand side exceeded the right-hand side, then some type in the pool (but close to t_h) would imitate type t_h.

Define

$$\hat{x} = \lim_{t \downarrow t_h} \mu(t). \tag{B13}$$

I claim that $\hat{x} > x_p$. Monotonicity (theorem 1) rules out $\hat{x} < x_p$. If $\hat{x} = x_p$, then $t \in T(x_p)$ would have an incentive to imitate some type slightly greater than t_h, thereby achieving a discrete improvement in status at the cost of an arbitrarily small decline in intrinsic utility.

Monotonicity (theorem 1) implies that no type chooses $x \in (x_p, \hat{x})$. I claim that, under the D1 criterion, $\phi(t_h, x) = 1$ for all $x \in (x_p, \hat{x})$. Note that if this claim is true, type t_h has an incentive to deviate to some x' slightly greater than x_p, thereby achieving a discrete improvement in status at the cost of an arbitrarily small decline in intrinsic utility. Consequently, by proving the claim, I introduce a contradiction and thereby establish the theorem.

Consider any $t < t_h$. Take $x = x_p$ and b such that $\lambda h(b) = H_p$, and apply claim 1. It follows that, for $x \in (x_p, \hat{x})$, $I(x, t) > I(x, t_h)$. Under the D1 criterion, this implies that $\phi(t, x) = 0$ for $x \in (x_p, \hat{x})$ and $t < t_h$. Now consider any $t > t_h$. Choose some t' such that $t > t' > t_h$. Note that $\mu(t') \geq \hat{x}$. Take $x = \mu(t')$ and b such that $\lambda h(b) = H(t')$, and apply claim 2. It follows that, for $x \in (x_p, \hat{x})$, $I(x, t) > I(x, t')$. Under the D1 criterion, this implies that $\phi(t, x) = 0$ for $x \in (x_p, \hat{x})$ and $t > t_h$.

For completeness, one must rule out the possibility that, for any $x \in (x_p, \hat{x})$, there is some t for which $I(x, t_h) > I(x, t)$. The preceding argument rules out $t < t_h$. It also demonstrates that $I(x, t)$ is strictly decreasing in t for $t > t_h$. Thus, if there exists such a t, it must be the case that

$$\lim_{t \downarrow t_h} I(x, t) > I(x, t_h). \tag{B14}$$

But this implies that

$$\lim_{t \downarrow t_h} U^*(t) > U^*(t_h), \tag{B15}$$

which in turn implies that type t_h would imitate some type t slightly greater than t_h. Thus the claim and the theorem are established. Q.E.D.

Proof of Theorem 4

I begin with some preliminary results.

LEMMA 1. $\gamma_l(t, x)$ is decreasing in t for $t < \phi_s(s)$, and $\gamma_h(t, x)$ is increasing in t for $t > 2 - \phi_s(2 - x)$.

Proof. Consider $t', t'' < \phi_s(x)$, with $t' < t''$. Then

$$\begin{aligned}\gamma_l(t', x) - \gamma_l(t'', x) &= \left[\gamma_l(t', \phi_s^{-1}(t'')) - \int_{\phi_s^{-1}(t'')}^{x} g'(z - t')dz\right] \\ &\quad - \left[\gamma_l(t'', \phi_s^{-1}(t'')) - \int_{\phi_s^{-1}(t'')}^{x} g'(z - t'')dz\right] \\ &= [\gamma_l(t', \phi_s^{-1}(t'')) - \lambda h(t')] \\ &\quad + \int_{\phi_s^{-1}(t'')}^{x} [g'(z - t'') - g'(z - t')]dz.\end{aligned} \tag{B16}$$

The first term on the third line is nonnegative, since otherwise the t' agents would imitate the t'' agents. The second term is strictly positive, since $g'(a) < 0$ and $g''(a) > 0$ for $a > 0$. A completely symmetric argument establishes the desired result for $\gamma_h(\cdot)$. Q.E.D.

LEMMA 2. Suppose that (x_p, t_l', t_h') is an equilibrium. Consider $(t_l'', t_h'') \neq (t_l', t_h')$ with $t_l'' \leq t_l' \leq t_h' \leq t_h''$. Then $\xi(t_l'', t_h'') < \xi(t_l', t_h')$.

