MORALLY-MOTIVATED SELF-REGULATION

David P. Baron

Northwestern University and Stanford University

December 2007

ABSTRACT

Some individuals and firms voluntarily mitigate the harmful consequences of their economic activities in situations in which they could free ride. In the context of a random matching model where citizens play a free-rider game, this paper examines the scope of self-regulation motivated by altruism or warm glow preferences. Moral preferences are represented as stronger among neighbors than among strangers, and those preferences may be unconditional or reciprocal. The focus is on the role of organizations in increasing self-regulation by addressing free-rider problems. Social label, certification, and enforcement organizations are considered, as are public regulation and social pressure applied by an activist NGO funded by voluntary donations by citizens. Social label and certification organizations can exist with reciprocal but not unconditional altruism, and they expand the scope of self-regulation, but their effect is limited. Enforcement organizations expand the scope of self-regulation for both unconditional and reciprocal altruism, and for-profit enforcement is more aggressive than non-profit enforcement. Public regulation can also overcome the free-rider problem, but self-regulation can crowd out the political support for public regulation, whereas it may not crowd out private organization.

Morally-Motivated Self-Regulation

David P. Baron[1]
Northwestern University and Stanford University
December 2007

I. Introduction

Individuals take a variety of voluntary actions to mitigate externalities, redistribute wealth, and provide public goods that benefit others. Some have reduced their environmental impact, contributed to relief efforts, supported environmental NGOs, bought fair trade products, and shown a willingness to pay a premium for green products, such as green electricity. Organizations have also been formed to facilitate voluntary measures. Individuals can purchase offsets for the carbon footprints of their households through Climate Trust, Atmosfair, and NativeEnergy. Air travelers can purchase carbon offsets through Expedia and Travelocity. Firms have established programs for environmental protection, sustainability, and the assurance of credence attributes of their products, and corporate social responsibility is increasingly embraced by firms. Google, for example, has pledged to become carbon neutral. Carbon trading is available on the Chicago Climate Exchange, working conditions in overseas factories are strengthened and monitored by the Fair Labor Association (FLA) and the Worker Rights Consortium, wealth is transferred to growers and local producers through fair trade organizations, and private organizations and charities provide education and medical care to address pandemics such as HIV/AIDS. These activities are examples of self-regulation – the private provision of public goods and private redistribution – as an alternative to public provision, regulation, and redistribution.

Self-regulation takes place outside the institutions of government and hence is in the realm of private rather than public politics.[2] Self-regulation can be motivated by a number of concerns. It could be motivated by self-interest as when a firm produces a green product because consumers are willing to pay a premium for it. A firm could also self-regulate to deter public regulation as in Lyon, Maxwell, and Hackett (2000) and Lyon and Maxwell (2004). A firm could also self-regulate to deter private politics, as in Baron (2007c) and Baron and Diermeier (2007) where self-regulation

[1] I thank Avinash Dixit and seminar participants at Columbia University, Northwestern University, Princeton University, and the University of Chicago for their helpful comments.

[2] Public politics takes place in the institutions of government, whereas private politics occurs outside, but often in the shadow, of those institutions. Private politics pertains to individual and collective action to influence the conduct of private agents, including oneself, as in the case of NGOs that apply social pressure to change the conduct of firms.

can lead an activist to select a different target for a boycott. Some self-regulation thus can be explained by private benefits.

This paper focuses instead on self-regulation motivated by moral concerns. The context is one in which citizens or firms can voluntarily self-regulate in the presence of incentives to free-ride. Free riding inhibits self-regulation, and citizens can mitigate the incentives to free-ride through the formation of private organizations or through public organizations. One public organization is government regulation, but private self-regulation can crowd out the political support for public regulation (Calveras, Ganuza, and Llobet 2007). The focus here is on private organizations as an alternative to public organization. Private organizations have an advantage, since they can be formed quickly and without majority approval.

One private organization considered screens using social labels that allow citizens with similar preferences to interact among themselves. A second type of private organization is a certification organization that provides future partners with information about the past conduct of citizens. A third type strengthens self-regulation by enforcing informal pledges to self-regulate. Both non-profit and for-profit enforcement organizations are considered. Finally, a society could rely on activists and NGOs to monitor the conduct of citizens and publicly disclose their failure to self-reulate. Although organizations can increase the scope of self-regulation, unorganized self-regulation could crowd out support for public regulation, affect the demand for private organizations, and affect donations to fund social pressure.

