On the Emergence of Social Conformity

Edward Cartwright

University of Kent

January 2005

Abstract

We consider a dynamic model of conformity that permits both a conformist and a non-conformist equilibrium. We provide conditions under which conformity can ‘invade' a population. More precisely, starting from a non-conformist equilibrium, we show that the conformity of an arbitrarily small proportion of the population can lead to the spread of conformism and the ultimate emergence of the conformist equilibrium. This occurs independently of whether or not the non-conformist equilibrium Pareto dominates the conformist equilibrium.

JEL Classification: C70, D01, C72.

Keywords: Conformity, best reply, coordination, norm.

Address for correspondence: Edward Cartwright, Department of Economics, Keynes College, University of Kent, Canterbury, Kent. CT2 7NP, UK. Tel: 44 (0)1227 823460. email: E.J.Cartwright@kent.ac.uk

1 Introduction

It has long been recognized that individuals are influenced by desires for feelings such as esteem, popularity or acceptance (see e.g. Jones 1984, Lindbeck 1997). These desires can lead to the emergence of a norm of behaviour to which individuals conform. Those who deviate from the norm face 'being punished' for 'not fitting in' or being seen as 'extreme'. Often, however, we see conformity existing in some choice situations but not in others; in one instance we may find very strong conformity with all individuals behaving in some common way, while in another, seemingly very similar situation, individuals may be found simply 'doing what they want' with little, if any, observed conformity.[1] This poses the question of why conformity may or may not arise in different contexts. It also feeds through to the general question of why conformity may arise at all. This paper considers a simple model to try and address these questions.

One explanation for why conformity may exist in some contexts but not others is that the relative strength of desires for popularity and esteem may vary in different contexts. This explanation is put forward by Bernheim (1994). In the model of conformity introduced in Bernheim an agent's payoff is a weighted sum of two components - intrinsic utility, determined by his actions, and esteem, determined by what 'type' others believe him to be. Conformity arises if agents are sufficiently motivated by esteem to sacrifice intrinsic utility in order to 'fit in' with the societal norm. Whether or not conformity arises depends on the relative weight given to esteem - if esteem has low weight then conformity does not result while the larger is the weight given to esteem the stronger is the conformity.

An alternative explanation for why conformity may arise in some contexts

[1] One example is dress code. In some departments nearly all male staff wear suits, in some nearly all wear casual wear and in others people 'wear what they want'. On a larger scale, in London male workers typically wear a dark suit while in Paris workers simply seem to 'wear what they like'.

and not others is offered by Akerlof (1980). Akerlof envisages a multiple equilibrium environment in which there is (i) an equilibrium in which a norm is established and equilibrium behavior is to conform to that norm (even if this means not maximizing intrinsic utility), or alternatively, (ii) an equilibrium where a norm is not established (or at least the norm is to just 'do what you want') and so individuals are not guided by fears of being seen as 'extreme'. Whether or not conformity arises in this multiple equilibrium setting is largely a matter of historical chance.

In this paper we propose a model, drawing, in particular, on the research of Bernheim (1994), that attempts to bring together these two differing explanations for why conformity may or may not arise. More particularly, an agent will have the choice between two strategies - to conform or not-conform. If an agent chooses not to conform then he behaves so as to maximize his intrinsic utility. If an agent conforms then he performs some 'focal action' and in so doing sacrifices intrinsic utility but is guaranteed to not feel 'the odd one out'. Crucially, we assume that if an agent conforms he accords others esteem depending on their behaviour with those choosing the 'focal action' given most esteem. Conversely, if an agent does not-conform he accords the same esteem to everyone irrespective of their behaviour. The consequence is a coordination game - if everybody chooses to not-conform then it is optimal to not-conform while if everybody conforms it is optimal to conform. In the framework of Bernheim (1994) the choice of conformity or non-conformity could be seen as a choice between giving a high or low weight to esteem in the utility function. Preferences and the importance of esteem are thus in a certain sense determined within the model.[2]

As always in coordination games its an open question what strategy an individual should choose. In this particular setting an individual has to weigh up whether he thinks a conformity norm will or will not exist or, in other words, what proportion of the population will or will not conform. Evolution-

[2] A related issue is whether individuals can choose emotions - see Elster (1998).

ary or learning models have shed much light on equilibrium selection issues (see e.g. Fudenberg and Levine 1998) and so we take this approach in the current paper. Individuals are modelled as interacting repeatedly over time and choosing whether or not to conform using a best reply rule;- that is, an agent conforms in this period if conforming would have maximized his payoff in the previous period. A seemingly robust result of the existing literature on best reply dynamics is how an equilibrium in risk dominant strategies is the most likely long run outcome. Taking our lead from this literature we can ask what is the risk dominant option or 'safer option' here - to conform or not-conform. To do this we envisage an initial state in which all except proportion ε of the population are playing not-conform, where ε is small. We ask whether conformity 'can invade'. That is, whether, over time the proportion of agents choosing to conform grows from ε to one. The converse would be to ask whether non-conformity can invade.

We find that both conformity and non-conformity can invade depending on the distribution of types in the population and the potential strength of the desire for esteem. Typically, however, we find that conformity is able to invade while non-conformity not. In particular, for some individuals, conforming is a relatively 'easy option' because the drop in intrinsic utility from conforming is small and so a little esteem is enough to compensate them. Once the proportion of agents conforming grows then esteem becomes of more significance and others, who have more extreme types and stand to loose a substantial intrinsic utility from conforming, will also choose to conform. In short, conformity spreads with the last individuals to conform being those with the most extreme types. We thus find conformity, or alternatively a high weight to esteem, can emerge in the population relatively easily; this may be the case even if all individuals would have a much higher payoff in the 'non-conformist' equilibrium.

It is interesting to relate our results to Akerlof (1980). Akerlof questioned whether social custom would be gradually eroded because it is costly for indi-

viduals to persist in the custom. In fact it is shown that custom can survive. Our analysis would strongly support this conclusion with the 'conformist' equilibrium proving vey stable and robust. Indeed, our analysis suggests that we can, in many instances, go even further by saying that conformity or social custom can not only survive but propagate, even if it is not in agents interests for it to do so. Indeed, it would seem the mere possibility of a norm existing could be enough to set off a self fulfilling prophecy whereby a conformity norm does evolve. To conform essentially appears the risk dominant strategy. This result can also be related to the literature questioning why conformity or norms exist; Elster (1989) surveys research attempting to justify social norms as 'optimal', or more formally, to show that a conformist equilibrium Pareto dominates a non-conformist equilibrium. Our analysis would suggest that the Pareto ranking of a conformist or non-conformist equilibrium is not critical in an explanation of why conformity can exist (even in a world of optimizers).

One important caveat to our results is explored further in Section 5.6. For the most part we assume a unique conformist equilibrium in the sense that agents can focus on some behaviour, say playing action a, and know that behaving in this way cannot leave them the odd one out. But suppose that individuals are uncertain what behaviour would become the norm in any conformist equilibrium - it may be action a_1 or it may be a_2. The 'risks' to conforming are now higher - if a person conforms, not only do his sacrifice intrinsic utility, but he also still risks being the 'odd one out' if he chooses the 'wrong' action. We show that it is less likely conformity can invade with this additional level of uncertainty. This perhaps points to one important reason that we observe conformity in some choice situations but not others. If there is an obvious potential focal point then our analysis would suggest conformity may arise but if such a focal point is not apparent maybe non-conformity is a more likely outcome.

It should be emphasized that we focus throughout on 'emotional' or 'so-

cial' conformity in that individuals conform to 'fit in' and receive esteem. Much of the previous literature on conformity has focussed on 'informational' conformity whereby individuals imitate successful or popular actions in the hope of obtaining a higher intrinsic utility (e.g. Bikhandani, Hirshleifer and Welch 1992, Juang 2001). Clearly conformity or herding can arise if the number of individuals choosing an action is seen as a signal of its relative payoff.[3] This leads to questions as to what extent conformity is the result of emotional or informational factors. It does appear that much of conformity can be explained by information factors (see Shiller 1995 and the discussion of the Asch experiments contained therein). It is, therefore, interesting, and perhaps surprising, to find, as we do in this paper, that conformity can arise purely based on emotional factors. If emotional and informational factors combine it is not hard to see why conformity may prevail.

We proceed as follows: Section 2 summarizes the model of conformity introduced in Bernheim (1994) and Section 3 completes the description of the model to be used in this paper. In Section 4 we present some results on the monotonicity of payoffs before providing our main results in Section 5 where we consider a dynamic model of choice. Section 6 concludes with some technical derivations left to the Appendix

2 Model of conformity

The model of conformity that we shall use was introduced and motivated by Bernheim (1994). This section will detail the parts of the model relevant to our work and we refer the reader to Bernheim for a much more complete description and discussion.

There is a continuum of agents. Each agent is assigned a type from set $T = [0, 2]$ and chooses an action from set $X = [0, 2]$. The distribution of types

[3] It is interesting to note that non-conformist or conformist equilibria could still prevail - if no one is choosing the same action then maybe there is no point in imitating. A non-conformist equilibrium would appear, however, very unstable in this context.

is described by a c.d.f. $F(\cdot)$ with corresponding p.d.f. $f(\cdot)$. It is assumed that $supp[f] = T$ or, in other words, that all taste types exist in the population. For simplicity we shall also assume that f is symmetric $f(1-v) = f(1+v)$ (or equivalently $F(t) = 1 - F(2-t)$) and that f is atomless.

An agent's *intrinsic utility* from playing action x when of type t is given by $g(x-t)$ where function $g(z)$ is symmetric $g(z) = g(-z)$, twice continuously differentiable, strictly concave and attains a maximum at $z = 0$. Action t is the *intrinsic bliss point* of an agent of type t.

An agent's type is private information. After an agent i has chosen an action other agents can form beliefs about his type. These beliefs can be summarized by an inference function $\phi(b, x)$ detailing the probability that an agent is perceived to be of type b if he has chosen action x. Note $\int_T \phi(b, x)db = 1$. An agent perceived to be of type b is accorded *esteem* $h(b)$ where h is twice continuously differentiable, strictly concave, symmetric $h(1+z) = h(1-z)$ and achieves a maximum at 1. If an agent is believed to be of type 1 he receives the highest esteem while an agent believed to be of type 0 or 2 receives the lowest esteem.

