The dynamics of costly norms[1]

Sam Jindani[2] and Peyton Young[3]

12 March 2025

Social norms that are costly for individuals are held in place because of social pressure to conform. Historical examples include duelling in the Europe and footbinding in China; contemporary examples include female genital cutting and wasteful consumption. Such norms exhibit varied dynamics. Some collapse suddenly after staying unchanged for centuries, while others erode gradually. Still others escalate gradually before collapsing. This paper develops a theory that explains these patterns in terms of different forms of social influence. The analysis can guide the design of policies aimed at mitigating or eliminating costly norms, and shows that some interventions can be counterproductive.

1 Introduction

Harmful or costly norms are pervasive in many societies and can lead to significant losses in welfare (Edgerton 1992; Bicchieri 2005; Belloc and Bowles 2013). Historical examples include duelling in Europe and footbinding in China; contemporary examples include female genital cutting (FGC) in parts of Africa and norms of conspicuous consumption, whose main purpose is to enhance social status, more or less everywhere.[4] Although adherence to such norms can be risky, injurious to one's health, or very expensive, the social pressure to conform is so powerful that people prefer to bear the cost.

The dynamics of such norms are strikingly diverse. Some persist for a long time and then die out suddenly, as was the case for duelling in the UK and footbinding

[1] We thank Sam Bowles, Jean-Paul Carvalho, Charles Efferson, and Yuya Kudo for helpful comments, and Shuo Yang for excellent research assistance.

[2] Department of Economics, National University of Singapore; `sam.jindani@nus.edu.sg`.

[3] Nuffield College, University of Oxford; `peyton.young@economics.ox.ac.uk`.

[4] On duelling, see Nye 1993; Hopton 2007; Banks 2012. On footbinding, see Mackie 1996; Ko 2007; Gates 2015. On FGC, see Shell-Duncan and Hernlund 2000; Shell-Duncan et al. 2011. On conspicuous consumption, see Bagwell and Bernheim 1996; Hopkins and Kornienko 2004; Hopkins 2023.

in China. Others transition to intermediate forms before disappearing, as was the case for duelling in France and may be the case for FGC practices in parts of Africa. Still others, such as norms of conspicuous consumption, tend to escalate gradually, sometimes followed by a sudden collapse. We develop a model that explains these patterns in terms of different forces of social interaction.

The theoretical framework builds on the social-interactions approach introduced by G. A. Akerlof (1980, 1997), Durlauf (1997), Brock and Durlauf (2001), and Blume and Durlauf (2003). In these models an individual's payoff consists of two components: an intrinsic or personal utility from taking the action, and a social utility that depends on the degree of disparity between one's own action and the actions taken by others. We consider a general formulation of the model, which allows more than two actions and does not assume symmetric influence between agents, as is common in much of the literature.[5] Moreover, while the prior theoretical literature has largely been concerned with the characterisation of equilibria and with long-run selection dynamics, the present paper focusses instead on short- and intermediate-run dynamics. By *short-run dynamics*, we mean whether or not the process converges to equilibrium from out-of-equilibrium conditions. As we will see, this issue is not trivial in social-interactions models. By *intermediate-run dynamics*, we mean the transition from one equilibrium to the next following exogenous shocks. We will see that the intermediate-run dynamics can be highly varied and depend crucially on the shape of the social-influence function, in ways that shed light on historical and contemporary examples.[6]

In general, the social utility of an action depends on the extent to which it differs from the actions of others. However, the direction of deviation matters and so do the particular interpretations placed on deviations. Taking a less costly action

[5] The social-interactions literature contributes to the more general programme of integrating social factors into game-theoretic modelling. On the theoretical side, see in particular Schelling 1978; Granovetter 1978; Kandori 1992; Fehr and Schmidt 1999; G. A. Akerlof and Kranton 2000, 2010; Bowles and Gintis 2002; Bowles 2004; Bicchieri 2005, 2016; Bénabou and Tirole 2006; Jackson 2008; Boyd, Gintis, and Bowles 2010; Goyal 2012, 2023; Young 2015; R. Akerlof 2016, 2017. There is also a significant empirical and experimental literature that attests to the importance of these social factors. See among others Glaeser, Sacerdote, and Scheinkman 1996; Glaeser and Scheinkman 2001; Henrich et al. 2001; Fehr and Gächter 2002; Mathew 2017; Centola et al. 2018; Enke 2019; Henrich 2020; Andreoni, Nikiforakis, and Siegenthaler 2021; Cao et al. 2021.

[6] In contrast, by *long-run dynamics*, we mean the analysis of stationary distributions traditionally undertaken in evolutionary game theory. For applications of evolutionary game theory to social norms, see among others Young 1993; Kandori, Mailath, and Rob 1993; Blume 1993; Skyrms 1996, 2003; Kandori and Rob 1998; Brock and Durlauf 2001; Blume and Durlauf 2003; Friedman and Ostrov 2008; Belloc and Bowles 2013. For general book-length treatments of evolutionary game theory see Weibull 1995; Samuelson 1997; Young 1998; Vega-Redondo 1996; Bowles 2004; Boyd and Richerson 2005.

than the norm may be interpreted as shirking and lead to shunning, mockery, or loss of status. Taking a more costly action than the norm may be interpreted as unnecessarily reckless or a sign of insecurity. These are situations where social interactions disincentivise deviations, but there is no reason to think they do so symmetrically in the direction of the deviation. In other situations, social interactions incentivise deviations toward more costly actions. For example, tipping the waiter more than the norm expresses exceptional generosity and enhances one's status; owning a luxury handbag or watch signals wealth. As we shall see, these differences have important implications for norm dynamics. We provide intuitive conditions under which norms will tend to collapse suddenly on the one hand and erode gradually on the other.

