# Peer Monitoring, Ostracism and the Internalization of Social Norms[☆]

Rohan Dutta[1], David K. Levine[2], Salvatore Modica[3]

**Abstract**

We study the consequences of endogenous social norms that overcome public goods problems by providing incentives through peer monitoring and ostracism. We examine incentives both for producers and for monitors. The theory has applications to organizational design - offering possible explanations for why police are rotated between precincts while professional organizations such as doctors are self-policing. It leads to a Lucas critique for experiments and "natural" experiments - a small level of intervention may be insufficient to produce changes in social norms while a high level of intervention may have a very different effect because it becomes desirable to change social norms. Finally we study the internalization of social norms - showing how on the one hand it makes it possible to overcome incentive problems that pure monitoring and punishment cannot, and on the other how it leads to an interesting set of trade-offs. We conclude with some discussion of cultural norms where norms are not established benevolently by a particular group for its benefit but established by others for their own benefit.

*Keywords:* one, two

[☆]First Version: October 14, 2017. We would like to thank Andrea Mattozzi. We gratefully acknowledge support from the EUI Research Council.

[*]Corresponding author David K. Levine, 1 Brooking Dr., St. Louis, MO, USA 63130

*Email addresses:* rohan.dutta@mcgill.ca (Rohan Dutta), david@dklevine.com (David K. Levine), salvatore.modica@unipa.it (Salvatore Modica)

[1]Department of Economics, McGill University

[2]Department of Economics, EUI and WUSTL

[3]Università di Palermo

## 1. Introduction

Empirically the importance of self-enforcing social norms in enabling groups to overcome public goods problems is well-established[4] and there is some theoretical work on the subject. In contrast to work in behavioral economics where pro-social behavior is seen as an intrinsic part of preferences, social norms are endogenous and adapt themselves to circumstances: this can be clearly seen in the cross-cultural experiments of Henrich et al (2001). For example, it is observed that Indonesian whale hunters have a problem that whales are large and infrequently caught. Hence incentives must be provided both for hunting and for sharing the catch. As a result a culture has developed in which status is determined by gift-giving. This is reflected in the laboratory: because a large gift grants large status to the giver, unlike other cultures, the Indonesian whale hunters choose in ultimatum bargaining games to give more than $1/2$ and to punish those who are overly generous.

Our model is one of individual behavior - Nash equilibrium with respect to selfish preferences - in which decisions are collective only in the sense that groups have the ability only to coordinate on a particular equilibrium that is mutually advantageous. In this theory we see pro-social behavior arising because there are penalties for anti-social behavior. We also consider the possibility that individuals find it personally advantageous to internalize social norms - resulting in apparently altruistic behavior despite the fact that individuals have no intrinsic preference for altruism.

Specifically we elaborate on the model of peer incentives introduced in Levine and Modica (2016) and used in Levine and Modica (2017) to study lobbying groups and in Levine and Mattozzi (2017) to study political parties. Here we focus on the issue of peer monitoring - an issue that also arises in the earlier work of Kandori, see Kandori (1992). We study a model of public goods production by group members who are monitored by other group members. We do so in an environment where monitoring is difficult in the sense that each producer is observed by at most one monitor. Both the producers and monitors face incentive problems: the producers because the group would like to induce them to take costly actions that provide a public benefit and the monitors because they have to choose whether to report accurately their noisy information about the actions they observe. After reports are received the group engages in individually valuable social activities and bad reports about a producer can be punished by ostracism. However: ostracism is costly for the punishers as well as punished. We assume that if there is substantial consensus over ostracism no individual can be decisive in preventing ostracism and the associated punishment costs. The same is not true for the monitor who by providing false reports can avoid bearing punishment costs.

Our basic hypothesis is that the group designs an incentive compatible mechanism for itself that is mutually beneficial for members. That is: we do not necessarily assume that social norms are left-over from some past meaningful equilibrium - we assume that groups can - at some cost - change social norms to reflect changed circumstances. In this direction we point to three pieces of evidence: the rapid change in social norms (measured in minutes) concerning the treatment of airplane hijackers that took place on September 11, 2001, the change in social norms (measured in

[4]Particularly see Olson (1965) and Ostrom (1990) among many others.

months) concerning public protest that took place in East Germany following the commitment by Gorbachev that military intervention in East Europe was off the table, and the rapid and organized change in social norms (following a debate that lasted over 12 years) that took place in Sweden when the change was made from left-side to right-side of road driving.[5] In these examples the incentives for change were large and in two of the cases took a substantial amount of time to come to fruition. The view we take is that there are fixed costs associated with introducing or changing social norms and we investigate both the nature of optimal social norms and the consequences of frictions for changes in social norms.

We study a number of issues. First, the cost of punishing the monitor depends on how socially close the monitor is to the producer. This results in a trade-off between the fact that close monitors face a greater incentive problem but are more likely to have accurate information. Hence organizations such as the police or military - or even corporations - may choose a strategy of rotating individuals between locations so that monitors are not likely to become close to producers. Similarly supervisor review may be used rather than peer review because supervisors are more socially distant than peers. Strategies of distant monitors make sense if the reduction in information from social distance is not too great. By contrast, for specialized professions such as doctors monitoring requires a great deal of expertise and close observation and so it may be worth paying the incentive price of social closeness in order to get accurate information - professional organizations, in other words, may be better off policing themselves as they generally do in practice. We comment also on how the optimality of social norms outside the laboratory may lead to the failure of procedures such as double-blind designed to reduce or eliminate possibility of outside influence.

Second, we study the implications of social norms for experimentation. One issue is the tradeoff between social benefit and the social cost of monitoring. Of particular interest is the impact of external incentives for public good provision - for example an outside agency subsidizes contributions to the public good. Naturally this tends to increase public goods output. However in the face of fixed costs for implementing non-trivial social norms external incentives can have a perverse effect as has been noted in the experimental literature: introducing a subsidy may reduce output. We note however: it always increases welfare. We also show that fixed costs can lead to a more general Lucas critique: interventions that are not sufficiently widespread - in the laboratory certainly, but even in "natural" experiments - may not lead to a change in social norms while a broader intervention may change social norms and so have different consequences. Hence interventions that appear effective in the "small" may not work "in the large" or vice versa.

Third, we examine the implications of the fact that social norms may be internalized. We model this by assuming that group members may specialize by investing in particular strategies. As a result of this costly investment, the chosen strategy provides a utility benefit when it is used. We assume moreover that investment in a publicly known social norm is less costly than inventing a personalized strategy. This has consequences somewhat similar to the notion of conformity or identity that has

[5]See the discussion in Levine (2012) for details of these three cases.

been studied by Benassy (1998) and Akerlof and Kranton (2000) although for rather more concrete reasons. We show that even a small a premium can have large effects: in particular social norms that are not feasible without internalization may be feasible with internalization and this may have disproportionate benefits. We also comment on the implications of internalization for behavioral economics and experimental work in the social sciences.

We conclude with some comments about social norms which are chosen by groups made up of individuals and cultural norms which are often imposed - for example - by parents on children.

We do not provide an extensive literature review in the introduction - rather we comment on the relevant literature in the context of our specific assumptions. Our basic model is a variation on the workhorse principal agent model: albeit one in which there is a monitor and costly punishments. The contribution of this paper is not in the fact that this variation is different from the many principal agent models that have been studied, but rather in the questions about social organization it enables us to address. We would highlight three things that we study that are new in the context of mechanism design:

1. In a social setting there is a tradeoff between the information available to the monitor and the malincentives of the monitor: close social ties with the producer on the one hand provide more accurate information, but on the other hand reduce the incentive of the monitor to tell the truth.

2. In a social setting there is a natural default: the group does not self-organize, there is no production of the public good, and there is no monitoring or incentives. We expect that there is some cost involved with agreeing to an alternative that involves incentives and ostracism and we are interested in the "stickiness" induced by this cost.

3. In a social setting it is natural that there is an advantage to conforming to and abiding by a widely used social norm. This induces an additional source of "stickiness" and creates tradeoffs in which internalization may either complement or substitute for incentives and ostracism.

## 2. The Base Model

We study a large group where monitoring is difficult in the sense that each production decision is observed by at most one other person. Specifically we study a setting where there is a continuum with a unit mass of pairs consisting of a producer and a monitor.[6] This means that there is a unit mass of producers and a unit mass of monitors, so that the population mass is two. The producer may use a unit of effort $e \in \{0,1\}$ to produce a public good at a cost of $ec$ where $c > 0$.[7] The value of the public good is such that if the fraction of pairs that produces ($e = 1$) is $\phi$ every individual in the group receives a benefit of $\phi V$. The monitor costlessly observes a noisy signal of the producer's choice: the monitor observes a signal $z \in \{0,1\}$ where with probability $\pi$ the signal is wrong $z \neq e$ and with probability $1 - \pi$ the signal is correct $z = e$ and $\pi < 1 - \pi$. Each monitor then makes a report $x \in \{0,1\}$. Following this a social interaction takes place. The population is rematched into

[6] For technical details on how a model like this works seeEllickson et al (1999).

[7] Production here is in the broad sense of pro-social behavior.

social subgroups of size $N \geq 4$ for a *social interaction* with the probability that a producer is in the same subgroup as her monitor (and vice versa) equal to $h$. Exactly one of the $N$ members of each subgroup - the presenter - is chosen with equal probability to learn an interesting story and may volunteer to share it with the subgroup. If the presenter chooses to share, the remaining $N-1$ members - the anonymous audience - observe the report by or about the presenter and a public randomizing device that enables the audience to coordinate their decisions concerning ostracism. Based on this, and this alone, they vote on whether the presentation will be allowed or whether the presenter should be ostracized. There is a number $1 < K < N-1$ such that the presenter is ostracized if and only if $K$ or more members of the audience vote against the presentation. If the presentation takes place it provides a value of $N$ to the presenter and $\beta N$ to each member of the audience where $0 \leq \beta \leq 1$ - so that the per-capita value of a presentation is $1 + \beta(N-1)$ in the social group.[8]The ostracism decision satisfies *audience anonymity* in that it is independent of the composition of the audience.

A *truthful strategy* is a choice of whether or not to produce as a producer, to send the message equal to the signal if a monitor, to always volunteer a story conditional on having one, and a rule for ostracizing the presenter as a function of the role (producer or monitor), the report and the public randomizing device. A *social norm* is a truthful strategy that if followed by everyone is a Nash equilibrium. Our assumption is that the group chooses the *optimal social norm* that maximizes the ex ante utility of any of the ex ante identical group members. We refer to this utility as the *social utility.*[9]

*Alternative Social Norms*

There are two types of social norm. One social norm - the *default norm* - is for no effort to be provided, all stories to be volunteered and to ostracize nobody. The social utility from this is the (per capita) utility from only the social interaction

$$U = 1 + \beta(N-1).$$

An alternative is to try to provide incentives for providing output, that is to try to construct a social norm that *implements production*, in which $e = 1$, the monitor tells the truth about the signal and all stories are volunteered. If it is possible to do this we say that *production can be implemented.* The only tool for implementation is ostracism. That is, by sacrificing some part of the social utility $U$ from the social interaction it may be possible to obtain $V$, the social utility from public good production. Notice that ostracism is costly for the monitor: if $h\beta > 0$ the monitor shares part of the cost of ostracizing the producer and so has incentive to let the producer off the hook. Hence ostracism must be used also to provide incentives for the monitor.

[8]If the presenter chooses not to share there is no story and everyone gets zero, but there is no reason for the presenter not to share since it can only increase her utility.

[9]Notice that by definition, social utility is a per capita value.

A key feature of the model is that any ostracism rule in which all members of the audience concur with each other is an equilibrium of that subgame: since $K < N-1$ no individual audience member can force the presentation to take place when everyone else votes against and since $K > 1$ no individual can block a presentation agreed to by everyone else. Hence potential social norms denoted by $s$ correspond to ostracism probabilities $p(x), q(x)$ for the producer and monitor as a function of the report $x \in \{0,1\}$. Notice that ostracizing one member of a pair imposes in expectation a cost of 1 on that person and a cost of $h\beta$ on the partner.

Since all pairs are identical, in a truthful equilibrium that implements production and where the presenter may be a producer or a monitor with the same probability, the per capita probability of ostracism is

$$\Pi(s) = (1/2)[\pi p(0) + (1-\pi)p(1)] + (1/2)[\pi q(0) + (1-\pi)q(1)].$$

The corresponding per capita expected social utility $W(s)$ is the per capita payoff from production $V$ minus the per capita cost of production which is half the producer cost of $c$ since only half of the population are producers, plus social utility from the social interaction minus the expected cost of ostracism: $W(s) = V - c/2 + (1 - \Pi(s))U$. In solving the problem of finding an optimal social norm it is useful to separate the problem of the optimal incentives for production from the problem of whether or not to produce at all. To this end note that $W(s) = U + V - [\Pi(s)U + c/2]$ and define the *cost of implementation* as $C(s) = \Pi(s)U + c/2$ - per capita production cost $c/2$ plus lost per capita utility $\Pi U$ which is the monitoring cost. If $s$ is to be an optimal social norm it must minimize implementation cost, and implementation will be optimal if and only if $V \geq \min_s C(s)$.

*Mechanism Design*

The model as presented is quite specific in the details of play. It is useful to take a more abstract approach and view the problem as one of mechanism design. That is, rather than the group coordinating on an equilibrium of the game that is optimal we may represent the group decision in the form of a principal who stands in for the group and focus on the problem of creating incentives for a representative producer/monitor pair. We reformulate the problem in terms of economic fundamentals: there is a producer who chooses $e \in \{0,1\}$ at a personal cost of $ce$ and utility of $(V - c/2)e$ to the principal, and a monitor who observes $z$ and reports $x$ to the principal. Based on the report the principal can choose to punish either or both of the two agents. Punishment of either one has a cost to the punished of 1, a cost to the partner of $h\beta$ and a cost to the principal of $U/2$. As this is a mechanism design problem, the designer - the principal - can precommit to punishment probabilities $p(x), q(x)$ as a function of the report of the monitor.

The game formulation is useful in that it is concrete and we see where the utilities of the producer, monitor and principal come from and how the ostracism decision can be decentralized in an incentive compatible way.[10] The mechanism design approach is useful in that it makes clear that

[10]The equivalence between a mechanism design problem and the choice of equilibrium in a game in which there is

many of the specific details of the game are not important: there are many other implementations of the mechanism design problem. It also enables us to see clearly which are the key feature of the economic environment: in particular which types of strategies are ruled out in the game formulation. Specifically audience anonymity means that a presenter cannot be ostracized based on who is in the audience. This has two key implications:

1. Collective punishment is ruled out. Because of the externality it is costly to be denied the opportunity to listen to a presentation. Consequently it would represent a punishment for members of the audience if the (innocent) presenter was ostracized. Notice that we do sometimes see collective punishment: for example the punishment of the family members of a terrorist. Here it is ruled out.
2. It is not possible to condition ostracism of a presenter based on whether the partner is in the audience. If it were possible would generally be desirable to do so: it would improve partner incentives if the presenter is punished only when the partner is not present. We do not know of examples where individuals are ostracized only when the person that ratted them out is not around.