Proof. I establish this lemma through a series of claims.

CLAIM 1. $\xi(t_l', t_h') \geq \max\{h(t_l'), h(t_h')\}$.

Since (x_p, t_l', t_h') is an equilibrium,

$$\lambda\xi(t_l', t_h') - \lambda h(t_l') \geq g(\phi_s^{-1}(t_l') - t_l') - g(x_p - t_l') \geq 0, \tag{B17}$$

where the second inequality follows from $t_l' \leq \phi_s^{-1}(t_l') \leq x_p$. A similar argument applies for $h(t_h')$.

CLAIM 2. Consider (t_l, t_h) with $\xi(t_l, t_h) \geq h(t_l)$. For all $t < t_l$, $\xi(t, t_h) > h(t)$. Likewise, if $\xi(t_l, t_h) \geq h(t_h)$, then for all $t > t_h$, $\xi(t_l, t) > h(t)$.

I shall prove this claim for t_l. The argument for t_h is symmetric. Note that

$$\begin{aligned}\xi(t, t_h) - h(t) &= \left[\frac{F(t_h) - F(t_l)}{F(t_h) - F(t)}\right]\xi(t_l, t_h) + \left[\frac{F(t_l) - F(t)}{F(t_h) - F(t)}\right]\xi(t, t_l) - h(t) \\ &> \left[\frac{F(t_h) - F(t_l)}{F(t_h) - F(t)}\right]h(t_l) + \left[\frac{F(t_l) - F(t)}{F(t_h) - F(t)}\right]h(t) - h(t) \\ &= \left[\frac{F(t_h) - F(t_l)}{F(t_h) - F(t)}\right][h(t_l) - h(t)] > 0.\end{aligned} \tag{B18}$$

The claim is thereby established.

CLAIM 3. Consider $t < 1$ such that $\xi(t, 2 - t) \geq h(t)$. For $t' < t$, $\xi(t', 2 - t') > h(t')$.

The proof of this claim is completely analogous to that of claim 2.

CLAIM 4. $\partial\xi(t_l, t_h)/\partial t_l > 0$ iff $\xi(t_l, t_h) > h(t_l)$, and $\partial\xi(t_l, t_h)/\partial t_h < 0$ iff $\xi(t_l, t_h) > h(t_h)$.

To prove this claim, simply note that

$$\frac{\partial \xi(t_l, t_h)}{\partial t_l} = f(t_l)[F(t_h) - F(t_l)]^{-1}[\xi(t_l, t_h) - h(t_l)]. \tag{B19}$$

The argument for t_h is analogous.

Now I prove the lemma. Without loss of generality, assume $t_l' \leq 2 - t_h'$ (when this does not hold, the argument is symmetric). There are three cases to consider.

Case 1

Suppose that $t_l'' \leq t_l' \leq 2 - t_h'' \leq 2 - t_h'$. Then

$$\xi(t_l'', t_h'') = \xi(t_l', t_h') + \int_{t_h'}^{t_h''} \frac{\partial \xi(t_l', t)}{\partial t} dt - \int_{t_l''}^{t_l'} \frac{\partial \xi(t, t_h'')}{\partial t} dt. \tag{B20}$$

By claims 1 and 2, $\xi(t_l', t) > h(t)$ for all $t \in (t_h', t_h'']$. By claim 4, this implies $\partial \xi(t_l', t)/\partial t < 0$ for all such t. Consequently, the second term on the right-hand side of (B20) is nonpositive (strictly negative if $t_h'' > t_h'$).

Since $\xi(t_l', t_h'') \geq h(t_h'')$ (as argued above) and $h(t_h'') \geq h(t_l')$ (since $t_l' \leq 2 - t_h''$), we have $\xi(t_l', t_h'') \geq h(t_l')$. By claim 2, $\xi(t, t_h'') \geq h(t)$ for all $t \in [t_l'', t_l')$. By claim 4, this implies $\partial \xi(t, t_h'')/\partial t > 0$ for all such t. Consequently, the third term on the right-hand side of (B20) is nonpositive (strictly negative if $t_l'' < t_l'$).