Moral preferences or warm glow preferences are necessary for both unorganized and organized self-regulation in the model. Moral preferences are other-regarding and take the form of altruism in which a citizen takes into account the benefits she provides to others when she self-regulates. Warm glow preferences (Andreoni 1988, 1990) are egoistic and reflect satisfaction from the action of providing the public good. Moral preferences thus pertain to the well-being of others, whereas warm glow preferences reflect the private goods aspect of an action. Altruistic and warm glow preferences can have the same representation and effect on behavior.

Moral preferences can be pure or impure and may be conditional on the actions of others and limited by socioeconomic distance. Generalized morality is a pure form of altruism in which a citizen's willingness to self-regulate is independent of how distant a trading partner is, where distance could be geographic or socioeconomic. With generalized morality a citizen self-regulates regardless of whether her matched partner is a neighbor or stranger. Whether the partner is a neighbor or a stranger matters with limited morality. As modeled here, the utility of a citizen from

providing a benefit to another citizen decreases the greater the socioeconomic distance between them. A citizen then may self-regulate when interacting with a neighbor but not when interacting with a stranger. The theory predicts that the scope of self-regulation is greater in close-knit than disparate societies and greater with generalized than limited morality.

Moral preferences can also be unconditional or conditional. Unconditional preferences are independent of the action of a partner, as in the case of pure altruism. The conditional preferences considered here are reciprocal in the sense that citizens care more about providing benefits to another citizen if the other citizen self-regulates than if he does not.[3] Reciprocal altruism results in a smaller scope of self-regulation in a heterogeneous citizenry and provides an opportunity for private organizations to expand self-regulation.

The form, in addition to the strength, of moral preferences matters for the extent to which organizations can help citizens overcome free-rider problems. Consider a green club or social label organization that is open to all citizens for a fee and sorts citizens with similar preferences so that they can interact with each other.[4] For example, citizens can purchase products produced by suppliers that comply with fair trade requirements. If in the model citizens have preferences exhibiting unconditional altruism, adverse selection prevents sorting; i.e., green citizens cannot interact only among themselves, since red citizens will also join the organization to free ride on the green citizens. If, however, citizens have preferences exhibiting reciprocal altruism, there exists a membership fee such that green citizens join the club and red citizens do not. This increases the self-regulation by green citizens.

Pledges to self-regulate can be enforced by non-profit organizations or by for-profit firms. In the model enforcement is self-imposed and takes the form of harm if a citizen fails to contribute to the public good. A non-profit organization maximizes the expected utility of those who avail themselves of its enforcement services, whereas a for-profit firm chooses its enforcement policy to maximize its profits with revenue from those citizens that use its services. Both non-profit and for-profit enforcement can increase the scope of self-regulation, but the policies chosen are different. For the case of enforcement on demand; i.e., enforcement voluntarily requested by citizens before they play the self-regulation game, the for-profit firm provides more aggressive enforcement than does the non-profit organization, and the price it charges for enforcement is higher than the fee

[3] Rabin (1998) discusses the economics and psychology of reciprocal altruism and related experimental evidence.

[4] Prakash and Potoski (2006)(2007) consider voluntary environmental organizations from a club perspective.

charged by the non-profit organization.

A certification organization provides information by certifying how a citizen played in the past and providing that information to future partners. A certification organization expands the scope of self-regulation with reciprocal altruism by inducing citizens with weaker moral preferences to pool with citizens with stronger moral preferences who self-regulate. The citizens with weaker moral preferences do so not because they gain from self-regulating but instead because they find themselves in a dilemma. If they do not self-regulate in accord with their single-period preferences, they will be identified as having weaker moral preferences and in the future will not be able to free ride on those with stronger moral preferences. This provides an explanation for self-regulation in which peer conduct matters not because of sociological influence but instead because the citizen will be identified as one who will free ride. In addition to inducing pooling by those with weaker moral preferences, citizens with stronger moral preferences separate from those with weaker moral preferences for a larger set of matches. This allows the citizens with stronger moral preferences to increase their self-regulation in the future.

Neither social label organizations nor certification organizations can exist with unconditional altruism, whereas with reciprocal altruism and heterogeneous moral preferences both can exist and expand self-regulation. The maximal scope of self-regulation, however, is bounded above by the self-regulation with unconditional altruism. Enforcement organizations, in contrast, expand self-regulation beyond that with unconditional altruism.

Public regulation is an alternative to unorganized self-regulation and the private organizations that support self-regulation. The demand for public regulation is greater in a heterogeneous society when moral preferences reflect reciprocal altruism than unconditional altruism. The expected gain from uniform public regulation is increasing in the benefits from self-regulation and in the strength of moral preferences. From a positive perspective regulation must be approved by a majority of citizens, and the political support for regulation can be crowded out by self-regulation.

