The Socially Optimal Level of Altruism

Richard Povey

New College and Wadham College, Oxford University

June 3rd 2011

(With thanks to Kevin Roberts, Godfrey Keller and Alan Beggs.)

Introduction

[W]hen altruism improves static non-cooperative outcomes, it lessens the severity of credible punishments. An altruist may well be perceived as a "softy" and his threats may not be taken seriously.

[Bernheim & Stark, 1988]

[T]he most efficient way to provide low payoffs, in terms of incentives to cheat, is to combine a grim present with a credibly rosy future.

[Abreu, 1986]

Intrinsic and Extrinsic Incentives

- **Intrinsic Incentives** - Altruistic preferences provide an *intrinsic* motivation for individuals to exhibit altruistic behaviour.
- **Extrinsic Incentives** - Punishment systems provide an *extrinsic* motivation.
- Often it is empirically difficult to distinguish between the two (e.g. enlightened self-interest in the repeated prisoners' dilemma) [Hammond, 1975].
- These two forms of incentives represent alternative "social technologies" that can potentially be used to achieve socially beneficial outcomes, but which can interfere with one another in a perverse manner. The moral preferences and institutions which have evolved in human society represent a particular "policy mix" which may (or may not) be socially optimal.

A Word on Idiom

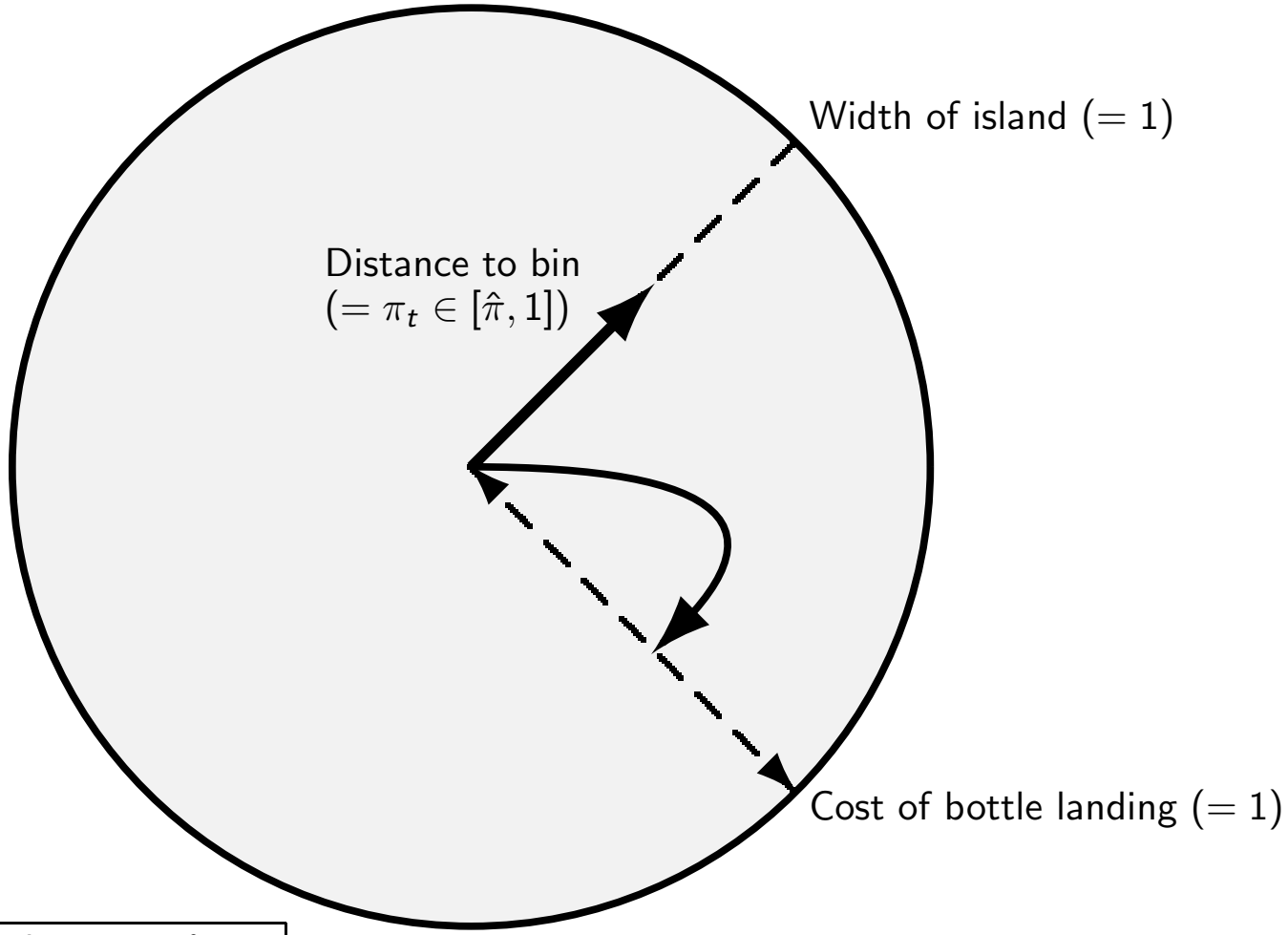
The sequential punishment model presented in this paper is intended as a highly abstract and stylized representation of social interaction, rather than as a realistic model of a specific situation. A simple "parable" can often help with the intuition. Models with a similar idiom include:

- Robinson Crusoe economy [Ruffin, 1972].
- Samuelson's "chocolate pension game" [Samuelson, 1958].
- Diamond's model of fiat money in a "coconut economy" [Diamond, 1984].

So, in that spirit, a desert island parable seems appropriate...

The Island Parable

Individuals (who have been on the island long enough to set up a "back garden") finish off a cold beer one at a time and must decide whether to walk to the bin or just throw their bottle into another individual's garden:



Coefficient of altruism - $\theta < 1$

Discount factor - $0 \leq \delta < 1$

Players' Preferences

- Let $f_{i,t}$ be the **felicity** of player i in period t.
  - Felicity represents "private utility" from "economically fundamental" goods.
- In period t, player t moves so as to maximize his **expected social utility** u_t, discounted looking forward:

  Utility thus includes broader "moral preferences". This is of course only one among a number of alternative ways to define altruism. The advantage is that it enables us to simplify away from any "multiplier effects". Not to say that these do not exist and are not important in the real world, but in the sequential punishment model we wish to focus on the role of punishment and its interaction with altruistic preferences in as clean and simple an environment as possible.

$$E_\pi\left[u_t\right]|_{\pi_1 \ldots \pi_t} = E_\pi\left[\sum_{j=t+1}^{\infty}\left(\delta^{j-t}\left(f_{t,j} + \theta\sum_{k\neq t}^{\infty} f_{k,j}\right)\right)\right]\Bigg|_{\pi_1 \ldots \pi_t}$$