Since one of these inequalities must be strict, combining them gives $\xi(t_l'', t_h'') < \xi(t_l', t_h')$.

Case 2

Suppose that $t_l'' \leq 2 - t_h'' \leq t_l' \leq 2 - t_h'$. Then

$$\begin{aligned} \xi(t_l'', t_h'') = \xi(t_l', t_h') &+ \int_{t_h'}^{2-t_l'} \frac{\partial \xi(t_l', t)}{\partial t} dt \\ &- \int_{2-t_h''}^{t_l'} \frac{\partial \xi(t, 2-t)}{\partial t} dt - \int_{t_l''}^{2-t_h''} \frac{\partial \xi(t, t_h'')}{\partial t} dt. \end{aligned} \tag{B21}$$

Arguing exactly as in case 1, we see that the second term on the right-hand side of (B21) is nonpositive, and $\xi(t_l', 2 - t_l') \geq h(t_l')$. From claim 3, this implies that $\xi(t, 2 - t) > h(t)$ for all $t \in [2 - t_h'', t_l')$. By claim 4, $\partial \xi(t, 2 - t)/\partial t > 0$ for such t. Thus the third term is nonpositive. Since $\xi(2 - t_h'', t_h'') \geq h(2 - t_h'')$, one establishes that the fourth term is nonpositive exactly as for the third term in (B20). Thus $\xi(t_l'', t_h'') < \xi(t_l', t_h')$ (again, the strict inequality follows because one of the terms just mentioned must be strictly negative).

Case 3

Suppose that $2 - t_h'' \leq t_l'' \leq t_l' \leq 2 - t_h'$. The proof is completely analogous to that for case 2.

Since these three cases are exhaustive, the lemma is proved. Q.E.D.

Now I prove the theorem. Suppose that there are two equilibria, (x_p, t_l', t_h') and (x_p, t_l'', t_h''). There are two cases to consider.

Case A

Suppose that $t_l' \leq t_l'' < t_h'' \leq t_h'$ (where at least one of the weak inequalities must be strict). Then, by lemmas 1 and 2,

$$\xi(t_l'', t_h'') > \xi(t_l', t_h'), \tag{B22}$$

$$\gamma_l(t_l'', x_p) \leq \gamma_l(t_l', x_p), \tag{B23}$$

and

$$\gamma_h(t_h'', x_p) \leq \gamma_h(t_h', x_p), \tag{B24}$$

where the inequality in (B23) is strict if $t_l'' > t_l'$ and the inequality in (B24) is strict if $t_h'' < t_h'$. Suppose without loss of generality that $t_l'' > t_l'$. Then

$$\gamma_l(t_l'', x_p) < \gamma_l(t_l', x_p) \leq \lambda\xi(t_l', t_h') < \lambda\xi(t_l'', t_h''). \tag{B25}$$

But since $t_l'' > 0$, this contradicts the assumption that (x_p, t_l'', t_h'') is an equilibrium (specifically, eq. [25] is violated).

Case B

Suppose that $t_l' < t_l'' < t_h' < t_h''$. Then, since $t_l'' > 0$ and $t_h' < 2$,

$$\begin{aligned} \lambda\xi(t_l'', t_h'') &= \gamma_l(t_l'', x_p) < \gamma_l(t_l', x_p) \leq \lambda\xi(t_l', t_h') \\ &= \gamma_h(t_h', x_p) < \gamma_h(t_h'', x_p). \end{aligned} \tag{B26}$$

But this contradicts the assumption that (x_p, t_l'', t_h'') is an equilibrium (specifically, eq. [26] is violated).

Note that cases A and B are exhaustive. Consequently, there is at most one equilibrium. Q.E.D.