An alternative to private self-regulation organizations and public regulation is rely on NGOs to put social pressure citizens to self-regulate. NGOs monitor conduct and can report to the public on citizens who fail to self-regulate, which can result in harm to a reputation, a brand, or self-esteem. NGOs are funded by voluntary donations by citizens, and donations face a free-rider problem of their own. Preferences that reflect reciprocal altruism are sufficient, however, to overcome the collective action problem. Monitoring and the threat of harm increase self-regulation. Private organizations and NGOs have advantages over public regulation because they can be formed

without majority approval, and the incentive to establish the organizations likely to be crowded out by unorganized self-regulation. They may, however, not be able to fully-resolve free-rider problems.

To explore the scope of self-regulation, an abstract rather than descriptive model is used. The model is based on the random matching model developed by Dixit (2003b)(2004) to examine the effect of distance and contract enforcement on trade. Here citizens play a self-regulation game and have morally-based preferences. The model is related to that of Tabellini (2007), who considers a version of Dixit's model in which people are in a prisoners' dilemma and experience guilt if they do not cooperate. The equilibria in his paper have properties related to those in Section II in this paper. He also considers a two-period overlapping generations version in which parents embed their children's welfare in their own and can transmit values or norms to their children.

Levy and Razin (2007) provide a theory in which a religious organization arises endogenously when people have heterogeneous beliefs about being punished if they defect in a prisoners' dilemma game. The emergence of the religious organization relies on the prisoners' dilemma exhibiting strategic complements, and if that is not the case, as in the basic self-regulation game, the organization does not emerge. Their result is analogous to that for the social label organization considered in Section V, where the organization arises with reciprocal but not with unconditional altruism.

The next section introduces the basic model with unconditional moral preferences, and Section III considers reciprocal altruism and characterizes its impact on the scope of self-regulation. Section IV considers from both normative and positive perspectives public regulation as an alternative to self-regulation. Section V considers private organizations that increase self-regulation by addressing free-rider problems. These include social label organizations formed by citizens to facilitate self-selection, a certification organization, and enforcement both by a non-profit organization and a profit-maximizing firm. Section VI considers social pressure on citizens to self-regulate, where the social pressure is applied by an activist funded by voluntary donations by citizens. Conclusions are offered in the final section.

II. Generalized and Limited Morality

A. The Basic Model

1. Matching and the Free-Rider Problem

The approach in this paper is to investigate self-regulation in a static rather than repeated setting. This may be thought of as corresponding to changing circumstances in which long-run behavior is not relevant or where convergence requires a large number of repetitions. The basic

model is constructed with few population characteristics to focus on preference-induced conduct. The agents in the model can include individuals and firms, both of which will be referred to as citizens. Self-regulation by individuals may involve the mitigation of an environmental externality, charitable contributions, and the provision of a local public good. Firms can also self-regulate based on moral preferences. For example, a firm could provide unobservable credence attributes of its product that consumers cannot learn through search, experience, or consumption. Such attributes could include the conditions under which a product is produced, including any unregulated environmental externalities associated with production, how well workers are treated and paid, and whether it is made from sustainable inputs. These actions could be due to the preferences of shareholders, as in Baron (2007a)(2007c) and Graff Zivin and Small (2005). Morally-based preferences could also reflect the preferences of the managers of the firm when there is separation of ownership from control, as considered in Baron (2007b). A firm, for example, could undertake costly actions under the framework of corporate social responsibility that benefit a community or targeted recipients.[5]

Consider a society in which citizens are uniformly distributed on a circle with circumference $2L$.[6] The parameter L may be thought of as how disparate is the society, so the more factionalized the society the larger is L. Each citizen is randomly matched with another citizen at a socioeconomic distance y with probability density $\eta(y) = \frac{\alpha e^{-\alpha y}}{2(1-e^{-\alpha L})}, \alpha>0$. Figure 1 illustrates the set-up with citizens A and B matched at a distance y. The distance y could be geographic and hence distinguish between neighbors and strangers, or it could be socioeconomic where citizens are differentiated by culture, language, class, and the like. If the citizens are firms, distance could pertain to a product or technology space. Firms in the chemical industry are closer to oil companies in technology space than they are to information technology firms. Similarly, the technology used by real estate brokers is closer to that of the information technology industry than to the technology used in the pharmaceutical industry.

[5] Siegel and Vitaliano (2007) found that firms producing credence goods were more likely to practice corporate social responsibility than firms producing search goods. They argue that this finding is consistent with a product differentiation strategy, but it is also consistent with the concept of self-regulation.