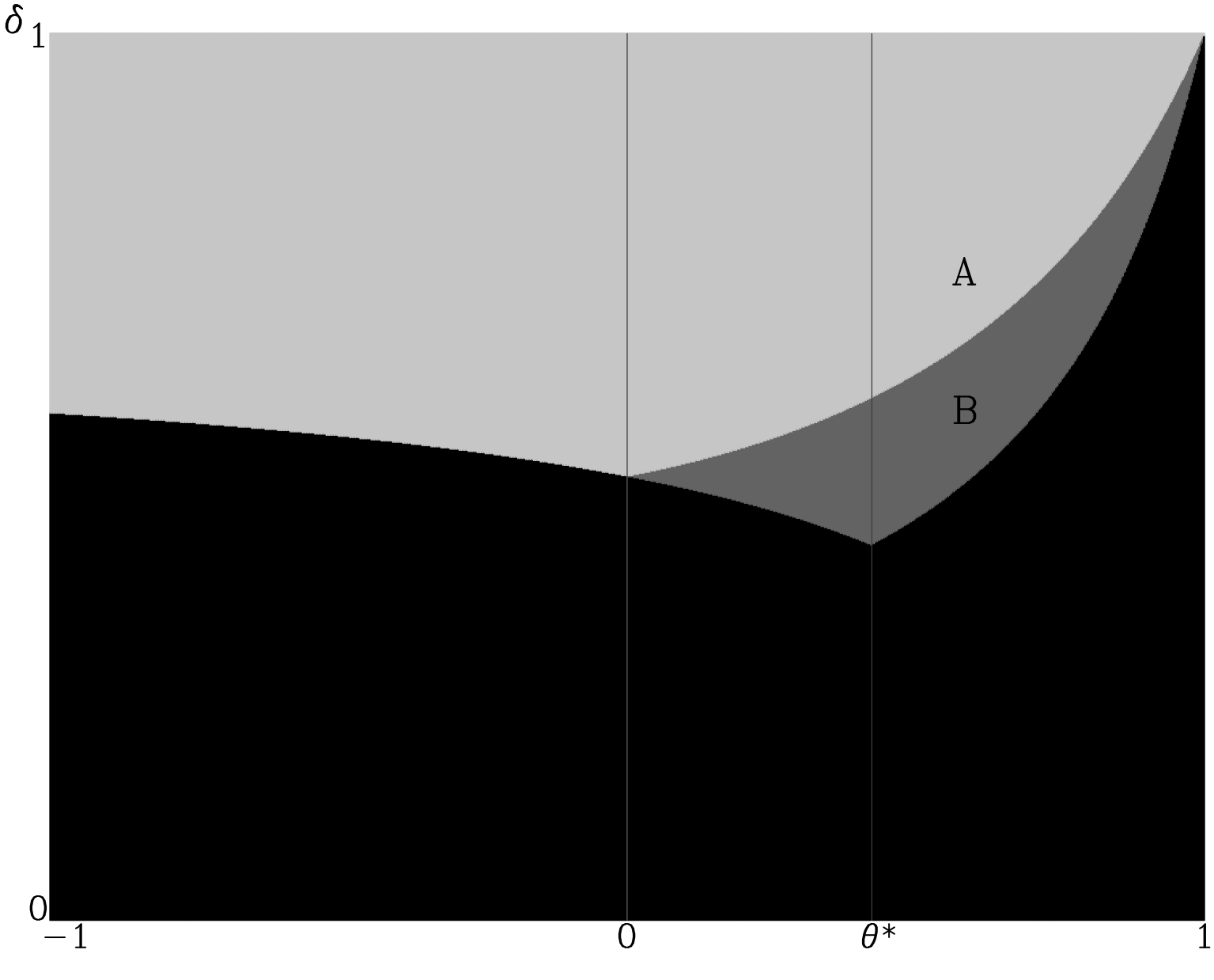
Three Effects

- **Temptation Effect** - Individuals with higher altruism are less *tempted* to inflict harm upon another individual for their own gain. (This is the main benefit from higher altruism.)
- **Willingness Effect** - Individuals with higher altruism are less *willing* to punish another individual for a previous misdemeanour by inflicting harm upon them. (This is a cost to higher altruism.)
- **Severity Effect** - Individuals with higher altruism also find some kinds of punishment less severe. In particular, if a fine was imposed, and some or all of the revenue is redistributed to another individual whose felicity has some weight in the utility function of the individual we are trying to punish, then any given size of fine is less severe for the punishee. (Another cost to higher altruism.)