An agents utility if he chooses x, is of type t and beliefs are ϕ is,

$$u(x, t, \phi) = g(x - t) + \lambda \int_T h(b)\phi(b, x)db \tag{1}$$

where λ is an index of the *weight attached to esteem*. As can be seen an agents utility is a weighted sum of intrinsic utility and esteem where the higher is λ the higher is the weight accorded to esteem.

2.1 Signalling Equilibria

We make use of the standard definition of a signalling equilibrium and refine the set of equilibria using the D1 criterion (see Bernheim 1994 and Fudenberg and Tirole 1998 for explanations). The D1 criterion effectively states that on observing an action x that occurs with zero probability, in the candidate

signalling equilibrium, it will be inferred that action x was played by the agent who had the most incentive to play it. This D1 criterion has explicit implications for the model considered here and we will explain these below.

Bernheim (1994) derives the set signalling equilibria satisfying the D1 criterion.[4] There are two distinct types of signalling equilibrium explained below and illustrated in Figure 1. [Specific examples will be considered in the body of the paper].

For small λ (or, more explicitly any $\lambda \leq \overline{\lambda}$ for some $\overline{\lambda}$) one obtains *fully separating equilibria* that can be characterized by a function $\phi_s(x)$ where $\phi(b,x) = 1$ if $b = \phi_s(x)$. An agent who plays x is believed to be of type $\phi_s(x)$ and consequently an agent of type t plays $\phi_s^{-1}(t)$. The function ϕ_s is 'symmetric' in the sense that $\phi_s(2-x) = 2 - \phi_s(x)$. Note that agents of different types choose different actions and hence the equilibrium is fully separating.

For large λ (or more explicitly any $\lambda \geq \overline{\lambda}$) one obtains *equilibria with incomplete separation* in which agents with types close to one choose the same action. Thus, there is conformity with agents of differing types playing a common action. More formally, given beliefs ϕ and an action x let $t_l(\phi,x) = \min\{b : \phi(b,x) > 0\}$ and $t_h(\phi,x) = \max\{b : \phi(b,x) > 0\}$. A signalling equilibrium with incomplete separation has the property that there exists a unique $x_p \in [0,2]$ where $t_l(\phi,x_p) \neq t_h(\phi,x_p)$. Further, $t_l(\phi,x_p) \leq 1 \leq t_h(\phi,x_p)$ and equilibrium beliefs satisfy,

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

As in a fully separating equilibria, an agent with type $t \notin [t_l(\phi,x_p), t_h(\phi,x_p)]$ plays action $\phi_s^{-1}(t)$ and beliefs satisfy $\phi(b,x) = 1$ if $b = \phi_s(x)$. In an equilibrium with incomplete separation the D1 criterion has bite: there exists a

[4] We state here the important features for our analysis - for a complete description and explanation see Bernheim (1994).

set of actions $\overline{X}$ that occur with zero probability according to the signalling equilibrium (see Figure 1b) and, applying the D1 criterion, it will be inferred that an agent who plays $x \in \overline{X}$ is of either type $t_l(\phi, x_p)$ or $t_h(\phi, x_p)$.

When $\lambda \leq \overline{\lambda}$ there is the unique signalling equilibrium as detailed above. When $\lambda > \overline{\lambda}$ there will be multiple equilibria with incomplete separation each equilibrium 'centered' around a different x_p; there will, however, be a unique equilibrium with incomplete separation centered around $x_p = 1$ (Bernheim 1994). In the following we shall give special attention to this equilibrium. Indeed, we will think of action 1 as being a focal point in the sense that any conformist equilibrium would be 'centered around' action 1. This appears a reasonable assumption to make given that agents do want to be perceived as of type 1. In Section 5.6 we do consider relaxing this assumption.

3 Conform or not conform

Instead of modelling agents as choosing an action and belief function taking λ as given we shall model agents as choosing between two strategies - to *conform* or *not conform*. We can think of this as agents as choosing between two different signalling equilibria corresponding to $\lambda = 0$ and $\lambda = \lambda^* > 0$.[5] In the $\lambda = 0$ case esteem is of no consequence and so it is a situation where agents are 'free to choose whatever they want'. If $\lambda > 0$ then agents will seek esteem and conformity may result. In choosing between conform or not-conform agents are essentially choosing the level of λ and therefore the relative importance of esteem.

We characterize a signalling equilibrium by the triple (α, ξ, ϕ) where function α maps types to actions, ξ maps actions to esteem and ϕ represents beliefs. More precisely, $\alpha(t)$ is the action chosen by an agent of type t, $\xi(x)$ is the esteem accorded to an agent who plays x and, as above, $\phi(b, x)$ is the

[5] One could consider a choice between λ_H and λ_L where $\lambda_H > \lambda_L \geq 0$; setting $\lambda_L = 0$, as we do, significantly simplifies the notation and analysis without altering the main conclusions.

probability an agent who plays x is believed to be of type b. When $\lambda = 0$ there exists the trivial and unique signalling equilibrium we denote $(\alpha^0, \xi^0, \phi^0)$ where each agent chooses his internal bliss point and the utility from esteem is zero. When $\lambda = \lambda^* > 0$ there exists (as explained above) a unique signalling equilibrium centered around 1 that we denote by $(\alpha^*, \xi^*, \phi^*)$.

Taking the two signalling equilibria $(\alpha^0, \xi^0, \phi^0)$ and $(\alpha^*, \xi^*, \phi^*)$ consider a game in which agents have two *strategies* - to conform (C) or to not conform (N). These strategies can be explained:

To not conform (N): If the agent is type t he chooses action $a = t$ and accords all other agents esteem 0.

To conform (C): If the agent is type t he chooses action $\alpha^*(t)$ and accords other agents esteem according to function ξ^*.

If we let c denote the proportion of the population who choose to conform the payoff function of an agent of type t can be defined,

$$\begin{aligned} U(N,t,c) &= g(0) + c\lambda^*\xi^*(t) \\ U(C,t,c) &= g(t-\alpha^*(t)) + c\lambda^*\xi^*(\alpha^*(t)). \end{aligned} \tag{2}$$

This simple framework captures the choice that an agent may face in not knowing whether conformity will or will not prevail in the population. He has the option of choosing to not conform, play his internal bliss point and treat other agents equally. Or, expecting that others will conform he can conform himself and give esteem to other agents accordingly. We emphasize how an agent receives esteem even if he chooses not to conform. This seems reasonable - an agent who plays N may feel 'left out' or the 'odd one out' if he subsequently finds that a large proportion of other agents conformed. Conversely, an agent who conforms does not receive esteem from agents who choose not to conform.

The combinations of c and t for which $U(N,t,c) = U(C,t,c)$ will prove important in the following. Thus, we introduce a *threshold function* $c^*(t)$ mapping the set of types into the unit interval where $U(N,t,c^*(t)) = U(C,t,c^*(t))$ for all t. It can easily be verified that a unique $c^*(t)$ exists for all $t \neq 0,1,2$ and furthermore if $c \lesseqgtr c^*(t)$ then $U(N,t,c) \gtreqless U(C,t,c)$.

A *strategy profile* s will be a function mapping the set of types to the set of strategies where $s(t)$ denotes the strategy played by agents of type t.[6] The strategy profile in which all types of agent conform will be denoted $\overrightarrow{C}$ and the strategy profile in which all types of agent do not conform will be denoted $\overrightarrow{N}$. We make use of the standard definition of a Nash equilibrium and so both $\overrightarrow{C}$ and $\overrightarrow{N}$ are Nash equilibria.

Note the two different forms of equilibria we now have in the model (i) the Nash equilibria that result from the game in which agents choose between strategies N and C and (ii) the signalling equilibria that result given strategy profiles $\overrightarrow{C}$ and $\overrightarrow{N}$ and are equilibria relative to the scenario where agents choose actions taking λ as given. This distinction should always be apparent but we will make use of the Nash equilibrium versus signalling equilibrium terminology to help clarify. Our analysis will focus on the agents choice between N and C.

4 Monotonicity of Payoffs

This section will address the question of whether agents prefer the 'conformist' Nash equilibrium $\overrightarrow{C}$ giving behavior $(\alpha^*, \xi^*, \phi^*)$ or the 'non conformist' Nash equilibrium $\overrightarrow{N}$ giving behavior $(\alpha^0, \xi^0, \phi^0)$. We can say little on this question in absolute terms as this will depend on whether esteem is a positive or negative addition to utility. We can, however, say something about *relative* preferences. Our first result shows that the closer is an agents type to one then the higher is his payoff in the conformist equilibrium. Note

[6] We shall assume that all players of the same type play the same strategy.

that in the non-conformist equilibrium every agent receives a payoff of $g(0)$. Thus, relatively speaking, agents with types near to one prefer the conformist equilibrium.

Before stating our first result we introduce some simplifying notation. We introduce a function u^* mapping the set of types into the real line where $u^*(t) := u(a^*(t), t, \phi^*)$ for all t. Thus, $u^*(t)$ is the payoff of an agent of type t in the (conformist) signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$.

Proposition 1: Consider signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ with equilibrium beliefs ϕ^*. For any $t, t' \in [0, 2]$ we have $u^*(t) > u^*(t')$ if and only if $|1 - t| < |1 - t'|$. [More generally, if $(\alpha^*, \xi^*, \phi^*)$ is a signalling equilibrium with incomplete separation around x_p^* then $u^*(t) > u^*(t')$ for any $t, t' \in [0, x_p^*]$ if and only if $t > t'$ and, similarly, for any $t, t' \in [x_p^*, 2]$ if and only if $t < t'$.].

Proof: Let $t_l := t_l(\phi^*, x_p^*)$ and $t_h := t_h(\phi^*, x_p^*)$. Consider the region $[t_l, t_h]$. (For a signalling equilibrium with complete separation this region does not exist and so the step can be skipped.) Any two agents with types $t, t' \in [t_l, t_h]$ receive exactly the same payoff from esteem. It follows that differences in their payoffs will reflect the differences between $g(t - x_p^*)$ and $g(t' - x_p^*)$. Given the properties of function g we obtain the desired relationship.