Besides audience anonymity, implementations in which the ostracism decision is taken after the presenter is known but before the audience is known lead to the same principal agent problem.[11]

We see also that other details of the game are less relevant. For example: we have assumed that only one individual in a social interaction has a story to tell. It might be that everyone has a story to tell and there is limited time to tell stories. In this case ostracism can be accomplished by replacing the offender with a non-offender. The consequence of this is that the social cost of ostracism $U/2$ is reduced - this extreme case is equivalent to $\beta = 0$ and $U = 1$ - only the ostracized suffers. As long as each member of the group makes some social contribution that has value to other group members we would not expect this extreme case. In a similar vein there might be repeated social interactions and opportunities for ostracism. From the mechanism design problem we see that these details are not important: it is the cost to the ostracized, the partner of the ostracized and the social cost that matters. Note also that we have required symmetry: that ostracism be based on roles and not on names. But it is easy to see that because individuals are in fact symmetric there would be no gain to trying to punish individuals based on their names.

## 3. Cost Minimizing Social Norms

We first characterize cost minimizing social norms. Subsequently we turn our attention to the implications for optimal social norms.

---

voting over the mechanism is standard: see for example Fudenberg, Levine and Maskin (1994).

[11] In the implementation at hand meetings are indistinguishable, so must all have the same $p(x), q(x)$ There are implementations in which in addition to random assignments to meetings signals are sent to distinguish between meetings. For incentives, however, the only thing that matters is the expected $p(x), q(x)$. All meeting must satisfy the constraint $p(x), q(x) \leq 1$. If this constraint can be satisfied in expected value then it can be satisfied by assigning all meetings the expected value.

**Theorem 1.** *If and only if the* implementation condition

$$\frac{c}{(1-2\pi)(1-h^2\beta^2)} \leq 1$$

*is satisfied can production be implemented. In the cost minimizing social norm producers who are reported to have taken the bad action ($x = 0$) are ostracized with probability $p(0) = P$ and monitors who report the good action ($x = 1$) are ostracized with probability $q(1) = Q$ and there is no other ostracism. The ostracism probabilities are*

$$P = \frac{c}{(1-2\pi)(1-h^2\beta^2)}, \quad Q = h\beta P,$$

*and the cost of implementation is*

$$C = \left[\frac{U}{2}\frac{\pi + (1-\pi)h\beta}{(1-2\pi)(1-h^2\beta^2)} + \frac{1}{2}\right] c.$$

We prove this result at the end of the section: first we comment on what it means and some comparative statics.

- The implementation condition is crucial: if it is satisfied then any sufficiently valuable public good, that is, large enough $V$, will be produced, and if it is not no public good will be produced no matter how valuable.

- The size of the social subgroup $N$ is the only parameter that has no effect on the implementation condition. It effects only the social utility of interaction $U$.

- The implementation condition clearly implies that $c < 1$: that is, the cost of producing the good cannot be greater than the cost of being ostracized.

- If the implementation condition is satisfied then it follows that the implementation cost $C \leq N + 1/2$. That is, as we vary the parameters (other than $N$) so as to approach the boundary where the implementation condition fails implementation cost remains bounded and does not rise towards infinity.

- Fixing the other parameters if $h, \beta$ are sufficiently close to one the implementation condition fails. The reason for this is a feedback effect: a bigger punishment for the producer implies a bigger punishment for the monitor. The feedback effect is that the latter reduces the incentive for the producer to produce: by not producing she can reduce the probability the monitor is punished for sending a good report. A high degree of social interaction ($h$) combined with a strong externality ($\beta$) makes this feedback effect very strong and consequently implementation becomes impossible.

- The only case in which implementation cost is zero is if both $\pi = 0$ so that the producer is not punished due to erroneous signals and $h\beta = 0$ so that the monitor does not bear any of

the cost of punishing the producer. Notice that if $\pi = 0$ and $h\beta > 0$ we must punish the monitor for good reports even though that is the only kind submitted and they are known to be true: the reason is that the monitor must be willing to report the producer if the producer deviates from equilibrium.

- The only way to get the monitor to tell the truth is to make her indifferent between the two reports. There is no mechanism or social norm in which the monitor strictly prefers to tell the truth. We will return to this point subsequently.
- Malicious gossip is valued in the sense that a monitor is less likely to be ostracized for filing a bad report.
- The cost of implementation is proportional to $c$ the incentive to cheat on the social norm. This is a common result in peer monitoring models: for example Levine and Modica (2016) or Levine and Mattozzi (2017).

*Proof.* As ostracism is socially costly we should ostracize only as needed for incentive compatibility: in particular if the producer is reported to produce there should be no ostracism ($p(1) = 0$) and if the monitor files a bad report there should be no ostracism ($q(0) = 0$): a formal proof can be found in Lemma 1 in the Web Appendix. Hence the problem is reduced to choosing the probability of ostracizing the producer for a bad report $P = p(0)$ and the monitor for a good report $Q = q(1)$. For any individual the cost of being ostracized is 1 and the cost of having the partner ostracized is $h\beta$.

The incentive constraint for a monitor is simple. If it is to be optimal to tell the truth the monitor must be indifferent between reporting 0 and 1. If a good report ($x = 1$) is filed the monitor is excluded with probability $Q$. If a bad report ($x = 0$) is filed there is probability $P$ that the producer is ostracized. Hence the incentive constraint for a monitor is

$$Q = h\beta P.$$

When a producer chooses to contribute there is a direct cost $c$ but also an indirect cost due to monitoring. The monitor receives a bad signal with probability $\pi$ resulting in a probability $P$ that the producer is ostracized and a good signal with probability $1 - \pi$ resulting in a probability $Q$ that the monitor is ostracized. Hence the overall expected cost to the producer of contributing is $c + (\pi P + (1 - \pi)Qh\beta)$.

When a producer chooses not to contribute there is no longer a direct cost, but there still remain the indirect costs. This is $(1 - \pi)P + \pi Qh\beta$.

The incentive constraint for the producer is that the the cost of contributing must be no greater than the cost of not contributing: this can be written as

$$P - h\beta Q \geq \frac{c}{1 - 2\pi}.$$

Since punishment is costly this constraint must hold with equality.

If we plug in the incentive constraint for the monitor $Q = hP\beta$ into the incentive constraint for the producer we find the probability with which the producer must be excluded

$$P = \frac{1}{1-h^2\beta^2}\frac{c}{1-2\pi}.$$

From this we see that if the punishment is feasible $P \leq 1$ if and only if the condition given in the theorem holds. Since $Q = hP\beta$ and $h, \beta \leq 1$ it follows that if $P$ is feasible, so is $Q$.

Finally we compute the minimum implementation cost. We have

$$\begin{aligned}\Pi &= (1/2)\left[\pi P + (1-\pi)Q\right] = (1/2)\left[\pi + (1-\pi)h\beta\right]P \\ &= \frac{1}{2}\frac{1}{1-h^2\beta^2}\frac{\pi + (1-\pi)h\beta}{1-2\pi}c\end{aligned}$$

giving the desired expression. □

## 4. Tradeoffs

### *4.1. Rotation and Expertise*

It is important to recognize that while $\beta$ and $h$ enter into the social cost and implementability in a symmetric way (as a product) they measure rather different things. The factor $\beta$ reflects the strength of ties within the social network within which ostracism takes place: it reflects the density of the social network, the value of interactions in the network and the degree of friendship among network members. By contrast $h$ represents the extent of social interaction between monitors and producers as opposed to general network members. While $\beta$ is basically a fixed aspect of the network the interaction $h$ between monitors an producers can potentially be changed to reduce implementation costs.

The example of peer versus supervisor evaluation helps to fix thoughts. In the literature on personnel management a great deal of attention is on whether peer or supervisor evaluation is superior for rewarding and punishing employees. The two systems differ in $h$: generally speaking we expect that peers interact with each other and supervisors interact with each other, but interactions between the two groups is less common - in other words peer evaluation has high $h$ and supervisor evaluation has low $h$. In some instances peers and supervisors are actively discouraged from interacting: for example in the military officers clubs used to be common to encourage officers to socialize with one another but not with enlisted ranks. Supervisor evaluation, then, is superior for delivering low $h$ - and among the deficiencies noted with peer supervision is the problem of "friendship" - that is, high $\beta$ combined with high $h$. The problem is that low $h$ generally results in higher $\pi$: indeed there is data - see for example Kraut (1975)[12] that indicates that peer evaluation

[12]Most studies in this literature look only at the correlation between peer and supervisor rating or the within group correlation of rankings ("reliability"). Kraut (1975), by contrast, looks at peer and supervisor evaluations made at the

is substantially more accurate than supervisor evaluation.

Peer versus supervisor evaluation is only one of several means of weighting monitoring more in the direction of "outsiders" (low $h$) as against "insiders" (high $h$). In the police, for example, the system of rotation in which police officers are periodically moved between precincts to deliberately break social ties is often used to reduce corruption - meaning among other things, greater willingness of the police to monitor locals because of reduced social ties as well as greater willingness of police to monitor each other, also because of reduced social ties. Similarly, the use of outside management consultants is a method of avoiding the social ties that comes from evaluation done by insiders. In both cases a common complaint is that the effectiveness of monitoring is reduced as the monitors have less interaction with and knowledge of the producers.

We observe that the use of outsiders is quite different in different organizations. For example, as noted, in the police the use of outsiders for monitoring is relatively common and encouraged by supervisor evaluation and by rotation. By contrast in professions such as the medical profession monitoring is done almost entirely by insiders. Notice also that there is a substantial public goods aspect of production: lack of effort by doctors results in poorer patient outcomes, reduced demand for the services of doctors and less income for all doctors. Similarly with police corruption. Hence in both cases the insiders have incentives to self-organize to police themselves.

To study more clearly the trade-off between $\pi$ and $h$ let us assume a trade-off of the form $\pi = f(h)$ where $f$ is twice continuously differentiable with $f'(h) < 0$ and $f"(h) > 0$ so that increasing $h$ first raises monitoring accuracy $1 - \pi$ substantially, but eventually less so. The next theorem is proven in the Web Appendix.

**Theorem 2.** *Let $C(h, \pi)$ denote the least cost of implementation if the implementation condition is satisfied and $\infty$ otherwise. If there exists a $h, f(h)$ such that the implementation condition is satisfied then there is a unique minimum of $C(h, \pi)$ subject to $\pi = f(h)$ and the optimum satisfies*

*1. $h$ is decreasing, $\pi$ increasing in $\beta$*

*2. if $h_f, \pi_f$ are the solutions of the cost minimization problem and $\overline{f}$ satisfies $\pi_f = \overline{f}(h_f)$ and greater signal sensitivity than $f$ in the sense that $|\overline{f}'(h)| > |f'(h)|$ then $h_{\overline{f}} > h_f$ and $\pi_{\overline{f}} < \pi_f$.*

As an application we consider two occupations: surgeon and police officer. Surgeons require a high level of specialized knowledge - more than a decade of specialized training[13] - while police officers require less than a year.[14] We interpret that to mean that the sensitivity of $f$ to $h$ is much greater for surgeons than for police officers - outsiders are unlikely to have the specialized knowledge needed to evaluate "surgical output" while it is not so difficult for an outsider to evaluate "police output." The social network of surgeons is also sparser than that of police officers in the sense that there are many fewer of them: some 50,000 surgeons in the US[15] versus some 750,000 sworn police

end of a four week training course and shows that peer evaluation is a far better predictor of subsequent promotions.

[13] https://study.com/articles/Surgeon_Career_Summary_and_Required_Education.html

[14] http://work.chron.com/long-train-cop-21366.html

[15] https://www.statista.com/statistics/209424/us-number-of-active-physicians-by-specialty-area/; there is some

officers.[16] We take this as an indication that good friends of police officers are more likely to be among other police officers than good friends of surgeons among other surgeons: that is that $\beta$ is likely to be lower for surgeons than police officers. Hence the theorem tells us that we should see higher $h$ for surgeons than for police officers. Indeed the evidence suggests this is the case: as noted the police use supervisor evaluation and rotation to achieve a relatively low $h$ while surgeons are self-policing.[17]

*4.2. Alternative Monitoring Technologies*

For simplicity we have considered a benchmark monitoring technology where there there is exactly one monitor per producer. The broad scope of different types of monitoring is beyond the scope of this paper and has been discussed elsewhere in the literature on mechanism design (Cremer and McLean (1988), Rahman (2012)) but we want to briefly indicate some of the issues that arise and the implications for the simple model.

An alternative to the one-for-one model is to assume that there is a fraction of monitors randomly assigned to a fraction of producers. Hence a producer may have no monitors, one monitor, or many monitors, randomly determined. In the monitoring technology it is necessary to specify who knows what about monitoring and when: does the producer know how many monitors there are when making the production decision? Are the monitors aware there are other monitors when making their report? Are the signals they observe of producer behavior correlated, and if so how? When a monitor is a presenter does the audience know they are facing a monitor or can the monitor lie and pretend not to have witnessed anything? From this it should be clear that there is a broad array of models. There are two general lessons that can be drawn from the mechanism design literature. First: secrecy about the presence of monitors is valuable - incentive compatibility is easier if agents are not aware of each other. For example, if a producer knows there are no monitors then the producer cannot be given incentive to produce - see Rahman (2012) for some additional discussion of secrecy. By contrast it is useful if the "audience" - the mechanism designer - knows who the monitors are - else additional incentive constraints must be introduced to induce monitors to reveal themselves. Second: multiple monitors alleviates the incentive problem for monitors because it is possible to compare the reports of different monitors. In order to make effective use of this, however, it must be that the signals observed should be correlated - see Cremer and McLean (1988) for some additional discussion.

For illustrative purposes, consider two extreme examples: very few monitors so that the number of monitors can as a good approximation be taken to be either zero or one, with the producer unaware of whether a monitor is present, and very many monitors all of whom observe exactly the

work on social networks of doctors, see for example, West et al (1999), but for a very limited set of doctors and there does not appear to be comparable data for police officers.

[16] https://en.wikipedia.org/wiki/Law_enforcement_in_the_United_States

[17] The form of ostracism in the case of surgeons appears to be primarily in referring patients to other surgeons: see Kinchen et al (2004) and Sarsons (2017) who document that perceived medical skill is the most important factor in surgeon referrals and that bad surgical events lead to reduced referrals.

same signal: our benchmark case lies between these two extremes. In both cases we assume that the audience is aware of whether a monitor is present. In the former case let $\eta$ be the probability that a monitor is present to witness a production decision. The only effect this has on the model is to change the incentive constraint for the producer which now becomes

$$\eta P = \frac{1}{1-h^2\beta^2}\frac{c}{1-2\pi},$$

since there is now only a chance $\eta$ of being monitored and consequently facing ostracism. The implementability condition is consequently harder to satisfy since larger ostracism probabilities are needed for incentive compatibility - but if the implementability condition is satisfied the implementation cost does not change: larger punishments are used, but less frequently so that their cost remains the same.

Now consider the case of a number of monitors who observe exactly the same signal. We can now ostracize all monitors with probability one for disagreement: if all tell the truth, then all will strictly prefer to tell the truth. Since on the equilibrium path all tell the truth, all agree, and so in equilibrium there is no punishment of monitors. In other words with multiple monitors observing the same signal there is no incentive problem for monitors. This is equivalent to $h = 0$.