Notation

- δ - Discount factor.
- θ - Coefficient of altruism.
- $\pi_t \in [\hat{\pi}, 1]$ - Benefit from harming / punishing in period t (randomly distributed between $\hat{\pi}$ and 1).
- ψ - Punishment path.
- ψ_k - Trigger level at point k.
- $\bar{\psi}$ - Flat punishment path / trigger level.
- $\bar{\psi}^*, \psi^*$ - Optimal flat / generic punishment path.
- θ^* - Socially optimal level of altruism - Enables efficient equilibrium to be sustained for largest possible range of δ.
- δ^* - Lowest possible value of δ for which the socially efficient outcome can be sustained. (Corresponds to θ^*.)

Overview - Socially Efficient Equilibria



A - Supportable using Nash-reversion path ($\psi = \bar{\psi} = \theta$)

B - Supportable using optimal path ψ^*

Solution Concepts

- **Folk Theorem** - [Aumann & Shapley, 1992] [Rubinstein, 1979] [Fudenberg & Maskin, 1986] For any given θ, as $\delta \longrightarrow 1$, the socially efficient outcome becomes supportable. We are interested here, however, in what happens as $\theta \longrightarrow 1$ for any given $\delta < 1$.
- **Optimal Penal Codes** - [Abreu, 1988] Abreu's framework of optimal penal codes in the form of punishment paths provides a natural framework that can be adapted to analyse socially efficient equilibria in the sequential punishment model.
- **Renegotiation Proofness** - [Farrell & Maskin, 1989] [Benoit & Krishna, 1993] We assume that society is able to avoid the temptation to let malefactors "off the hook". Thus we stick with subgame perfection rather than further refining the equilibrium criterion.

Constraints

$$U_k(\psi) \equiv \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - 1) g(\pi) d\pi\right] \quad (1)$$

$$V_k(\psi) \equiv \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - \theta) g(\pi) d\pi\right] \quad (2)$$

$$\lambda_k(\psi) \equiv \left(\frac{\delta}{1-\delta}\right) V_k(\psi) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \psi_k - \theta \quad (3)$$

$$\kappa(\psi) \equiv -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \theta - 1 \quad (4)$$

$$\phi(\psi) \equiv -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) \quad (5)$$

- ψ^* maximizes $\phi(\psi^*)$ such that $\forall_k : \lambda_k(\psi^*) \geq 0$.
- The socially efficient outcome is sustainable iff $\kappa(\psi^*) \geq 0$.

Results

Theorem

If $\theta \leq \delta + (1-\delta)\hat{\pi}$ *then* $\psi^* = \bar{\psi}^* = \hat{\pi}$, *otherwise* $\bar{\psi}^* = \theta - (1-\theta)\frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1} g(\pi)d\pi$.

$$\kappa\left(\bar{\psi}^*\right) = -\frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1}(\theta\pi - 1)g(\pi)d\pi + \theta - 1 \tag{6}$$

$$\frac{d\kappa}{d\theta} = 1 - \frac{\delta}{1-\delta}\left(\int_{\bar{\psi}^*}^{1}\pi g(\pi)d\pi + \left(1-\theta\bar{\psi}^*\right)g(\bar{\psi}^*)\frac{d\bar{\psi}^*}{d\theta}\right) \tag{7}$$

Theorem

As $\theta \longrightarrow 1^-$, $\theta \longrightarrow 1^-$, $\kappa\left(\bar{\psi}^*\right) \longrightarrow 0^-$.

Theorem

If $\delta > \frac{1}{1+\bar{\pi}}$ *then* $\kappa\left(\psi^*\right) \longrightarrow \infty$ *as* $\theta \longrightarrow -\infty$.