Consider now the region $[0, t_l]$ and two agents i and i' with types t and t' where $t_l \geq t > t'$. Suppose that agent i' is earning at least as high a payoff as i; that is, $u^*(t) \leq u^*(t')$. Suppose also that $t \leq \alpha^*(t')$. If agent i were to play action $x = \alpha^*(t')$ he would receive the same esteem as agent i' and (given that $\alpha^*(t') \geq t > t'$) a strictly higher intrinsic utility than i'. Thus, by playing $\alpha^*(t')$ agent i would receive a payoff strictly greater than $u^*(t')$ contradicting that i was maximizing his payoff by choosing $\alpha^*(t)$. Thus, suppose that $t > \alpha^*(t')$. If agent i was to play action $x = t$ he would receive strictly more esteem than i' and the maximum possible intrinsic utility of $g(0)$. Thus, i would receive a payoff strictly greater than $u^*(t')$ but this again contradicts that i was maximizing his payoff by choosing $\alpha^*(t)$. Therefore, $u^*(t) > u^*(t')$.

A symmetric argument holds for interval $[t_h, 2]$.

We next note that, given the symmetry assumed, $u(t, a^*, \phi^*) = u(2 - t, a^*, \phi^*)$ for all t (when $x_p^* = 1$). It thus remains to put the three intervals considered above together. Note, however, that a condition of signalling equilibrium is that agents on the margin, i.e. agents with types t_l and t_h, should be indifferent between choosing x_p^* and $\phi_s^{-1}(t_l)$ or, respectively, $\phi_s^{-1}(t_h)$. It can now be seen that the statement of the Proposition follows.■

Having shown that a differential exists in $u^*(t)$ depending on how close is t to 1 a natural question is to ask what factors could influence the size of this differential. One factor is λ^* but we leave discussion of varying λ^* until Section 5.5. A second factor is the form of the distribution of agents over types as given by F. Our second result shows that the signalling equilibrium payoffs cannot decrease the more concentrated is the distribution of types around $t = 1$. In particular, if we compare two distributions over types F^* and F' where $F^*(t) < F'(t)$ for all $t \in [0, 1]$ then, keeping λ^* fixed, we can compare payoffs between the two respective 'conformist' signalling equilibria $(\alpha^*, \xi^*, \phi^*)$ and (α', ξ', ϕ'). We find that equilibrium payoffs must be at least as high under equilibrium $(\alpha^*, \xi^*, \phi^*)$ as (α', ξ', ϕ'). Further, if $(\alpha^*, \xi^*, \phi^*)$ is a signalling equilibrium with incomplete separation then, for those agents playing 1, payoffs are strictly higher under $(\alpha^*, \xi^*, \phi^*)$.

Proposition 2: Let F^* and F' represent two *distinct* distributions over the set of types where $F^*(t) \leq F'(t)$ for all $t \in [0, 1]$ and let $(\alpha^*, \xi^*, \phi^*)$ and (α', ξ', ϕ') be the corresponding signalling equilibria. Then,

$$u^*(t) \geq u'(t) \tag{3}$$

for any t and, further, if $(\alpha^*, \xi^*, \phi^*)$ is a signalling equilibrium with incomplete separation then (generically) the inequality in (3) is strict for any agent of type t where $\alpha^*(t) = 1$.[7]

[7] Where $u'(t) = u(\alpha'(t), t, \phi')$ for all t.

Proof: We begin by noting that ϕ_s is independent of the distribution over types (Bernheim 1994). One immediate consequence is that $u^*(t) = u'(t)$ for any agent of type t where $\alpha^*(t) = \alpha^{'}(t) = \phi_s^{-1}(t)$. It remains, therefore, to consider types t for which $\alpha^*(t)$ or $\alpha'(t)$ equal one.

Let $t_l^* := t_l(\phi^*, 1)$ and $t_l' := t_l(\phi', 1)$.[8] Thus,

$$u\left(1, t_l^*, \phi^*\right) = u\left(\phi_s^{-1}\left(t_l^*\right), t_l^*, \phi^*\right) \tag{4}$$

and

$$u\left(1, t_l', \phi'\right) = u\left(\phi_s^{-1}\left(t_l'\right), t_l', \phi'\right). \tag{5}$$

We conjecture that $t_l^* \leq t_l'$. If not, $t_l^* > t_l'$ implying that $\xi^*(1) > \xi'(1)$. Consequently, given (5), $u\left(1, t_l', \phi^*\right) > u\left(\phi_s^{-1}\left(t_l'\right), t_l', \phi^*\right)$. This leads to the desired contradiction. We next conjecture that $\xi^*(1) \geq \xi'(1)$. Suppose not and consider an agent of type t_l^*. If $\xi^*(1) < \xi'(1)$ then $u\left(1, t_l^*, \phi^*\right) < u\left(1, t_l^*, \phi'\right)$ but given (4) this would imply $u\left(1, t_l^*, \phi'\right) > u\left(\phi_s^{-1}\left(t_l^*\right), t_l^*, \phi'\right)$ contradicting that $t_l^* \leq t_l'$. Given that $\xi^*(1) \geq \xi'(1)$ it is now simple to see that $u^*(t) \geq u'(t)$ for all t.

Suppose $\xi^*(1) > \xi'(1)$. For any agent of type t where $\alpha^*(t) = \alpha^{'}(t) = 1$ it is trivial that $u^*(t) > u'(t)$. For an agent i of type $t > t_l^*$ where $\alpha^*(t) \neq \alpha^{'}(t)$ we note that i could obtain payoff $u'(t)$ if beliefs are ϕ^* by playing $\phi_s^{-1}(t)$; that he does not do so implies $u\left(1, t, \phi^*\right) > u\left(\phi_s^{-1}\left(t\right), t, \phi^*\right) = u'\left(t\right)$. Thus, the inequality in (3) is strict for any agent of type t where $\alpha^*(t) = 1$.■

Applying Proposition 2 we obtain our final result of this section: the difference in payoffs (in the conformity signalling equilibrium) between those agents of types 1 and types 0 and 2 cannot decrease if the distribution of types is more concentrated around 1. Specifically, using the same notation as Proposition 2 and and letting $\Delta(t) \equiv u^*(t) - u'(t)$:

Proposition 3: For any two distributions over the set of types F^* and

[8] In the case of a fully separating signalling equilibrium set t_l^* or t_l' to equal 1.

F' where $F^*(t) \leq F'(t)$ for all $t \in [0,1]$ it is the case that $\Delta(1) \geq \Delta(0)$. Furthermore, generically, $\Delta(1) > \Delta(0)$ if $(\alpha^*, \xi^*, \phi^*)$ is an equilibrium with incomplete separation and $t_l^*(\phi^*, 1) > 0$.

Proof: Let $C \equiv \lambda^*(\xi^*(1) - \xi'(1))$. From Proposition 2 and its proof we know that $C \geq 0$ and:

1. If both $(\alpha^*, \xi^*, \phi^*)$ and (α', ξ', ϕ') are fully separating equilibria then $\Delta(t) = 0$ for all t.

2. If both $(\alpha^*, \xi^*, \phi^*)$ and (α', ξ', ϕ') are equilibria with incomplete separation where $t_l^* = t_l' = 0$ then $\Delta(t) = C$ for all t.

3. If $(\alpha^*, \xi^*, \phi^*)$ is an equilibrium with incomplete separation and $0 < t_l^*$ then (i) $\Delta(t) = 0$ for all $t \leq t_l^*$ and (ii) $\Delta(t) = C$ for all $t \in [t_l', 1]$.■

In summary, we see that in relative terms agents with types near 1 prefer the conformist equilibrium. We also see that the 'more concentrated' is the distribution of types towards $t = 1$ then the higher are payoffs in the conformist equilibrium and, potentially, the larger is the payoff differential between agents of type $t = 1$ and those with more extreme types. Intuitively these results would suggest that agents with types near 1 are more likely to conform. They would also suggest that a 'conformist equilibrium' is more likely to occur the higher the proportion of agents with types near 1. The dynamic model of the following section allows us to question this in more detail.

5 A dynamic model of choice

To capture the choice that agents face between conforming and non-conforming we consider a model of learning. Agents interact over an indefinite number of time periods $\tau = 0, 1, 2, ...$ A strategy profile s^τ details the strategies

chosen where $s^\tau(t)$ is the strategy chosen by agents of type t in period τ.[9] Relative to strategy profile s^τ let c^τ denote the proportion of agents playing C. There exists an initial strategy profile s^0 and in subsequent periods each agent is assumed to choose the strategy that would have maximized his payoff in the previous period. That is, agent i chooses C in period τ if and only $U(C, t, c^{\tau-1}) > U(N, t, c^{\tau-1})$ (and where we assume some appropriate tie breaking rule). This behavior gives rise to a deterministic dynamic process through which the strategy profile evolves to $s^1, s^2, ...$ The dynamic will be recognised as a best reply dynamic as much studied in the literature (see e.g. Fudenberg and Levine 1998)

In the spirit of the literature on contagion and the emergence of convention in coordination games (see e.g. Young 1993 and Morris 2000) we question whether conformity or non-conformity can 'invade' a population. It is trivial that both $\overrightarrow{C}$ and $\overrightarrow{N}$ are Nash equilibrium and thus absorbing states of the dynamic. Let s^ε denote a strategy profile in which proportion ε of the population play C and proportion $1-\varepsilon$ play N. We say that *conformity can invade a population* if for any $\varepsilon > 0$ there exists strategy profile s^ε such that the dynamic with initial state s^ε converges to $\overrightarrow{C}$. Similarly we say that *non conformity can invade a population* if for any $\varepsilon > 0$ there exists strategy profile $s^{1-\varepsilon}$ such that the dynamic with initial state $s^{1-\varepsilon}$ converges to $\overrightarrow{N}$. Note it is not possible that both conformity and non-conformity can invade a population.