Proof of Theorem 5

First I establish the existence of $\epsilon > 0$ such that $[1 - \epsilon, 1 + \epsilon] \subset X^*$. For each value of x such that $2 - \tilde{x} \leq x \leq 1$ and b with $1 \geq b \geq \phi_s(x)$, define $\tau_l(b, x)$ as the solution τ, $0 \leq \tau \leq x$, to

$$\lambda h(b) = \gamma_l(\tau, x). \tag{B27}$$

If no solution exists, then $\tau_l(b, x) = 0$; $\tau_l(b, x)$ simply inverts the function $\gamma_l(\cdot, x)$. Likewise, for each value of x such that $2 - \tilde{x} \leq x \leq 1$ and b with $1 \geq b \geq \phi_s(2 - x)$, define $\tau_h(b, x)$ as the solution τ, $2 \geq \tau \geq x$, to

$$\lambda h(2 - b) = \gamma_h(\tau, x). \tag{B28}$$

If no solution exists, then $\tau_h(b, x) = 2$. It is easy to verify that $\gamma_k(\cdot)$ is continuous for $k = l, h$. Combining this with lemma 1 and the monotonicity of $h(\cdot)$, one can show that $\tau_k(\cdot)$ is single-valued and continuous.

Note that $\phi_s(2 - x) \geq \phi_s(x)$. So on $1 \geq b \geq \phi_s(2 - x)$, we can define the function

$$\xi^*(b, x) \equiv \xi(\tau_l(b, x), \tau_h(b, x)). \tag{B29}$$

It is easy to check that the equilibrium conditions (24), (25), and (26) are equivalent to

$$h(b^*) = \xi^*(b^*, x_p). \tag{B30}$$

Then t_l and t_h are, respectively, $\tau_l(b^*, x)$ and $\tau_h(b^*, x)$. We know that $h(1) > \xi^*(1, x_p)$ (since $\tau_l(1, x) < 1$). Moreover, the relevant functions are continuous. Therefore, a sufficient condition for existence is

$$h(\phi_s(2 - x_p)) \leq \xi^*(\phi_s(2 - x_p), x_p). \tag{B31}$$

Consider the case of $x_p = 1$. It is easily verified by direct substitution that

$$\tau_l(\phi_s(1), 1) = \phi_s(1) \tag{B32}$$

and

$$\tau_h(\phi_s(1), 1) = 2 - \phi_s(1). \tag{B33}$$

From equations (B32) and (B33), it follows that

$$h(\phi_s(1)) < \xi(\phi_s(1), 2 - \phi_s(1)) = \xi^*(\phi_s(1), 1). \tag{B34}$$

So (B31) is satisfied *strictly* for $x_p = 1$. By continuity, central pooling equilibria exist for x_p in a neighborhood of one. Thus there exists ϵ with the desired property.

Now suppose that there is a central pooling equilibrium (x_p, t_l, t_h) with $x_p \geq \bar{x} \geq 1$. Since $\xi(t_l, t_h) < h(1)$, for all $t \leq 1$ we have

$$g(\phi_s^{-1}(t) - t) + \lambda h(t) \geq g(\bar{x} - t) + \lambda h(1) > g(x_p - t) + \lambda \xi(t_l, t_h). \tag{B35}$$

Since $t_l \leq 1$, we must therefore have

$$\gamma_l(t_l, x_p) > \lambda \xi(t_l, t_h), \tag{B36}$$

which contradicts (25). A symmetric argument implies that there is no central pooling equilibrium (x_p, t_l, t_h) with $x_p \leq 2 - \bar{x}$. Q.E.D.

Proof of Theorem 6

Suppose that $\mu_l = x_p = \mu_h$. Then

$$\xi(t_l, t_h) = \gamma_l(t_l, x_p) = h(t_l). \tag{B37}$$

The first equality follows from the fact that the equilibrium is interior. The second equality follows from the fact that $\phi_s^{-1}(t_l) = \mu_l = x_p$. A similar argument establishes that

$$\xi(t_l, t_h) = h(t_h). \tag{B38}$$

Consequently, $h(t_l) = h(t_h)$, so $t_l = 2 - t_h$. But $\xi(t_l, 2 - t_l) > h(t_l)$ for $t_l < 1$, which is a contradiction.

Now I prove that, for some $\delta > 0$, if $x_p \in [1 - \delta, 1 + \delta]$, then $\mu_h > x_p > \mu_l$. In the proof of theorem 5, I established the existence of some neighborhood around $x_p = 1$ for which (B31) holds as a strict inequality. Choose δ such that $[1 - \delta, 1 + \delta]$ lies in that neighborhood.