Results

$$\frac{\delta^*}{1-\delta^*} = \frac{1-\theta^*}{1-\theta^*\bar{\pi}} \tag{8}$$

$$\theta^* = \delta^* + (1-\delta^*)\hat{\pi} \tag{9}$$

Theorem

$0 < \theta^* < 1$ *and* $0 < \delta^* < \frac{1}{1+\bar{\pi}}$

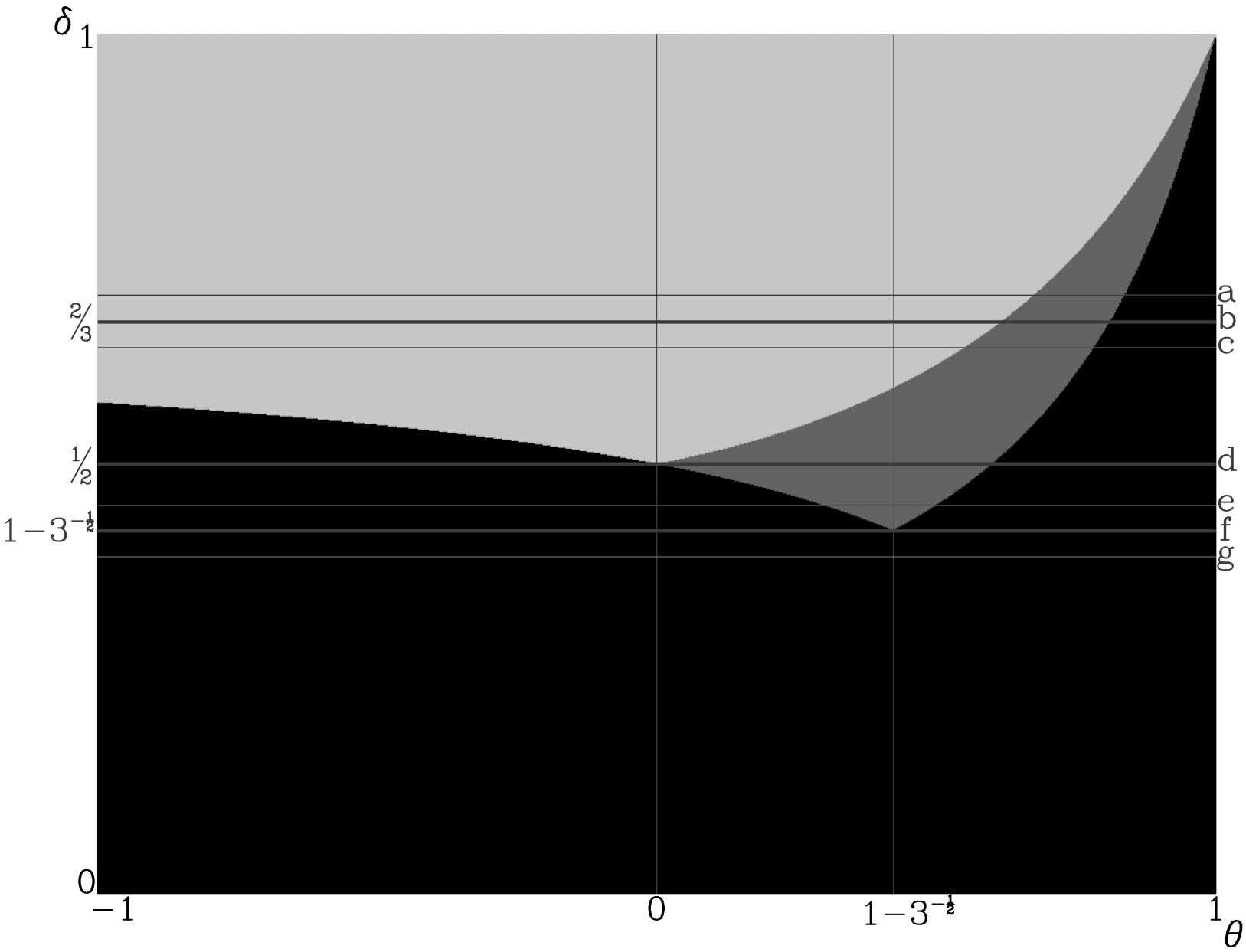
Theorem

As $\theta \longrightarrow 1^-$, $\theta \longrightarrow 1^-$, $\kappa(\psi^*) \longrightarrow 0^-$.