We also show how this framework can shed light on policy interventions. We consider bans and taxes on harmful behaviour, interventions that affect the social-influence function, and policies aimed at convincing subgroups to switch to less costly behaviours. We show that such interventions can have unintended, counterproductive effects. This highlights the importance of understanding the nature of social preferences when designing policies.[7]

Our contribution is thus threefold. First, we study a general form of the social-interactions model, allowing for more than two actions, asymmetric social influence, and both disincentivised and incentivised deviations. Second, we analyse the dynamic behaviour of the process both in the short and intermediate run. Third, we examine how this theoretical framework can be brought to bear on policy.

2 A model of costly social norms

There is a finite set of actions $A = \{0, 1, 2, \ldots, n\}$ and m agents. Let $\delta = 1/m$. A *state* $p = (p_0, p_1, \ldots, p_n)$ specifies the proportion of agents choosing each action. Let $\Delta = \{p \in \mathbb{R}_+^{n+1} : \sum_{i \in A} p_i = 1\}$ denote the n-dimensional simplex and $\tilde{\Delta}$ denote the set of feasible states; that is,

$$\tilde{\Delta} = \{p \in \Delta : p_i \in \{0, \delta, 2\delta, \ldots, 1\} \text{ for each } i \in A\}. \tag{1}$$

[7] Papers that study policy interventions against costly social norms include Shell-Duncan et al. 2011; Platteau, Camilotti, and Auriol 2018; Efferson, Vogt, and Fehr 2020. A related paper to the present one is our study of FGC practices in Somalia (Gulesci, Jindani, La Ferrara, Smerdon, Sulaiman, and Young 2024).

An agent's utility from choosing an action has two components: personal and social. Let $c_i \geq 0$ be the *personal cost* of taking action $i \in A$, which for simplicity we will assume is the same for all agents. Assume that costs are distinct and ordered so that $c_0 < c_1 < \cdots < c_n$, and normalise $c_0 = 0$.

The *social utility* of an action represents the influence exerted by other members of society. This influence may be external: agents shun, express disapproval towards, or break off relations with those whose actions differ from theirs. Or it may be internal: agents feel shame or inadequacy when they violate a norm.

We shall assume that social influence disincentivises the choice of *less costly* actions. The gentleman who refuses a challenge to a duel may be mocked and shunned by his peers. Likewise, the billionaire who doesn't own a super-yacht is made to feel inferior, and the family that chooses not to cut their daughter is stigmatised. By contrast, *more costly* actions may be either disincentivised or incentivised. On the one hand the gentleman who insists on using pistols when swords are called for or standing at fewer paces than the norm may be seen as reckless and hot-headed, thereby suffering a loss of esteem. On the other hand, the billionaire who owns the largest super-yacht feels superior and enjoys high social status.

We capture these different forces through a *social-influence function*, which depends on the proportion of agents choosing each action in the current state. An agent exerts social influence on another agent to a degree that depends on the difference between the costs of the two agents. The *social utility* of an agent playing action $i \in A$ in state $p \in \tilde{\Delta}$ is

$$-\sum_{j \in A} p_j s(c_j, c_i), \tag{2}$$

where $s : \mathbb{R}^2_+ \to \mathbb{R}$ satisfies:

1. $s(c, c) = 0$ for every $c \geq 0$; and
2. if $c \leq c' \leq c''$ or $c \geq c' \geq c''$, then $|s(c, c'')| \geq |s(c, c')|$ and $|s(c, c'')| \geq |s(c', c'')|$.

The second condition is a monotonicity condition. The interpretation is that an agent choosing action j exerts social influence proportional to $s(c_j, c_i)$ on an agent choosing i. Note that $s(c_j, c_i)$ may be positive or negative and is greater (in absolute value) the more different i and j are. Moreover, social influence may be asymmetric, in the sense that we allow $s(c, c') \neq s(c', c)$. This specification is more general than much of the literature, which typically assumes that social influence is not only symmetric but also depends only on the absolute distance between actions $|c_j - c_i|$.

For the sake of simplicity, assume that personal and social utilities are separable. Thus the total utility of an agent playing action i in state p is

$$u_i(p) = -c_i - \lambda \sum_{j \in A} p_j s(c_j, c_i), \tag{3}$$

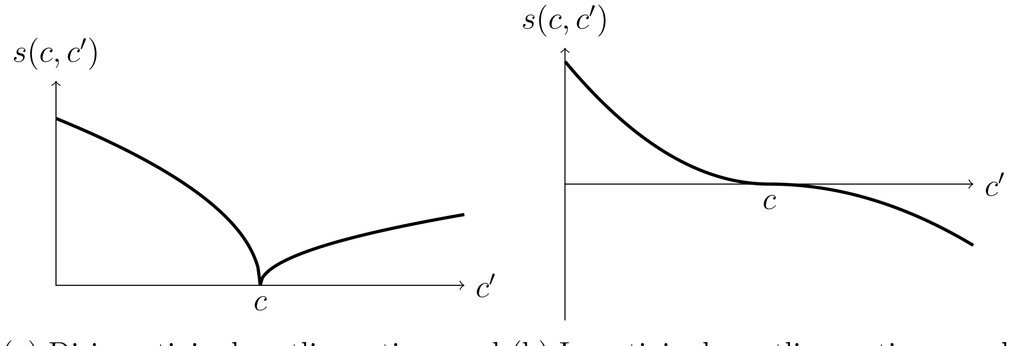
where $\lambda \geq 0$ captures the importance of social relative to personal utility.

Let p^i denote the state in which all agents play action i; such a state is *homogenous*. For $i \neq j$, define the unit-switching vector $e^{ij} \in \mathbb{R}^n$ as follows: $e^{ij}_i = -\delta$, $e^{ij}_j = \delta$, and $e^{ij}_k = 0$ for $k \neq i, j$. Thus if the current state is p and an agent switches from i to j, the new state is $p + e^{ij}$. It will be convenient to let $e^{ii} = (0, \ldots, 0)$ for all i. A state is *stable* if it is a strict Nash equilibrium. Formally, for all i such that $p_i > 0$ and for all j, $u_i(p) > u_j(p + e^{ij})$. (The unit-switching vector accounts for the fact that an agent's switching changes the state.) A *norm* is a homogeneous, stable state. For simplicity, we will also refer to $i \in A$ as a norm if p^i is stable. Let $A^* = \{i \in A : p^i \text{ is stable}\}$ denote the set of norms.

