INCENTIVES AND PROSOCIAL BEHAVIOR

Roland Bénabou (Princeton University)

Jean Tirole (IDEI Toulouse)

Hammamet, May 26, 2005

INTRODUCTION

- ✓ Participation in costly, low-individual-benefit activities (vote, volunteer, give to charitable organizations, help strangers, join rescue squad, risk life,…)

- ✓ Cannot be explained by sole presence of individuals with other-regarding preferences:

  (a) *Crowding out effects*

  [e.g., Festinger-Carlsmith 59, Deci-Ryan 85, Gneezy-Rustichini 00a,b, many papers by Frey, Fehr and co-authors.]

  (b) *Social glory and shame attached to good and selfish deeds*

  [codes of honor, shame; conspicuous donations; pressure (Batson 98, Freeman 97).]

  (c) *Self-image concerns*

  [Adam Smith 1776; Dana et al 03a,b, Murnighan et al. 01 on evidence in dictator games]

Approach

Model of prosocial behavior that combines:

- ✓ Heterogeneity in individuals' degrees of altruism and greed.
- ✓ Concern for social reputation and self-respect: people
  - want to signal to others desirable traits (generosity, fairness, public spirit, lack of greed, identity attributes,…),
    [Veblen; Leibenstein 50, Bernheim 84, Pesendorfer 95, Denrell 98…]
  - strive to maintain a certain view of "what kind of person" they are. Cognitive approach (retrospective justification): individuals use their past behavior as "diagnostic" of their deep preferences.
    [Bem 72, Quattrone-Tversky 84, Bodner-Prelec 03, Bénabou-Tirole 04.]

Yields:

- ✓ Analysis of how the three motives for prosocial behavior interact, and how this balance is affected by changes in the material and informational *environment* (contribution in time vs. money; observability, memorability…) and in *policies* (power of incentives, disclosure of rewards…).
- ✓ *Information-based approach to crowding out*:
  Rewards or punishments spoil reputational value of good deeds by creating (self) doubt as to their underlying motivation

"An intrinsically motivated person is deprived of the chance of displaying his or her own interest and involvement in an activity when someone else offers a reward, or orders him/her to do it" [Frey-Jegen 02.]

In line with "overjustification effect" in psych. [e.g., Lepper et al 73]

- ✓ *Equilibrium analysis (endogenizing rewards), welfare.*

Two (informational) crowding-out mechanisms

	Impact on individual's perception of task or self	Self- or social signaling
Mechanism	Conveyance of bad news about nature of task, its payoffs, or individual's ability	"Intrinsically motivated individual deprived of the chance of displaying his or her own interest and involvement in activity" [Frey-Jegen 2002]
Example	Rewarding child for task, school...	Blood donation
Impact of reward	• Limits immediate reinforcement. • Crowds out future re-engagement	Immediate
	"Intrinsic and Extrinsic Motivation" (*RES* 03)	"Incentives and Prosocial Behavior"

Outline

- Model.
- Heuristics of image-spoiling effect of rewards.
- Crowding out: signal-extraction & the overjustification effect
- Social norms of socially and personally acceptable behaviors (are individual decisions strategic complements or substitutes?)
- Design of contracts: reward, confidentiality, welfare.
- Sponsor competition.

I. THE MODEL

- Agents choose their participation in some prosocial activity

  [provision of public good, contribution to a worthy cause, engaging in friendly action, refraining from imposing a negative externality, voting,...]

- Each selects participation level a: discrete (0/1) or continuous.
- Cost $C(a)$, monetary reward $y \cdot a$ $\quad (y \gtrless 0)$.
- Incentive rate may reflect a proportional subsidy or tax on a.

  Exogenous for now.

❑ *Direct benefits from pro (or anti) social activity:*

$$(\underbrace{v_a}_{\text{intrinsic}} + \underbrace{v_y y}_{\text{extrinsic / material}})\, a - C(a)$$

An individual's preference type or "identity" $\boldsymbol{v} = (v_a, v_y)$ is drawn from continuous distribution.

❑ *Social esteem / self-esteem: reputational concerns*

$$x\,[\gamma_a E(v_a|a,y) - \gamma_y E(v_y|a,y)], \quad \text{with} \quad \gamma_a \geq 0 \quad \text{and} \quad \gamma_y \geq 0.$$

wants to appear / see himself as prosocial

wants to appear / see himself as disinterested

probability that behavior is observed by others (social signaling) or remembered (self signaling). Also length of record, number of people who hear of it, etc.

Letting $\mu_a \equiv x\,\gamma_a$ and $\mu_y \equiv x\,\gamma_y$, overall utility is:

$$(v_a + v_y y)a - C(a) + \mu_a E(v_a|a,y) - \mu_y E(v_y|a,y).$$

which each agent will maximize over a, taking into account how his behavior will be interpreted (in equilibrium).

Two interpretations of the model:

- *Social signaling concerns:* better reputation may allow the individual to be matched with more desirable partners, e.g. Gintis et al. (2001) or Seabright (2003), or have pure consumption value.

- *Self-signaling - identity concerns:* self-image may have affective value (people enjoy feeling generous or disinterested; see e.g., Akerlof and Dickens (1987), Köszegi (2000), Landier (2000), Bodner and Prelec (2003)), instrumental value (providing motivation to undertake and persevere in long-term tasks or social relationships; see, e.g., Carrillo and Mariotti (2000), Bénabou and Tirole (2002)), or both. Manipulating it will require *imperfect awareness / imperfect recall of own preferences / motives* in any case.

II. THE IMAGE-SPOILING EFFECT OF REWARDS: BASIC INTUITIONS

- Binary decision: $a = 0$ (cost = 0) or $a = 1$ (cost c_a).