It is useful to point out one case where the benchmark model fits extremely well: that of spousal abuse. Here the abuser is the "producer" and output means "lack of abuse" while the group is society as a whole. The monitor - the abused spouse - interacts frequently with the abuser (high $h$) and the externality is strong (high $\beta$). Hence we should expect as we see in practice that abused spouses are reluctant to report abuse, reluctant to file charges, reluctant to testify in court, and indeed reluctant to admit to anyone that abuse has taken place.[18]

We now consider two applications of these ideas.

*Poverty and Public Goods.* Consider an urban slum versus a rural village in a developing nation. In the former case $\pi$ is large since nobody knows anybody else - the implication is that - regardless of $h$ - the implementability condition is not likely to be satisfied and we are unlikely to see the production of public goods in urban slums - indeed we see are piles of trash.

By contrast in a rural village $\pi$ is likely to be quite small since everyone knows everyone else. However, as we have noted, there are two types of transactions: those that are likely to be seen by many people and those that are unlikely to be seen except perhaps by one. As we have observed, in the former case $h$ is effectively zero: all the matters is that $\pi$ is small meaning the implementability condition is likely to be satisfied and indeed the cost of implementation to be low. Hence we are likely to see public goods production where it is easy to see if people contribute. This is consistent

[18]Note that in this example the "benefit" $V$ of not being abused may accrue primarily to the monitor. This is different than the implementation in the base game, but it is easy to see that the incentives are those of the mechanism design problem. Producer incentives are the same because the producer does not get $V$ and monitor incentives are the same because the production decision is taken before the reporting decision is made. We have ruled out the possibility of the monitor committing to a reporting rule: if this were possible in this example the monitor would wish to do so.

with the evidence - for example from Ostrom - about public goods such as water projects which are carried out and enforced through ostracism. By contrast where monitoring is difficult $h$ and $\beta$ are likely to be quite large, so the implementability condition is likely to fail despite the fact that $\pi$ is very low. Overall we expect rural villages to be like urban slums for public goods where production is hard to observe: in neither case would we expect to see any. A case in point may be the cheating of outsiders or tourists - one-on-one transactions with outsiders in a shop, hotel, or restaurant are difficult to observe. There is a public good element: cheating strangers gives the village a bad reputation so that outsiders are unlikely to come in the first place. Notice that modern technology has made it easier to monitor one-on-one transactions for hotels and restaurants - on line review services such as Trip Advisor not only allow tourists to avoid places they are likely to be cheated, but allow villagers to observe that a particular individual is engaging in cheating. Hence we predict that better online information should lead to social norms that discourage the cheating of outsiders - for example this should be the case in hotels and restaurants for which online reviews are readily available, but not for shops (jewelry, souvenirs, clothing, art).

*Double-Blind in the Laboratory.* In studies particularly of social preferences such as the dictator game (see for example, Tisserand et al (2015)) it is believed that participants behave altruistically to make a good impression on others - in particular the experimenter. In an effort to eliminate this a double-blind treatment is often used in which neither the other participants nor the experimenters can tell who did and did not donate money.

We wish to propose a rather different interpretation of behavior and motives. Specifically: we believe that what participants are "worried" about is violating a social norm from outside the laboratory (be "generous") and getting caught. They are assured that their behavior in the laboratory will not be "leaked" to the outside world ("what happens in Vegas stays in Vegas"). In the literature it is generally assumed that these representations are believed. We do not believe these representations are true, nor do we believe that participants believe this is the case. We have two reasons for this doubt:

1. Mistakes happen. If hackers can obtain confidential and damaging emails from Yahoo, what are the chances the experimental records are so secure that they will never leak to the outside world?

2. Even if identities are protected - for example through double-blind - there is a long history of deception in experiments by psychologists who have systematically lied to their subjects. What, for example, is to keep a deceptive experimenter from using a secret camera to record supposedly confidential placement of money into an envelope?

In order to model the possibility of the leaking of confidential information we use as above a model in which there is only a chance $\eta$ of being monitored (the probability of a leak) and assume that $h = 0$ since monitor incentives are not relevant when there is a public release of information. We argue that while through instructions, design, and reputation, the perceived value of $\eta$ (the one that matters) may be made small, it is unlikely to be made zero. Subjects - rightly - have some concern that if they behave selfishly in the laboratory word of this will get back to their friends

outside the laboratory and they will then have an unfortunate reputation for behaving badly when they think nobody is looking.

How does the theory figure into this? Consider an effort to reduce perceived $\eta$ through instructions, design and the like. The theory says that a reduction in $\eta$ that is not sufficiently great will simply raise the probability of ostracism but have no effect on behavior. By contrast if $\eta$ is reduced enough the implementability condition will fail and social norm will call for selfishness - that is, if $\eta$ is small enough that punishment is inadequate to deter deviation from the social norm then the social norm awards no punishment at all. This means that we should observe that modest efforts to reduce $\eta$ should have little or no effect, while strenuous effects to reduce $\eta$ may have a substantial effect. We offer this as a possible explanation for the following stylized fact about dictator experiments (see for example, Tisserand et al (2015)): many studies have found that double-blind has no effect, while a few find that it substantially reduces giving. We suggest that the difference between these studies lies in the extent to which $\eta$ was successfully reduced: where the effort was modest we would expect no effect, but where the effort was strenuous we would expect an effect. That is: it is not "double-blind" versus "single-blind" that matters - it is how persuasive the double-blind is that matters.

*Implications for Measurement.* Two of the key parameters are $h$ and $\beta$. For implementability they appear symmetrically as a product $h\beta$. The per capita benefit of social interaction $U$ depends on $\beta$ but not on $h$. Be this as it may - the two parameters measure rather different things.

The parameter $h$ represents the frequency of social interaction between monitor and producer: in our discussion of surgeons versus police we indicated how data on network density can give an idea of this frequency. The is a substantial literature on measurement of network ties - see for example Jackson (2010) - that is also relevant for measuring this frequency.

By contrast $\beta$ measures the value of social interaction. Here again measurement is possible: the value of a social network as Facebook, whose market cap of 521 billion US dollars,[19] gives an indication of the monetary value that individuals place on social interaction. Worldwide annual revenue from telecommunications services is a bit over a trillion US dollars[20]. Although this includes business interactions the bulk of the spending appears to be on social interactions. To put these numbers in some kind of perspective worldwide film and movie revenue[21] - commercial entertainment as opposed to social interaction - is about 286 billion US dollars per year. We also have more direct data: for example, that certain types of people spend five hours per day on the phone.[22]

[19] January 12, 2017.
[20] https://www.statista.com/statistics/268628/worldwide-revenue-from-telecommunications-services-since-2005/
[21] https://www.statista.com/topics/964/film/
[22] Andrews et al (2015)

## 5. Sticky Social Norms, Experimental Economics, and a Lucas Critique

So far we have studied what can be described as frictionless social norms: social norms are adopted to maximize social utility. In practice it is neither instantaneous nor costless for large groups to discuss and agree on social norms, and there is always the option of simply settling on the default equilibrium - agreeing to nothing and "letting nature take its course", each individual following their own personal interest without monitoring and ostracism. The existence of frictions has implications for experimental and empirical economics. In particular, interventions - changes in incentives for producing public goods - will have a different effect depending on whether or not they are sufficient to overcome the "friction" of changing social norms. This can lead to perverse consequences where incentives designed to encourage public goods production instead reduce it because they displace peer monitoring and ostracism. Moreover, small scale interventions - either in the laboratory, in the field, or as measured in a natural experiment - may be insufficient to change social norms and so may provide misleading guidance about large scale interventions which are sufficient to change social norms. This latter is a kind of Lucas critique: when social norms are endogenous data generated with fixed social norms does not tell us about what happens when social norms change in response to policy.

### *5.1. Cost Versus Benefit with Subsidies*

We consider a setting where different levels of public goods production are possible and analyze the effect of a well-informed outside party that attempts to provides incentives for public good production. We first study the frictionless case, then introduce a friction in the form of a fixed cost for agreeing to a non-trivial social norm.

So far we have considered a simple choice: produce or do not produce. Now we extend the analysis to a choice of production level or quality $0 \leq \theta \leq \Theta$. We suppose that the cost of producing at the level $\theta$ is equal to $\theta$[23] and that the value of public good produced is $V\theta$. Given a social norm of $\theta$ the individual producer may choose to produce $e\theta$ where $e \geq 0$; if $e = 1$ production is $\theta$ and the norm is followed, otherwise it is not. We continue to assume a simple signalling technology with just two signals $z \in \{0, 1\}$ which we think of as meaning "bad, the social norm was not followed" and "good, the social norm was followed." Specifically if $e = 1$, that is, the producer follows the social norm, then with probability $1 - \pi$ the signal is 1 and with probability $\pi$ the signal is 0. If $e \neq 1$ then with probability $1 - \pi$ the signal is 0 and with probability $\pi$ the signal is 1. With this structure it is clear that if the producer chooses not to follow the social norm the optimal deviation is to produce 0 since the chances of being punished are the same for any deviation. A more refined signalling technology must include at least this incentive constraint, but might have additional incentive constraints - for example, if small deviations are less likely to be detected than large deviations.

[23] For simplicity we assume through this section that $\Theta$ is sufficiently small that the implementability condition is satisfied at $\Theta$.

The analysis of the simple model is straightforward. As we are going to hold fixed the monitoring and ostracism technology, it is convenient at this point to define the *cost coefficient of public good production* (as found in Theorem 1)

$$\mu = \frac{U}{2}\frac{\pi + (1-\pi)h\beta}{(1-2\pi)(1-h^2\beta^2)} + \frac{1}{2}$$

and observe that the group is maximizing $V\theta - \mu\theta^2$;[24] the optimal social norm is given by $\theta = V/(2\mu)$ with corresponding social utility $V^2/(4\mu)$.[25] This is strictly positive so it is always better to implement production rather than use the default social norm. Other than the obvious result that $\theta$ is strictly decreasing in $\mu$ there is not much to see here.

We now introduce direct incentives for public goods production. Specifically, we suppose that the cost of production is defrayed by a subsidy $r\theta$ that is taken from the value of the public good $V\theta$. Notice that each producer produces $V$ both for herself and for her monitor, so that it is possible to provide a producer with a subsidy of up to $2V$. Hence $0 \leq r \leq 2V$.

In this example the outsiders have better information than the group as they directly observe $\theta$ while group members see only a noisy signal.[26] Now the optimal deviation for an individual is no longer to produce 0, rather it is to maximize utility net of punishment $r\theta - \theta^2$, that is, to produce $r/2$ and receive a utility net of punishment of $r^2/4$.

Notice we have used the cost $c$ to measure two different things: one is the direct cost of production, which here is $d = \theta^2$. Previously we also used it to represent the utility difference for the producer (net of punishment) between following the social norm and deviating to the best alternative. Up until now that has simply been the savings in production cost, so $c = d = \theta^2$. This is no longer the case: rather if the producer produces the social norm of $\theta$ the individual utility net of punishment is $r\theta - \theta^2$,while the utility from the best alternative is $r^2/4$ so that now $c = r^2/4 - (r\theta - \theta^2)$. It is this quantity that figures in the producer incentive constraint

$$P = \frac{c}{(1-2\pi)(1-h^2\beta^2)}.$$

Hence the (per capita) social utility from implementing production norm $\theta$ is

$$\begin{aligned} U + V\theta - d/2 - (\mu - 1/2)c &= U + V\theta - \theta^2/2 - (\mu - 1/2)(r^2/4 - (r\theta - \theta^2)) \\ &= U + (V + (\mu - 1/2)r)\theta - \mu\theta^2 - (\mu - 1/2)r^2/4. \end{aligned}$$

[27] In the web appendix we prove

[24] The value of following the norm $\theta$ is $U + V\theta - \mu\theta^2$, we ignore the constant in the maximum problem.

[25] We assume that $\Theta > V/(2\mu)$ otherwise there is a boundary solution.

[26] For example, the IRS generally knows more about income and tax payments than friends and neighbors.

[27] In the (LHS) expression for the cost minimizing (per capita) social utility, $U$ is as before the per capita social utility without any ostracism or production, $V\theta$ is the per capita value of the public good, $d/2$ is the expected per

**Theorem 3.** *It is always optimal to implement production, at the level*[28]

$$\theta = (V + (\mu - 1/2)\, r)\, /(2\mu)$$

*which as expected is increasing in $r$. The social utility advantage of implementing least cost production over the default equilibrium where $r/2$ is produced is*

$$G(r) \equiv (V + (\mu - 1/2)r)^2/(4\mu) + (1-\mu)r^2/4 - Vr/2$$

*which satisfies $G(0) = V^2/(4\mu)$, $G(2V) = 0$ and $G'(r) < 0$ for all $r < 2V$.*

*Fixed Cost.* There is an intrinsic asymmetry between the default social norm and implementing production. The former is completely decentralized: the group need not do any organization, members are simply left on their own to optimize. By contrast implementing production requires agreement over a production level and enforcement scheme. It is natural to think that this sort of central organization has a fixed cost $F > 0$ associated with it.[29] If this is the case then implementation of production will take place only if the gain in utility from implementation over the default social norm exceeds the fixed cost. From Theorem 3 we can see that this leads to a possible non-monotonicity with respect to the subsidy level $r$:

**Corollary 1.** *Let $\overline{r}$ be the positive solution of $G(r) = F$ if one exists, 0 otherwise. Note that for $F$ sufficiently small a positive solution always exists and that $\overline{r} < 2V$. Then for $r < \overline{r}$ it is optimal to implement production and output is $\theta = (V + (\mu - 1/2)\, r)\, /(2\mu)$ while for $r > \overline{r}$ the default social norm is optimal and output is $r/2 < (V + (\mu - 1/2)\, r)\, /(2\mu)$. Hence for $\overline{r} = 0$ output increases in $r$ while for $\overline{r} > 0$ output increases in $r$ up to $\overline{r}$, drops discontinuously, and then increases again. In either case social utility is always increasing in $r$.*

We note that the discontinuity can be so severe that output just above $\overline{r}$ may be less than at $r = 0$. To see this, at $r = 0$ output is $V/(2\mu)$ while just above $\overline{r}$ it is roughly $\overline{r}/2$. Hence if $\overline{r} < V/\mu$ output at $r = 0$ is greater. Since $G(0) = V^2/(4\mu)$ as $F$ approaches this value from below we see that $\overline{r} \to 0$ so must eventually fall below $V/\mu$. On the other hand at $r = 2V$ production $r/2$ is higher than at $r = 0$ (since $V > V/2\mu$ follows from $\mu > 1/2$).

Concerning welfare: we see the not surprising fact that it is always desirable to increase the level of subsidy in order to reduce costly ostracism - despite the fact it may lower output it always increases social utility. Indeed: we see that if possible the externality should be entirely internalized by setting $r = 2V$ in which case production is $r/2$ and the first best is obtained.

---

capita direct cost of production since only half the population are producers and $(\mu - 1/2)c$ is the expected per capita cost due to ostracism.

[28] Again assumed to be smaller than $\Theta$.