It seems useful to begin with some examples to demonstrate that conformity and non-conformity can invade a population. We do this in the next section. Having worked through these examples we then proceed to the general case doing so in three stages. We begin by looking at what we shall call a *complete conformity* case in which signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ has the property that $\alpha^*(t) = 1$ for all $t \in [0, 2]$; that is all agents choose action 1. We then consider a *partial conformity* case where $(\alpha^*, \xi^*, \phi^*)$ is a signalling

[9] For simplicity we assume that players of the same type always play the same strategy

equilibrium with incomplete separation but $t_l > 0$. Finally, we consider the case where $(\alpha^*, \xi^*, \phi^*)$ is a fully separating equilibrium.

5.1 Examples

We illustrate with a *spherical case* $g(z) = -z^2$ and $h(b) = -(1-b)^2$. For simplicity we set

$$\lambda^* = \frac{1}{\xi^*(1)+1} \tag{6}$$

where $\xi^*(1) = \int_{[0,2]} h(b)f(b)db$. Condition (6) implies that in the conformist equilibrium $(\alpha^*, \xi^*, \phi^*)$ an agent of type 0 or 2 is indifferent between actions 0 or 1 and respectively actions 2 or 1. Thus, signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ will have the property that all agents play action 1. Note that any agent not playing 1 is perceived to be of either type 0 or 2.

Simplifying equations (2) we get $U(N,t,c) = -c\lambda^*$ and $U(C,t,c) = -(1-t)^2 + c\lambda^*\xi^*(1)$ for all t implying a threshold function

$$c^*(t) = (1-t)^2. \tag{7}$$

In period $\tau+1$ an agent will play C if and only if $c^\tau \geq c^*(t)$. Thus, an agent i who has type t plays C in period $\tau+1$ if and only if $t \in \left[1-c^{\tau\frac{1}{2}}, 1+c^{\tau\frac{1}{2}}\right]$ .Those agents playing C will thus have types in some closed interval around $t=1$. For conformity to spread requires those players with types just outside this interval to change to conformity. Thus, for conformity to invade we require that $\int_t^{2-t} f(b)db = 1-2F(t) > c^*(t)$ for all $t \in (0,1)$. Conversely, for non-conformity to invade requires $c^*(t) > 1-2F(t)$ for all $t \in (0,1)$. If $1-2F(t) > c^*(t)$ for some t and $1-2F(t') < c^*(t')$ for some t' then neither conformity or non-conformity can invade. Ploting $c^*(t)$ against $1-2F(t)$ provides, therefore, a convenient visual check on whether or not conformity can invade and this will prove useful in the following.

Whether or not conformity can invade will clearly depend on the dis-

tribution over types $F(\cdot)$. We provide some examples, illustrated in Figure 2:

Conformity invades: Let $f(t) = \frac{1}{2}$. Looking at Figure 2a we observe that $c^*(t) < 1 - 2F(t)$ and so conformity can invade. To illustrate how this conclusion could be derived more explicitly consider initial strategy profile s^ε where agents with types $t \in [1-\varepsilon, 1+\varepsilon]$ play strategy C and all other agents play N. In period 1, we observe that the proportion playing c increases to $\varepsilon^{\frac{1}{2}}$. Generalizing, in period τ all those agents with types $t \in \left[1-\varepsilon^{\frac{1}{2^\tau}}, 1+\varepsilon^{\frac{1}{2^\tau}}\right]$ play C implying that $c^\tau = \varepsilon^{\frac{1}{2^\tau}}$. Thus, c^τ converges to 1.

Limit case: Let $f(t) = 1 - t$ over $[0, 1]$.[10] Here $1 - 2F(t)$ and $c^*(t)$ coincide basically implying stalemate. Given any initial state the proportion of agents playing C will be the same in period 1 onwards: $U(C, t, c^1) \geq U(N, t, c^1)$ for all those agents who playing C in period 1 while $U(N, t, c^1) > U(C, t, c^1)$ for all those agents who play N.[11] Thus, depending on the initial state, the final proportion of agents playing C could be anything between 0 and 1.

Non-conformity invades: Let $f(t) = 2$ for $t \in [0, 0.1]$ and $f(t) = \frac{20}{27} - \frac{20}{27}t$ for $t \in [0.1, 1]$. From Figure 2c we observe that $c^*(t) > 1 - 2F(t)$ and so non-conformity can invade. This is possible because of the large proportion of agents with 'extreme' types.

Stability of conformity and non-conformity: Let the distribution over types be bimodal with peaks at 0.5 and 1.5. Specifically, let $f(t) = 0.01$ for $t \in [0, 0.49]$ and $t \in [0.51, 1]$ while $f(t) = 24.51$ for $t \in [0.49, 0.51]$. It can be easily verified that neither conformity or non-conformity can invade the

[10] This does not formally satisfy the assumption that $supp[f(\cdot)] = T$ but this is clearly not important in this example.

[11] We perhaps need to be a little more precise about the tie breaking rule to be conclusive but the logic of the example should be clear.

population.[12] In short the strategy played by those agents with types in the range $[0.49, 0.51]$ and $[1.49, 1.51]$ determines the strategy that holds in the population.

5.2 Complete conformity

In proceeding to the general case we begin by assuming signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ has the property that $\alpha^*(t) = 1$ for all $t \in [0, 2]$; that is all agents choose action 1. In many ways this is the most interesting case given the clear contrast between the conformity and non-conformity Nash equilibria. Note that the examples in the previous section fell into this case.

We saw in Sections 4 and 5.1 that whether or not conformity can invade will depend on the distribution of agents over types. Our main result, Proposition 4, provides a bound on the distribution over types sufficient for conformity to be able to invade. This result shows that if the distribution over types is unimodal then (recalling we assume the distribution in symmetric) conformity can invade. This seems a relatively mild condition justifying the claim that it is 'relatively easy' for conformity to invade in the complete conformity case. We also, highlight, as demonstrated in the proof of Proposition 4, that conformity invades in a wave emanating from agents with types near 1 and spreading to those with 'more extreme' types.

Proposition 4: If $F(t) \leq \frac{1}{2}t$ for all $t \in [0, 1]$ and signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ has the property that $\alpha^*(t) = 1$ for all $t \in [0, 2]$ then conformity can invade.

The proof highlights a number of interesting aspects and proceeds in three stages. First, we obtain an analogue of Proposition 1. In signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ the payoff of an agent of type t is $g(1-t) + \lambda^* \int_T h(b)f(b)db$

[12] It may, however, be that strategies C and N 'survive' in the population in the sense that we obtain an equilibrium where a positive proportion of the population are playing each strategy.

for all t. Noting that all players receive the same esteem and recalling the properties of $g(\cdot)$ it is immediate that the threshold function $c^*(t)$ is monotonic over $t \in [0,1]$; that is $c^*(t) < c^*(t')$ for any $t, t' \in [0,1]$ where $t > t'$. Thus, conformity can only invade 'in a wave spreading from those agents with types near 1'.

Second, we provide an analogue of Proposition 2. If conformity can invade when the distribution over types is F' then, ceteris paribus, it can invade when the distribution over types is F^* and $F^*(t) \leq F'(t)$ for all $t \in [0,1]$. Let $(\alpha^*, \xi^*, \phi^*)$ and (α', ξ') be the respective signalling equilibria (given λ^*) and $c^*(t)$ and $c'(t)$ the threshold functions. We note that $\xi^*(1) \geq \xi'(1)$ while $\xi^*(t) = \xi'(t)$ for any $t \neq 1$. From equations (2) we can therefore observe that $c^*(t) \leq c'(t)$ for all t.

Third, we show that conformity can invade if $f(t) = \frac{1}{2}$ for all t. Note that in signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ agents of type 0 and 2 must be at least as well off playing action 1 as 0 or 2. Thus,

$$\lambda^* \geq \frac{g(0) - g(1)}{\xi^*(1) - h(0)}.$$

Therefore, from (2),

$$U(C, t, c) \geq U(N, t, c) \text{ if } c \geq \frac{g(0) - g(1-t)}{g(0) - g(1)}. \tag{8}$$

This implies that conformity can invade if,

$$\int_t^{2-t} f(t) \geq \frac{g(0) - g(1-t)}{g(0) - g(1)}$$

or Using the symmetry of f if,

$$F(t) \leq \frac{1}{2} - \frac{1}{2}\left[\frac{g(0) - g(1-t)}{g(0) - g(1)}\right] \tag{9}$$

for all $t \in [0,1)$. If $f(t) = \frac{1}{2}$ then $F(t) = \frac{1}{2}t$. But, given that $g(\cdot)$ is concave, $g(1-t) \geq g(1) + t[g(0) - g(1)]$ for all $t \in [0,1]$. This implies,

$$\frac{1}{2}t \leq \frac{1}{2}\frac{g(1-t) - g(1)}{g(0) - g(1)} = \frac{1}{2} - \frac{1}{2}\frac{g(0) - g(1-t)}{g(0) - g(1)}$$

for all $t \in [0,1]$. Thus, conformity can invade. It can be seen that without being more specific about $g(\cdot)$ the bound provided by $f(t) = \frac{1}{2}$ is tight.

5.3 Partial Conformity

We turn now to the case where $0 < t_l^*(\phi^*, 1) < 1$ and so agents with types $t \in [t_l^*, 2 - t_l^*]$ play action $\alpha^*(t) = 1$ and an agent of type $t \notin [t_l^*, 2 - t_l^*]$ plays $\alpha^*(t) = \phi_s^{-1}(t)$. Thus, there is conformity in the sense that (when $\lambda = \lambda^*$) agents who do not have type 1 play action 1. There is, however, not complete conformity in that some agents choose not to play 1.