  Participation iff:

$$\underbrace{v_a - c_a}_{\text{intrinsic}} + \underbrace{v_y y}_{\textit{extrinsic}} + \underbrace{R(y)}_{\textit{reputational}} \geq 0$$

where

$$R(y) \equiv \mu_a [E(v_a \mid 1,y) - E(v_a \mid 0,y)] - \mu_y [E(v_y \mid 1,y) - E(v_y \mid 0,y)].$$

a) No reward, $y = 0 \Rightarrow$ participates if

$$v_a \geq c_a - R(0) \equiv v_a^*.$$

$\Rightarrow$ nothing is learned about v_y, and participation occurs for intrinsic motivation above a certain threshold.

So reputation for those who participate is

$$E\left(v_a \mid v_a \geq v_a^*\right) \equiv M^+\left(v_a^*\right)$$

and for those who do not

$$E\left(v_a \mid v_a \leq v_a^*\right) \equiv M^-\left(v_a^*\right)$$

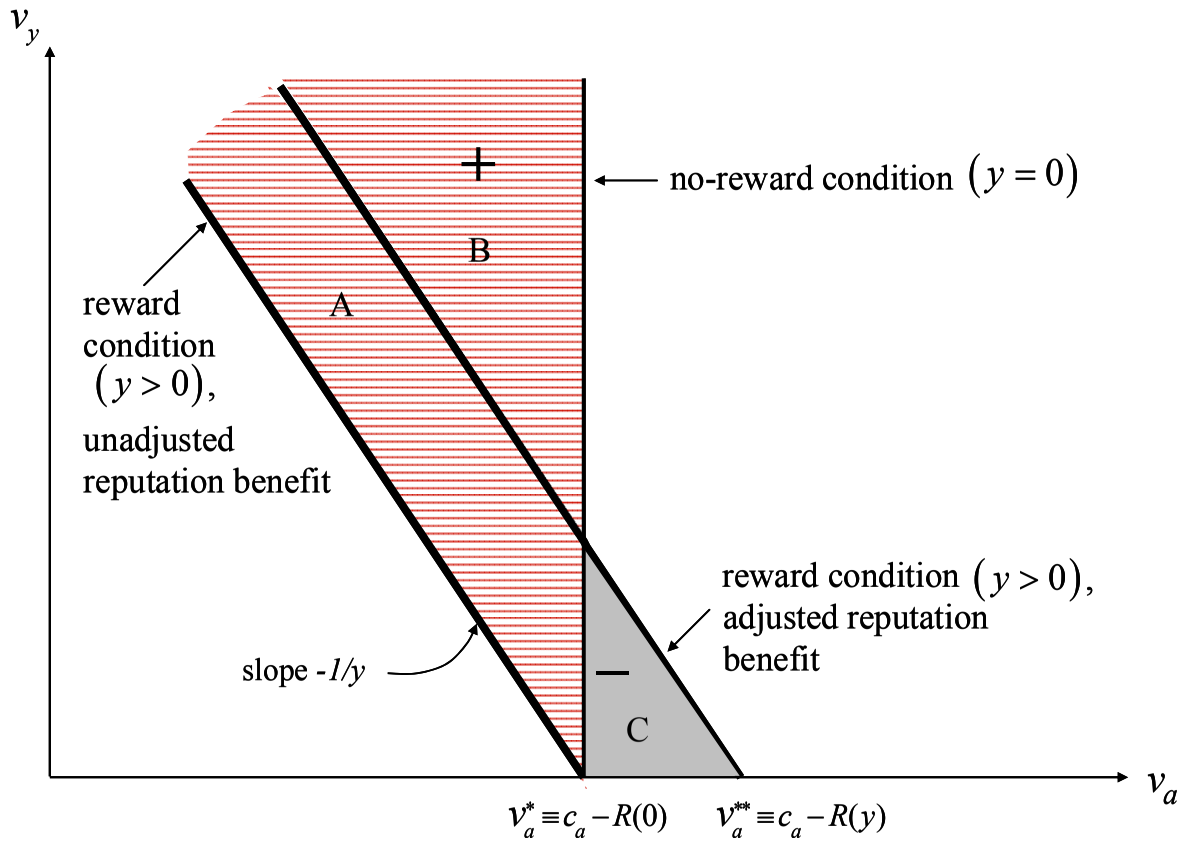
$\Rightarrow$ threshold for participation v_a^* (when interior) is defined by

$$\Psi(v_a^*) \equiv v_a^* + \underbrace{\mu_a\left[M^+\left(v_a^*\right) - M^-\left(v_a^*\right)\right]}_{R(0)} = c_a,$$

$M^+(v_a)$ governs *honor,* while $M^-(v_a)$ governs *stigma.* Both $\nearrow$ in v_a.

b) With reward $y > 0 \Rightarrow$ participates if

$$v_a + v_y y \geq c_a - R(y)$$



- First step: ignoring changes in inference
  - New participants have lower v_a: honor from $a = 1$ declines, stigma from $a = 0$ increases ⇒ net effect a priori ambiguous
  - New participants are greedy types ⇒ adverse reputational effect.

- Then: *equilibrium reputation adjusts*, from *R(0)* to *R(y)*.

- If *R* declines, more greedy types attracted but more prosocial ones repelled. Overall *supply may decline* if area *C* > area *B*.

A first result on the reputational impact of rewards

Proposition *Let* $\Psi' \geq 0$ *and assume the lower bound for* v_y *is 0. Then, if either*

- $\mu_y = 0$*: reputation bears on* v_a *only,*

or

- (v_a , v_y) are *independent or negatively affiliated,*

the introduction of a reward lowers the reputational value of participation: for all $y > 0$, $R(y) < R(0)$.

Intuition: negative correlation between v_a and $v_y \Rightarrow$ beliefs about prosocial orientation and greed tend to be updated in opposite directions $\Rightarrow$ agents who contribute only in response to $y > 0$ must pay a "double dividend" in lost reputation.