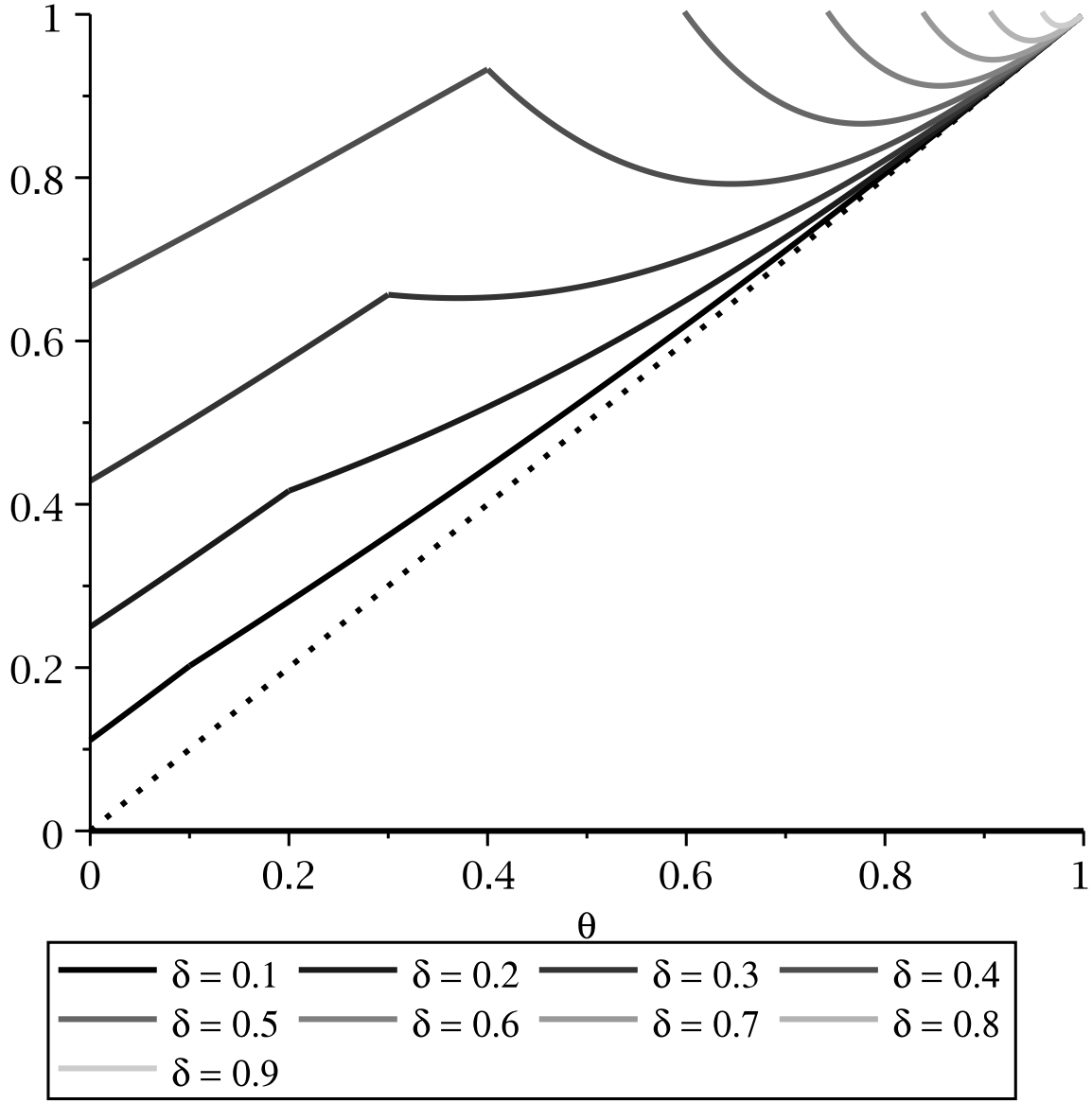
Theorem

If $\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$ *then* $\psi^* = \bar{\psi}^*$.