[29] The fixed cost might well depend on the size of the group: for example Levine and Modica (2017) assume it is proportional to group size. Here we are keeping the size of the population fixed and normalized to 2 - although of course in practice it depends on whether we are talking about 100% of the population of New Jersey or the United States.

*Incentives and Experiments.* Gneezy and Rustichini A (2000) show that indeed introducing modest incentives can lead to the discouragement of the activity it is designed to promote: they show that introducing modest fine for being late to pick up children at a day-care resulted in more parents picking up their children late. This is consistent with the theory here (the case $F$ slightly smaller than $G(0)$ discussed above), the public good being in this case is the welfare of the school: that prior to the fine lateness was punished by a social norm among parents; and that with the incentive provided by (avoidance of) the fine it was no longer worth implementing a non-trivial social norm and consequently lateness increased.[30]

There is a broader literature documenting that introducing incentives can reduce effort: for example, Gneezy and Rustichini B (2000) show in the laboratory that introducing small incentives in the laboratory can lead to a collapse in production (while larger incentives do not). Our fixed-cost theory cannot account for what happened in the laboratory: that is, in a laboratory setting where participants come for a few hours at most and do not interact with each other in any substantial way, we cannot argue that introducing incentives somehow caused the collapse of a social norm from outside the laboratory.

By contrast there is a reasonably well regarded "behavioral" theory that does explain what happened inside the laboratory (it is by no means inconsistent with the theory of endogenous social norms): it is that while participants want to adhere to social norms from outside the laboratory their actions corresponding to the social norm may not have the same meaning inside the laboratory. A clear and careful model of this can be found in Benabou and Tirole (2006): the principal (the experimenter) by means of the incentives provided affects what effort signals about the agent's identity. In the absence of incentives participants find the signaling value of the task sufficiently high (they seem prosocial) and provide effort accordingly. When small incentives are introduced this signaling value is diminished as greater effort may now be attributed to greed. The participants then conclude the task is not worth much and so provide little effort.

The "behavioral theory" can also explain what happened at the day-care center, and so perhaps being able to kill two ducks with one stone we should prefer this explanation over the collapse of social norms explanation. However, the behavioral explanation to us seems less plausible in the day-care setting. If a fine is introduced for lateness should parents take this as an indication that really lateness is not so bad and they should therefore be late more often? Or should parents take it as an indication that really lateness is bad, that too many parents are being late, and that therefore they should make an effort to be late less often? Moreover - while the fines were introduced (and removed) in the most mysterious way possible - did not parents talk to one another and to teachers about the issue of lateness both before and after fines were introduced? Did not parents have a relatively good idea how inconvenient it was for the school when they were late before fines were

[30]Our theory does not explain why when the fine was removed parents continued being late - however, the data after the fine was removed is very short in duration so we cannot say whether in a few more weeks or months lateness began to drop. In general we expect the frictions (and time) to agree to a non-trivial social norm to be greater than that needed to revert to the default.

introduced?

While it is tempting to conclude that what happened at the day-care center is the same as what happened in the laboratory and to prefer a unified explanation - it appears to us that the behavioral explanation accounts well for what happened in the laboratory but poorly for what happened at the day-care while the collapse of social norms theory accounts well for what happened at the day-care but poorly for what happened in the laboratory. Regardless, the theories differ in one key implication. In the information revelation story no matter how small the incentive introduced production will drop. In the fixed cost of social norms theory production will drop only if the incentives introduced are large enough (but not too large). So according to the fixed cost theory if we were to redo the day-care experiment with fines set low enough to provide little incentive for good behavior then we should not see an increase in lateness. We note that these issues have continued importance - for example in the use of incentives (or non-use) for promoting blood donations - see, for example, Meyer and Tripodi (2017).

*5.2. Intervention in the Large and in the Small*

Previously we looked at an intervention where the interests of the intervenor were aligned with those of the group. We wish now to look at an intervention where the interests of the intervenor are opposed to that of the group: an intervention by outsiders for whom the public good is a public bad. For example, the group is attempting to collude for monopoly profit of political power that is to the disadvantage of everyone else and the "everyone else" perhaps through legal or state action intervene to provide incentives.

To keep things simple we return to the simple case where there are just two levels of output - that is, $e \in \{0, 1\}$ and $\theta = 1$. The intervention takes place with probability $\lambda$ and consists of a binding offer to the producer of paying $R$ for the action 0; we now assume that the cost of this payment is borne entirely by the outsiders.

In this setting we can distinguish between *discriminatory* and *non-discriminatory* social norms. A discriminatory social norm conditions the ostracism probabilities on whether or not an intervention took place; a non-discriminatory social norm is a one-size fits all, the same rules for ostracism apply regardless of whether or not there was an intervention. The producer always observes whether the intervention was made and and accordingly chooses whether to produce or not. If intervention occurrence were his private information then only non-discriminatory norms are feasible; but we do not assume this is the case. Rather we assume that there is some - possibly small - fixed cost $f > 0$ to the group of implementing a discriminatory norm. To take an example: it strikes us as unlikely that, although the cost would be small, it would be worthwhile to form a broad social consensus to the specify how individuals should behave in the - highly unusual - circumstance that they should be in an experimental laboratory confronted with a particular social science experiment. The point is that while the cost might be small, the probability of the particular intervention might be small as well. This idea is akin to that found in the incomplete contracting literature such as Hart and Moore (1988). In the web appendix it is shown that

**Theorem 4.** *Suppose $R$ is small and that $V$ is large in the sense that*

$$R \leq (1-h^2\beta^2)(1-2\pi) - c$$

*and*

$$V > \left[\frac{U}{2}\frac{\pi+(1-\pi)h\beta}{(1-h^2\beta^2)(1-2\pi)} + \frac{1}{2}\right](c+R).$$

*Then there is $0 < \underline{\lambda} \leq \overline{\lambda} < 1$ such that:*

*For $\lambda \in [0, \underline{\lambda})$ the optimal social norm is non-discriminatory and production takes place only when there is no intervention;*

*For $\lambda \in (\underline{\lambda}, \overline{\lambda})$ the optimal social norm is discriminatory and production always takes place;*

*For $\lambda \in (\overline{\lambda}, 1]$ the optimal social norm is non-discriminatory and production always takes place.*

Some intuition for this result is the following. With a non-discriminatory norm in which production is "blown off" when there is intervention it is unfortunately the case that the monitoring and ostracism costs needed to incentivize production the remainder of the time must be paid anyway. This only makes sense if the probability of intervention is low. On the other hand if the intervention probability is close to 1 then the only way to get a decent amount of output is to bite the bullet and provide incentives for production in the face of intervention.

The key element here is the lower bound $\underline{\lambda} > 0$: if the intervention is less likely than this then what will be observed is production if and only if there is no intervention: the intervention "works." Unfortunately as soon as the intervention is more widely used - that is $\lambda > \underline{\lambda}$ - production always takes place and the intervention stops "working." If small $\lambda$ corresponds to a social experiment (inside or outside the laboratory) - or a natural experiment that is on a small scale - the data conclusively shows that the intervention works. This, however, is misleading - it works because it is on too small a scale to change social norms - ostracism probabilities remain unchanged. As soon as the intervention is adopted on a broad scale it becomes optimal to change the social norm, raising ostracism probabilities - and the intervention no longer works.

We should also note the obvious fact that if $V$ is small enough or $R$ large enough then the intervention will work for all $\lambda$.

*Wages and Employment.* Economists generally think of wages and employment being determined competitively in markets in which firms and workers are small players. The larger social groups to which these small players belong, by contrast, can have substantial monopoly or monopsony power. We see this explicitly in the case of trade unions - which indeed exercise monopsony power by enforcing social norms such as "do not work too hard" through the mechanism of peer monitoring and ostracism. In our setting "not working too hard" is a public good because it enables the group to exercise its monopsony power.

Consider the field experiment of Gneezy and List (2006) in which they paid some solicitors a fixed bonus above the market wage and others not. They discovered that initially those with the bonus increased their effort, but over the entire course of the experiment did about the same

amount of work per unit of pay as those without the bonus. This is consistent with a social norm in which the wages per unit of effort are part of a social norm: solicitors do the amount of work per pay as called for by the social norm regardless of whether the money is paid as a piece rate or a lump sum.

We make two observations about the consequences of stickiness. First, if wages and employment are determined by social norms then the presence of fixed costs of changing social norms will have a dynamic effect very similar to that of the menu cost model of Calvo (1983): changes will occur only when economic circumstances change enough to make it worthwhile to pay the fixed cost of changing the norm. Second, the Lucas critique may apply to empirical work studying changes in market conditions.

A particular case where the Lucas critique may be relevant is in the continuing controversy over the employment effect of the minimum wage. Consider studies by two labor economists, both John Bates Clarke medal winners: David Card and Kevin Murphy. Card and Krueger (1994) provide evidence that changes in the minimum wage have little effect on employment, while Deere, Murphy and Welch (1995) provide evidence that the minimum wage has a substantial effect on employment. These two studies use rather different evidence: Card and Krueger (1994) use a natural experiment comparing the effect of minimum wage change in one state against nearby states where the minimum wage did not change. Deere, Murphy and Welch (1995) examine the before and after effect of a change to the federal minimum wage. If employment is determined by social norms then changes in a single state may represent a small intervention insufficient to change social norms: as from the above result, ostracism probabilities remain fixed and if intervention occurs the public good is not provided, workers work harder and get the outside benefit represented by the increased wage, and there is little employment effect. On the contrary changes in the federal minimum wage may represent a large intervention sufficient to change social norms so there would be a substantial employment effect. This raises the issue of whether studies of the minimum wage might need to look more closely at the extent to which social networks and social norms play a role in determining employment.

## 6. Internalization

We turn now to the notion of internalizing a social norm: by this we mean that individuals enforce the social norm on themselves feeling guilt for violating a social norm - or pride in adhering to it. Consistent with our objective of endogenizing behavior that has been taken to be exogenous in behavioral economics we adopt a simple model of internalization. We assume that individuals can internalize any strategy - that, roughly speaking, they can invest in learning a rule of behavior. In the current stylized simple setting, learning a behavior rule may seem relatively easy - produce or not produce? tell the truth or not tell the truth? - although the correct ostracism probabilities are perhaps not so trivial - but real social norms deal with a much broader array of more complicated interactions. Indeed, social norms may encompass entire codes of conduct in the sense of Block and Levine (2016) or secret handshakes as in Robson (1990). We also refer to the literature on

automata and the complexity of strategies (see for example Abreu and Rubinstein (1988) and the more recent literature on competition Gale and Sabourian (2005)) - a literature that implicitly or explicitly supposes that it is more difficult to implement complicated strategies than simple ones. Once a strategy has been chosen and invested in, our assumption is that the investor receives a benefit from adhering to it.[31] In other words the utility function changes so that doing the things you have learned to do well brings utility. This makes sense also from the perspective of the habit formation literature (see for example, Constantanides (1990), Campbell and Cochrane (1999) or Boldrin, Christiano and Fisher (2001)) - in a broader sense learning to do something often makes it more pleasurable. Just as the utility of fine wine increases with experience, so may the benefit of altruistic giving.

Internalization that is completely neutral between different strategies or based entirely on considerations of complexity - as we shall see - does not change anything in a continuum game of the type we are studying where commitment has no value. The key to our results is that we allow it to be less costly to invest in a strategy that everyone else is using than to invent a new strategy. This gives rise to a mild preference for conformity similar to that studied by Benassy (1998) and Akerlof and Kranton (2000). It makes sense, think for example of one important social norm - a language. It is clearly much easier to learn a language that everyone else is using than to invent an entirely new language[32] - there is a substantial network externality in learning a social norm or code of conduct.

We turn next to the specific details of the model. The aim of the section is to characterize the cost minimizing social norm. The issue is whether production can be implemented and what are the ostracism probabilities needed to support it.

*A Model of Investment in Social Norms.* The mechanism design problem (and the induced game) we studied above effectively begins with the group/principal choosing and announcing a pure strategy $\sigma$ called the social norm. We now assume that after this announcement and before matching, production and monitoring takes place - in particular before one's identity as producer or monitor becomes known - individuals may choose to invest (or specialize) in a pure strategy $s$ of their choice. We assume that the cost of this investment is $a \geq 0$ if the strategy chosen $s = \sigma$, is the social norm, and a greater amount $a + \Gamma$ if the strategy chosen $s \neq \sigma$, is not the social norm, where $\Gamma \geq 0$. The choice of investment is known only to the investor.[33]

[31] If the cost of investment is positive it would not make sense to invest in a strategy if the only consequence was to feel guilt for not following it. However failure to receive a benefit is equivalent to a loss, so the loss of benefit may indeed be the same as guilt.

[32] It would of course be useless as well.

[33] In the bargaining literature (see Schelling (1956), Muthoo (1996) and Dutta (2012)) commitment is assumed to be observable. That literature focuses on the strategic advantage of commitment when there are small numbers and an unobservable commitment is useless. Here we assume the commitment is unobservable: if it were observable there would be an additional channel of punishment - ostracism could be based on failing to invest in the social norm. We wish to keep the punishment channel the same as in the base model so we keep the commitment unobservable. The observable case (with noisy signals of investment) is similar to the model of codes of conduct studied in Block and Levine (2016).

The consequence of investing in a strategy is that the investor gets utility from using it: specifically if $s$ is chosen and the terminal node is consistent with $s$ the investor receives a bonus of $B \geq 0$ for following the prescribed strategy. Without loss of generality we assume $a + \Gamma \leq B$ - that is investing in any strategy and following it is profitable - since otherwise any announced social norm would be trivially followed for lack of alternatives.[34] By *internalization* we mean that individuals choose to invest in the social norm.[35] Notice that the group/principal should never choose a social norm that will not be internalized: it would always be better to announce as the social norm the equilibrium strategy chosen by members. So the group problem is what norm to internalize, which in our setting means whether to implement production and if so how small the ostracism probabilities can be.

The two parameters $B, \Gamma$ describing internalization possibilities play a rather different role. The parameter $B$ can be viewed as a value of commitment: how much better is it to stick to a strategy once invested in? Notice that in other settings this commitment value could be useful for the purpose of commitment: here because players are infinitesimal, commitment by individual group members is not useful. Here the commitment value is useful only to the mechanism designer.

By contrast the parameter $\Gamma$ may be thought of as the benefit of conformity: how much cheaper is it to invest in a strategy because everyone else is using it? This is akin to the desire for conformity in models such as Benassy (1998) or Akerlof and Kranton (2000), albeit for rather more concrete reasons. We do not imagine it is so much that group members have an innate desire to be like other groups members - although this could be the case - but rather than in practice strategies are complicated objects that are not easy to learn and there are network externalities. We think it is easier to learn a strategy that has been discussed, announced, and that other people are adopting than to invent your own. As noted this can easily be seen in the case of language: it is obviously far easier to learn a language that everyone else is using than to make up your own grammar.

*Essential versus Inessential Indifference.* When we solved the cost minimization problem without the possibility of internalization the solution involved several forms of indifference. The producer is indifferent between producing and not producing, the monitor is indifferent between reporting 0 and 1 and the audience members are indifferent to ostracizing or not ostracizing. The first indifference - that of the producer - is inessential in the sense that we could punish a bit more for a bad signal and the producer would strictly prefer to produce. The indifference of the monitor is essential in the sense that if the monitor is not indifferent between reporting 0 and 1 the monitor will not tell the truth. Similarly: it is weakly dominant for audience members to vote not to ostracize: it is only because of indifference that they are willing to do so: again this is an essential form of indifference.