One interesting aspect of the partial conformity case is how, if conformity can invade, it does not invade 'in a wave' emanating from those agents with types near to 1 and spreading 'smoothly' to those agents with more 'extreme' types. More formally, the threshold function $c^*(t)$ will not be monotonic over $t \in [0,1]$. This can be seen by directly comparing an agent of type $t^+ := t_l^* + \delta$ to an agent of type $t^- := t_l^* - \delta$ where $\delta > 0$ is small. Given that δ is small $\xi^*(t^+) \simeq \xi^*(t^-)$ and so $U(N, t^+, c) \simeq U(N, t^-, c)$ for any c. For δ small enough $u(\phi_s^{-1}(t^-), t^-, \phi^*) \simeq u(1, t^+, \phi^*)$ and so $U(C, t^+, c) \simeq U(C, t^-, c)$ when $c = 1$. Note that $U(C, t^+, c) = g(1 - t^+) + c\lambda^*\xi^*(1)$ and $U(C, t^-, c) = g(\phi_s^{-1}(t^-) - t^-) + c\lambda^* h(t^-)$. Immediately we can see that $\xi^*(1) > h(t^-) \simeq h(t_l^*)$. Thus, given that $\lambda^* = \frac{g(\phi_s^{-1}(t_l^*) - t_l^*) - g(1 - t_l^*)}{\xi^*(1) - h(t_l^*)} > 0$ we see that $g(\phi_s^{-1}(t^-) - t^-) \simeq g(\phi_s^{-1}(t_l^*) - t_l^*) > g(1 - t_l^*) \simeq g(1 - t^+)$. We find therefore that $U(C, t^+, c) < U(C, t^-, c)$ for any $c < 1$. It follows that $c^*(t^-) < c^*(t^+)$. An agent of type t^+ may, therefore, choose to conform when an agent of type t^- does not even though $|1 - t^-| < |1 + t^-|$. This will be illustrated in the following examples.

We give three examples that demonstrate that conformity can invade in the case of partial conformity but 'not as easily' as when there is strong conformity. In all examples (using the spherical case) we set $f(t) = \frac{1}{2}$ for $t \in [0, 2]$ and so we know conformity would invade if λ^* was sufficiently high for complete conformity. As in Section 5.1, we compare $c^*(t)$ and $1 - 2F(t)$ to obtain a visual picture of whether or not conformity can invade.It should be noted, however, that the non-monotonicity of $c^*(t)$ means that more care has to be taken in interpreting the figures (than in Section 5.1) as we shall see.

Conformity can invade: Set $\lambda^* = 1.25$. As detailed in the Appendix one can solve for $\phi_s(x)$, t_l^* and consequently the threshold function $c^*(t)$. Figure 3 plots $c^*(t)$ and $1 - 2F(t)$. The non-monotonicty of $c^*(t)$ is clearly apparent. We can see, however, that conformity can invade.

Conformity cannot invade: Set $\lambda^* = \frac{1}{2}$. Figure 4 plots the threshold function $c^*(t)$ and also $1 - 2F(t)$. We can see that conformity cannot invade in this case. Neither, it should be noted, can non-conformity and so both $\overrightarrow{C}$ and $\overrightarrow{N}$ are stable.

Conformity can invade (just): Set $\lambda^* = 1$. Figure 5 plots $c^*(t)$. Also plotted is $1 - 2F(t)$ for $1 \geq t \geq 0.2973$. The non-monotonicity of the threshold function really comes into play in this example. Conformity can invade. It will spread until those agents with types $t \in [0.2973, 1]$ are playing C. It will then continue to spread but in two waves rather than one. In particular, $t_l = 0.1932$ and we can see that agents with type $t_l = t_l + \delta$ will be some of the last agents to switch to conformity while agents with types $t_l - \delta$ will switch to conformity relatively early

5.4 Fully Separating

For completeness we consider the final possibility that signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ is a fully separating equilibrium. For our purposes this is the least interesting case as there is no observed conformity (when $\lambda = \lambda^*$) with agents of different types choosing different actions.

We demonstrate, through an example, that conformity (and non-conformity) cannot invade 'as easily' in this no conformity setting as in the strong or partial conformity settings considered above. The example considered is the spherical example with $\lambda = 0.25$.[13] Figure 6 plots the threshold function $c^*(t)$ and $1 - 2F(t)$ for the case where $f(t) = \frac{1}{2}$. It is immediately apparent that conformity cannot invade. Indeed for conformity to invade would require a distribution of types highly concentrated around 1. Note also that non-conformity cannot invade either and could only do so if the distribution of types was highly concentrated at the extreme types. In short both $\overrightarrow{C}$ and $\overrightarrow{N}$ are stable.[14]

Perhaps the most interesting thing we can learn from the fully separating case is the inability of non-conformity to invade. Recall than in a fully separating equilibrium $(\alpha^*, \xi^*, \phi^*)$ beliefs are such that $\xi^*(t) = h(t)$ for all t; that is, despite an agent playing a different action to his bliss point others will correctly perceive his type. The conformist equilibrium appears particularly 'pointless or difficult to explain' if this is the case. Consequently we might have expected conformity to erode and non-conformity to be able to invade (Akerlof 1980). This is not the case. In the more general context our analysis points to two broad possibilities: either conformity can invade or neither conformity or non-conformity can invade. In either case the conformist equilibrium $\overrightarrow{C}$ is stable. In other words it appears very difficult for conformity

[13] This is the largest λ consistent with a fully separating signalling equilibrium.

[14] It can be observed that the threshold function is monotonic over $t \in [0, 1]$ in this example. That would appear to be a general property of the spherical example. Intuitively, however, there seems little reason to believe that this is a general property for any g and h functions.

to erode if established. This complements the results of Akerlof (1980).

5.5 Increasing the weight attached to conformity

>From the discussion so far we might conjecture that conformity is 'more likely' to invade (i) the more concentrated is the distribution of types around $t = 1$ and (ii) the higher is λ^*. In fact, things are not quite this simple. For example, the non-monotonicity of the threshold function $c^*(t)$ means that a, ceteris paribus, increase in the concentration of the distribution of types around 1 may actually imply conformity cannot invade where it could before.[15] The effect of increasing λ^* is also ambiguous as we now discuss.

In treating the complete conformity case we can obtain the following simple result,

Proposition 5: Let signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ have the property that $\alpha^*(t) = 1$ for all $t \in [0, 2]$ when $\lambda^* \equiv \lambda'$. If conformity can invade then conformity can also invade, ceteris paribus, for any $\lambda^* > \lambda'$.

The proof is trivial once we note that the threshold value is $c^*(t) = \frac{g(0)-g(1-t)}{\lambda^*(\xi^*(1)-h(t))}$. Thus, increasing λ^* decreases $c^*(t)$ for all t. Note that increasing λ^* can imply going from a case where conformity cannot invade to one where it can invade. For example, in Section 5.1 when λ^* was as given in (6) we saw that conformity could not invade if $f(t) = 1 - t$; a marginal increase in λ^* would mean that conformity could invade.

Once we move beyond the complete conformity case the effects of increas-

[15] Consider, for instance, Figure 5 illustrating the case where $\lambda^* = 1$, $f(t) = 0.5$ and $t_l = 0.1932$. Suppose that we marginally change the distribution over types by lowering $f(t)$ for $t \in [t_l - \delta, t_l]$ and increasing $f(t)$ for $t \in [t_l, t_l + \delta]$ for some $\delta > 0$. Informally, we can see that this means conformity may no longer be able to invade. Formally, we have to recognize that t_l would change and recalculate the signalling equilibrium. The conclusion, however, that conformity may no longer be able to invade is robust.

ing λ^* are more ambiguous. Recall the general expression

$$U(C,t,c) \geq U(N,t,c) \text{ if and only if } c \geq \frac{g(0)-g(\alpha^*(t)-t)}{\lambda^*(\xi^*(\alpha^*(t))-\xi^*(t))}. \tag{10}$$

A preliminary question is to ask whether increasing λ^* decreases the threshold value $c^*(t)$. In general, increasing λ^* may increase $c^*(t)$, as can be seen from comparing Figures 3 to 6. For example, looking at Figures 3 and 4 we see that $c^*(0.15)$ is greater when $\lambda^* = 1.25$ than when $\lambda^* = 0.5$. This observation confirms that there is no simple relationship between λ^* and whether conformity can invade.

We may hope to be able to say more by focussing on the partial conformity case and those agents with types near 1. It is clear that if conformity invades then it does so initially in a wave emanating from $t = 1$. So, if increasing λ^* lowers $c^*(t)$ for those agents with types near 1 this gives at least some evidence that increasing λ^* allows conformity to invade. Consider a signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ with incomplete separation. For a non-empty set of types $\overline{T} \equiv [\overline{t}, 2-\overline{t}]$ $(1 \neq \overline{T})$ we have that,

$$c^*(t) = \frac{g(0)-g(1-t)}{\lambda^*(\xi^*(1)-h(t_l^*))}. \tag{11}$$

That is, $\xi^*(t) = h(t_l)$ for an agent with type $t \in \overline{T}$ implying that if he chooses his bliss point he will be perceived as being of type t_l. Note that $\xi^*(t) < h(t_l^*)$ for agents with types $t \in [t_l, \overline{t})$ (see Figure 1b). To question the sign of the derivative of $c^*(t)$ with respect to λ^* (for $t \in \overline{T}$) we proceed in two stages: (i) the effect that changing λ^* has on t_l^* and then (ii) the effect that a change in t_l^* has on $\xi^*(1)$ and $h(t_l^*)$.

If an agent of type t_l^* is indifferent between playing action 1 and action $\phi_s^{-1}(t)$ this implies that,

$$g(t_l^* - \phi_s^{-1}(t)) + \lambda^* h(1-t_l^*) = g(1-t_l^*) + \lambda^* \xi^*(1).$$

Let $\lambda' > \lambda^*$ and let $\overline{\phi}_s$ denote the corresponding beliefs function. It can be shown (see the Proof of Theorem 2 in Bernheim (1994)) that $g(t_l^* - \overline{\phi}_s^{-1}(t)) \leq g(t_l^* - \phi_s^{-1}(t))$. This, and that $\xi^*(1) > h(1 - t_l^*)$ implies that

$$g(t_l^* - \overline{\phi}_s^{-1}(t)) + \lambda' h(1 - t_l^*) < g(1 - t_l^*) + \lambda' \xi^*(1).$$

Consequently, we see that as λ^* increases t_l^* must decrease; that is, a higher proportion of agents conform in choosing action 1.