Illustration - Uniform Distribution of Benefit, $\hat{\pi} = 0$



Optimal Second-Best Equilibria



Conclusion

[I]n comparison with a situation wherein altruism is absent altogether, the prevalence of just some altruism could result in Pareto inferior outcomes. Hence, if the formation of altruism may not only fail to do any good but may actually make things worse whereas the formation of sufficiently high levels of altruism is almost always beneficial,...a troubling discontinuity arises: to the extent that the formation of altruism is like the rising of bread dough (i.e. it has to be gradual) groups yearning to build up their social stock of altruism may have to endure Paretial deterioration before experiencing Paretial gains. Perhaps one reason why a great many societies consist of self-interested economic men and women rather than altruistic economic men and women has to do with this nonmonotonicity.

[Stark, 1989]

Conclusion

- This discontinuity continues for high levels of altruism. (Although the second-best equilibrium becomes perfectly efficient as $\theta \longrightarrow 1$, the first-best equilibrium breaks down.)
- **Optimal Penal Codes** in subgame perfect equilibria provide a richer framework to explore these effects. The sequential punishment model is sufficiently simple to allow a complete characterization of the sustainable socially optimal equilibria.
- **Renegotiation Proofness** - Speculatively, this equilibrium refinement could further enrich these effects. If individuals are **benevolent** $(\theta \geq 0)$ then they will want to renegotiate away from equilibria which harm another individual. On the other hand, if individuals are **malevolent** $(\theta < 0)$ then they will not be willing to renegotiate away from a punishment path since they all *enjoy* watching another individual being punished.
- **Asymmetric Information** - Another intuitive reason why imperfect altruism may be socially optimal - further research.

References I

Abreu, Dilip (1986).

"Extremal Equilibria of Oligopolistic Supergames".
Journal of Economic Theory, 39, 191–225.

Abreu, Dilip (1988).

"On the Theory of Infinitely Repeated Games with Discounting".
Econometrica, 56(2), 383–396.

Aumann, Robert J. and Shapley, Lloyd S. (1992).

"Long Term Competition - A Game Theoretic Analysis".
UCLA Economics Working Papers 676, UCLA Department of Economics.

Benoit, Jean-Pierre and Krishna, Vijay (1993).

"Renegotiation in Finitely Repeated Games".
Econometrica, 61(2), 303–23.

Bernheim, Douglas B. and Stark, Oded (1988).

"Altruism within the Family Reconsidered: Do Nice Guys Finish Last?".
The American Economic Review, 78(5), 1034–1045.

Diamond, Peter (1984).

"Money in Search Equilibrium".
Econometrica, 52(1), 1–20.

Farrell, Joseph T. and Maskin, Eric S. (1989).

"Renegotiation in Repeated Games".
Games and Economic Behaviour, 1(1), 327–360.

References II

Fudenberg, Drew and Maskin, Eric (1986).

"The Folk Theorem in Repeated Games with Discounting or with Incomplete Information".
Econometrica, 54(3), 533–54.

Hammond, Peter (1975).

Charity: Altruism or Cooperative Egoism?
In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 115–131). Russell Sage Foundation, New York.

Rubinstein, Ariel (1979).

"Equilibrium in Supergames with the Overtaking Criterion".
Journal of Economic Theory, 21(1), 1–9.

Ruffin, Roy J. (1972).

"Pollution in a Crusoe Economy".
The Canadian Journal of Economics, 5(1), 110–118.

Samuelson, Paul A. (1958).

"An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money".
Journal of Political Economy, 66, 467.

Stark, Oded (1989).

"Altruism and the Quality of Life".
The American Economic Review, 79(2), 86–90.