The dynamics are defined as follows. Time is discrete. At the start of each time period, an agent is chosen at random to revise her action and chooses a best response to the current state (if there are multiple best responses, she chooses one at random with uniform probability).

In what follows, we will see that the sign and shape of the social-influence function play an important role in determining the dynamics of the process. As discussed above, social interactions typically disincentivise less costly actions. To capture this, we assume that $s(c, c') > 0$ whenever $c' < c$. Social interactions may either disincentivise or incentivise more costly actions. The former case corresponds to $s(c, c') \geq 0$ whenever $c' > c$, as shown in figure 1a. The latter case corresponds to $s(c, c') < 0$ whenever $c' > c$, as shown in figure 1b. Note that the model allows for s to be discontinuous, although this isn't illustrated in figure 1.

Consider the case where deviations take on a symbolic dimension, so that even small deviations incur a substantial social cost. For example, if someone requests to duel at twelve paces instead of the conventional ten, he may be accused of cowardice and be treated almost as badly as if he had refused to duel altogether. Similarly, in the case of conspicuous consumption, ordinal comparisons can be at least as important as cardinal comparisons: if someone owns a super-yacht that is only a few feet longer than someone else's, they still own the biggest super-



(a) Disincentivised costlier actions and weakening social influence

(b) Incentivised costlier actions and strengthening social influence

Figure 1: Different shapes of the social-influence function.

yacht. In such cases we assume that, for every $c < c'$, $\frac{|s(c,c')|}{c'-c}$ and $\frac{|s(c',c)|}{c'-c}$ are strictly increasing in c and strictly decreasing in c', provided they are nonzero.

We say that social influence *weakens* relative to intrinsic cost. This is illustrated in figure 1a. One can think of this as a concavity of social influence, although the formal conditions are more general.

In other cases, small deviations incur relatively little social pressure. For example, in the case of FGC in Somalia, the less severe form of cutting known as Sunna is increasingly seen as an acceptable alternative to the more traditional Pharaonic form, so that a woman with Sunna may suffer relatively little social consequences; at the same time, a woman who is not cut at all may be severely ostracised. Another example is the case of duelling in France, where, in the nineteenth century, swords became popular again, replacing pistols and leading to a sharp fall in fatalities. Around this time, swords became viewed as an honourable, traditional way of conducting duels, and drawing a small amount of blood was considered sufficient to resolve the dispute.[8] In these and related cases, for every $c < c'$, $\frac{|s(c,c')|}{c'-c}$ and $\frac{|s(c',c)|}{c'-c}$ are strictly decreasing in c and strictly increasing in c', provided they are nonzero. We say that social influence *strengthens* relative to intrinsic cost. This is illustrated in figure 1b.

[8] These changes had the effect of drastically reducing fatality rates. More than a third of duels in the early nineteenth century in France ended in the death of one of the duellists; by the second half of the nineteenth century, it was two percent. In *A tramp abroad*, published in 1880, Mark Twain wrote that 'Much as the modern French duel is ridiculed by certain smart people, it is in reality one of the most dangerous institutions of our day. Since it is always fought in the open air, the combatants are nearly sure to catch cold.'

3 Disincentivised deviations

We now turn to an analysis of the model when deviations in both directions are disincentivised. The dynamics of the process can be complex. To see why, suppose we start from a highly heterogenous state in which some agents are at highly costly actions and others at relatively uncostly ones. Depending on the social-influence function, an updating agent's best response could be costly, uncostly, or intermediate. In particular, it could be an action that is not chosen by any other player. Moreover, the best response might depend on the current action of the updating player, because different updating players face slightly different distributions of actions. Nevertheless, theorem 1 below establishes that when all deviations are disincentivised, the process always reaches a norm from any starting state.[9] Later, we will see that under different assumptions, convergence may not occur.

Theorem 1. *If all deviations are disincentivised, the process converges to a norm from any initial state.*

Proof. Let p be the initial state; suppose it is not a norm. Let

$$\bar{u} = \max_{\{i : p_i > 0\}} u_i(p) \tag{4}$$

$$\bar{A} = \arg\max_{\{i : p_i > 0\}} u_i(p). \tag{5}$$

Thus, $\bar{u}$ is the maximal utility at p and $\bar{A}$ is the set of actions that achieve the maximal utility. $\underline{A} = \{i : p_i > 0\} \setminus \bar{A}$ be the set of actions that are played and do not achieve the maximal utility. We argue there is positive probability of reaching either a norm or a state with strictly higher maximal utility in at most $m - 1$ periods.

First, suppose p is homogeneous. Let i be the action played. Since p is not a norm, some agent can weakly increase their payoff by switching to a best response $j \neq i$. If this strictly increases their payoff, then we are done. Suppose it leaves their payoff unchanged. Switching to $j > i$ from p^i strictly decreases the agent's payoff, so it must be the case that $j < i$ and thus $s(c_i, c_j) > 0$. But then in the following period a second agent could switch from i to j and strictly increase her payoff.

[9] Note that if $n = 2$ or if social influence is symmetric, then the game is a potential game and convergence is assured (Monderer and Shapley 1996); however, in general this is not the case.

Second, suppose $\bar{A}$ is not a singleton. Let $i, j \in \bar{A}$ be such that $i < j$. Then $s(c_j, c_i) > 0$, and hence some agent can strictly increase their payoff by switching from j to i.