III. THE OVERJUSTIFICATION EFFECT AND CROWDING OUT

- Continuous $a \in$ R, cost $C(a)=ka^2/2$.
- Both $\boldsymbol{v}$ and $\boldsymbol{\mu}$ may vary across individuals; private information.
- Normal distribution:

$$\mathbf{v} \equiv \begin{pmatrix} v_a \\ v_y \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{v}_a \\ \overline{v}_y \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ay} \\ \sigma_{ay} & \sigma_y^2 \end{bmatrix} \right), \quad \overline{v}_a \gtrless 0, \quad \overline{v}_y > 0,$$

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_a \\ \mu_y \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{\mu}_a \\ \overline{\mu}_y \end{pmatrix}, \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \right), \quad \overline{\mu}_a \geq 0, \quad \overline{\mu}_y \geq 0,$$

with $(\boldsymbol{v},\boldsymbol{\mu})$ independent. Optimal decision for agent with type $(\boldsymbol{v},\boldsymbol{\mu})$:

$$v_a + v_y y + R'(a;y) = C'(a),$$

where

$$R'(a;y) \equiv \mu_a \frac{\partial E(v_a \mid a;y)}{\partial a} - \mu_y \frac{\partial E(v_y \mid a;y)}{\partial a}.$$

External or internal observer's inference problem: from action a, knows only *the sum of the three sources of motivation.*

$$v_a + v_y y + R(a; y) = C'(a)$$

Signal-extraction problem: key intuitions

- In trying to infer intrinsic motivation v_a, extrinsic part $v_y y$ acts as a source of (possibly correlated) noise, which amplifies with reward y.

  => trying to foster prosocial behavior by increasing y will tend to crowd out reputational motivation $R(a,y)$.

  Classical "overjustification effect" of rewards (or their absence).

- When agents also differ in their degrees of image-concern **μ,** the reputational incentive $R(a,y)$ is itself a further source of (endogenous) noise in inferring v_a or v_y. Wonder whether a is done for appearances.

  => trying to foster prosocial behavior by making glory and shame more observable / public (scaling up the **μ**'s) is self-limiting: also has a negative feedback on $R(a,y)$.

  "*Overjustification effect" of publicity praise/blame (or its absence).*

(1)

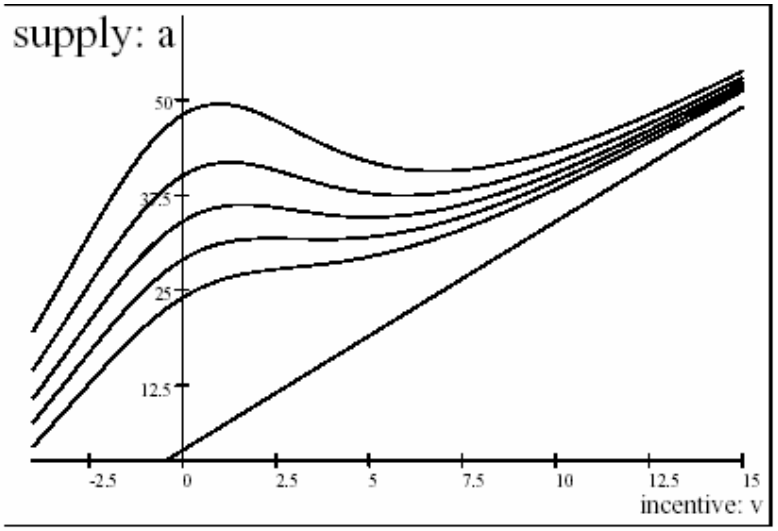


Figure 2a: varying μ_a (with $\mu_y = 0$). The straight line corresponds to $\mu_a = 0$ (no reputation concern)

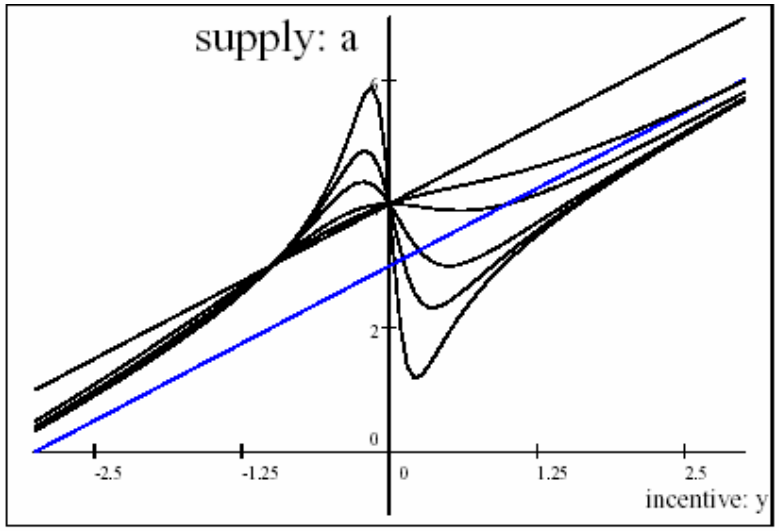


Figure 2b: varying $\theta = \sigma_y/\sigma_a$ (with $\mu_a = 0$). The bottom straight line corresponds to $\mu_y = 0$ (no reputation concern), the top one to $\theta = 0$ (standard one-dimensional signaling model)

(2) **Proposition** ***(small net incentives and signal-reversal).*** *1) Small rewards or punishments are counterproductive, $\bar{a}'(0) < 0$, whenever*

$$\frac{\bar{v}_y}{k} < \bar{\mu}_a \left(\frac{\sigma_{ay}}{\sigma_a^2} \right) + \bar{\mu}_y \left(\frac{\sigma_y^2 - 2\sigma_{ay}^2/\sigma_a^2}{\sigma_a^2} \right).$$

2) Let $\bar{\mu}_y > 0$ and assume that v_a and v_y are uncorrelated, or more generally not too correlated. Then, as $\sigma_a/\sigma_y = 1/\theta$ becomes small, the slope of the supply function at $y = 0$ tends to $-\infty$.