One way to see that an indifference is essential is to perturb the model. Take first the issue of

[34] In the bargaining literature (see Muthoo (1996)and Dutta (2012)) it is assumed that investing in a strategy and following it yields zero profit. This is in order to isolate the effect of the strategic advantage of commitment.

[35] Let $a(s)$ denote the investment cost. If we start with a model with investment cost $a'(s)$, benefit $B'$ and guilt $G'$ for violating the strategy and we assume that $a'(s) < B'$ this is equivalent to the model here with $a(s) = G' + a'(s), B = G' + B'$.

monitoring. Suppose that the monitor may choose whether or not to observe the signal and that there is a positive cost for observing the signal. In this case production cannot be implemented: if we make the monitor indifferent between $0, 1$ so willing to tell the truth, it is always better to report either one of the two and not pay the monitoring cost.

Take second the issue of voting. Suppose that there is a small but positive probability that unanimity is required for ostracism. Again production cannot be implemented: it is strictly dominant for each audience member to vote against ostracism.

Assuming that $B, \Gamma > 0$ makes all indifference inessential. If implementation of production is possible and we take the cost minimizing incentives as in Theorem 1 then every group member strictly prefers to invest in the social norm, strictly prefers to produce, strictly prefers to report the truth, and strictly prefers to vote for ostracism when the social norm calls for it and against it when it does not. Adding a small monitoring cost or probability that unanimity is required for ostracism no longer changes things.

*Fixed and Adjustment Costs.* We have seen that fixed costs are important for understanding how social norms do or do not adjust. In addition to the fixed costs of organization, internalization also creates a type of adjustment cost. In a dynamic setting sticking with the current internalized social norm has an advantage over choosing a new social norm which will require that the investment cost be paid a second time.

*Conformity and Ostracism: Complements or Substitutes?*

Our main result characterizes the cost minimizing social norm. The main point is that if $\Gamma$ is not too small internalization can enable production where monitoring alone is not enough. In this case conformity and peer pressure are complements. However for larger $\Gamma$ as we will explain the two become substitutes. In the web appendix we show

**Theorem 5.** *Suppose $B, c > 0$. Define*

$$\begin{aligned} \overline{P} &= \frac{c}{(1-2\pi)(1-h^2\beta^2)} \\ \underline{P} &= \overline{P} \cdot \max\left\{0,\, 1-(1+(1-2\pi)h\beta)\,\frac{B}{c},\, (1-\frac{B}{c})(1-h^2\beta^2)\right\}. \end{aligned}$$

*If $\underline{P} > 1$ implementation of production is not possible. If $\underline{P} \leq 1$ then there exists $(1+\pi)B/2 \geq \overline{\Gamma} \geq \underline{\Gamma} \geq 0$ such that production can be implemented if and only if $\Gamma \geq \underline{\Gamma}$ where $\underline{\Gamma} = 0$ if and only if $\overline{P} \leq 1$. If $B \geq c$ then $\overline{\Gamma} = c/2, \underline{P} = 0$ and for $\Gamma \geq \overline{\Gamma}$ there is* complete internalization: *production is implemented without ostracism that is $P = Q = 0$. If $\Gamma \geq \underline{\Gamma}$ the cost minimizing internalized social norm implements production and, for generic parameter values, ostracism probabilities are given by unique continuous piecewise linear functions $\hat{P}(\Gamma)$ and $\hat{Q}(P)$ with the following properties:*

*1. $\overline{P} \geq \hat{P}(\Gamma) \geq \underline{P}$ is defined on $[\underline{\Gamma}, B]$, is strictly decreasing on $[\underline{\Gamma}, \overline{\Gamma}]$; $\hat{P}(\Gamma) = \underline{P}$ for $\Gamma \geq \overline{\Gamma}$ and $\hat{P}(\underline{\Gamma}) = \min\{1, \overline{P}\}$;*

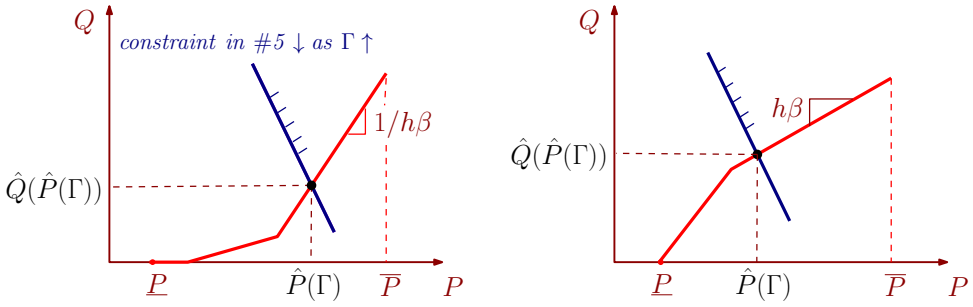
*2. $h\beta P \geq \hat{Q}(P) \geq 0, h\beta P - B$ is defined for $P \geq \underline{P}$ and is non-decreasing with $\hat{Q}(\overline{P}) = h\beta\overline{P}$;*

*3. [lower function] If $\pi \leq .333$ then for $\hat{P}(\underline{\Gamma}) < P < \hat{P}(\overline{\Gamma})$ the producer strictly prefers to produce, and $\hat{Q}(P)$ is convex, non-increasing in $B$;*

*4. [upper function] If $\pi \geq .382$ then for $\hat{P}(\underline{\Gamma}) < P < \hat{P}(\overline{\Gamma})$ the monitor strictly prefers to tell the truth, and $\hat{Q}(P)$ is concave, non-decreasing in $B$;*

*5. For $\underline{\Gamma} \leq \Gamma \leq \overline{\Gamma}$ we have $(\pi h\beta - (1-2\pi))\hat{P}(\Gamma) - (\pi - (1-2\pi)h\beta)\,\hat{Q}(\hat{P}(\Gamma)) = 2\Gamma - c$. In addition there is a $\overline{B}$ such that for $B \geq \overline{B}$ the cost minimizing solution $\hat{P}$,$\hat{Q}$ is independent of $B$.*

Figure 6.1: Ostracism probabilities. Case $\pi \leq .333$ illustrated on the left, case $\pi \geq .382$ on the right




The ostracism probabilities in cases 3 and 4 are depicted in Figure 6.1. The result has a number of cases with key implications. If $B = 0$ then $\underline{P} = \overline{P}$ and the implementability condition is the same as in Theorem 1. If the value of commitment $B > 0$ then $\underline{P} < \overline{P}$ but again if the latter is smaller than 1 we get the implementability condition of Theorem 1 - production can be implemented with $\Gamma = 0$. If on the other hand $\underline{P} > 1$ we are in the case where $B$ is insufficiently large to allow implementation. As $B$ grows larger the situation improves, and for $B \geq c$ we see that $\underline{P} = 0$, and even if $\overline{P} > 1$ for $\Gamma$ not too small production can be implemented - in fact with $\underline{P} = 0$ properties *1* and *2* above imply that for $\Gamma \geq \overline{\Gamma}$ we have $\hat{P}(\Gamma) = \hat{Q}(\hat{P}(\Gamma)) = 0$ and we achieve complete internalization, where no monitoring or ostracism is needed.

If $\underline{\Gamma} > 0$ without internalization production cannot be implemented. Indeed, for small $\Gamma < \underline{\Gamma}$ production cannot be implemented, but for large enough $\Gamma \geq \underline{\Gamma}$ internalization enables the implementation of production. In other words: internalization can enable production where monitoring alone is not enough. But notice: internalization and monitoring are not substitutes - they complement each other. Without any benefit from conformity ($\Gamma = 0$) and consequently no internalization production cannot be implemented, and without monitoring production cannot be implemented. Yet with both together production can be implemented.

Once $\Gamma \geq \underline{\Gamma}$ internalization begins to substitute for monitoring: we see that $\hat{P}(\Gamma), \hat{Q}(\hat{P}(\Gamma))$ are both decreasing functions of $\Gamma$ so that greater value of conformity reduces the need for monitoring and ostracism. Eventually when $\Gamma = \overline{\Gamma}$ further increases in the value of conformity have no effect and only increases in the value of commitment $B$ can further reduce monitoring costs. As $B$ and $\Gamma$ are increased eventually $B > c$ and $\Gamma > c/2$ we achieve complete internalization with no punishment or incentives needed at all.

Notice that the benefits of being able to implement may be disproportionate: even if $B, \Gamma$ are

quite small if they enable implementation the gain is on the order of the value of production $V$ which can be very large.

Finally, while internalization and monitoring can be complements or substitutes what about $B$ and $\Gamma$? Notice that these two are not independent in the sense that we have assumed $a+\Gamma < B$ and as $a \geq 0$ we call $\Gamma \leq B$ *feasible*. We observe that $\overline{\Gamma} \leq (1+\pi)B/2 \leq (3/4)B$ so that there is a range of feasible $\Gamma \geq \overline{\Gamma}$ for $a$ small enough. For $\Gamma$ in this range the theorem says that the cost minimizing solution is independent of $\Gamma$. Once $\overline{\Gamma}$ is reached increases in $\Gamma$ are not helpful. Similarly once $\overline{B}$ is reached further increases in $B$ are not helpful. In addition no matter how large is $B$ if $\Gamma$ is small then by continuity the solution must be close to that of the basic problem described in Theorem 1. Furthermore as $B \to 0$ we have $\underline{P} \to \overline{P}$ and from property *2* this implies $\hat{Q}(\hat{P}(\Gamma)) \to h\beta\overline{P}$ so that for small $B$ the solution must also be close to that that of the basic problem described in Theorem 1. Hence we see that if $\Gamma$ is small large $B$ helps very little and that if $B$ is small large $\Gamma$ helps very little. In other words: both $B$ and $\Gamma$ need to be large for internalization to be useful - neither on its own is enough. Roughly speaking $B$ and $\Gamma$ are complements.

*Producer or Monitor?.* The value of commitment $B$ and benefit of conformity $\Gamma$ play different roles in the theory. Higher commitment value $B$ loosens the incentive constraints for both producer and monitor. The benefit of conformity $\Gamma$ are more like credits that can be spent either on the producer or the monitor: if either one gains too much from deviating then it will be better to suffer the loss of $\Gamma$ and invest in a strategy other than the social norm. Intuitively the expected cost of ostracizing the producer $P$ is proportional to the probability of the bad signal $\pi$ while the expected cost of ostracizing the monitor $Q$ is proportional to the probability of the good signal $1-\pi$. Hence we would expect that for small $\pi$ we should spend our $\Gamma$ credits on the monitor trying to reduce $Q$ while if $\pi$ is large we should spend our $\Gamma$ credits on the producer trying to reduce $P$. Indeed this is the case. What is surprising is that $\pi$ can be relatively large in absolute terms and we should still focus on the monitor: in Theorem 5 the condition $\pi \leq .333$ (recall that $\pi < 0.5$) implies that it is $Q$ that is minimized. Another way to say this is that we emphasize internalization for the monitor and provide strong incentives to the producer to produce. Indeed the producer strictly prefers to produce, while if $\Gamma$ is large enough to permit it the monitor is indifferent to lying. In the second case $\pi \geq .382$ the signal is noisy and we emphasize internalization for the producer and provide strong incentives to the monitor not to file good reports. Which case is relevant depends in general on $h\beta$ as well as $\pi$ but is surprisingly insensitive to $h\beta$: that matters only in the range $.333 \leq \pi \leq .382$.

*In the Laboratory.* As $c \to 0$ we get complete internalization - public good production without monitoring and ostracism. This is relevant to laboratory experiments as $c$ in that setting is quite low. This implies that in the laboratory we are likely to measure internalized social norms even if we successfully eliminate outside influence. This raises yet another Lucas critique of laboratory experiments, since we can measure only existing social norms, but not how social norms might change. We should mention work such as that of Rand, Greene and Nowak (2012) that studies the

impact of the amount of time allowed for decisions in the laboratory. A robust finding is that when little time is allowed for a decision people are more generous than when they are allowed more time. The authors' interpretation is that with little decision time individuals impulsively apply a "daily life" norm of cooperation, while upon reflection the logic of self-interest prevails. Our theory suggests another possibility: that the social norm of cooperation is followed in both cases, but reflection may indicate that the norm does not actually require so much generosity. We can additionally make sense of another, less discussed aspect of the experimental result, which is the higher variance of contributions in the delayed decision case: the observed variability may reflect the uncertainty about the social norm prevailing in the laboratory context. In other words, given more time you know better that you do not know what exact norm applies. We observe incidentally that this finding contrasts with the opposite one usually observed in isolated decision contexts, where longer decision time makes choice variance decrease and decision converge to the optimal individual choice (as in the time pressure model axiomatized in Cerreia-Vioglio et al (2018)).

The Rand, Greene and Nowak (2012) data also suggest an extension of the model to allow social norms where "proper" behavior can depend on expectations of a signal about which limited information might be available. This would also enable us to incorporate Benabou and Tirole (2006) type behavior.

## 7. Conclusion

We conclude by indicating how the ideas in this paper fit into the broader literature of behavioral economics and cultural norms: we propose indeed that the theory of endogenous social norms may form an important link between these ideas.

*Social Norms or Psychology?.* When social norms are endogenous, especially if they are internalized, it is not easy to distinguish - in the laboratory for example - between social norms and intrinsic preferences. The issue is somewhat controversial as writers such as Bowles et al (2003) and Roemer (2015) point to evidence that while incentives seem important for providing public goods, incentives seem less important for providing costly punishments. In particular, without internalization, the incentives for monitors and for ostracism are weak - relying as they do on essential indifference. In practice social norms often incorporate repeated rounds: a rule of "punish violators and if you fail to do so you are a violator yourself" can be found even in such places as written constitutions for prison gangs as documented in Skarbek (2014), and Levine and Modica (2016) provide theoretical results in this direction. However, as the current theory shows, with internalization the need for multiple rounds of auditing is mitigated.

While writers such as Bowles et al (2003) and Roemer (2015) point to evolutionary reasons why punishment might be "hard-wired" and while we do not doubt that small children do not need to be taught to punish the theft of a toy, social norms must - and do - specify punishment levels scaled to the nature of the offense, the benefit of deviating, and the chances of getting caught. The vast array of social norms we see across time and location indicate to us that most likely they are endogenous reflecting circumstances even if they do tap into intrinsic preferences for "revenge."

It is instructive here to think of the custom of tipping service providers: commonplace in the US and UK, but rare in Italy. In Italy it works rather the other way around: not only is there no tipping but repeat customers get a discount - kind of a negative tip. In the US and UK there is a definite social sanction for not tipping. Other people at your table as well as the waiters may sneer at you - indeed you may be explicitly told not to return. We would argue that these are not just customs, but rather are based on the need for incentives. With low waiter turn-over both within restaurants and within communities social norms among waiters can support good service and tipping is not needed - this is the situation in Italy. With high waiter turn-over and waiters not tied to the local community it is difficult for social norms to support good service, and so tipping is a needed incentive.