Finally, suppose $\bar{A}$ is a singleton and p is not homogeneous. Let $i^* \in \bar{A}$. If there is some $i \in \underline{A}$ such that switching from i to i^* yields strictly more than $\bar{u}$, then we are done. Moreover, an agent playing $i \in \underline{A}$ can achieve at least $\bar{u}$ by switching to i^*, so if i^* is not a best response, then we are done. Therefore suppose that, for all $i \in \underline{A}$, i^* is a best response and switching to i^* yields exactly $\bar{u}$. Suppose the agents playing actions in $\underline{A}$ successively switch to i^*. The utility of i^* remains $\bar{u}$. If at any point some $j \neq i^*$ becomes a best response, then an agent can achieve strictly more than $\bar{u}$ and we are done. If not, we must reach the homogeneous state p^{i^*} in at most $m - 1$ periods. Moreover, since no other action is a best response, p^{i^*} must be a norm.

We have established that for any state that is not a norm, there is a positive probability of reaching either a norm or a state with strictly higher maximal utility in at most $m - 1$ periods. Since there are finitely many states, there is a global maximum utility and a minimum increment in the maximal utility, so the maximal utility cannot continue increasing indefinitely. It follows that, from any state, the probability of reaching a norm in finite time is one. □

Theorem 1 implies that, in the short run, the process will converge to a norm. In the intermediate run, changes in parameters can lead to transitions between norms. For example, historians have attributed the demise of duelling to societal changes that took place at the time of the industrial revolution, including increases in life expectancy, economic changes, and changes around norms of masculinity. We are thus interested in the dynamics of the process following a change in parameters that makes a norm unstable. We refer to this as the *intermediate-run dynamics.* Proposition 1 below characterises these dynamics when social influence weakens and shows that norms will tend to collapse suddenly.

Proposition 1. *Suppose that all deviations are disincentivised and social influence weakens relative to intrinsic cost. Then, starting from any homogeneous but unstable state, the process collapses directly to the uncostly norm.*

Proof. Let p^i be the state. By assumption, i is not stable so i is not a best response.

Consider $j > i$. Since $s(c_i, c_j) \geq 0$ and $c_j > c_i$, we have

$$
\begin{align}
& -c_i > -c_j - \lambda \frac{m-1}{m} s(c_i, c_j) \tag{6} \\
\implies & u_i(p^i) > u_j(p^i + e^{ij}), \tag{7}
\end{align}
$$

so $j > i$ is not a best response at p^i.

Consider $j < i$. We have

$$
\begin{align}
u_j(p^i + e^{ij}) - u_i(p^i) &= c_i - c_j - \lambda \frac{m-1}{m} s(c_i, c_j) \tag{8} \\
&= (c_i - c_j) \left(1 - \lambda \frac{m-1}{m} \frac{s(c_i, c_j)}{c_i - c_j} \right). \tag{9}
\end{align}
$$

Under the assumption that social influence weakens, the right-hand side is decreasing in $j < i$. Hence action 0 is the unique best response at p^i.

Let $l \in \{0, 1, \ldots, m-1\}$ and suppose 0 is the unique best response at $p = p^i + le^{i0}$. Consider state $p' = p^i + (l+1)e^{i0}$. Fix $j \neq 0$. If $j > i$, then

$$
\begin{align}
& u_i(p') - u_j(p' + e^{ij}) \nonumber \\
& \quad = c_j - c_i + \lambda \left(\frac{l+1}{m} (s(c_0, c_j) - s(c_0, c_i)) + \frac{m-l-2}{m} s(c_i, c_j) \right) \tag{10} \\
& \quad > 0, \tag{11}
\end{align}
$$

where the inequality follows from the fact that $c_j > c_i$, $s(c_0, c_j) \geq s(c_0, c_i)$, and $s(c_i, c_j) \geq 0$. Hence $j > i$ is not a best response at p' for an agent choosing i. Similarly,

$$
\begin{align}
& u_i(p' + e^{0i}) - u_j(p' + e^{0j}) \nonumber \\
& \quad = c_j - c_i + \lambda \left(\frac{l}{m} (s(c_0, c_j) - s(c_0, c_i)) + \frac{m-l-1}{m} s(c_i, c_j) \right) \tag{12} \\
& \quad > 0. \tag{13}
\end{align}
$$

Hence $j > i$ is not a best response at p' for an agent choosing 0.

Next, consider $j \leq i$. We have

$$\begin{aligned}
&u_0(p' + e^{i0}) - u_j(p' + e^{ij}) \\
&\quad = c_j - c_0 + \lambda \left(\frac{m-l-2}{m}(s(c_i, c_j) - s(c_i, c_0)) + \frac{l+1}{m} s(c_0, c_j) \right) && (14) \\
&\quad = u_0(p + e^{i0}) - u_j(p + e^{ij}) + \lambda\delta(s(c_i, c_0) - s(c_i, c_j) + s(c_0, c_j)), && (15)
\end{aligned}$$

where $\delta = 1/m$. Since 0 was the unique best response at p, we have $u_0(p + e^{i0}) > u_j(p + e^{ij})$. The fact that $s(c_0, c_j) \geq 0$ and $s(c_i, c_0) \geq s(c_i, c_j)$ implies that $s(c_i, c_0) - s(c_i, c_j) + s(c_0, c_j) \geq s(c_i, c_0) - s(c_i, c_j) \geq 0$. Hence $0 < j \leq i$ is not a best response for an agent choosing i.

Similarly, we have

$$\begin{aligned}
&u_0(p') - u_j(p' + e^{0j}) \\
&\quad = c_j - c_0 + \lambda \left(\frac{m-l-1}{m}(s(c_i, c_j) - s(c_i, c_0)) + \frac{l}{m} s(c_0, c_j) \right) && (16) \\
&\quad = u_0(p) - u_j(p + e^{0j}) + \lambda\delta(s(c_i, c_0) - s(c_i, c_j) + s(c_0, c_j)) && (17) \\
&\quad > 0. && (18)
\end{aligned}$$

Hence $0 < j \leq i$ is not a best response for an agent choosing 0.

Thus 0 is the unique best response at p' for all agents. It follows that 0 is a unique best response at $p^i + le^{i0}$, for $l = 0, 1, \ldots, m$. Hence 0 is chosen in every period, as required. □

Sudden collapse is consistent with the historical dynamics of duelling in the UK, where, having endured for centuries, the practice disappeared in a matter of decades in the mid-nineteenth century. Similarly, footbinding in China disappeared within the span of a single generation at the turn of the twentieth century.