3) Suppose that participation entails a unit opportunity cost with monetary value w. Then $\bar{a}'(w) < 0$ and $\bar{a}'(w) \to -\infty$ under the conditions stated in (1) and (2) respectively.

(2) Variability of image concerns ((μ_a , μ_y) differ).

- *New form of overjustification effect*: good actions come to be suspected of being image-motivated.

Individual behavior more noisy measure of true preferences (v_a , v_y).

- Implications for the effects of prominence, memorability and publicity

[medals, honorific titles and diplomas, naming rights, non-anonymous donations, public praise and blame (e.g., in schools), televised arrests, pillory…]

Increase in x ⇔ homothetic increase in (μ_a , μ_y).

→
- direct amplifying impact
- dampening effect: observers increasingly ascribe behavior to image concerns.
- Example where reputational contribution to aggregate supply grows only like $x^{1/3}$.

IV. HONOR, STIGMA, AND SOCIAL NORMS

- What makes a behavior socially or morally unacceptable is often the very fact that "it is just not done". But in other times, other places: "everyone does it".
  [choosing surrender over death, not going to church, not voting, divorce, welfare dependency, minor tax evasion, conspicuous modes of consumption,....]
- People contribute more when they know / see that others do [public goods, fundraising, voting; helping strangers, Salvation Army exps.]
- Often explained / modeled by postulating some a priori complementarity in preferences / payoffs (e.g., between v_a and $\bar{a}$: untargeted "reciprocity", "social norm'', etc.)
- In fact: *complementarities arise endogenously* from the interplay of honor and shame.
- More generally: when are individuals' contributions strategic complements or substitutes?

- Focus on reputation over social orientation: $v_y \equiv 1$ known, whereas v_a is unknown: density g on $[v_a^- , v_a^+]$.

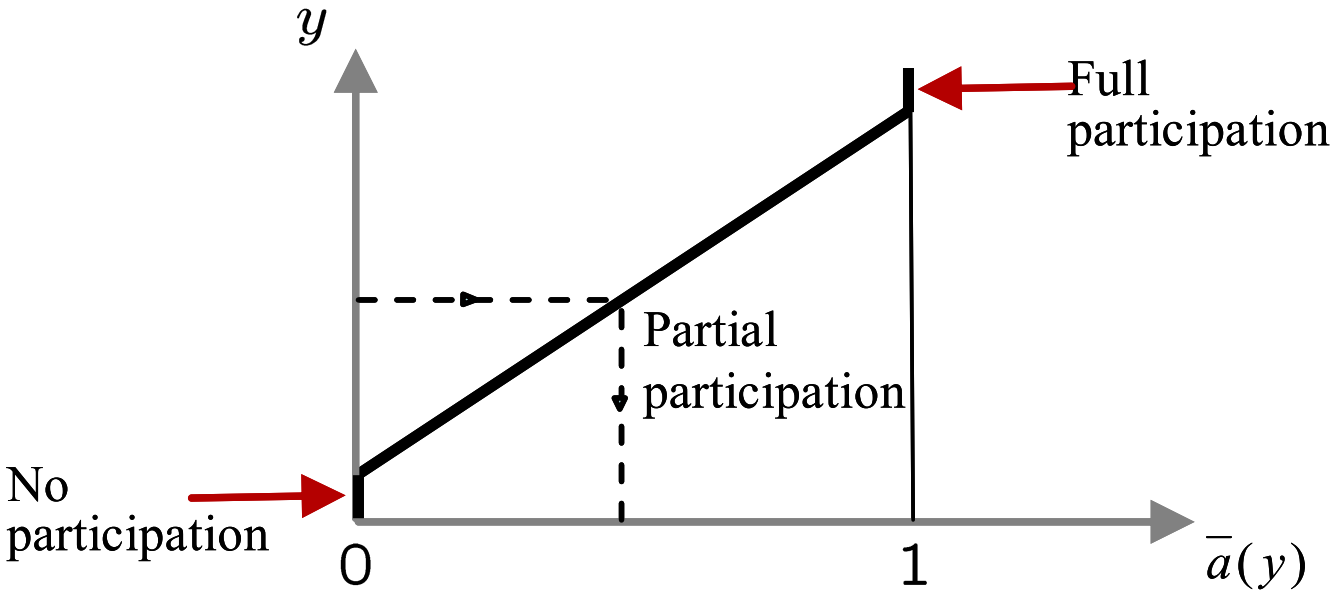
  - An agent participates if

$$v_a \geq c_a - y - R(y) \equiv v_a^*(y)$$

  - Participation threshold v_a^* is defined by comparing the net cost $c_a - y$ with the reputation-adjusted return $\Psi(v^*_a)$, where

$$\Psi(v_a) \equiv v_a + \mu_a[\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a)] \equiv v_a + \mathcal{R}(v_a).$$

Proposition 1 *If Ψ is increasing ($\mathcal{R}$'>-1) the equilibrium is unique. If Ψ is decreasing ($\mathcal{R}$'<-1) or non-monotonic, there is a range of rewards y over which multiple equilibria coexist.*