[6] Ellison (1993) and Eshel, Samuelson, and Shaked (1998) consider complete information, repeated games with random matching in which players are distributed on a circle and can provide local public goods that benefit only immediate neighbors. Eshel, Samuelson, and Shaked allow players to choose to be an altruist or an egoist based on comparing among neighbors the average payoffs to each type. They characterize the limiting distribution of types and conclude that players are primarily altruists. Ellison shows that although in the limit players play the risk dominant equilibrium the rate of convergence can be sensitive to the matching model. Convergence is rapid when players are matched only with their neighbors.

The higher is the parameter α in the density $\eta(y)$ the greater is the probability that a match is local, so a higher α can be interpreted as reflecting how close-knit is a society. The matching may be thought of as representing the everyday activities of citizens, and it is more likely that those activities involve citizens who are close to rather than far from each other. A citizen is thus more likely to interact with a neighbor than a stranger. Alternatively, if the matches are determined by some (unmodeled) search activities, it is more likely that citizens find more desirable rather than less desirable matches.[7]

The matched citizens interact resulting in a surplus for each. This interaction could involve a trade or other economic transaction and is assumed to be governed by the law and hence has no risk of default or cheating. Since both citizens gain, they will interact, so their surplus can be suppressed. Associated with the interaction is an opportunity for both citizens to self-regulate by providing a local public good or redistributing wealth, both of which will be referred to as contributing. The local public good could be the mitigation of a portion of a harmful externality associated with the interaction. In the case of private redistribution the redistribution could benefit some unmodeled recipient, such as the disadvantaged, microfinance borrowers. victims of disease, or children in developing countries. Contributing is assumed to have a cost $c>0$ and may provide benefits $b\geq 0$ to the contributor and to the citizen with whom she is matched.[8] The benefit could be from a reduction in pollution emitted by an upstream producer, a firm and employees improving safety in a factory, a firm adopting best practices in disposing of toxic waste, or a neighborhood clean-up project. The citizens are assumed to have an incentive to free-ride; i.e., $c>b$. The aggregate benefits ($2b$) need not be greater than the cost, so there may be no dilemma for some matches. To simplify the analysis and allow a focus on symmetric equilibria, the costs and benefits are assumed to be the same for both citizens. Heterogeneity is incorporated by letting the strength of moral preferences differ among citizens. More generally, the costs and benefits from contributing can differ for the two citizens, as illustrated in the example presented in Section II.A.3.

The basic free-rider problem could be resolved through contracts, but the situations considered are assumed to be either noncontractable or require costly enforcement. In particular, the actions in the self-regulation game are not observable. Instead, the focus is on the influence of moral preferences on self-regulation.

2. Preferences

[7] Dixit provides an interpretation of this formulation as resulting from search activities.

[8] The model can be extended to public goods that benefit all citizens, as considered in Section II.D.

A citizen is assumed to have preferences for costs and benefits, and she also may have other-regarding preferences regarding to the effect of her actions on the well-being of her matched partner. Alternatively, she may have warm glow preferences for the act of self-regulating. In the former case, preferences are altruistic, and in the latter case the citizen cares about the private good aspect of her action. To simplify the exposition, the term altruism will be used to encompass both altruistic and warm glow preferences, and both will be represented by the same expression.[9] Andreoni and Miller (2002) conclude from experiments that most subjects exhibited altruism and those who did behaved in accord with revealed preference theory. Hence, their revealed preferences could be represented by a utility function.

Altruistic preferences may be stronger the closer the trading partner is to the citizen, since she may care more about those who are closer to her; e.g., she cares more about neighbors than strangers (Banfield 1958). In a dictator game experiment Bohert and Frey (1999) found that dictators' offers to other players were decreasing in the social distance between the players, where social distance corresponded to identifiability and familiarity. Similarly, in a voluntary public goods experiment Keser and van Winden (2000) found that contributions were greater and free-riding less among those who interacted repeatedly than among those who were strangers. La Ferrara (2003) provides an overlapping generations model of credit in a "kin group" and found support for the model from data from Ghana. Credit terms were better (e.g., no interest) and default rates were lower for intra-kin loans and for households that contributed funds for lending in the past. These results are consistent with the importance of socioeconomic distance and also with the importance of reciprocity as considered in Section III.