If instead social influence strengthens relative to intrinsic cost, then the norm can gradually erode. To illustrate, consider an example with three agents and three actions, as shown in figure 2. (Each dot represents an agent. For the sake of concision, we omit the details of the calculations.) Initially, the costliest action is a norm. Then an exogenous shock makes social utility less important relative to intrinsic utility – that is, λ decreases – destabilising the norm. Because social influence strengthens, switching to the uncostly action incurs a relatively large penalty, so that agents switch to the intermediate action. Once enough agents have switched to the intermediate action, the uncostly action becomes a best response, and the process converges to the uncostly norm. These dynamics are

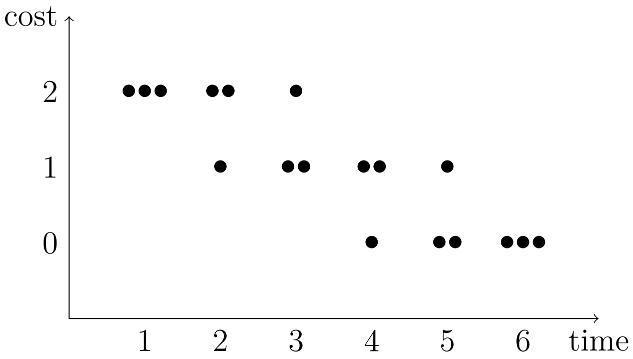


Figure 2: When social influence strengthens, gradual erosion can occur.
Notes: $m = 3$, $n = 2$, $c_i = i$, and $s(c, c') = (c' - c)^2$. $\lambda = 2$ initially but then decreases to $\lambda' = 1$.

consistent with duelling in France and FGC in Somalia.

The next result establishes that when social influence strengthens, the process will tend to erode gradually.

Proposition 2. *Suppose all deviations are disincentivised and social influence strengthens relative to intrinsic cost. Then, starting from any homogenous but unstable state, the process transitions to the next most costly norm.*

Proof. Let p^i be the current state for some unstable $i > 0$ and let $i^* \geq 0$ be the largest norm such that $i^* < i$.

Note that at p^i, any $j > i$ is not a best response. Moreover, this remains true as long as agents switch to actions below i. Hence no action above i will be chosen in any future period.

Suppose $i^* > 0$ and let $j < i$ and $k > i^*$. Since i^* is a norm, $u_j(p^{i^*} + e^{i^*j}) < u_{i^*}(p^{i^*})$. Note that

$$u_j(p^{i^*} + e^{i^*k} + e^{i^*j}) = u_j(p^{i^*} + e^{i^*j}) - \frac{\lambda}{m}(s(c_k, c_j) - s(c_{i^*}, c_j)) \text{ and} \tag{19}$$

$$u_{i^*}(p^{i^*} + e^{i^*k}) = u_{i^*}(p^{i^*}) - \frac{\lambda}{m} s(c_k, c_{i^*}). \tag{20}$$

Since social influence strengthens, we have

$$(c_k - c_j)s(c_k, c_{i^*}) < (c_k - c_{i^*})s(c_k, c_j) \text{ and} \tag{21}$$

$$(c_k - c_j)s(c_{i^*}, c_j) < (c_{i^*} - c_j)s(c_k, c_j). \tag{22}$$

Adding the two inequalities yields

$$s(c_k, c_{i^*}) + s(c_{i^*}, c_j) < s(c_k, c_j). \tag{23}$$

It follows that j is not a best response at $p^{i^*} + e^{i^* j}$. Arguing by induction, j is not a best response for any p such that $p_k = 0$ for each $k < i^*$. Hence no $j < i^*$ will be chosen in any future period.

Thus, since the only norm between i^* and i is i^* and the process must converge to a norm by theorem 1, the process converges to i^*. □

As shown in the proof of proposition 2, when social influence strengthens, it satisfies

$$s(c, c') + s(c', c'') < s(c, c'') \tag{24}$$

for every $c'' < c' < c$. We call this the *reverse triangle inequality* (Gulesci et al. 2024). The intermediate action c' is a good *social substitute* for both extreme actions, in the sense that the degree of social influence between c and c' and between c' and c'' is small compared to the degree of influence between c and c''. The reverse triangle inequality is a necessary condition for a gradual transition to occur, whereas proposition 2 shows that strengthening social influence is a sufficient condition for transitions to be gradual.

4 Incentivised costly deviations

We now turn to the case of conspicuous consumption (Veblen 1899), where an additional motive must be brought to bear, namely the desire to outdo others.[10] In this case $s(c, c') < 0$ whenever $c' > c$. This changes the dynamics of the model significantly. Whereas the previous case created a motive to coordinate, the present case is the opposite: it creates an incentive for agents to choose different actions from others. Now the process may not converge to a norm; in fact, it may cycle.

To take a concrete example, suppose one billionaire buys a super-yacht in order to boast about it to others. The others may now feel inadequate, and buy yachts in order to regain status. Indeed, in order to feel superior, they may buy bigger yachts. At the same time, they may not wish to buy too large a yacht: doing so might expose them to accusations of excess or suspicions of insecurity. The norm may in this way gradually escalate.[11] As yachts become more and more

[10] Conspicuous consumption is defined as 'expenditure on or consumption of luxuries on a lavish scale in an attempt to enhance one's prestige' (OED).

[11] In the past twenty years, the average length of a luxury yacht has crept up from 60 to 80 metres. According to a Silicon Valley CEO, until recently 'a fifty-metre boat was considered a good-sized boat. Now that would be a little bit embarrassing' (Osnos 2022).