Choose some $x_p \in [1 - \delta, 1 + \delta]$, and suppose without loss of generality that, contrary to the theorem, $\mu_l = x_p$. Arguing exactly as before, we have $\xi(t_l, t_h) = h(t_l)$. There are now three cases to consider.

Case 1

Suppose that $x_p < 1$. Then $\phi_s(2 - x_p) > \phi_s(x_p) = t_l = h^{-1}(\xi(t_l, t_h))$. (Note that throughout I follow the convention that $h^{-1}(z) \leq 1$; obviously, since $h(\cdot)$ is symmetric around one, the inverse is not defined uniquely in the absence of this convention.) Consequently, all t strictly prefer $(x_p, h(\phi_s(2 - x_p)))$ to $(x_p, \xi(t_l, t_h))$. But t_h weakly prefers $(2 - \phi_s^{-1}(2 - t_h), h(t_h))$ to $(x_p, h(\phi_s(2 - x_p)))$ (by construction of $\phi_s(\cdot)$) and therefore cannot be indifferent between $(x_p, \xi(t_l, t_h))$ and $(2 - \phi_s^{-1}(2 - t_h), h(t_h))$ as required by equilibrium.

Case 2

Suppose that $x_p = 1$. It is easy to show that the equilibrium must be symmetric in this case ($t_h = 2 - t_l$). Then plainly $\xi(t_l, t_h) = \xi(t_l, 2 - t_l) > h(t_l)$, which is a contradiction.

Case 3

Suppose that $x_p > 1$. By symmetry, $\xi(t_l, t_h) = \xi(2 - t_h, 2 - t_l)$; moreover, the fact that (x_p, t_l, t_h) is an equilibrium implies that $(2 - x_p, 2 - t_h, 2 - t_l)$ is an equilibrium. Since $2 - x_p < 1$, it follows that

$$h[h^{-1}(\xi(2 - t_h, 2 - t_l))] = \xi^*[h^{-1}(\xi(2 - t_h, 2 - t_l)), 2 - x_p], \qquad \text{(B39)}$$

so

$$\xi(t_l, t_h) = \xi^*[h^{-1}(\xi(t_l, t_h)), 2 - x_p] = \xi^*(t_l, 2 - x_p). \qquad \text{(B40)}$$

Now I reason as follows:

$$\begin{aligned} h(\phi_s(x_p)) &= h(\phi_s(\phi_s^{-1}(t_l)) = h(t_l) \\ &= \xi(t_l, t_h) = \xi^*(t_l, 2 - x_p) = \xi^*(\phi_s(x_p), 2 - x_p). \end{aligned} \qquad \text{(B41)}$$

Define $\hat{x} = 2 - x_p$; $\hat{x} < 1$, and I have shown that

$$h(\phi_s(2 - \hat{x})) = \xi^*(\phi_s(2 - \hat{x}), \hat{x}). \qquad \text{(B42)}$$

So by definition $\hat{x} \notin [1 - \delta, 1 + \delta]$. But then $x_p \notin [1 - \delta, 1 + \delta]$, which is a contradiction. Q.E.D.

References

Akerlof, George A. "A Theory of Social Custom, of Which Unemployment May Be One Consequence." *Q.J.E.* 94 (June 1980): 749–75.

Bagwell, Laurie S., and Bernheim, B. Douglas. "Veblen Effects in a Theory of Conspicuous Consumption." Manuscript. Princeton, N.J.: Princeton Univ., May 1993.

Banerjee, Abhijit. "A Simple Model of Herd Behavior." Manuscript. Princeton, N.J.: Princeton Univ., December 1989.

Banerjee, Abhijit, and Besley, Timothy. "Peer Group Externalities and Learning Incentives: A Theory of Nerd Behavior." John M. Olin Discussion Paper no. 68. Princeton, N.J.: Princeton Univ., December 1990.

Banks, Jeffrey S. "A Model of Electoral Competition with Incomplete Information." *J. Econ. Theory* 50 (April 1990): 309–25.

Bernheim, B. Douglas. *The Vanishing Nest Egg: Reflections on Saving in America.* New York: Twentieth Century Fund, Priority Press, 1991.