If t_l^* falls then clearly so do both $\xi^*(1)$ and $h(t_l^*)$. From (11) we can see that if $\frac{d}{dt_l^*}\xi^*(1) < \frac{d}{dt_l^*}h(t_l)$ then an increase in λ^* does feed through into a decrease in $c^*(t)$ for all $t \in \overline{T}$. Recalling that,

$$\xi^*(1) = \frac{1}{F(2 - t_l^*) - F(t_l^*)} \int_{t_l^*}^{2-t_l^*} h(b) f(b) db.$$

we see that,

$$\frac{d}{dt_l}\xi^*(1) = \frac{2f(t_l^*)}{1 - 2F(t_l^*)} \left(\xi^*(1) - h(t_l^*)\right).$$

In general $\frac{d}{dt_l^*}\xi^*(1)$ may, therefore, be less than or equal to $\frac{d}{dt_l^*}h(t_l^*)$ and so, once again, definitive results are not apparent. We can, however, say more for specific examples. For instance, we can provide some evidence for the property apparent in Sections 5.2-4 that increasing λ^* increases the likelihood that conformity can invade in the spherical case when $f(t) = \frac{1}{2}$ for all t. For this example,

$$\frac{d}{dt_l^*}\xi^*(1) = \frac{1}{1 - t_l^*}\left((1 - t_l^*)^2 - \frac{1}{3}(1 - t_l^*)^2\right) = \frac{2}{3}(1 - t_l^*).$$

If we note that $\frac{d}{dt_l}h(t_l^*) = 2(1 - t_l^*)$ then we see that increasing λ^* decreases $c^*(t)$ for all $t \in \overline{T}$.

5.6 Multiple Conformity equilibria

In the introduction we briefly talked of two possible levels of uncertainty for agents: (1) whether an action will become a norm and, (2) if so, which action. So far, we have concentrated on the first of these by assuming that the conformity equilibrium is centered around action 1. We briefly here make some comments on the second possibility.

Suppose that when $\lambda = \lambda^*$ there are two 'focal' signalling equilibria - $(\alpha^*, \xi^*, \phi^*)$ centered on x_p^* and (α^{**}, ξ^{**}) centered on x_p^{**}. For example, it may be that $\alpha^*(t) = 0.9$ for all t while $\alpha^{**}(t) = 1.1$ for all t. We can now think of there as being three strategies:

To not conform (N): If the agent is type t he chooses action $a = t$ and accords all other agents esteem 0.

To conform to x_p^* (C^*): If the agent is type t he chooses action $\alpha^*(t)$ and accords other agents esteem according to function ξ^*.

To conform to x_p^{**} (C^{**}): If the agent is type t he chooses action $\alpha^{**}(t)$ and accords other agents esteem according to function ξ^{**}.

As before, we can ask whether conformity or non-conformity can invade. Whether or not conformity can invade will depend on how conformity begins to become established in the population. If we consider a population where proportion $1 - \varepsilon$ are playing strategy N and ε are playing strategy C^* then the analysis would essentially be identical to that considered in the rest of the paper. An alternative approach is to suppose that proportion ε of the population is conforming but there may be uncertainty over whether agents are choosing C^* or C^{**}.

To make things more concrete suppose that proportion $1 - 2c$ of agents are playing N, proportion c are playing C^* and proportion c are playing C^{**}. Further suppose that $\xi^*(t) = h(0)$ for all $t \neq x_p^*$ and $\xi^{**}(t) = h(0)$ for all

$t \neq x_p^{**}$. Finally, let $\xi^*(x_p^*) = \xi^{**}(x_p^{**}) \equiv A$. The payoff of an agent of type $t \neq x_p^*, x_p^{**}$ will be given by,

$$\begin{aligned} U(N,t,c) &= g(0) + 2c\lambda^* h(0) \qquad (12) \\ U(C^*,t,c) &= g(t-\alpha^*(t)) + c\lambda^* \left(h(0)+A\right) \\ U(C^{**},t,c) &= g(t-\alpha^{**}(t)) + c\lambda^* \left(h(0)+A\right). \qquad (13) \end{aligned}$$

An agent of type t will have a clear preference between C^* and C^{**} so let $c^*(t)$ denote the threshold value c for which an agent of type t is indifferent between conforming and not-conforming. Given that $h(0) < A$ we can see that the threshold value just described will be less than the threshold value used in the rest of the paper when there was a unique conformity equilibrium. In other words, it is 'more difficult' for conformity to invade when there are multiple conformity equilibrium than when there is a unique conformity equilibrium.

The intuition for the above result is that an agent who conforms may still be perceived as 'extreme' by others because he has conformed on a different action to them. To provide a simple example consider a person not knowing what to wear to a party. His intrinsic preference is to wear jeans and a tee-shirt. He considers it possible, however, that a 'suit norm' or 'smart casual' norm may exist. Suppose the party invite made allusion to smart casual. The individual then faces a straight choice - to where jeans and feel 'physically comfortable' but risk feeling 'social uncomfortable' or to where smart casual and feel 'physically uncomfortable' but to be guaranteed feeling 'social comfortable'; in this case he may well choose to conform and where smart casual. By contrast, suppose there is no indication what people will wear. The costs and benefits of wearing jeans are the same. The costs and benefits of smart casual have changed - he will still feel 'physically uncomfortable' but now also runs the risk of feeling 'social uncomfortable' if everyone else wears a suit; he may well take 'the safe option' and wear jeans.

6 Conclusions

We have considered a model of conformity permitting both a conformist and non-conformist equilibrium. It has been demonstrated that the non-conformist equilibrium is often unstable. In particular, if a small proportion of the population conforms or acts in accordance with some norm then conformity can spread through the population to the point where everyone is conforming. This spread of conformity happens independently of whether or not the resulting conformist equilibrium Pareto dominates the non-conformist equilibrium. Our analysis thus serves as a partial explanation for why conformity can arise.

As highlighted in Section 5.6 the spread of conformity, as modelled, is reliant on there being some focal action that serves as a point around which conformity can arise. If there are multiple potential norms then the spread of conformity may be held up. Essentially, conforming becomes a more risky option if you cannot be sure to conform on the same actions as others. This suggests that an important determinant of whether or not conformity exists in a particular setting will be the existence or otherwise of a clear focal action to act as a norm. This is a question that could be pursued further. For example, understanding the process by which conformity can arise if there are multiple potential norms may also shed light on how norms can come to change or evolve over time.

As already remarked this paper has focussed exclusively on a social or emotional reason for conformity. This can be contrasted with much of the literature dealing with herding and cascades or informational reasons for conformity. An interesting avenue for future research would seem to be to try and see how these two different reasons for conformity integrate together. For example do social and information conformity positively or negatively reinforce each other in bringing about conformity. Social conformity (based on some action a) may, for instance, act as a barrier to achieving efficiency through informational conformity (at some point b). The seeming ease with

which social conformity arises in our analysis would suggest that this is possible.

7 Appendix

For the spherical example we derive the signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$ when $\lambda = \lambda^*$. Indifference curves for a type t agent are give by,

$$(t-x)^2 + \lambda(1-b)^2 = C$$

where x is action and b is inferred type. In equilibrium indifference curves must be tangent to $\phi_s(x)$ and beliefs must be self fulfilling implying,

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

As pointed out by Bernheim, (14) is equivalent to the following linear dynamic system:

$$\begin{bmatrix} \frac{dt}{dv} \\ \frac{dx}{dv} \end{bmatrix} = \begin{bmatrix} x-t \\ \lambda(1-t) \end{bmatrix}.$$

Working through one obtains the differential equation $x'' + x' + \lambda x = \lambda$. This has roots $-\frac{1}{2} \pm \frac{1}{2}(1-4\lambda)^{\frac{1}{2}}$. Clearly the roots are real if and only if $\lambda \leq \frac{1}{4}$.

When $\lambda = \frac{1}{4}$ one gets the general solution $x = C_1 e^{-\frac{1}{2}v} + C_2 v e^{-\frac{1}{2}v} + 1$ and $t = 1 - 4x'$. Appropriate initial conditions are $x(0) = t(0) = 0$ giving particular solution,

$$\begin{aligned} x &= -e^{-\frac{1}{2}v} - \frac{v}{4}e^{-\frac{1}{2}v} + 1 \\ t &= 1 - e^{-\frac{1}{2}v}\left(1 + \frac{v}{2}\right). \end{aligned}$$

Let $d \equiv (4\lambda - 1)^{\frac{1}{2}}$. When $\lambda > \frac{1}{4}$ one gets the general solution $x = e^{-\frac{1}{2}v}\left(C_1 \cos\frac{d}{2}v + C_2 \sin\frac{d}{2}v\right) + 1$. Using $t = 1 - \frac{1}{\lambda}x'$ one gets $t = 1 + \frac{1}{2\lambda}e^{-\frac{1}{2}v}$

$\big((C_1 - C_2 d)\ \cos\frac{d}{2}v + \ (C_2 + C_1 d)\sin\frac{d}{2}v\big)$. Using initial conditions $x(0) = t(0) = 0$ one obtains the particular solution,

$$\begin{aligned} x &= 1 + e^{-\frac{1}{2}v}\left(\frac{2\lambda - 1}{d}\sin\frac{d}{2}v - \cos\frac{d}{2}v\right) \\ t &= 1 + e^{-\frac{1}{2}v}\left(\frac{1}{2\lambda}\left(\frac{2\lambda - 1}{d} - d\right)\sin\frac{d}{2}v - \cos\frac{d}{2}v\right). \end{aligned}$$

Plugging in $d = 1$ when $\lambda = \frac{1}{2}$, $d = 2$ when $\lambda = 1.25$ and $d = \sqrt{3}$ when $\lambda = 1$ one gets the desired solution.