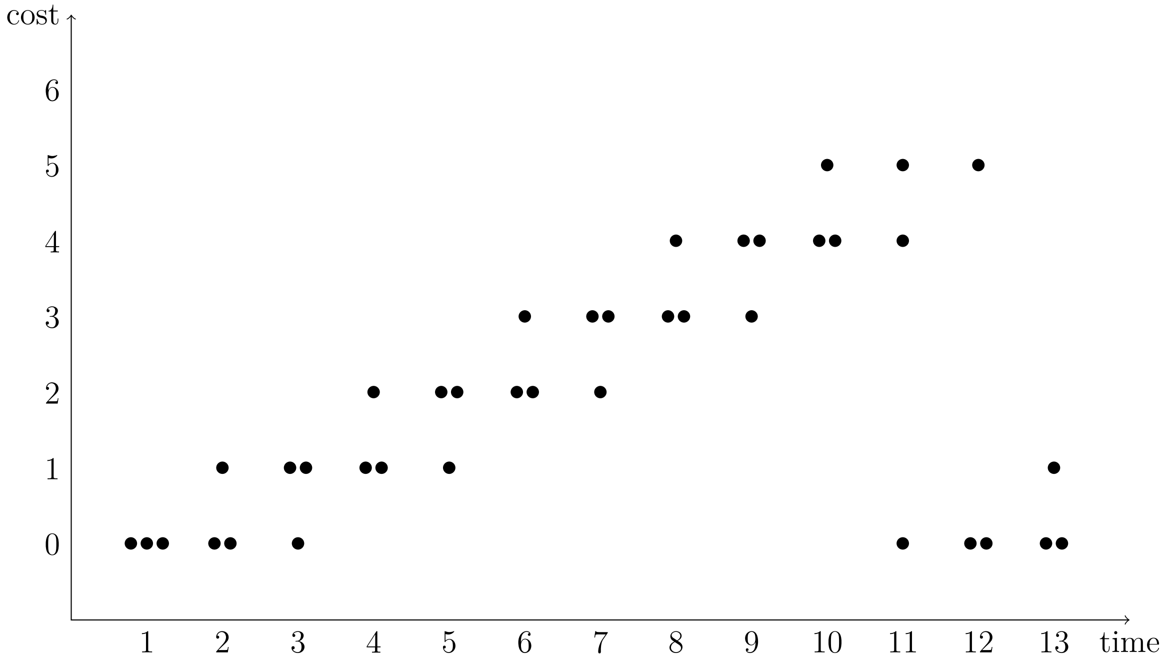


Figure 3: When costly deviations are incentivised, the process can cycle.
Notes: $m = 3$, $n = 6$, $c_i = i$, $\lambda = 1$, and $s(c, c') = 6$ if $c' < c$ and $s(c, c') = -2$ if $c' > c$.

extravagant, some may decide that it is no longer worth staying in the race. As more people give up their yachts, the use of yachts as a status marker may collapse entirely.

To illustrate the mechanics of the model, consider an example with three agents and seven actions, as shown in figure 3. Initially, all agents are at the uncostly action. Because costly deviations are rewarded, agents have an incentive to switch to action 1. Once the first agent does so, it exerts pressure on the remaining agents, creating an incentive for them to also switch. Once a second agent switches to 1, the third agent has an incentive to jump over the other two. The process continues in this fashion, with costs gradually escalating. Eventually, the cost of keeping up with the other agents becomes so high that an agent will prefer to switch all the way down to the uncostly action. Once this happens, the others will follow. In period 13, the process is at the same state it was at in period 2, and will continue to cycle from there on.

Thus the process displays gradual escalation followed by sudden collapse, which matches the qualitative dynamics of certain norms of conspicuous consumption. For instance, in Renaissance Europe, it was fashionable for wealthy women to wear platform shoes known as chopines (Bossan 2012; Riello and Rublack 2019). The height of the shoes symbolised the status of the wearer and allowed her to tower over others.[12] Over time, chopines gradually increased in height until they became almost unwearable: some were over fifty centimetres tall and wearers required a servant to help them walk. Higher chopines entailed a higher monetary cost, both because the shoes themselves were costlier and because the wearer would need longer dresses. Eventually, in the seventeenth century, chopines suddenly fell out of fashion and heels became popular instead.

Note that the example was chosen so that the cycling dynamics are relatively simple, and in particular that there is a unique path from any state in the cycle. In general, however, cycles may be significantly more complex and paths may not be unique.

5 Policy interventions

Harmful social norms are of great concern to governments and other institutions and are the targets of policy interventions. We now show how our model can be used to analyse such interventions. Under different assumptions, different policies can be more or less effective, or even have unintended consequences. This highlights the importance of understanding the underlying process and its dynamics when formulating policy.

From the Renaissance onwards, duelling was banned in most of Europe and North America and condemned by the Church. Luxury goods were sometimes taxed at a higher rate than other good and sometimes banned altogether under what are known as sumptuary laws. Such bans and taxes have the effect of changing the intrinsic costs of different actions. To illustrate how the model can be brought to bear on this, consider the example of duelling. Suppose there are three possible actions: duelling with pistols, duelling with swords, and not duelling. They are labelled 2, 1, and 0, respectively, so duelling with pistols is costlier than duelling with swords. Both types of duelling are illegal, but the sentence for duelling with swords is lesser. The current norm is to duel with swords, so the government is considering matching the sentences for duelling with pistols, hoping to eliminate the norm altogether. However when costly deviations are incentivised – because duelling is associated with bravery, say – this could backfire. To model

[12] In Shakespeare's *Hamlet*, the prince greets one of the players and comments that 'your ladyship is nearer to heaven than when I saw you last, by the altitude of a chopine' (2.2.427).

this, suppose $\lambda = 1$,

$$s(c, c') = \begin{cases} \sqrt{c - c'} & \text{if } c' \leq c \\ -\frac{\sqrt{c' - c}}{2} & \text{otherwise,} \end{cases} \tag{25}$$

and $c_2 = 4/5$. Suppose that $c_1 = 1/5$ initially, but that $c_1' = 3/5$ after the introduction of harsher sentences. Assume m is large. Then one can check that 1 is a norm before the introduction of harsher sentences, but that afterwards the process starting at 1 converges to 2. That is, the policy intended to eliminate a less dangerous form of duelling causes people to switch to a more dangerous form of duelling. Intuitively, this is because social pressure weakens relative to intrinsic cost, so that as c_1 increases the relative incentive to outdo others increases.