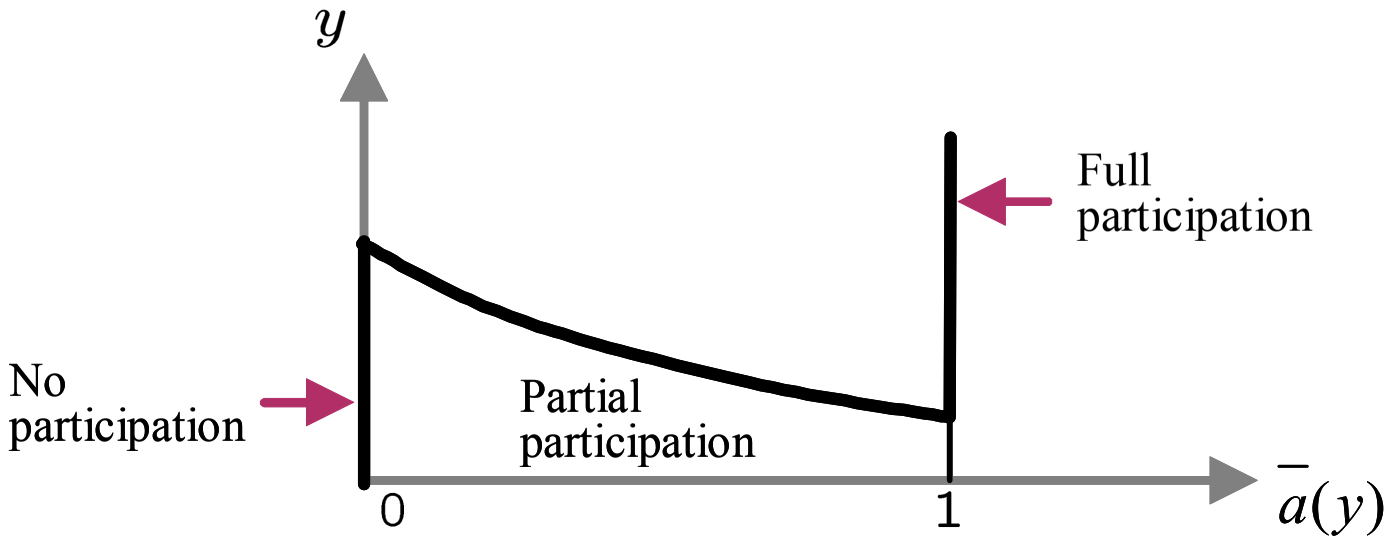


Examples:

$g(v_a) = 2v_a$

on $[0,1]$

$\mu_a < 3/2$

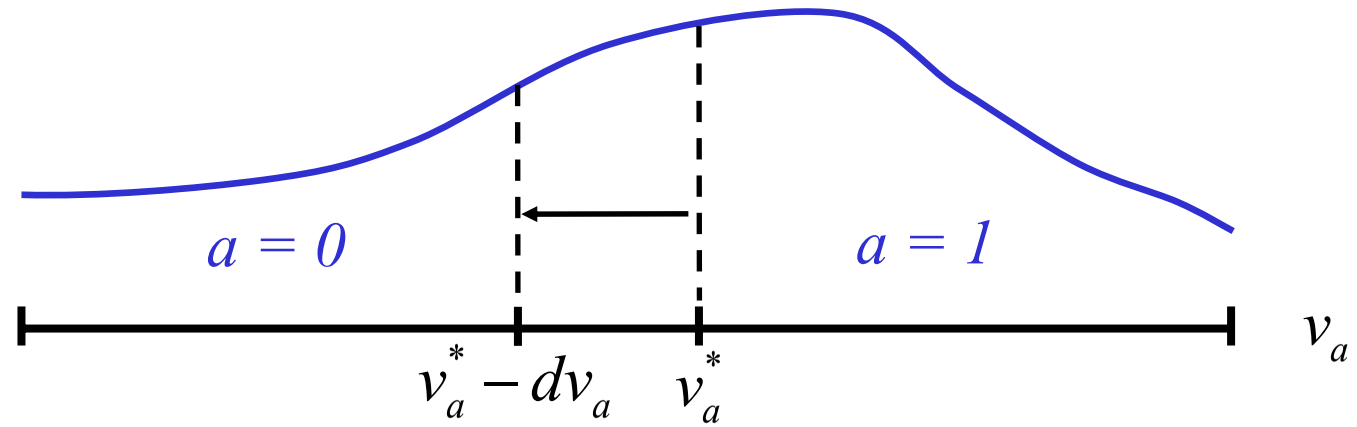


$g(v_a) = 2v_a$

on $[0,1]$

$\mu_a > 6$

Complementarity / substitutability: intuition



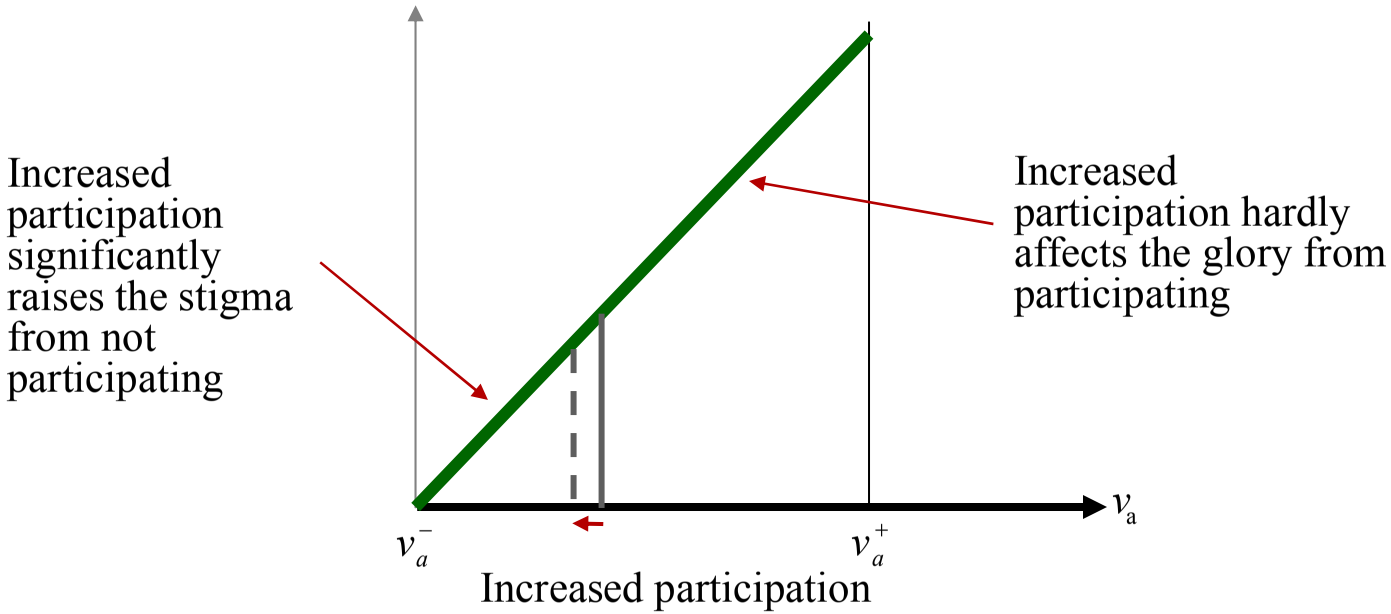
- When $\mathcal{R}' = \mu_a (M^+ - M^-)' < 0,$ a wider participation worsens the pool of abstainers more than that of contributors ⇒ the *stigma* from abstention rises *faster* than the *honor* from participation declines.
- If $\mathcal{R}' < -1,$ net increase in reputational pressure is strong enough that agents in $[v_a^* - dv_a, v_a^*]$, who initially abstained, now feel compelled to contribute ⇒ further increases participation and confines abstention to an even worse pool, etc., pushing towards corner solutions.
- When $-1 < \mathcal{R}' < 0,$ complementarity is weak enough that marginal agents still prefer to stay out ⇒ stability; a fortiori when there is *susbtitutability*, $\mathcal{R}' > 0$.

Sources of strategic complementarity $(\mathcal{R}'/\mu_a = (M^+ - M^-)' < 0)$

Intuition: factors that *accentuate stigma* from $a = 0$ and /or *dampen glory* from $a = 1$ facilitate the emergence of *SC*, and endogenous social norms. Vice versa for *SS*.

- Applies when "most people" have high v_a and only a few "bad apples" have a low value.

Proposition *(Jewitt (2004)): if* v_a *has increasing density,* $\mathcal{R}' < 0$. *Conversely,* $\mathcal{R}' > 0$ *for decreasing density.*



➢ *Possibility of participation "without merit".*

With probability δ, an individual is "forced" to contribute (constraint, strong extrinsic incentives), or finds it unusually cheap. Such circumstances are unobservable by others.

$\Rightarrow$ $M^{NP}(v_a) = M^{-}(v_a)$ unchanged, but $M^{+}(v_a)$ is replaced by a weighted average $M^{P}(v_a;\delta)$ of prior mean $\overline{v}_a$ and $M^{+}(v_a)$.

Proposition *If SC occurs for some δ, it occurs for all $\delta' > \delta$.*

➢ *Excuses*

With probability δ, an individual faces circumstances (unobservable to others) that prevent participation.

Same principle, but now stigma rather than merit is "diluted".

Proposition *If SS occurs for some δ, it occurs for all $\delta' > \delta$.*

V. EQUILIBRIUM CONTRACTS

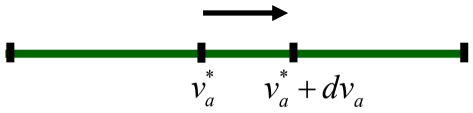
1. Monopoly sponsor sets reward: $y^m \equiv \arg\max_y \{(B-y)\bar{a}(y)\}$

Proposition *Assume v_a unknown, $v_y \equiv 1$ known, and $\Psi' > 0$. A monopoly sponsor may offer contributors a reward that is too high from the point of view of social welfare.*

- $v_y = 1 \Rightarrow$ aggregate welfare = sponsor's profit + individuals' surpluses.
- $\Psi' > 0 \Rightarrow$ unique cutoff $v_a^* = v_a^*(y)$ and upward-sloping $\bar{a}(y) = 1 - G(v_a^*(y))$.
- Marginal $\searrow$ in y from $y^m \Rightarrow$ negligible effect on sponsor's profits.
- Consumers: net gain if

$$\underbrace{\left(\frac{g(v_a^*)}{\Psi'(v_a^*)}\right)}_{\text{\# of agents who stop contributing}} \underbrace{\left(\mu_a\left[M^+(v_a^*) - M^-(v_a^*)\right]\right)}_{\text{reputational loss of each one}} > \underbrace{1 - G(v_a^*)}_{\text{\# of inframarginal contributors who get smaller reward}}$$

total reputation lost by marginal switchers = total reputation gained by inframarginal contributors and non-contributors (martingale property of beliefs). Monopolist internalizes the first, but not the second.



2. Competition::

$$y = B$$

➡ competition often reduces welfare.

3. Fee confidentiality?

❑ Suppose the sponsor can choose between two policies:

- *confidentiality*: only the agent knows the level of y offered (participation remains publicly observable)
- *public disclosure* of the terms of the contract.

Proposition (i) It is optimal for the sponsor to publicly disclose the fee.

(ii) SC ($\mathcal{R}'<0$) $\Rightarrow$ the sponsor offers a higher fee $y^D > y^C$ and elicits a higher participation under disclosure than under confidentiality. Public announcement of y^D is credible (renegotiation-proof).

(iii) SC ($\mathcal{R}'>0$) $\Rightarrow$ announced fee $y^D < y^C$ is lower under disclosure, and so is participation. But requires commitment ability: if secret renegotiation is feasible, the reward will be increased to y^C.