*Social versus Cultural Norms.* The issue of intrinsic versus endogenous preferences is mitigated when we observe that as well as social norms there are cultural norms. The key difference is that while individuals choose social norms cultural norms are generally derived at a young age from others, especially parents. Our view is that while there is a difference in degree between social and cultural norms - cultural norms require a much larger investment and have a much greater value of commitment - they are part of the same theory as that of social norms. To us the key lies in the idea of investment in strategies: once we recognize that strategies are invested in and that different strategies may have different investment costs we should recognize as well that these investments can be subsidized by interested parties. Hence: it is easier to learn the language spoken by your parents in your home; parents explicitly teach social norms to their children taking on themselves part of the investment cost. States invest in public schooling which in part teach social norms - we think perhaps in terms of the Madrasas of the Taliban: the receipt of some valuable human capital in exchange for learning the social norms valuable to the Taliban. All schools work this way, however - Bowles and Gintis (1976) documented the teaching of social norms in US schools; everyone who has read history post K-12 recognizes the substantial element of national myth taught in school; and of course interest groups fight over curriculum precisely because they want to promote particular social norms. In economics there has been a tendency to view schooling through the lenses of human capital acquisition - and we agree that schooling is not mere signalling but teaches valuable skills. We should recognize, however, that those skills are a subsidy for learning social norms - if we examine the history of public education we observe that it originated in Scotland and that the valuable skill of high literacy was taught for religious reasons - to promote a social norm.

This idea of the strategic choice of social norms is not new: the idea of social norms that may be acquired horizontally (from peers) or vertically (from parents) has been used by Bisin and Verdier (2001) and Bisin and Verdier (2005) among others to examine the evolution of institutions. The model they use of costly efforts by parents to influence the social norms of their children is compatible with the view here: we think our theory adds an extra dimension to their analysis by emphasizing the endogenous nature of the social norms that are promoted. Their analysis complements ours as well because it deals with the endogenous formation of groups, a topic which we ignore in this paper.

In this context we should mention as well the possibility of competing social norms. In practice people may belong to several groups. This may not matter to the extent that social norms are incomplete and deal only with behavior relevant to that group: for example, the social norm of economists deals with how many papers one should referee, but not how often one should attend religious services, while religious social norms deal with the latter but not generally the former. On the other hand there can be competing social norms - for example a Catholic doctor who has a patient wanting an abortion. This raises a complex set of issues that we have studied in part in Dutta, Levine and Modica (2018).

**References**


Abreu, D. and Rubinstein, A. (1988), "The structure of Nash equilibrium in repeated games with finite automata," *Econometrica* 1259-1281.

Akerlof, George A., and Rachel E. Kranton (2000) "Economics and identity," *The Quarterly Journal of Economics* 115(3): 715-753.

Andrews, Sally, David A. Ellis,Heather Shaw and Lukasz Piwek (2015): "Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use," *PLOSOne*.

Bénabou, Roland, and Jean Tirole (2006): "Incentives and prosocial behavior," *The American Economic Review* 96(5): 1652-1678.

J.P. Bénassy (1998): "Conformism and multiple sycophantic equilibria", in P. Howitt and A. Leijonhufvud (eds), *Money, Markets and Method*, Edward Elgar.

Bisin, A., and Verdier, T. (2001):"The economics of cultural transmission and the dynamics of preferences," *Journal of Economic theory* 97(2): 298-319.

Bisin, A., and Verdier, T. (2005): "Cultural transmission," The New Palgrave Dictionary of Economics.

Boldrin, M., Christiano, L. J., and Fisher, J. D. (2001): "Habit persistence, asset returns, and the business cycle," *American Economic Review*: 149-166.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001): "In search of homo economicus: behavioral experiments in 15 small-scale societies," *The American Economic Review* 91(2): 73-78.

Block, J. I., and Levine, D. K. (2016): Codes of conduct, private information and repeated games," *International journal of game theory*, 45: 971-984.

Bowles, S., and Gintis, H. (1976): Schooling in capitalist America (Vol. 57). New York: Basic Books.

Calvo, G. A. (1983): "Staggered prices in a utility-maximizing framework," *Journal of monetary Economics* 12(3): 383-398.

Cerreia-Vioglio, S., F. Maccheroni, M. Marinacci and A. Rustichini (2018): "Multinomial logit processes and preference discovery: inside and outside the black box", mimeo

Campbell, J. Y. and Cochrane, J. H. (1999): "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of political Economy*, 107: 205-251.

Card, D., and Krueger, A. B. (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review* 84(4): 772-793.

Constantinides, G. M. (1990): "Habit formation: A resolution of the equity premium puzzle," *Journal of political Economy*, 98: 519-543.

Cremer, J. and McLean, R. P. (1988): "Full extraction of the surplus in Bayesian and dominant strategy auctions." *Econometrica* 56: 1247-1257.

Deere, D., Murphy, K. M., and Welch, F. (1995): "Employment and the 1990-1991 minimum-wage hike," *The American Economic Review* 85(2): 232-237.

Dutta, Rohan (2012): "Bargaining with Revoking Costs," *Games and Economic Behavior*, 74: 144-153.

Dutta, Rohan, David K. Levine and Salvatore Modica (2018): "Damned if You Do and Damned if You Don't: Two Masters," mimeo EUI.

Ellickson, Bryan, Birgit Grodal, Suzanne Scotchmer and William R. Zame: "Clubs and the Market", *Econometrica* 67: 1185-1217

Fudenberg, Drew, David Levine and Eric Maskin (1994): "The Folk Theorem with Imperfect Public Information," *Econometrica* 62(5): 997-1039.

Gale, D and Sabourian, H. (2005): "Complexity and competition," *Econometrica*, 73: 739-769.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. (2003): "Explaining altruistic behavior in humans," *Evolution and Human Behavior*, 24: 153-172.

Gneezy, U., and List, J. A. (2006): "Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments," *Econometrica* 74(5): 1365-1384.

Gneezy, U., and Rustichini, A. (2000): "A fine is a price," *The Journal of Legal Studies* 29(1): 1-17.

Gneezy, U., and Rustichini, A. (2000): "Pay enough or don't pay at all," *The Quarterly Journal of Economics* 115(3): 791-810.

Hart, O., and Moore, J. (1988). "Incomplete contracts and renegotiation," *Econometrica* 56(4): 755-785.

Jackson, M. O. (2010): *Social and Economic Networks*, Princeton university press.

Kandori, M. (1992): "Social norms and community enforcement," *The Review of Economic Studies* 59(1): 63-80.

Kinchen, K. S., Cooper, L. A., Levine, D., Wang, N. Y., and Powe, N. R. (2004): "Referral of patients to specialists: factors affecting choice of specialist by primary care physicians," *The Annals of Family Medicine* 2(3): 245-252.

Kraut, A. I. (1975): "Prediction of managerial success by peer and training-staff ratings," *Journal of Applied Psychology* 60: 14.

Levine, David K. (2012): *Is behavioral economics doomed?: The ordinary versus the extraordinary* Open Book Publishers.

Levine, D. K. and A. Mattozzi (2017): "Voter Turnout with Peer Punishment," EUI

Levine, David and Salvatore Modica (2016): "Peer Discipline and Incentives within Groups", *Journal of Economic Behavior and Organization* 123: 19-30

Levine, David and Salvatore Modica (2017): "Size, Fungibility, and the Strength of Lobbying Organizations", *European Journal of Political Economy* 49: 71-83

Meyer, Christian Johannes and Tripodi, Egon, Sorting into Incentives for Prosocial Behavior (October 24, 2017). Available at SSRN: https://ssrn.com/abstract=3058195

Muthoo, Abhinay (1996): "A Bargaining Model Based on the Commitment Tactic," *Journal of Economic Theory* 69: 134-152.

Olson Jr., Mancur (1965): *The Logic of collective action: public goods and the theory of groups*, Harvard Economic Studies.

Ostrom, Elinor (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.

Rahman, David (2012): "But Who Will Monitor the Monitor?", *American Economic Review* 102(6): 2767-2797.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012): "Spontaneous giving and calculated greed,"

*Nature*, 489: 427-430.

Robson, A. J. (1990): "Efficiency in evolutionary games: Darwin, Nash and the secret handshake.," *Journal of theoretical Biology*, 144: 379-396.

Roemer, John (2015): "Kantian optimization: An approach to cooperative behavior," *Journal of Public Economics* 127(C): 45-57

Sarsons, H. (2017): "Interpreting Signals: Evidence from Doctor Referrals," Working Paper.

Schelling, Thomas C. (1956): "An Essay on Bargaining," *The American Economic Review* 46(3): 281-306.

Skarbek, D. (2014): "The social order of the underworld: How prison gangs govern the American penal system," O*xford University Press.*

Tisserand, J. C., Cochard, F., and Le Gallo, J. (2015): "Altruistic or Stategic Considerations: A Meta-Analysis on the Ultimatum and Dictator Games," Besançon: CRESE, Université de Franche-Comté.

West, E., Barron, D. N., Dowsett, J., and Newton, J. N. (1999): "Hierarchies and cliques in the social networks of health care professionals: implications for the design of dissemination strategies," *Social science & medicine* 48(5): 633-646.

**Web Appendix**

*Cost Minimizing Social Norms*

**Lemma 1.** *As ostracism is socially costly we should ostracize only as needed for incentive compatibility: in particular if the producer is reported to produce there should be no ostracism ($p(1) = 0$) and if the monitor files a bad report there should be no ostracism ($q(0) = 0$).*

*Proof.* The incentive constraint for a monitor is simple. If it is to be optimal to tell the truth the monitor must be indifferent between reporting 0 and 1. If a good report ($x = 1$) is filed the monitor is excluded with probability $q(1)$ and the producer is ostracized with probability $p(1)$. If a bad report ($x = 0$) is filed the corresponding probabilities are $q(0)$ and $p(0)$. Hence the incentive constraint for a monitor is $q(1) + h\beta p(1) = q(0) + h\beta p(0)$ or $q(1) - q(0) = h\beta\left(p(0) - p(1)\right)$.

When a producer chooses to contribute there is a direct cost $c$ but also an indirect cost due to monitoring. The monitor receives a bad signal with probability $\pi$ which leads to the producer and monitor being ostracized with probabilities $p(0)$ and $q(0)$, and a good signal with probability $1 - \pi$ and the producer and monitor are ostracized with probabilities $p(1)$ and $q(1)$. Hence the overall expected cost to the producer of contributing is $c + \pi\left(p(0) + h\beta q(0)\right) + (1 - \pi)\left(p(1) + h\beta q(1)\right)$. When a producer chooses not to contribute there is no longer a direct cost, but there still remain the indirect costs. This is $(1 - \pi)\left(p(0) + h\beta q(0)\right) + \pi\left(p(1) + h\beta q(1)\right)$. The incentive constraint for the producer is that the the cost of contributing must be no greater than the cost of not contributing: this can be written as

$$\left(p(0) + h\beta q(0)\right) - \left(p(1) + h\beta q(1)\right) \geq \frac{c}{1 - 2\pi}.$$

or

$$\left(p(0) - p(1)\right) - h\beta\left(q(1) - q(0)\right) \geq \frac{c}{1 - 2\pi}.$$

Using the monitor's incentive constraint we can rewrite the producer's constraint as

$$\left(p(0) - p(1)\right)\left(1 - h^2\beta^2\right) \geq \frac{c}{1 - 2\pi}$$

Since all ostracism is socially costly, at the optimum we must therefore have $p(1) = 0$. Further the monitor's constraint can be written as

$$p(0) = \frac{q(1) - q(0)}{h\beta}.$$

In turn the producer's constraint can be written as

$$\left(q(1) - q(0)\right)\frac{\left(1 - h^2\beta^2\right)}{h\beta} \geq \frac{c}{1 - 2\pi}.$$

This in turn means that at optimum $q(0) = 0$. □

*Monitoring/Interaction Tradeoffs*

**Theorem.** [2 in the text] *Let $C(h,\pi)$ denote the least cost of implementation if the implementation condition is satisfied and $\infty$ otherwise. If there exists a $h, f(h)$ such that the implementation condition is satisfied then there is a unique minimum of $C(h,\pi)$ subject to $\pi = f(h)$ and the optimum satisfies*

*1. $h$ is decreasing, $\pi$ increasing in $\beta$*

*2. if $h_f, \pi_f$ are the solutions of the cost minimization problem and $\overline{f}$ satisfies $\pi_f = \overline{f}(h_f)$ and greater signal sensitivity than $f$ in the sense that $|\overline{f}'(h)| > |f'(h)|$ then $h_{\overline{f}} > h_f$ and $\pi_{\overline{f}} < \pi_f$.*

*Proof.* We first show that the isocost curves

$$\left[\frac{U}{2}\frac{1}{1-h^2\beta^2}\frac{\pi+(1-\pi)h\beta}{1-2\pi}+\frac{1}{2}\right]c = C$$

are downward sloping and concave. Write

$$\frac{\pi+(1-\pi)h\beta}{(1-h^2\beta^2)(1-2\pi)} = 2\frac{(C/c)-1/2}{U} = D$$

and rationalize this to

$$\pi+(1-\pi)h\beta - D(1-h^2\beta^2)(1-2\pi) = 0.$$

Rearrange to isolate $\pi$

$$\begin{aligned}\pi(1-h\beta)+h\beta - D(1-h^2\beta^2)+2\pi D(1-h^2\beta^2) &= 0\\ \left(1-h\beta+2D(1-h^2\beta^2)\right)\pi &= D(1-h^2\beta^2)-h\beta\end{aligned}$$

to solve for the isocost curve

$$\pi = \frac{D(1-h^2\beta^2)-h\beta}{1-h\beta+2D(1-h^2\beta^2)}.$$

We may now directly compute the derivative

$$\frac{d\pi}{dh} = -\frac{\beta+2Dh\beta^2}{1-h\beta+2D(1-h^2\beta^2)}+\frac{\left[D(1-h^2\beta^2)-h\beta\right]\left[\beta+4Dh\beta^2\right]}{\left[1-h\beta+2D(1-h^2\beta^2)\right]^2}$$

and using $D(1-h^2\beta^2)-h\beta = \left(1-h\beta+2D(1-h^2\beta^2)\right)\pi$ write this as

$$\begin{aligned}\frac{d\pi}{dh} &= -\frac{\beta+2Dh\beta^2}{1-h\beta+2D(1-h^2\beta^2)}+\frac{\pi\left[\beta+4Dh\beta^2\right]}{1-h\beta+2D(1-h^2\beta^2)}\\ &= -\frac{\beta(1-\pi)+2Dh\beta^2(1-2\pi)}{1-h\beta+2D(1-h^2\beta^2)} < 0\end{aligned}$$

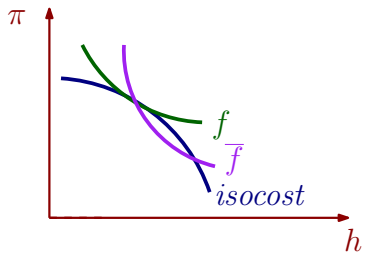
showing that the isocost curves are downwards sloping. Moreover the numerator is strictly increasing in $h$ and the denominator strictly decreasing so the derivative has a negative slope - which is to say that the isocost curves are strictly concave.