In the 1990s, religious leaders in Somalia proclaimed that the form of FGC known as Sunna, which was less severe than the traditional form known as Pharaonic, was compatible with religious obligations. One effect this intervention might have had is to reduce the degree of social pressure exerted against agents who chose Sunna. As discussed in detail by Gulesci et al. (2024), such a policy could lead to the eradication of the norm entirely via a *stepping-stone transition*, but it could also lead to the intermediate action becoming the new norm.

In the case of footbinding in China, anti-footbinding societies, whose members pledged not to bind their daughters' feet and not to marry their sons to women with bound fee, contributed to the demise of the norm (Mackie 1996). This raises two questions: First, how many people need to be convinced to abandon a norm before others follow and it dies out? Second, if there are intermediate forms of the norm, is it better to convince people to directly abandon the norm, or to first convince people to switch to an intermediate form before convincing them to abandon the norm altogether? To answer the first question, if i and j are norms, one can check that the number of players who have to switch from i to j in order for j to become a best response is

$$r_{ij} = \left\lceil \frac{\lambda(m-1)s(c_i, c_j) - m(c_i - c_j)}{\lambda(s(c_i, c_j) + s(c_j, c_i))} \right\rceil . \tag{26}$$

As one would expect, this number depends on c_i, c_j, and $s(c_i, c_j)$. It is decreasing in c_i and increasing in c_j and $s(c_i, c_j)$. It is also inversely related to $s(c_j, c_i)$. This is because as agents switch to j, they exert social influence on those remaining at i; a high $s(c_j, c_i)$ thus makes it easier to transition to j.

The answer to the second question depends on the model specification. For example, suppose there are three actions at the process starts at the costly norm

p^2. In order to eradicate the norm, the policymaker could either convince r_{20} agents to switch from 2 to 0, or first convince r_{21} agents to switch from 2 to 1, wait until the process converges to 1, and then convince r_{10} agents to switch from 2 to 1. Assume $c_i = i$, $\lambda = 1$, m is large, and

$$s(c, c') = \begin{cases} 2\sqrt{c - c'} & \text{if } c \geq c' \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Then one can verify that $r_{20} < r_{21} + r_{10}$, so a direct approach is preferable in this case. However, if instead

$$s(c, c') = \begin{cases} \frac{5}{4}(c - c')^2 & \text{if } c \geq c' \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

then $r_{20} > r_{21} + r_{10}$, so in this case an indirect approach may be more effective. Thus the direct approach is more effective when social influence weakens relative to intrinsic cost, while the indirect approach is more effective when social influence strengthens.

6 Conclusion

This paper has developed a model of costly social norms that accounts for the rich dynamics observed empirically. In the case where social influence disincentivises costly deviations, the model can give rise to collapse or gradual decline, depending on the shape of the social-influence function. In the opposite case, where social influence incentivises costly deviations, the model can give rise to cycles of gradual escalation followed by collapse. The model also sheds light on policy interventions. Understanding the structure of social preferences is crucial, since interventions that are effective in some cases might be ineffective or even counterproductive in others.

References

Akerlof, G. A. 1980. 'A theory of social custom, of which unemployment may be one consequence'. *Quarterly Journal of Economics* 94: 749–775.

———. 1997. 'Social distance and social decisions'. *Econometrica* 65: 1005–1027.

Akerlof, G. A., and R. E. Kranton. 2000. 'Economics and identity'. *Quarterly Journal of Economics* 115: 715–753.

———. 2010. *Identity economics: How our identities shape our work, wages, and well-being.* Princeton: Princeton University Press.

Akerlof, R. 2016. '"We thinking" and its consequences'. *American Economic Review* 106: 415–419.

———. 2017. 'Value formation: The role of esteem'. *Games and Economic Behavior* 102: 1–19.

Andreoni, J., N. Nikiforakis, and S. Siegenthaler. 2021. 'Predicting social tipping and norm change in controlled experiments'. *Proceedings of the National Academy of Sciences* 118.

Bagwell, L. S., and B. D. Bernheim. 1996. 'Veblen effects in a theory of conspicuous consumption'. *American Economic Review* 86: 349–373.

Banks, S. 2012. *Duels and duelling.* Oxford: Shire.

Belloc, M., and S. Bowles. 2013. 'The persistence of inferior cultural-institutional conventions'. *American Economic Review* 103: 93–98.

Bénabou, R., and J. Tirole. 2006. 'Incentives and prosocial behavior'. *American Economic Review* 96: 1652–1678.

Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms.* Cambridge: Cambridge University Press.

———. 2016. *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford: Oxford University Press.

Blume, L. E., and S. N. Durlauf. 2003. 'Equilibrium concepts for social interaction models'. *International Game Theory Review* 5: 193–209.

Blume, L. E. 1993. 'The statistical mechanics of strategic interaction'. *Games and Economic Behavior* 5: 387–424.

Bossan, M.-J. 2012. *The art of the shoe.* New York: Parkstone International.

Bowles, S. 2004. *Microeconomics: Behavior, institutions, and evolution.* Princeton: Princeton University Press.

Bowles, S., and H. Gintis. 2002. ‘Prosocial emotions’. Working paper, June.

Boyd, R., H. Gintis, and S. Bowles. 2010. ‘Coordinated punishment of defectors sustains cooperation and can proliferate when rare’. *Science* 328: 617–620.

Boyd, R., and P. J. Richerson. 2005. *The origin and evolution of cultures.* Oxford: Oxford University Press.

Brock, W. A., and S. N. Durlauf. 2001. ‘Discrete choice with social interactions’. *Review of Economic Studies* 68: 235–260.

Cao, Y., B. Enke, A. Falk, P. Giuliano, and N. Nunn. 2021. ‘Herding, warfare, and a culture of honor: Global evidence’. Working paper.