Intuition

- Under public disclosure (but not confidentiality), strategic complementarity creates a "bandwagon effect" that raises the slope of the supply curve $\Rightarrow$ makes announcing a higher fees profitable. Ex-post, agents will insist on receiving them.

- Strategic substitutability has the converse effect on supply $\Rightarrow$ lower announced fees. But in this case both the sponsor and the participants would agree to increase them ex-post, if they could do so secretly.

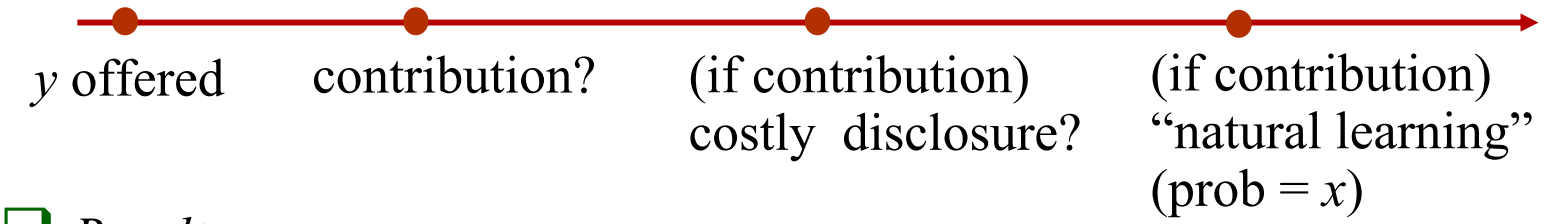
Examples

- Charitable organization contracting with public figures, celebrities, etc. for their participation to a benefit concert, fundraising gala, etc.
- University inviting speakers to give big public lectures.

4. Conspicuous vs. anonymous generosity

People often react with disapproval when someone tries to buy social prestige by revealing how generous, disinterested, etc., they are. Conversely, the most admired contributions and sacrifices are the anonymous ones.

- *Modeling:* probability x that others find out without disclosure; probability 1 if disclosure.



- *Results:*

(1) Strategic complementarities in disclosure (stigma from real or perceived non-participation increases when others disclose)

→ multiple "disclosure norms".

(2) If μ_a unknown as well, disclosure carries a stigma (signals individual image-conscious), which may discourage disclosure.

➢ Welfare–reducing competition

Introduce non-price competition. Show that free entry may reduce welfare, as leads to sponsors screening contributors in inefficient ways.
Illustrations: religions and sects competing in asceticism, rituals, sacrifice. Marathons, walks etc. associated to fundraising for worthy causes.

❑ Modelling

- no entry cost ($k = 0$)
- $v_y \equiv 1$, whereas $v_a = v_a^H$ (prob. ρ) or $v_a = v_a^L$ (prob. 1-ρ)
- non-monetary cost of contributing is c_a, unless sponsor demands a (verifiable) "sacrifice" ⇒ becomes, for low and high type respectively

$$c_a^L >> c_a^H > c_a$$

- Sacrifice = pure deadweight loss; only benefit for the sponsor is to help screen the agents, because it is less costly for the more motivated. Will assume c_a^L large enough that low type is never willing to sacrifice.

Proposition *A monopoly sponsor who wants both types to contribute does not screen contributors inefficiently. By contrast, competing sponsors may require high- valuation individuals to make costly sacrifices that represent pure deadweight losses, thereby reducing total welfare.*

Intuition: non-price screening imposes a negative externality on low-type agents, the cost of which a monopolist serving whole market (with two prices) must fully bear, but which competitive sponsors do not internalize.

Screening by requiring costly sacrifices:

(a) inflicts a deadweight loss $c_a^H - c_a$ on the high type, which sponsor must somehow pay for;

(b) boosts the high type's reputation and lowers that of the low type.

❑ *Monopolist* serving both types appropriates reputational gain (via lower reward) but must compensate for reputational loss. Martingale property => net effect of (b) on his profits and on welfare is zero, leaving only (a) => requiring sacrifices not profitable.

❑ *Free entry*

- v_y known => agent's choice of y has no reputational consequence => price competition will drive all sponsors to offer B.
- By requiring costly sacrifice, entrants can now attract the high types away from competitors who impose no such requirement, leaving low-type (or their sponsors) with the resulting reputational loss. Unlike the monopolist, they do not internalize this negative externality.
- This "cream-skimming" leads to an equilibrium where
  - all active sponsors offer a reward of B
  - a proportion ρ require sacrifice and serve the high-types;
  - remaining $1-\rho$ require only the normal contribution c_a and serve only the low types.
- *Welfare:* can show that both types of agents are better off under competition than under monopoly. The sponsors (or their beneficiaries) however, must necessarily lose more than all agents gain. Indeed:

- total participation remains unchanged (both types still behave prosocially)

- the same is true of average reputation (by the martingale property),

- rewards are pure transfers.

- there is now, however, a deadweight loss of $\rho(c_a^H - c_a)$,

corresponding to the sacrifices made by the high-types to separate.

=> *competition unambiguously reduces welfare.*

SUMMARY

3 motivations for prosocial behaviors:

- altruistic
- material self-interest
- social or self image concerns.

➢ Interactions between them + response to different environments are key

➢ Altering rewards or visibility *changes the meaning* attached to prosocial or antisocial behavior, and feeds back onto reputational incentive to engage in it.

Four main themes

❑ *Rewards and punishments*

- Spoiling effect can result in crowding out.
- Sponsor may announce low rewards, and offer higher ones in private.
- Contributors may refrain from turning down rewards by fear of signaling a high image consciousness.

❑ *Publicity and disclosure*

- Some prominence / memorability encourages prosocial behavior.
- Too much may backfire (increase in signal-to-noise ratio when heterogeneity in image consciousness).