For the first result we do a direct computation. Let

$$g(h\beta) \equiv -\frac{(1-\pi)+2Dh\beta(1-2\pi)}{1-h\beta+2D(1-h^2\beta^2)}$$

so that the first order condition for the minimum is $f'(h) = \beta g(\beta h)$. Hence $dh/d\beta = [g(\beta h) + \beta h g'(\beta h)/[f"(h) - \beta^2 g'(\beta h)] < 0$ and $d\pi/d\beta = f' \cdot dh/d\beta > 0$.

For the second result we see that with a convex constraint and quasi-concave objective function increasing sensitivity by the weak axiom of revealed preference must move the solution $h, \pi$ down and to the right - see the figure below.




□

*Production Subsidies*

**Theorem.** [3 in the text] *It is always optimal to implement production, at the level*[36]

$$\theta = (V + (\mu - 1/2)\, r)\, /(2\mu)$$

*as expected increasing in $r$. The utility advantage of implementing least cost production over the default equilibrium where $r/2$ is produced is*

$$G(r) \equiv ((V + (\mu - 1/2)r)^2/(4\mu) + (1-\mu)r^2/4 - Vr/2$$

*which satisfies $G\,(0) = V^2/(4\mu)$, $G(2V) = 0$ and $G'(r) < 0$ for all $r < 2V$.*

*Proof.* Solving the first order condition for maximizing social utility gives the optimal $\theta$ to implement. Plugging in, the corresponding social utility is

$$\begin{aligned} &U + (V + (\mu - 1/2)r)\theta - \mu\theta^2 - (\mu - 1/2)r^2/4 \\ =&U + (V + (\mu - 1/2)r)^2/(2\mu) - (V + (\mu - 1/2)r)^2/(4\mu) - (\mu - 1/2)r^2/4 \\ =&U + (V + (\mu - 1/2)r)^2/(4\mu) - (\mu - 1/2)r^2/4. \end{aligned}$$

[36] Again assumed to be smaller than $\Theta$.

By contrast the social utility from the default equilibrium where $r/2$ is produced is $U + V\theta - \theta^2/2 = U + Vr/2 - r^2/8$, so the utility gain of the cost minimizing implementation of production over the default equilibrium is

$$\begin{aligned} G(r) &= (V + (\mu - 1/2)r)^2/(4\mu) - (\mu - 1/2)r^2/4 - Vr/2 + r^2/8 \\ &= (V + (\mu - 1/2)r)^2/(4\mu) + (1 - \mu)\, r^2/4 - Vr/2 \end{aligned}$$

which is the expression given in the Theorem.

Plugging in we get $G(0) = V^2/4\mu$, $G(2V) = 0$ and differentiating with respect to $r$ we find

$$\begin{aligned} G'(r) &= (\mu - 1/2)(V + (\mu - 1/2)r)/(2\mu) - (\mu - 1)r/2 - V/2 \\ &= \left[(V + (\mu - 1/2)r)\,(\mu - 1/2) - (\mu^2 - \mu)r - \mu V\right]/(2\mu) \\ &= \mu V - V/2 + (\mu - 1/2)^2\, r - (\mu^2 - \mu)r - \mu V/(2\mu) \\ &= -(V/2 - r/4)/(2\mu) \end{aligned}$$

which is negative for $r < 2V$. It follows that $G(r) > 0$ for $r < 2V$ so it is indeed better to implement production rather than use the default equilibrium. □

**Corollary.** [1 in text] *Let $\overline{r}$ be the positive solution of $G(r) = F$ if one exists, 0 otherwise. Note that for $F$ sufficiently small a positive solution always exists and that $\overline{r} < 2V$. Then for $r < \overline{r}$ it is optimal to implement production and output is $\theta = (V + (\mu - 1/2)\, r)/(2\mu)$ while for $r > \overline{r}$ the default social norm is optimal and output is $r/2 < (V + (\mu - 1/2)\, r)/(2\mu)$. Hence for $\overline{r} = 0$ output increases in $r$ while for $\overline{r} > 0$ output increases in $r$ up to $\overline{r}$, drops discontinuously, and then increases again. In either case social utility is always increasing in $r$.*

*Proof.* Indeed in the two ranges we have

$$\begin{aligned} &\frac{d}{dr}\left(U + (V + (\mu - 1/2)r)^2/(4\mu) - (\mu - 1/2)r^2/4\right) \\ =&2(V + (\mu - 1/2)r)(\mu - 1/2)/(4\mu) - (\mu - 1/2)r/2 \\ \propto&(V + (\mu - 1/2)r)(\mu - 1/2) - (\mu - 1/2)\mu r \\ =&(V - r/2)(\mu - 1/2) > 0 \end{aligned}$$

and

$$\frac{d}{dr}\left(U + Vr/2 - r^2/8\right) = (V - r/2)/2 > 0.$$

□

*Frequency of Intervention*

**Theorem.** [4 in the text] *Suppose $R$ is small and that $V$ is large in the sense that*

$$R \leq (1 - h^2\beta^2)(1 - 2\pi) - c$$

*and*

$$V > \left[\frac{U}{2}\frac{\pi + (1-\pi)h\beta}{(1-h^2\beta^2)(1-2\pi)} + \frac{1}{2}\right](c+R)$$

*Then there is* $0 < \underline{\lambda} \leq \overline{\lambda} < 1$ *such that:*

*For* $\lambda \in [0, \underline{\lambda})$ *the optimal social norm is non-discriminatory and production takes place only when there is no intervention;*

*For* $\lambda \in (\underline{\lambda}, \overline{\lambda})$ *the optimal social norm is discriminatory and production always takes place;*

*For* $\lambda \in (\overline{\lambda}, 1]$ *the optimal social norm is non-discriminatory and production always takes place.*

*Proof.* The first condition can be written as

$$\frac{1}{(1-h^2\beta^2)(1-2\pi)}(c+R) \leq 1$$

which is the implementability condition when there is intervention. If this fails then there will never be production when there is intervention. Define the cost of implementation when there is intervention

$$C_R \equiv \left[\frac{U}{2}\frac{\pi + (1-\pi)h\beta}{(1-h^2\beta^2)(1-2\pi)} + \frac{1}{2}\right](c+R) - \frac{R}{2}.$$

This is the cost to bear for production to always take place; incentives are set to be strong enough to produce with intervention, hence more than enough to produce without it. The corresponding social utility is $U + V - C_R$ - it is what the group gets with a non-discriminatory norm unconditionally inducing production. Since utility in the default equilibrium (no production, no ostracism) conditional on intervention is $U + R/2$ the condition that implementation be preferred to the default when there is intervention is $V - C_R > R/2$ which is the second condition. That is this second condition rules out the case where the group prefers the default to implementation when there is intervention - in which case it would choose not to implement for large $\lambda$. In addition the fact that the group prefers implementation to the default when there is intervention implies that if the fixed cost for discrimination is paid it is best always to produce.

Under these assumptions there are three candidates for optimal social norms: discriminatory and always implement, non-discriminatory and always implement production and non-discriminatory and produce only when there is no intervention. The third case is the new one, so we discuss it first. To do so, we must figure out what social utility is when there is intervention. In this case there is no production so benefit is not $U + V$ but $U + R/2$; and the producer is punished with probability $1 - \pi$ and the monitor with probability $\pi$, where since the norm is non-discriminatory and the producer must produce without intervention, the punishment probabilities are given as before by

$$P = \frac{c}{(1-h^2\beta^2)(1-2\pi)}, Q = h\beta P.$$

In conclusion the resulting utility $U_R$ is given by

$$\begin{aligned} U_R =& U + R/2 - (U/2)\left((1-\pi)P + \pi Q\right) \\ =& U + R/2 - (U/2)\left(1 - \pi + \pi h\beta\right)P \\ =& U + \frac{R}{2} - \frac{U}{2}\frac{1-\pi+\pi h\beta}{(1-h^2\beta^2)(1-2\pi)}c. \end{aligned}$$

Define also the cost of implementation when there is no intervention as

$$C_0 = \left[\frac{U}{2}\frac{\pi + (1-\pi)h\beta}{(1-h^2\beta^2)(1-2\pi)} + \frac{1}{2}\right]c.$$

Notice that $C_0 < C_R$. We can now compute the utility of each of the three alternatives:

| type of norm | production | social utility | at $\lambda = 0$ | at $\lambda = 1$ |
|---|---|---|---|---|
| discriminatory | always | $U + V - (1-\lambda)C_0 - \lambda C_R - f$ | $U + V - C_0 - f$ | $U + V - C_R - f$ |
| non-discriminatory | always | $U + V - C_R$ | $U + V - C_R$ | $U + V - C_R$ |
| non-discriminatory | only when non-intervention | $(1-\lambda)(U + V - C_0) + \lambda U_R$ | $U + V - C_0$ | $U_R$ |

Notice that in all cases social utility is linear in $\lambda$ so picking the best of three linear functions immediately shows that there are at most three intervals in which different social norms can be optimal. As to what those intervals are, the result for $\underline{\lambda}$ now follows from observing that when $\lambda = 0$ non-discriminatory produce only when non-intervention is strictly best, while the upper range follows from the fact that $V - C_R > R/2$ while from above $U_R < U + R/2$. The existence of an intermediate range for $f$ sufficiently small can be seen from the fact that when $f = 0$ the discriminatory social norm of always producing is strictly best except when $\lambda \in \{0, 1\}$. □

*Internalization*

Our goal is to characterize the conditions under which production can be implemented and the degree of internalization and monitoring involved in implementation that minimizes cost. Recall that the implementation of production requires a producer to produce and a monitor to report truthfully. This of course requires the ostracism probabilities to satisfy the relevant incentive constraints. In particular, producers and monitors conditional on committing to the norm that implements production should not want to deviate. When the benefit of commitment is $B$ the incentive constraint for the producer to produce is that the gain from not producing should be no greater than $B$, that is $c - (1-2\pi)(P - h\beta Q) \leq B$ (this we know from previous computations). For the monitor the incentive constraint is that the difference in utility between the good and bad report should not exceed $B$, that is $-B \leq h\beta P - Q \leq B$. Also, the *ex ante* (before the agent finds out his role) expected utility from following the best alternative strategy cannot exceed the extra cost $\Gamma$ in investing in such a strategy, which we call the *overall utility constraint*.

Before proving Theorem 5 we establish a preliminary lemma.

**Lemma 2.** *With internalization, cost minimization requires the directional constraints:* $c - (1 - 2\pi)(P - h\beta Q) \geq 0$ *and* $h\beta P - Q \geq 0$ .

*Proof.* First take the case where the directional (but not incentive) constraint fails for the monitor, so that $-B \leq h\beta P - Q < 0$. The incentive constraint holds for the producer so $c - (1-2\pi)(P - h\beta Q) \leq B$. Since the best alternative strategy involves deviating to 0 as a monitor, the overall utility constraint is $(1/2)\left[\max\{0, c-(1-2\pi)(P-h\beta Q)\} - (1-\pi)(h\beta P - Q)\right] \leq \Gamma$. Not too much lower $Q'$ continues to satisfy $c - (1-2\pi)(P - h\beta Q') \leq B$,

$$(1/2)\left[\max\{0, c-(1-2\pi)(P-h\beta Q')\} - (1-\pi)(h\beta P - Q')\right] \leq \Gamma$$

and $h\beta P - Q' \leq B$. Hence lower $Q'$ is incentive compatible but yields a lower cost.

Second take the case where the directional constraint fails for the producer so that $c - (1-2\pi)(P - h\beta Q) < 0$. Suppose the incentive constraint holds for the monitor and the directional constraint holds strictly so $0 < h\beta P - Q \leq B$. Since the best deviation for the producer is not to deviate, the overall utility constraint is $(1/2)\left[\pi(h\beta P - Q)\right] \leq \Gamma$. Not too much lower $P'$ satisfies $-B \leq h\beta P' - Q \leq B$, $0 < (1/2)\left[\pi(h\beta P' - Q)\right] \leq \Gamma$ and $c - (1-2\pi)(P' - h\beta Q) \leq B$. Hence lower $P'$ is incentive compatible but yields a lower cost.

Finally, take the case where the directional constraint fails for the producer so that $c - (1-2\pi)(P - h\beta Q) < 0$ and the directional constraint holds with equality for the monitor so $h\beta P - Q = 0$. Consider lowering $P$ slightly to $P'$ while also lowering $Q$ to $Q'$ so that the monitor continues to satisfy $h\beta P' - Q' = 0$ Since the best deviation for the producer is not to deviate and the monitor remains indifferent, the overall utility gain from deviating is 0, so the overall utility constraint is certainly satisfied. Not too much lower $P', Q'$ satisfies $c - (1-2\pi)(P' - h\beta Q') \leq B$. Hence the lower $P', Q'$ is incentive compatible but yields a lower cost. □

**Theorem.** [5 in the text] *Suppose* $B, c > 0$. *Define*

$$\begin{aligned} \overline{P} &= \frac{c}{(1-2\pi)(1-h^2\beta^2)} \\ \underline{P} &= \overline{P} \cdot \max\left\{0,\, 1-(1+(1-2\pi)h\beta)\,\frac{B}{c},\, (1-\frac{B}{c})(1-h^2\beta^2)\right\}. \end{aligned}$$

*If* $\underline{P} > 1$ *implementation of production is not possible. If* $\underline{P} \leq 1$ *then there exists* $(1+\pi)B/2 \geq \overline{\Gamma} \geq \underline{\Gamma} \geq 0$ *such that production can be implemented if and only if* $\Gamma \geq \underline{\Gamma}$ *where* $\underline{\Gamma} = 0$ *if and only if* $\overline{P} \leq 1$. *If* $B \geq c$ *then* $\overline{\Gamma} = c/2, \underline{P} = 0$ *and for* $\Gamma \geq \overline{\Gamma}$ *there is* complete internalization: *production is implemented without ostracism, that is* $P = Q = 0$. *If* $\Gamma \geq \underline{\Gamma}$ *the cost minimizing social norm is given by, for generic parameter values, unique continuous piecewise linear functions* $\hat{P}(\Gamma)$ *and* $\hat{Q}(P)$ *with the following properties:*

*1.* $\overline{P} \geq \hat{P}(\Gamma) \geq \underline{P}$ *is defined on* $[\underline{\Gamma}, B]$, *is strictly decreasing on* $[\underline{\Gamma}, \overline{\Gamma}]$, $\hat{P}(\Gamma) = \underline{P}$ *for* $\Gamma \geq \overline{\Gamma}$ *and* $\hat{P}(\underline{\Gamma}) = \min\{1, \overline{P}\}$;

*2.* $h\beta P \geq \hat{Q}(P) \geq 0, h\beta P - B$ *is defined for* $P \geq \underline{P}$ *and is non-decreasing with* $\hat{Q}(\overline{P}) = h\beta\overline{P}$;

*3. [lower function] If* $\pi \leq .333$ *then for* $\hat{P}(\underline{\Gamma}) < P < \hat{P}(\overline{\Gamma})$ *the producer strictly prefers to produce and* $\hat{Q}(P)$ *is convex, non-increasing in* $B$;

*4. [upper function] If $\pi \geq .382$ then for $\hat{P}(\underline{\Gamma}) < P < \hat{P}(\overline{\Gamma})$ the monitor strictly prefers to tell the truth and $\hat{Q}(P)$ is concave, non-decreasing in $B$;*

*5. For $\underline{\Gamma} \leq \Gamma \leq \overline{\Gamma}$ we have $(\pi h\beta - (1-2\pi))\hat{P}(\Gamma) - (\pi - (1-2\pi)h\beta)\,\hat{Q}(\hat{P}(\Gamma)) = 2\Gamma - c$.*

*In addition there is a $\overline{B}$ such that for $B \geq \overline{B}$ the cost minimizing solution $\hat{P}$,$\hat{Q}$ is independent of $B$.*

*Proof.* From Lemma 2 we know that the directional constraints $c - (1-2\pi)(P - h\beta Q) \geq 0$ and $h\beta P - Q \geq 0$ must be satisfied. When these are satisfied, the remaining part of the incentive constraints are $c - (1-2\pi)(P - h\beta Q) \leq B$ and $h\beta P - Q \leq B$. Finally, there is the *overall utility constraint* that the *ex ante* expected utility from following the best alternative strategy cannot exceed the extra cost $\Gamma$ in investing in such a strategy. Given the directional constraints the relevant overall utility constraint is $(1/2)\left[c - (1-2\pi)(P - h\beta Q) + \pi(h\beta P - Q)\right] \leq \Gamma$. There are in addition the probability constraints $0 \leq P, Q \leq 1$. Subject to these constraints, the goal is to minimize punishment cost $(U/2)(\pi P + (1-\pi)Q)$.