Altruistic preferences thus are represented by a utility $xe^{-\eta y}, \eta \geq 0$, where y is the socioeconomic distance to the matched partner and x is a parameter. The parameter η reflects the degree of limited morality with $\eta = 0$ corresponding to pure or generalized altruism and $\eta \to \infty$ corresponding to no altruism or warm glow preferences. Consequently, lower values of η correspond to stronger moral preferences. In the case of a local public good the parameter x could equal the benefits b provided to the matched partner. In the case of private redistribution if c represents a charitable contribution to a disadvantaged person located at y, then $x \geq c$ (and $b = 0$) corresponds to a contribution benefitting the recipient by at least as much as the cost of the contribution. In the

[9] The distinction between altruistic and warm glow preferences is less a philosophical one and more one of positive implications. Andreoni has shown that if citizens have altruistic preferences government provision of public goods financed by lump-sum taxes crowds out personal giving that funds public goods but does not do so with warm glow preferences. In the present paper there is no government provision, so public crowding out does not occur.

case of warm glow preferences x could differ from both c and b.

The utility from altruism may be independent of the action of the partner, or it may depend on that action. For example, a citizen may abandon her altruistic or warm glow preferences if her partner is not expected to reciprocate in self-regulating. Reciprocal altruism could be represented in a number of ways. Levine (1998) represents it through preferences in which a citizen is "more altruistic to an opponent who is more altruistic toward them." Rabin (1993, p. 1282) considers a concept of fairness in which "people are willing to sacrifice their own material well-being to help those who are being kind." He represents this by a "kindness function" that depends on strategies and beliefs. Here, reciprocal altruism is conditional only on the (anticipated) actions of the match partners. The utility from altruism is specified as $\theta x e^{-\eta y}$, $\theta \in [0, 1)$ when the partner does not reciprocate, where the parameter θ indexes the extent to which preferences are unconditional with $\theta = 0$ corresponding to pure reciprocal altruism. Initially, preferences are assumed to be independent of whether the partner reciprocates, i.e., unconditional altruism, and then reciprocal altruism is considered in Section III. Moral preferences thus can be generalized or limited and can be unconditional or reciprocal, as illustrated in Figure 2.

The basic self-regulation game is presented in Figure 3. A strategy S is a mapping from the match distance to the action set $\{C, N\}$, where C represents contributing and N represents not contributing. The timing in the game is that nature first draws a match for each citizen, and then the matched pairs simultaneously choose their actions. Initially, information is assumed to be complete. The equilibrium concept is Nash, and symmetric equilibria are considered. The game is played only once, so citizens have no opportunity to develop a reputation.

If a citizen contributes, her utility U_C with unconditional altruism is given by

$$U_C = B - c + x e^{-\eta y}, \tag{1}$$

where $B = 2b$ is the benefits when the other player also contributes and $B = b$ otherwise. If the citizen does not contribute, her utility is $U_N = B$, where $B = b$ if the other player contributes and $B = 0$ otherwise.

In addition, assume that $b - c + x e^{-\eta L} < 0$, so that a citizen with limited morality does not prefer to contribute for all matches. Similarly, assume that $b - c + x > 0$, so a citizen prefers to contribute when matched with a citizen at her own location.

3. An Example

The model is formulated as symmetric to simplify the analysis and expose clearly the intuition

underlying the equilibria. The symmetric model can be viewed as a special case of a more general self-regulation model, an example of which is presented here. Consider a credence good problem in which an individual has preferences regarding the working conditions in factories in developing countries and in particular prefers that a living wage is paid. The wage paid is not observable by the individual. The individual has a demand for a product and is matched with a firm that produces the product in a factory in a developing country. The firm j can pay a living wage (C^j) or pay a market wage (N^j), and the individual i can either buy the product of firm j (C^i) or buy another product (N^i) produced with a market wage, which is assumed to yield a normalized utility of 0. If the individual chooses C^i, she receives a benefit b_i, which could be 0, from buying the product from firm j and pays a price premium c_i for the product. The individual's moral preferences are represented by a utility $x_i e^{-\eta^i y}$ if it buys the product and a living wage is paid and a utility $\theta_i x_i e^{-\eta^i y}$, where θ_i could be 0 if the firm j does not pay a living wage. If the individual buys the product, the firm receives the price premium, so $b_j = c_i$. The cost of the living wage is $c_j > b_j$. When the firm pays a living wage, it may have warm glow preferences $x_j e^{-\eta^j y}$ from the good feelings of the individual about the firm and $\theta_j x_j e^{-\eta^j y}, \theta_j = 0$, if the individual does not buy the product, where θ_j could be 0. When the firm pays a living wage, the individual receives a benefit $\tilde{b}_j = \tilde{b}_j(c_j)$, which could represent spillover benefits when workers receive higher wages. When the individual buys the product, the firm may receive a benefit $\tilde{b}_i$ from (possible) consumer loyalty. The utility U_{C^i} of the individual then is

$$U_{C^i} = \begin{cases} b_i + \tilde{b}_j - c_i + x_i e^{-\eta^i y} & \text{if } C^j \\ b_i - c_i + \theta_i x_i e^{-\eta^i y} & \text{if } N^j, \end{cases}$$

and the utility U_{N^i} is

$$U_{N^i} = \begin{cases} \tilde{b}_j & \text{if } C^j \\ 0 & \text{if } N^j. \end{cases}$$