Finally, if an agent of type t_l^* is indifferent between playing action 1 and action $\phi_s^{-1}(t_l^*)$ this implies that,

$$-(t_l^* - \phi_s^{-1}(t_l^*))^2 - \lambda(1 - t_l^*)^2 = -(1 - t_l^*)^2 + \lambda\xi^*(1)$$

where

$$\xi^*(1) = \frac{1}{F(2 - t_l^*) - F(t_l^*)}\int_{t_l^*}^{2-t_l^*} -(1-b)^2 f(b)db.$$

Thus, $\xi^*(1) = -\frac{(1-t_l^*)^2}{3}$ when $f(t) = \frac{1}{2}$ for all t. One can now find t_l^* and fully characterize the signalling equilibrium $(\alpha^*, \xi^*, \phi^*)$. We do this using numerical search. When $\lambda^* = 1.25$ and $f(t) = 0.5$ we obtain $t_l^* = 0.0761$ and $\xi^*(1) = -0.2845$. When $\lambda^* = 1$ and $f(t) = 0.5$ we obtain $t_l^* = 0.1932$ and $\xi^*(1) = -0.2170$. Finally, when $\lambda^* = 0.5$ and $f(t) = 0.5$ we obtain $t_l^* = 0.6907$ and $\xi^*(1) = -0.0319$.

Figure 1a: A fully separating signalling equilibrium:

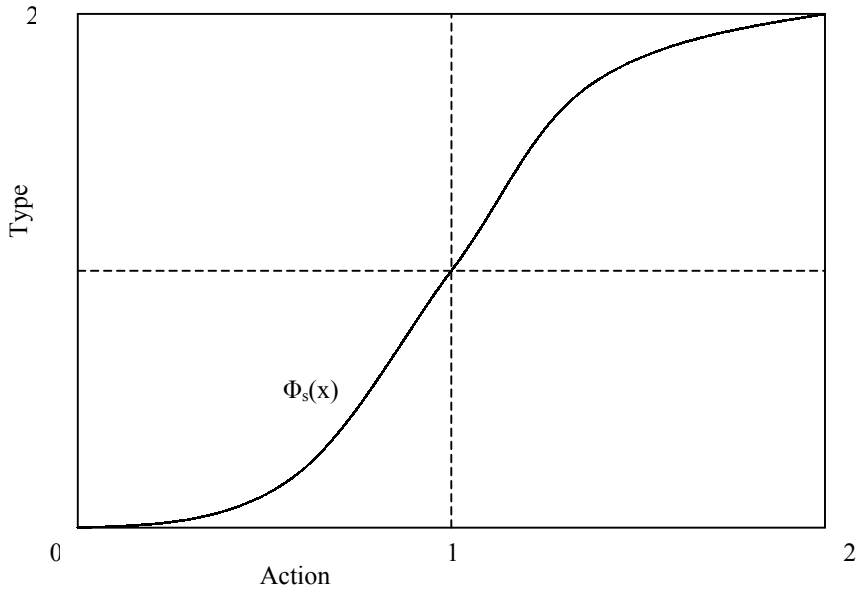


Figure 1b: A signalling equilibrium with incomplete separation where $x_p = 1$.

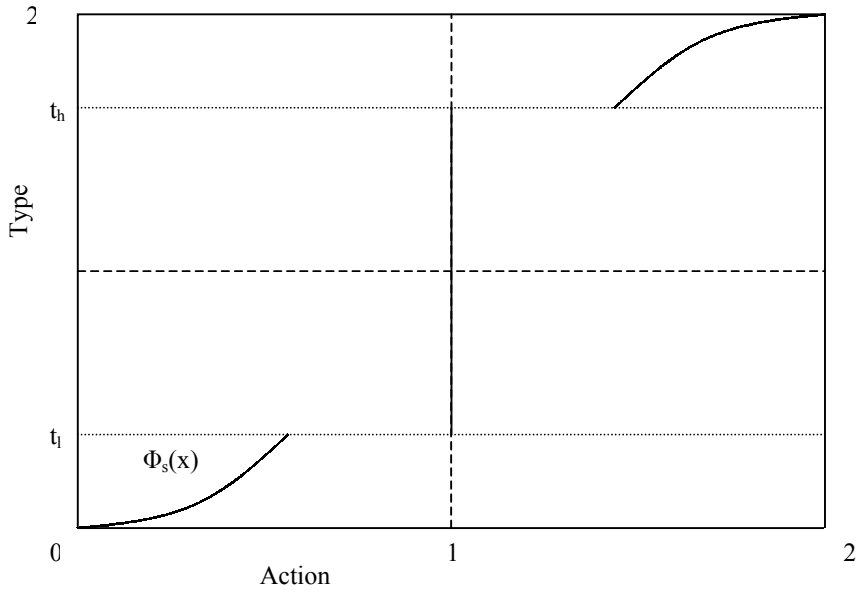


Figure 2a: c*(t) and 1 - 2F(t) plotted against t when f(t) = 0.5

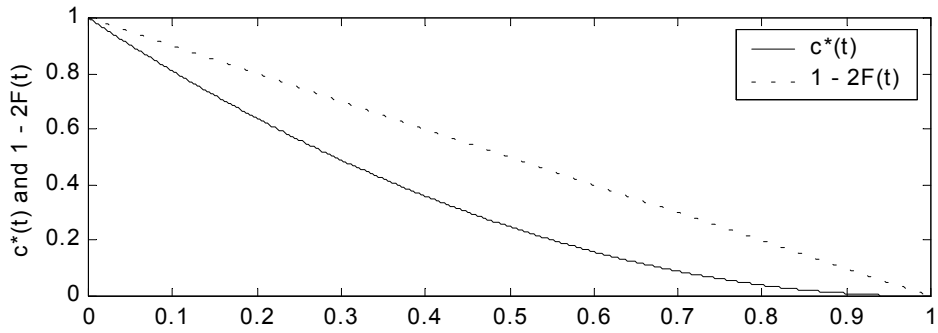


Figure 2b: c*(t) and 1 - 2F(t) plotted against t when f(t) = 1 - t

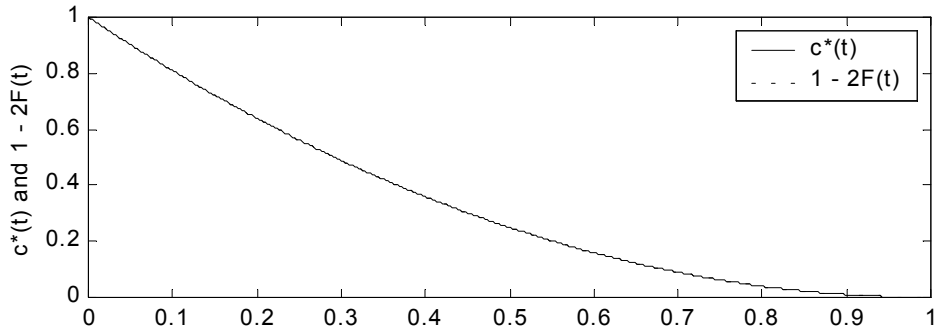


Figure 2c: c*(t) and 1 - 2F(t) plotted against t when many agents with extreme types

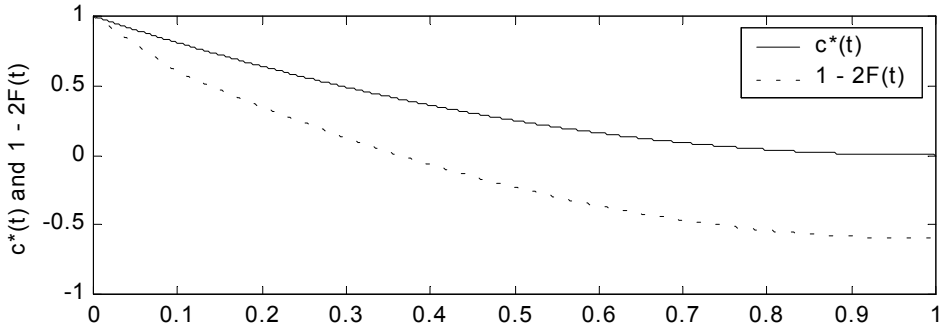


Figure 2d: c*(t) and 1 - 2F(t) plotted against t when f is bimodal

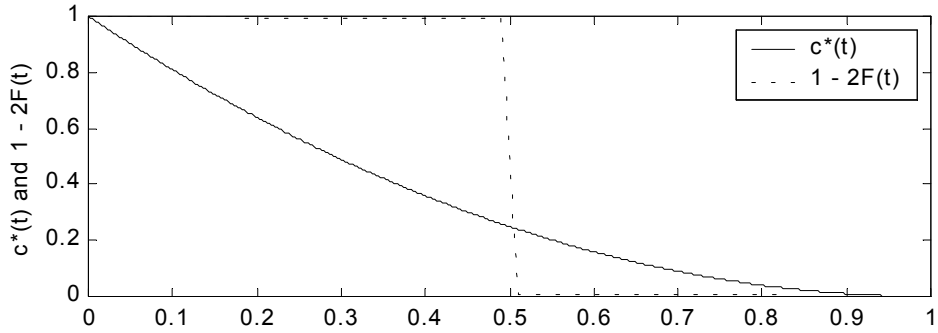


Figure 3: c*(t) and 1 - 2F(t) plotted against t when f(t) = 0.5 and lambda* = 1.25