———. "A Theory of Conformity." Manuscript. Princeton, N.J.: Princeton Univ., April 1993.

Besley, Timothy, and Coate, Stephen. "Understanding Welfare Stigma: Resentment and Statistical Discrimination." Manuscript. Princeton, N.J.: Princeton Univ., 1990.

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.

Blumer, Herbert G. "Fashion." In *International Encyclopedia of the Social Sciences,* vol. 5, edited by David L. Sills. New York: Macmillan, 1968.

Cho, In-Koo, and Kreps, David M. "Signaling Games and Stable Equilibria." *Q.J.E.* 102 (May 1987): 179–221.

Cho, In-Koo, and Sobel, Joel. "Strategic Stability and Uniqueness in Signaling Games." *J. Econ. Theory* 50 (April 1990): 381–413.

Cole, Harold L.; Mailath, George J.; and Postlewaite, Andrew. "Social Norms, Savings Behavior, and Growth." *J.P.E.* 100 (December 1992): 1092–1125.

Conlisk, John. "Costly Optimizers versus Cheap Imitators." *J. Econ. Behavior and Organization* 1 (September 1980): 275–93.

Courant, Richard, and John, Fritz. *Introduction to Calculus and Analysis.* 2 vols. New York: Wiley, 1974.

Fershtman, Chaim, and Weiss, Yoram. "Social Status, Culture and Economic Performance." Manuscript. Tel-Aviv: Tel-Aviv Univ., 1992.

Frank, Robert H. "The Demand for Unobservable and Other Nonpositional Goods." *A.E.R.* 75 (March 1985): 101–16.

Fudenberg, Drew, and Tirole, Jean. *Game Theory.* Cambridge, Mass.: MIT Press, 1991.

Glazer, Amihai, and Konrad, Kai A. "A Signalling Explanation for Private Charity." Economics Paper no. 90-92-35. Irvine: Univ. California, August 1992.

Green, Jerry, and Laffont, Jean-Jacques. "Competition on Many Fronts: A Stackelberg Signaling Equilibrium." *Games and Econ. Behavior* 2 (September 1990): 247–72.

Ireland, Norman. "On Limiting the Market for Status Signals." Manuscript. Warwick: Univ. Warwick, February 1992.

Jones, Stephen R. G. *The Economics of Conformism.* Oxford: Blackwell, 1984.

Kandori, Michihiro; Mailath, George J.; and Rob, Rafael. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica* 61 (January 1993): 29–56.

Katz, Michael L., and Shapiro, Carl. "Technology Adoption in the Presence of Network Externalities." *J.P.E.* 94 (August 1986): 822–41.

Kohlberg, Elon, and Mertens, Jean-Francois. "On the Strategic Stability of Equilibria." *Econometrica* 54 (September 1986): 1003-37.

Kreps, David M. *A Course in Microeconomic Theory.* Princeton, N.J.: Princeton Univ. Press, 1990.

Leibenstein, Harvey. "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand." *Q.J.E.* 64 (May 1950): 183–207.

Lewis, Tracy R., and Sappington, David E. M. "Inflexible Rules in Incentive Problems." *A.E.R.* 79 (March 1989): 69–84.

Matsuyama, Kiminori. "Custom versus Fashion: Hysteresis and Limit Cycles in a Random Matching Game." Discussion Paper no. 940. Evanston, Ill.: Northwestern Univ., June 1991.

Mayo, Elton. *The Social Problems of an Industrial Civilization.* Boston: Harvard Univ., Grad. School Bus. Admin., 1945.

Riley, John G. "Informational Equilibrium." *Econometrica* 47 (March 1979): 331–59.

Ross, Lee; Bierbrauer, Gunter; and Hoffman, Susan. "The Role of Attribution Processes in Conformity and Dissent: Revisiting the Asch Situation." *American Psychologist* 31 (February 1976): 148–57.

Schelling, Thomas C. *The Strategy of Conflict.* Oxford: Oxford Univ. Press, 1960.

Veblen, Thorstein. *The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions.* London: Macmillan, 1899. Reprint. London: Unwin, 1970.