The utility of the firm is defined analogously:

$$U_{C^j} = \begin{cases} b_j + \tilde{b}_i - c_j + x_j e^{-\eta^j y} & \text{if } C^i \\ b_j - c_j + \theta_j x_j e^{-\eta^j y} & \text{if } N^i, \end{cases}$$

and

$$U_{N^j} = \begin{cases} \tilde{b}_i & \text{if } C^i \\ 0 & \text{if } N^i. \end{cases}$$

The basic model is a symmetric version of this example. As is clear from the following section, the asymmetries in the example result in differences in the boundaries of self-regulation. The same type of differences in the boundaries is considered in terms of the parameter η of the moral utility.

Organizations that enable citizens to expand their self-regulation can be thought of as providing the following functions. With a heterogeneous citizenry a social label organization can be thought of as sorting individuals and firms, as in the case of green club, according to the strength of their moral preferences. A certification organization can be thought of as an inspection agency that verifies the actions of individuals and firms and reports that information to future partners. An enforcement organization can be thought of as inspecting the factories and releasing the results to the public. Enforcement can be conducted by private firms or by non-profit organizations such as the FLA. Social pressure can be applied by NGOs that monitor whether a living wage is paid and the individual buys the product of that firm and reveal the results to the public. Public regulation could require that domestic firms ensure that their suppliers pay a living wage and that individuals only purchase from firms whose suppliers pay a living wage.

B. The Equilibrium with Unconditional Altruism

To characterize the equilibrium with unconditional, limited altruism, note that the self-regulation game is a dominant strategy game in which the utility difference $\Delta U = U_C - U_N = b - c + xe^{-\eta y}$ is independent of the partner's action. A citizen thus has a dominant strategy S^* of contributing $U_C - U_N \geq 0$ and choosing N otherwise; that is,

$$S^* = \begin{cases} C & \text{if } y \leq y^o \\ N & \text{if } y > y^o, \end{cases}$$

where

$$y^o = \frac{1}{\eta} \ln\Big(\frac{x}{c-b}\Big) \tag{2}$$

is the boundary for contributions. For a match at the boundary y^o the utility in (1) from contributing is $U_C = b$, which is the benefit received from the contributions of the partner. For a given y the equilibrium is unique.

A citizen thus contributes when matched with a partner no farther away than y^o and otherwise plays N.[10] For matches closer than y^o the limited morality is sufficient to overcome the incentive

[10] Tabellini considers a model in which a citizen experiences guilt if she chooses N but less quilt the more distant is the other citizen in the match. Let the disutility from guilt be additive and represented by $ge^{-\gamma y}$, and assume that g and γ are common knowledge. Then, the citizen plays C for a wider set of matches. That is, the boundary $y^o(g,\gamma)$ of contributions satisfies

$$b - c + xe^{-\eta y^o(g,\gamma)} + ge^{-\gamma y^o(g,\gamma)} \equiv 0$$

and is strictly increasing and strictly concave in g with $y^o(0,\gamma) = y^o$. The boundary is strictly decreasing in γ. Guilt has an effect similar to that of unconditional altruism in the sense that the scope of self-regulation increases with stronger (lower η) moral preferences.

to free ride, and all citizens contribute. For more distant matches the incentive to free ride prevails, and citizens do not contribute.

The stronger are moral preferences (lower η), the larger is the set of matches for which citizens contribute. Also, $\lim_{\eta\to\infty} y^o = 0$, so in the limit as η increases the scope of self-regulation goes to 0. Altruistic or warm glow preferences are thus necessary for self-regulation, but the impact is limited compared to that with generalized utility, which results in contributions for every match since $c - b + x > 0$. The boundary y^o is strictly convex in η, so more limited morality results in relatively smaller decreases in the scope of self-regulation, defined by $\frac{y^o}{L}$.

The boundary y^o is strictly increasing in b and x and strictly decreasing in c and η. Consequently, the more beneficial relative to its cost, or higher quality, is self-regulation (contributing), the greater is the scope of self-regulation.[11] Figure 4 illustrates the equilibrium and the comparative statics with respect to the quality of self-regulation and the strength of moral preferences.

Since the self-regulation game takes place after the match has been drawn, the boundary y^o is independent of the parameter α of the match probability and of the dispersion L in society. Conditional on the socioeconomic distance between them citizens with warm glow or unconditional altruistic preferences behave the same regardless of how dispersed the citizenry is. Consequently, the greater is the dispersion of the citizenry the lower is the scope of self-regulation in society.