Centola, D., J. Becker, D. Brackbill, and A. Baronchelli. 2018. ‘Experimental evidence for tipping points in social convention’. *Science* 360: 1116–1119.

Durlauf, S. N. 1997. ‘Statistical mechanics approaches to socioeconomic behavior’. In *The economy as an evolving complex system II,* 81–104. CRC Press.

Edgerton, R. B. 1992. *Sick societies: Challenging the myth of primitive harmony.* New York: Free Press.

Efferson, C., S. Vogt, and E. Fehr. 2020. ‘The promise and the peril of using social influence to reverse harmful traditions’. *Nature Human Behaviour* 4: 55–68.

Enke, B. 2019. ‘Kinship, cooperation, and the evolution of moral systems’. *Quarterly Journal of Economics* 134: 953–1019.

Fehr, E., and S. Gächter. 2002. ‘Altruistic punishment in humans’. *Nature* 415: 137–140.

Fehr, E., and K. M. Schmidt. 1999. ‘A theory of fairness, competition, and cooperation’. *Quarterly Journal of Economics* 114: 817–868.

Friedman, D., and D. N. Ostrov. 2008. ‘Conspicuous consumption dynamics’. *Games and Economic Behavior* 64: 121–145.

Gates, H. 2015. *Footbinding and women’s labor in Sichuan.* Abingdon and New York: Routledge.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman. 1996. ‘Crime and social interactions’. *Quarterly Journal of Economics* 111: 507–548.

Glaeser, E. L., and J. A. Scheinkman. 2001. ‘Measuring social interactions’. Chap. 4, edited by H. P. Young and S. N. Durlauf. Cambridge, MA: MIT Press.

Goyal, S. 2012. *Connections: An introduction to the economics of networks.* Princeton: Princeton University Press.

———. 2023. *Networks: An economics approach.* Cambridge: MIT Press.

Granovetter, M. 1978. ‘Threshold models of collective behavior’. *American Journal of Sociology* 83: 1420–1443.

Gulesci, S., S. Jindani, E. La Ferrara, D. Smerdon, M. Sulaiman, and H. P. Young. 2024. ‘A stepping stone approach to norm transitions’. Working paper.

Henrich, J. 2020. *The WEIRDest people in the world.* New York: Farrar, Straus / Giroux.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. 2001. ‘In search of homo economicus: behavioral experiments in 15 small-scale societies’. *American Economic Review* 91: 73–78.

Hopkins, E. 2023. ‘Cardinal sins? Conspicuous consumption, cardinal status, and inequality’. Working paper, December.

Hopkins, E., and T. Kornienko. 2004. ‘Running to keep in the same place: Consumer choice as a game of status’. *American Economic Review* 94: 1085–1107.

Hopton, R. 2007. *Pistols at dawn: A history of duelling.* London: Piatkus.

Jackson, M. O. 2008. *Social and economic networks.* Princeton: Princeton University Press.

Kandori, M. 1992. ‘Social norms and community enforcement’. *Review of Economic Studies* 59: 63–80.

Kandori, M., G. J. Mailath, and R. Rob. 1993. ‘Learning, mutation, and long run equilibria in games’. *Econometrica* 61: 29–56.

Kandori, M., and R. Rob. 1998. ‘Bandwagon effects and long run technology choice’. *Games and Economic Behavior* 22: 30–60.

Ko, D. 2007. *Cinderella’s sisters: A revisionist history of footbinding.* Berkeley and Los Angeles: University of California Press.

Mackie, G. 1996. ‘Ending footbinding and infibulation: A convention account’. *American Sociological Review* 61: 999–1017.

Mathew, S. 2017. ‘How the second-order free rider problem is solved in a small-scale society’. *American Economic Review* 107: 578–581.

Monderer, D., and L. S. Shapley. 1996. ‘Potential games’. *Games and Economic Behavior* 14: 124–143.

Nye, R. A. 1993. *Masculinity and male codes of honor in modern France.* Oxford: Oxford University Press.

Osnos, E. 2022. ‘The floating world’. *New Yorker.*

Platteau, J.-P., G. Camilotti, and E. Auriol. 2018. ‘Eradicating women-hurting customs: What role for social engineering?’ In *Towards gender equity in development,* edited by S. Anderson, L. Beaman, and J.-P. Platteau, 319–356. Oxford: Oxford University Press.

Riello, G., and U. Rublack, eds. 2019. *The right to dress: sumptuary laws in a global perspective, c. 1200–1800.* Cambridge: Cambridge University Press.

Samuelson, L. 1997. *Evolutionary games and equilibrium selection.* Cambridge, MA: MIT Press.

Schelling, T. C. 1978. *Micromotives and macrobehavior.* New York: W. W. Norton.

Shell-Duncan, B., and Y. Hernlund, eds. 2000. *Female ‘circumcision’ in Africa: Culture, controversy, and change.* Boulder, CO: Lynne Rienner Publishers.

Shell-Duncan, B., K. Wander, Y. Hernlund, and A. Moreau. 2011. ‘Dynamics of change in the practice of female genital cutting in Senegambia: testing predictions of social convention theory’. *Social Science & Medicine* 73: 1275–1283.

Skyrms, B. 1996. *Evolution of the social contract.* Cambridge: Cambridge University Press.

———. 2003. *The stag hunt and the evolution of social structure.* Cambridge: Cambridge University Press.

Veblen, T. 1899. *The theory of the leisure class: An economic study of institutions.* Macmillan.

Vega-Redondo, F. 1996. *Evolution, games, and economic behaviour.* Oxford: Oxford University Press.

Weibull, J. W. 1995. *Evolutionary game theory.* Cambridge, MA: MIT Press.

Young, H. P. 1993. ‘The evolution of conventions’. *Econometrica* 61: 57–84.

———. 1998. *Individual strategy and social structure.* Princeton: Princeton University Press.

———. 2015. ‘The evolution of social norms’. *Annual Review of Economics* 7: 359–387.