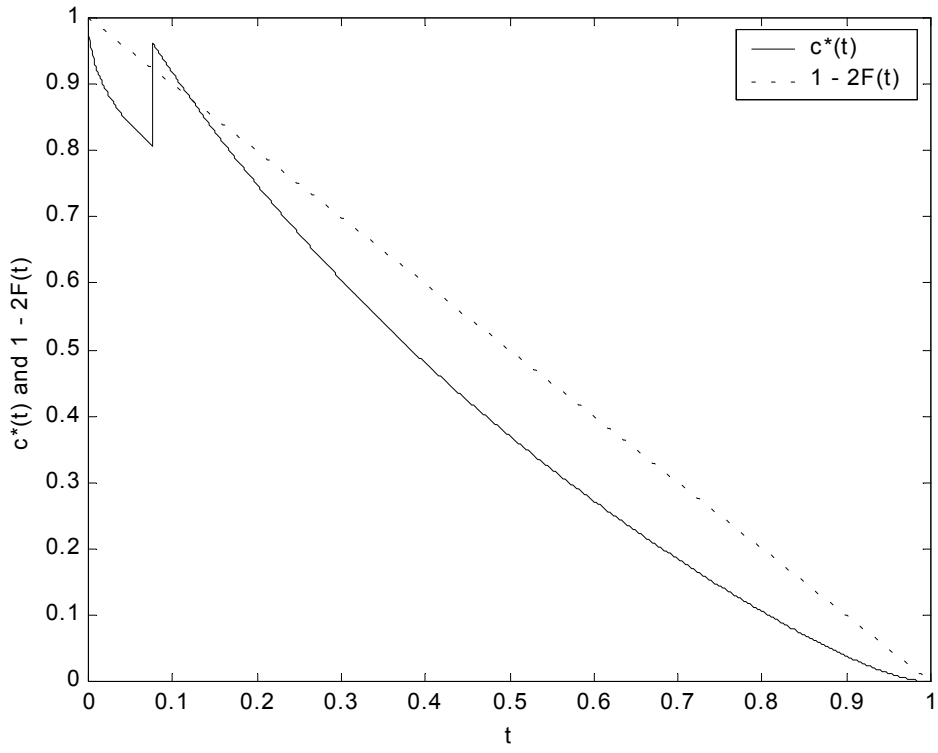


Figure 4: c*(t) and 1 - 2F(t) plotted against t when f(t) = 0.5 and lambda* = 0.5

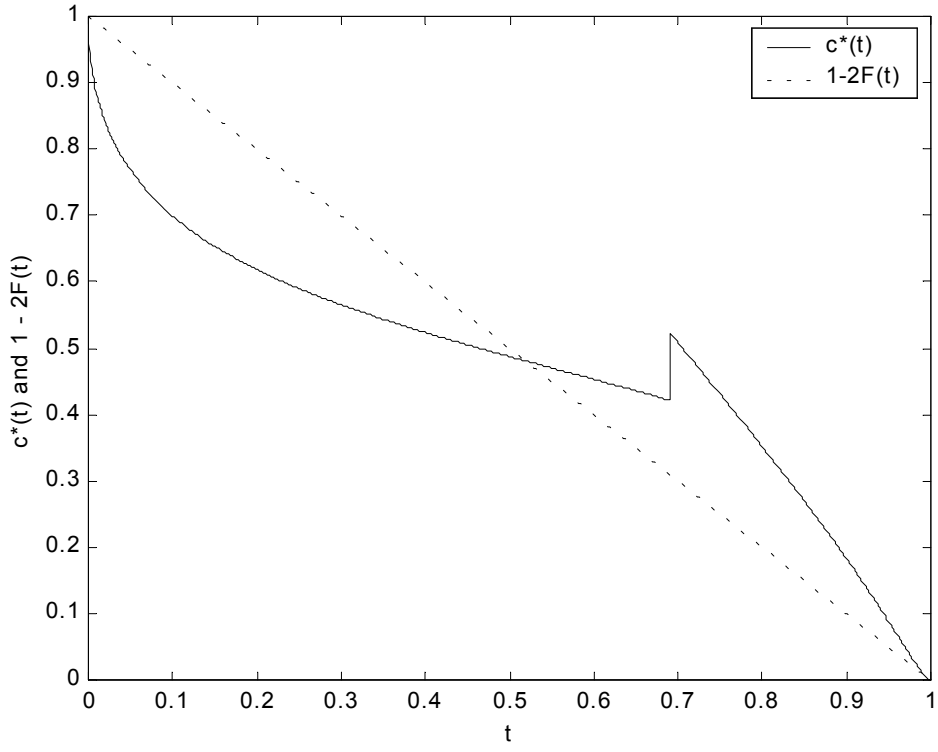


Figure 5: c*(t) and 1 - 2F(t) plotted against t when f(t) = 0.5 and lambda* = 1

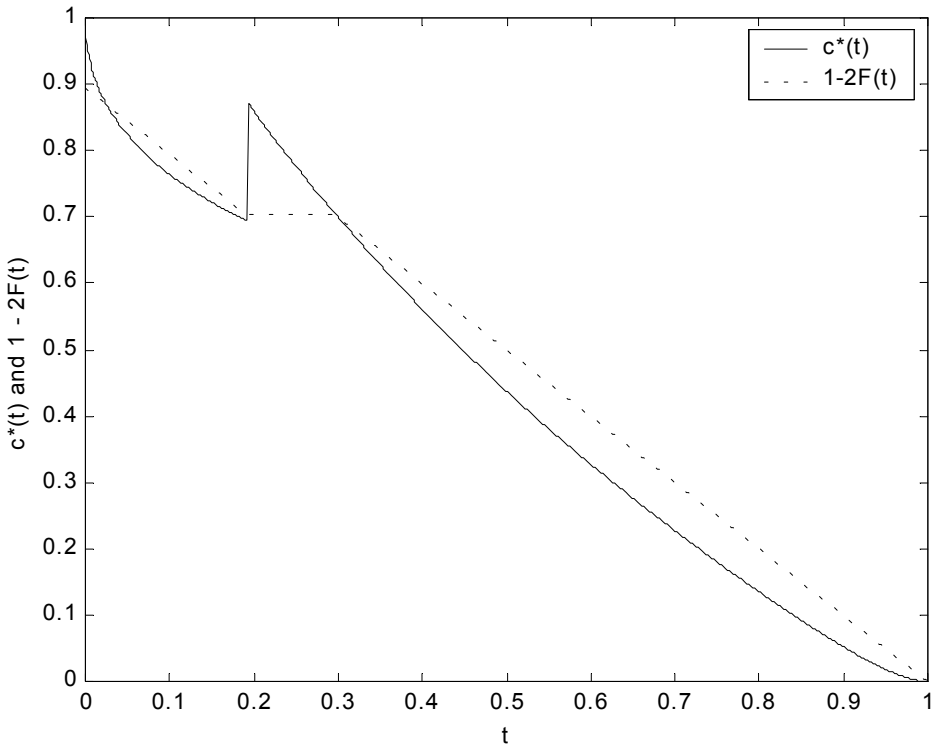
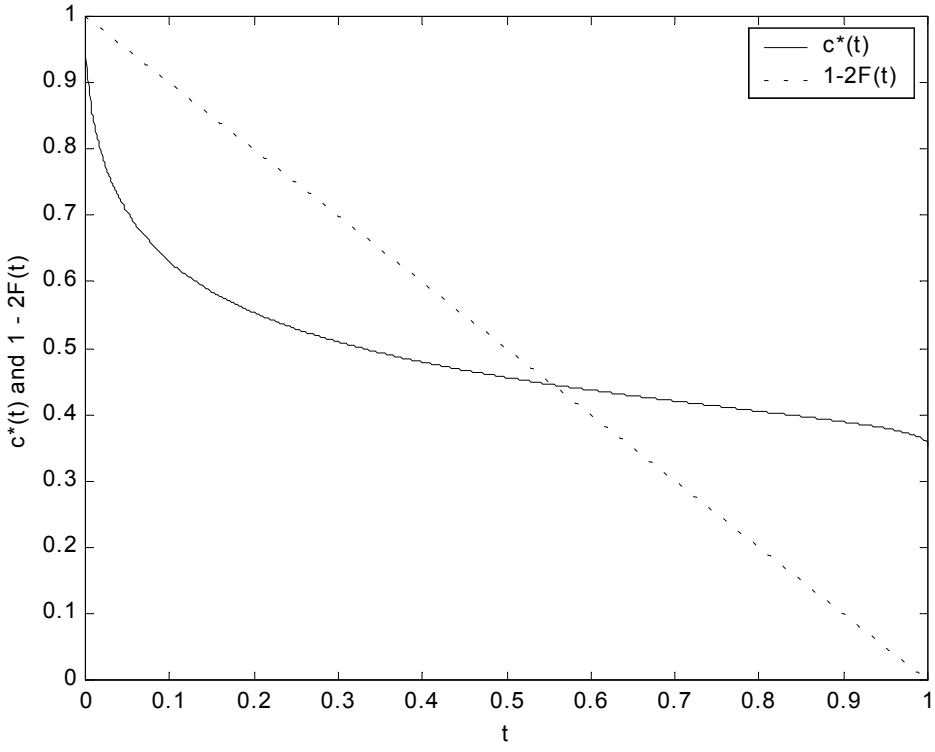


Figure 6: c*(t) and 1 - 2F(t) plotted against t when f(t) = 0.5 and lambda* = 0.25



References

[1] Akerlof, G. A., (1980) 'A Theory of Social Custom of which Unemployment May Be One Consequence', *Quarterly Journal of Economics* 94: 749-75.

[2] Bernheim, B. D., (1994) 'A Theory of Conformity', *Journal of Political Economy* 102: 841-877.

[3] Bikhandani, S., D. Hirshleifer and I. Welch, (1992) 'A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades', *Journal of Political Economy* 100: 992-1026.

[4] Elster, J., (1989) 'Social Norms and Economic Theory', *Journal of Economic Perspectives* 3: 99-117.

[5] Elster, J., (1998) 'Emotions and Economic Theory', *Journal of Economic Literature* 36: 47-74.

[6] Fudenberg, D. and D.K. Levine, (1998) *The Theory of Learning in Games*, MIT Press.

[7] Fudenberg, D. and J. Tirole, (1991) *Game Theory*, Cambridge, Mass. MIT Press.

[8] Jones, S. R. G., (1984) *The Economics of Conformism* Oxford. Blackwell.

[9] Juang, W., (2001) 'Learning from Popularity', *Econometrica* 69: 735-747.

[10] Lewis, D., (1967) *Convention: A Philisophical Study*, Cambridge, Mass: Harvard University Press.

[11] Lindbeck, A. (1997) 'Incentives and Social Norms in Household Behavior', *American Economic Review* 87: 370-377.

[12] Morris, S., (2000) 'Contagion', *Review of Economic Studies* 67: 57-78.

[13] Shiller, R. J., (1995) 'Conversation, Information and Herd Behavior', *American Economic Review* 85: 181-186.

[14] Young, P., (1993) 'Evolution of conventions', *Econometrica* 61: 57-84.