The ex ante expected utility EU^* of a citizen is

$$\begin{aligned} EU^* &= \int_0^{y^o} \Big(2b - c + xe^{-\eta y}\Big)\Big(\frac{\alpha e^{-\alpha y}}{1 - e^{-\alpha L}}\Big)dy \\ &= \frac{1 - e^{-\alpha y^o}}{1 - e^{-\alpha L}}(2b - c) + \Big(\frac{x\alpha}{\alpha + \eta}\Big)\frac{1 - e^{-(\alpha+\eta)y^o}}{1 - e^{-\alpha L}}. \end{aligned} \tag{3}$$

The first term is the expected utility from the benefits and costs of contributing, and the second term is the expected utility from altruism. The expected utility is increasing in α, since more matches are local, and decreasing in the dispersion L of society, since other citizens are farther away and hence contributions are made for fewer matches. The expected utility is increasing in the quality (higher b, lower c) of self-regulation and the strength (lower η) of moral preferences.

C. Heterogeneous Preferences

Citizens can differ in the extent to which they care about the well-being of other citizens or receive a warm glow from the act of contributing. The heterogeneity introduced in this section is in

[11] Also, the greater is the parameter x the greater is the scope of self-regulation. The limit as x decreases is zero contributions; i.e., $\lim_{x\to c-b} y^o = 0$.

the rates at which their unconditional altruism or warm glow preferences decline with socioeconomic distance. Suppose that citizens are one of two types with parameters $\eta_i, i = 1, 2, \eta_1 < \eta_2$, which are private, soft information that cannot be revealed to others.[12] Since the utility of a citizen for playing C versus N is independent of the action of her matched partner, the citizens of each type have dominant strategies. The dominant (Bayesian Nash equilibrium) strategy S_i^* of a citizen of type i is given by

$$S_i^* = \begin{cases} C & \text{if } y \leq y_i^o \\ N & \text{if } y > y_i^o, \end{cases} \tag{4}$$

where

$$y_i^o = \frac{1}{\eta_i} \ln\Big(\frac{x}{c-b}\Big), \ i = 1, 2. \tag{5}$$

The equilibrium with private information is thus qualitatively the same as the equilibrium in the case in which citizens are of one type. The same is true for any number of types. If some citizens are not altruists, let $\eta_2 \to \infty$, in which case $y_2^o = 0$.

In the equilibrium the type 2s free ride on the contributions of the type 1s. Because of unconditional altruism, however, this free riding does not affect the behavior of the 1s, since the utility $xe^{-\eta_1 y}$ results only from a citizen's own contributions. Figure 5 illustrates the equilibrium and identifies two levels of the free-rider problem. The first free-rider problem is that for matches $y > y_1^o$ both citizens have an incentive to free ride. The second free-rider problem is that for matches $y \in (y_2^o, y_1^o]$ the type 1s contribute and the type 2s free-ride on those contributions. With unconditional altruism the second free-rider problem has no effect on the choice of actions by the type 1s.

The free riding, however, has distributive effects. If the proportion of type 1 citizens is β, the expected utilities $EU_i^\beta, i = 1, 2$, are

$$EU_1^\beta = \frac{1}{1-e^{-\alpha L}}\Big[(2b-c)\Big(1-e^{-\alpha y_1^o}\Big) - (1-\beta)b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^o}\Big) + \frac{\alpha x}{\alpha+\eta_1}\Big(1-e^{-(\alpha+\eta_1)y_1^o}\Big)\Big] \tag{6}$$

and

$$EU_2^\beta = \frac{1}{1-e^{-\alpha L}}\Big[(2b-c)\Big(1-e^{-\alpha y_2^o}\Big) + \beta b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^o}\Big) + \frac{\alpha x}{\alpha+\eta_2}\Big(1-e^{-(\alpha+\eta_2)y_2^o}\Big)\Big]. \tag{7}$$

The term $-(1-\beta)b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^o}\Big)$ in (6) represents the loss to a type 1 from the possibility of being matched with a type 2 on the set $y \in (y_2^o, y_1^o]$, since for these matches the type 2s do not

[12] For example, a citizen cannot signal her type by moving first because that action is not observable.

contribute. The term $\beta b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^o}\Big)$ in (7) represents the gain to a type 2 from the possibility of being matched with a type 1 on that interval and free riding on her contribution.