- *Spillovers and social norms*
  - Multiple norms. Factors of strategic complementarity / substitutability.
  - Monopoly sponsor may offer rewards that are too generous.
- *Market for prosocial contributions*
  - Reversal of Bertrand competition.
  - Welfare-reducing competition.

BACKUP

Note $R(a;y)$ is independent of v_a and v_y, *conditionally on* $a \Rightarrow$ acts like a heteroskedastic shock, with mean and variance

$$\bar{R}(a;y) \equiv \bar{\mu}_a \frac{\partial E\left(v_a|a;y\right)}{\partial a} - \bar{\mu}_y \frac{\partial E\left(v_y|a,y\right)}{\partial a}$$

$$\Omega(a,y)^2 \equiv \left(\frac{\partial E(v_a|a,y)}{\partial a} \quad -\frac{\partial E(v_y|a,y)}{\partial a} \right) \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \begin{pmatrix} \frac{\partial E(v_a|a,y)}{\partial a} \\ -\frac{\partial E(v_y|a,y)}{\partial a} \end{pmatrix}.$$

Standard *signal-extraction* results: knowing $v_a + v_y y + R(a;y)$,

$$E\left[v_a|a;y\right] = \bar{v}_a + \rho(a,y) \cdot \left(C'(a) - \bar{v}_a - y \cdot \bar{v}_y - \bar{R}(a;y)\right),$$

$$E\left[v_y|a;y\right] = \bar{v}_y + \chi(a,y) \cdot \left(C'(a) - \bar{v}_a - y \cdot \bar{v}_y - \bar{R}(a;y)\right).$$

where

$$\rho(a,y) \equiv \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2}$$

$$\chi(a,y) \equiv \frac{y\sigma_y^2 + \sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2 + \Omega(a,y)^2},$$

• Intuition: posterior assessment of intrinsic motivation, $E[v_a|a;y]$, is a weighted average of the prior $\bar{v}_a$ and of the marginal cost of observed contribution $C'(a)$, *net of* the average extrinsic and reputational incentives to contribute at that level. Weights reflect variability and correlation of signals and noise.

• Equilibrium: solution to the system of two nonlinear differential equations in the expectations $E[v_a|a;y]$ and $E[v_y|a;y]$. Can solve it when the variance of the reputational incentive, $\Omega(a,y)^2$ is independent of a Means that either:

Case 1: all agents have the same reputational concerns (the covariance matrix of $\boldsymbol{\mu}$ is zero);

Case 2: the cost function $C(a)$ is quadratic (the two reputatipons are linear in a)

Here, will combine the two assumptions.

Quadratic costs: $C(a) = ka^2/2$

(1) Homogenous reputational concerns (same $(\bar{\mu}_a, \bar{\mu}_y)$ for all)

Solution for equilibrium is

$$a = \underbrace{\frac{v_a + yv_y}{k}}_{\text{direct incentive}} + \underbrace{\bar{\mu}_a \left(\frac{1}{1 + y^2\sigma_y^2/\sigma_a^2} \right) - \bar{\mu}_y \left(\frac{y\sigma_y^2/\sigma_a^2}{1 + y^2\sigma_y^2/\sigma_a^2} \right)}_{\text{reputational incentive } R(y)}$$

direct incentive
(intrinsic + extrinsic)

reputational incentive *R(y)*

- Reward *y* has usual direct effect but also acts like an increase in signal-to-noise ratio, depressing *R(y)*.
- Aggregate supply obtained by simply summing *a* across individuals.

We provide conditions for crowding out $(d\bar{a}/dy < 0)$ to happen for either small or intermediate incentives.

- *Quadratic costs:* let $C(a) = ka^2/2 \Rightarrow$ optimal action a is then

$$a = \frac{v_a + y \cdot v_y}{k} + \mu_a \rho(y) - \mu_y \chi(y),$$

Aggregate supply curve $\bar{a}(\ \ )$ is obtained by summing over all agents $\Rightarrow$ its *slope* is:

$$\bar{a}'(y) \overset{y}{=} \frac{\bar{v}_y}{k} + \bar{\mu}_a \rho'(y) - \bar{\mu}_y \chi'(y).$$

Corollary 2 (small incentives). *Under above assumptions,* $\bar{a}'(0) < 0$ *whenever*

$$\frac{\bar{v}_y}{k} < \bar{\mu}_a \left(\frac{\sigma_{ay}}{\sigma_a^2} \right) + \bar{\mu}_y \left(\frac{\sigma_y^2 - 2\sigma_{ay}^2/\sigma_a^2}{\sigma_a^2} \right).$$

– Relates to some of the experimental evidence on crowding out finds that these effects occur only for small rewards (e.g., Gneezy and Rusticchini (2000)).

– Note also that requires that $\sigma_{ay} \neq 0$, and is most likely to hold when $\sigma_{ay} > 0$.

– Sufficiently large incentives, on the other hand, always increase supply: as $y \to \pm\infty$, $\rho'(y)$ and $\chi'(y) \to 0$, so behavior is dominated by the direct, non-reputational effect.

• Is crowding out then only a "smalll-stakes" phenomenon? *No,* neither empirically (e.g., Fehr and Gächter (2000), Bohnet et al. (2001)) nor in model.

Menus of rewards?

Proposition

(i) No menu if agent's choice unobserved (or side-contract).

(ii) No menu if v_y is known, v_a unknown.

(iii) Continuous menu of rewards

$$Y(v_y) = \mu_y \log\ v_y + \text{cst}$$

is optimal when if v_a known while v_y is unknown, with a distribution satisfying the monotone hazard rate property.

3. Turning down rewards? Equilibria studied before remain equilibria when agent can turn down all or part of y, if:

- conditions (i) or (ii) apply
- even if v_y unknown, if there is enough heterogeneity in image concerns (μ_a, μ_y) : turning down the reward *signals high image-consciousness,* and therefore low intrinsic motivation.