The producer incentive and directional constraints can be written as

$$Q \leq \frac{1}{h\beta}\frac{B-c}{1-2\pi} + \frac{1}{h\beta}P \tag{CP}$$

$$Q \geq \frac{1}{h\beta}\frac{-c}{1-2\pi} + \frac{1}{h\beta}P, \tag{DP}$$

and those of the monitor as

$$Q \geq h\beta P - B. \tag{CM}$$

$$Q \leq h\beta P. \tag{DM}$$

The overall utility constraint can be rewritten as

$$(\pi h\beta - (1-2\pi))P - (\pi - (1-2\pi)h\beta)Q \leq 2\Gamma - c.$$

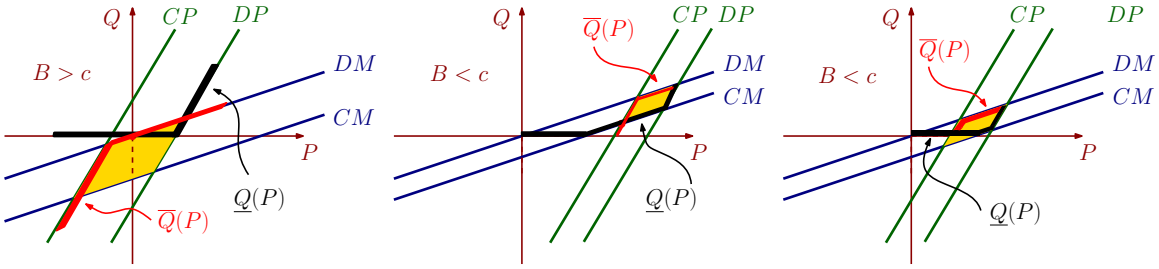
To reiterate: this is an LP problem with nine constraints and the objective is to minimize the ostracism probability $\pi P + (1-\pi)Q$.

We analyze the constraints and objective in $P, Q$ space. We start by organizing the two upper constraints on $Q$ - (CP) and (DM) - and the three lower constraints on $Q$ - $Q \geq 0$, (CM) and (DP). Denote the lower envelope of the two upper constraints (the minimum of the two binding functions of $P$) as the *upper function* $\overline{Q}(P)$: this is a concave strictly increasing function. Similarly denote the upper envelope of the three lower constraints as the *lower function* $\underline{Q}(P)$: this is a convex weakly increasing function. Both functions lie between $\beta hP$ and $\beta hP - B$.

Any feasible solution to the cost minimization problem must lie in between $\underline{Q}(P)$ and $\overline{Q}(P)$ - in fact we will show that the optimal solution lies on one of two. Before doing so observe that the two functions one being concave and one convex can intersect at at most two points. On the right they intersect where the directional constraints (DP) and (DM) are satisfied with equality: this is the solution without internalization - with $\Gamma = 0$, that is, at $P = \overline{P}$. Look next at the intersection

point of $\underline{Q}(P)$ and $\overline{Q}(P)$ on the left. The figure below illustrates.

Figure 7.1: The functions $\underline{Q}(P)$ and $\overline{Q}(P)$




The left panel corresponds to large $B$. In the other two panels $B$ is smaller (the distance between incentive and directional constraints is smaller then in the left panel). The central panel corresponds to a relatively larger $c$ compared to the right panel (the vertical intercept of the (CP) constraint is higher).

If the point where (CP) and (CM) are equalities has positive $P$ that is the intersection given by

$$\begin{aligned} P &= \frac{1}{(1-h^2\beta^2)}\left(-\frac{B-c}{1-2\pi} - h\beta B\right) \\ &= \overline{P}\cdot\left(1-\left[1+(1-2\pi)h\beta\right]\frac{B}{c}\right); \end{aligned}$$

if this is negative then also the $Q$-coordinate, given by $Q = h\beta P - B$, is negative so we move up along (CP) until we reach the horizontal axis; if the $P$-coordinate of that point, given by

$$\begin{aligned} P &= \frac{c-B}{1-2\pi} \\ &= \overline{P}\cdot(1-h^2\beta^2)(1-\frac{B}{c}), \end{aligned}$$

is positive, that is it; otherwise the intersection of $\underline{Q}(P)$ and $\overline{Q}(P)$ has $P = 0$. Denoting the intersection by $\underline{P}$ we have obtained the expression in the statement. Since the solution must lie between $\underline{P}$ and $\overline{P}$, if $\underline{P} > 1$ there is no feasible solution.

Suppose that $\underline{P} \leq 1$. Noting that $\underline{Q}(\underline{P}) = \overline{Q}(\underline{P})$ define the overall utility at the left intersection as $\underline{\Gamma} = (1/2)\left[c + (\pi h\beta - (1-2\pi))\underline{P} - (\pi - (1-2\pi)h\beta)\underline{Q}(\underline{P})\right]$. If $\Gamma \geq \underline{\Gamma}$ we see that the left intersection is the optimal solution, since all other feasible $P, Q$ must be weakly larger than $\underline{P}, \underline{Q}(\underline{P})$. In other words, $\hat{P}(\Gamma) = \underline{P}$ for $\Gamma \geq \underline{\Gamma}$ (*part of property 1*). Observe that the intersection is non-positive and $\underline{P} = 0$ if and only if (CP) is satisfied at $Q = P = 0$, that is, $B \geq c$. In this case the overall utility constraint reads $c \leq 2\Gamma$ so that if $B \geq c$ and $\Gamma \geq c/2$ we achieve complete internalization.

At the right end we observe that at $\overline{P}, h\beta\overline{P}$ the overall utility constraint at $\Gamma = 0$ holds with equality. So if $\overline{P} \leq 1$, then $\hat{P}(\Gamma) = \overline{P}$ and $\hat{Q}(\overline{P}) = h\beta\overline{P}$ is the optimal solution at $\Gamma = 0$, since no other $P, Q$ satisfies all the constraints (*part of property 2*). Hence $\underline{\Gamma} = 0$ if and only if $\overline{P} \leq 1$. In between we must examine the overall utility constraint $(\pi h\beta - (1-2\pi))P - (\pi - (1-2\pi)h\beta)Q \leq 2\Gamma - c$

and the objective function. Observe that the overall utility constraint shifts down in $P, Q$ space as $\Gamma$ increases. The table below shows the possibilities for the signs of coefficients in the constraint and the relative solutions (explained in the sequel):

| | $P$ coefficient | |
|---|---|---|
| $Q$ coefficient | $\pi h\beta - (1-2\pi) < 0$ | $\pi h\beta - (1-2\pi) > 0$ |
| $-\pi + (1-2\pi)h\beta < 0$ | depends | upper solution |
| $-\pi + (1-2\pi)h\beta > 0$ | lower solution | impossible |

Proof that the lower corner is impossible: if the $P$ coefficient is positive then$\pi h\beta > 1-2\pi$ implying $\pi > (1-2\pi)h\beta$ so that the $Q$ coefficient is negative.

If the $Q$ coefficient is negative we can write the overall utility constraint as

$$Q \geq \frac{-2\Gamma + c + (\pi h\beta - (1-2\pi))P}{\pi - (1-2\pi)h\beta};$$

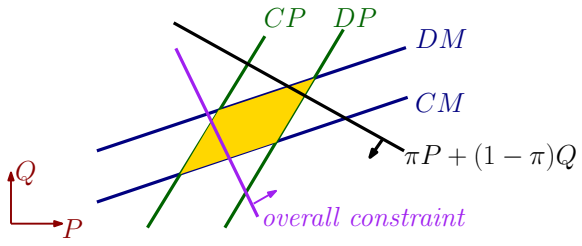
if in addition the $P$ coefficient is positive then the constraint is upward sloping, and $Q$ lies above. Moreover, the constraint is flatter than the upper function: to see this observe that at the intersection of (DM) and (DP) on the right side the constraint holds exactly with $\Gamma = 0$; as we move leftwards along the (DM) constraint the monitor remains indifferent and the (DP) constraint strictly favors not producing so that the expected overall utility gain from not producing is positive and the constraint (with $\Gamma = 0$) is violated. It follows that the solution, $\hat{P}(\Gamma), \hat{Q}(\hat{P})$ is the intersection of the upper function with the overall utility constraint binding, for all $\underline{\Gamma} \leq \Gamma \leq \overline{\Gamma}$, where $\underline{\Gamma} = (1/2)\left[c + (\pi h\beta - (1-2\pi))(\min\{1, \overline{P}\}) - (\pi - (1-2\pi)h\beta)\overline{Q}(\min\{1, \overline{P}\})\right]$.

If the $P$ coefficient is negative we can write the overall utility constraint

$$P \geq \frac{-2\Gamma + c - (\pi - (1-2\pi)h\beta)\, Q}{(1-2\pi) - \pi h\beta};$$

if in addition the $Q$ coefficient is positive then the constraint is upward sloping, $P$ lies to the right. Moreover, the constraint is steeper than the lower function: this follow from the fact that if we start as before at the intersection of (DP) and (DM) and we move downwards along the (DP) constraint the producer remains indifferent and the (DM) constraint strictly favors filing a good report so that the expected overall utility gain from not producing is positive and the constraint is violated (with $\Gamma = 0$). It follows that the solution, $\hat{P}(\Gamma), \hat{Q}(\hat{P})$ is the intersection of the lower function with the overall utility constraint binding, for all $\underline{\Gamma} \leq \Gamma \leq \overline{\Gamma}$, where $\underline{\Gamma} = (1/2)\left[c + (\pi h\beta - (1-2\pi))(\min\{1, \overline{P}\}) - (\pi - (1-2\pi)h\beta)\underline{Q}(\min\{1, \overline{P}\})\right]$.

What if both coefficients are negative? The constraint is negatively sloped and the feasible set lies above and to the right. If the constraint is steeper than the indifference curves the solution is the lower function; if it is flatter than the indifference curves the solution is the upper function. An illustration is given in the figure below:




The slope of the constraint in $(P, Q)$ space is

$$-\frac{-\pi h\beta + (1-2\pi)}{\pi - (1-2\pi)h\beta}$$

while the slope of the indifference curves is $-\pi/(1-\pi)$; so we get the lower function intersected with the binding overall utility constraint if

$$\frac{-\pi h\beta + (1-2\pi)}{\pi - (1-2\pi)h\beta} > \frac{\pi}{1-\pi}.$$

and the upper function intersected with the binding overall utility constraint if the inequality is reversed. This condition (since numerator and denominator on both sides are positive) is equivalent to

$$\begin{aligned}\pi^2 - \pi(1-2\pi)h\beta &< -(1-\pi)\pi h\beta + (1-\pi)(1-2\pi)\\ \pi^2 + 2\pi^2 h\beta &< \pi^2 h\beta + (1-\pi)(1-2\pi)\\ \pi^2 h\beta &< 1 - 3\pi + \pi^2.\end{aligned}$$

This is true for all $h\beta$ if it is true for $h\beta = 1$, that is to say $\pi < 1/3$. The reverse inequality is true for all $h\beta$ if it is true for $h\beta = 0$ which is to say $\pi^2 - 3\pi + 1 < 0$, which gives $\pi > (3 - \sqrt{9-4})/2 \approx 0.38197$.

We observe that if $\pi < 1/3$ then the coefficient on $P$ is $\pi h\beta - (1-2\pi) < -(1-3\pi) < 0$ so negative. On the other hand if $\pi > 1/3$ then the coefficient on $Q$ is $-\pi + (1-2\pi)h\beta \leq 1 - 3\pi < 0$ so negative. Note that the upper function is non-decreasing in $B$ while the lower function is non-increasing in $B$. Since the binding overall utility constraint is independent of $B$, this means that $\hat{Q}(P)$ is non-increasing in $B$ when $\pi < 1/3$ and non-decreasing in $B$ when If $\pi \geq .382$. This concludes our characterization of the functions $\hat{P}(\Gamma)$ and $\hat{Q}(P)$. *Properties 2, 3 and 4* follow from this characterization as does the first part of *property 1*, that $\overline{P} \geq \hat{P}(\Gamma) \geq \underline{P}$ is strictly decreasing on $[\underline{\Gamma}, \overline{\Gamma}]$.

The final part 5 of the theorem simply intersects the binding overall utility constraint with the upper or lower function as appropriate - we have already showed that this is indeed how we find the cost minimizing $P$.

Finally we wish to establish that $\overline{\Gamma} \leq (1+\pi)B/2$ and the existence of $\overline{B}$. The former follows

from the fact that the overall utility constraint is

$$(1/2)\left[c-(1-2\pi)(P-h\beta Q)+\pi(h\beta P-Q)\right]\leq\Gamma$$

while from the incentive constraints $c-(1-2\pi)(P-h\beta Q)\leq B$ and $h\beta P-Q\leq B$ so that the overall utility constraint cannot bind for $(1/2)(1+\pi)B<\Gamma$.

For $\overline{B}$ observe that if $B>c$ then the producer incentive constraint is always satisfied when the non-negativity and directional constraints are satisfied; and when the non-negativity and directional constraints are satisfied we have

$$\begin{aligned}h\beta P-Q&\leq h\beta[c/(1-2\pi)+h\beta Q]-Q\\&=h\beta c/(1-2\pi)-Q(1-h^2\beta^2)\leq h\beta c/(1-2\pi)\end{aligned}$$

so the monitor incentive constraint holds if $h\beta c/(1-2\pi)\leq B$. Hence when

$$\overline{B}=c\max\{1,\frac{h\beta}{1-2\pi}\}$$

for $B\geq\overline{B}$ the solution is determined entirely by the overall utility and directional constraints, none of which depend on $B$. □