Although heterogeneity and incomplete information about the preferences of other citizens have no effect on strategies, the expected scope $\frac{1}{L}(\beta y_1^o + (1-\beta)y_2^o)$ of self-regulation depends on the distribution of types. In the case considered here, the expected scope is increasing in β, since a greater proportion of citizens contribute for matches $y \in (y_2^o, y_1^o]$.

The results of this section are summarized in the following proposition.

Proposition 1: With unconditional moral preferences and a homogeneous citizenry self-regulation results only for matches with $y \in [0, y^o]$. The scope of self-regulation and the expected utility of citizens are increasing in the quality of self-regulation and the strength of moral preferences, and the expected utility is increasing in the localness of matches and decreasing in the dispersion in society. Heterogeneity of moral preferences results in equilibria with the same qualitative properties, and the actions of each type is unaffected by the proportion of types. The expected scope of self-regulation is increasing in the proportion of citizens with stronger moral preferences.

D. A Pure Public Good

The model can be extended to pure public goods where a contribution provides benefits to all citizens, as in the case of mitigating global warming. A pure public good could be represented in a number of ways. If a citizen cares about others in proportion to her match distance, the model is strategically equivalent to the local public goods model. To show this, suppose there is a finite number M of citizens. The expected utility EU_{Cj}^M of a citizen j if she contributes and all other citizens contribute is

$$EU_{Cj}^M = b - c + xe^{-\eta y} + \sum_{i \neq j}\Big(b + xe^{-\eta y}\Big).$$

The expected utility EU_{Nj}^M if she does not contribute is $EU_{Nj}^M = \sum_{i \neq j} b$, and the difference is

$$EU_{Cj}^M - EU_{Nj}^M = b - c + x^M e^{-\eta y}, \tag{8}$$

where $x^M = (M-1)x$. Then, (8) is the same as (1) with x^M replacing x.

If a citizen cares about others relative to their distance from her, the concept of a match distance is irrelevant. The expected utility of a citizen if she contributes is then

$$EU_C = b - c + \frac{bM\alpha}{\alpha + \eta}\Big(\frac{1 - e^{-(\alpha+\eta)L}}{1 - e^{-\alpha L}}\Big).$$

This is decreasing in η, increasing in b, and decreasing in c, as is the boundary of self-regulation in (2) and the expected utility in (3).

III. Reciprocal Altruism

A citizen may have conditional altruistic preferences. Altruism may extend only to other citizens who contribute, or warm glow preferences may extend only to the act of providing benefits to someone who deserves them; i.e., who earns them by also contributing. Such preferences can be represented as reciprocal altruism in the sense that firms have altruistic preferences but only when others also contribute.[13] Reciprocity pertains to actions, so a citizen must have beliefs about whether her trading partner will contribute. Since information is complete in the basic model with a homogeneous citizenry, a citizen understands which action her partner will take.

Reciprocal preferences transform the basic self-regulation game from a dominant strategy game to a coordination game. This introduces both complexity and opportunities. The complexity arises because of multiple equilibria, and the opportunity is for organizations to expand the scope of self-regulation. Multiple equilibria result because the utility difference between playing C and N depends on the action of the partner. Reciprocal (or conditional) altruism also provides an opportunity for an organization to affect the scope of self-regulation, as considered in Section V. Both the complexity and the opportunity arise because with reciprocal altruism the self-regulation game has strategic complements.

To characterize the equilibria with reciprocal altruism, let $\delta = \delta(y)$ denote the probability that the partner at a match distance y plays C. The difference in the expected utilities from playing C rather than N then is

$$EU_C - EU_N = b - c + (\delta + \theta(1-\delta))xe^{-\eta y},$$

so the boundary of self-regulation is defined by

$$y^r(\delta;\theta) = \begin{cases} 0 & \text{if } (\delta + \theta(1-\delta))x \leq c - b \\ \frac{1}{\eta}\ln\Big(\frac{(\delta + \theta(1-\delta))x}{c-b}\Big) & \text{if } (\delta + \theta(1-\delta))x > c - b. \end{cases} \tag{9}$$

If $y^r(0;\theta)>0$, the unique equilibrium for matches $y \in [0, y^r(0;\theta)]$ is for a citizen to play C, since she has a dominant strategy as in Section II. Similarly, for $y>y^r(1;\theta) = y^o$ the dominant strategy equilibrium is (N, N), since even if the partner plays C, a citizen cannot gain from contributing.

[13] Tabellini considers reciprocity similar to that considered here, but his basic model assumes strategic complements, so the qualitative properties of his equilibria are unchanged by reciprocity. His formulation of reciprocity corresponds to shame as considered in Section VI and results in a larger maximal scope of self-regulation. Reciprocity here results in a strictly smaller scope of self-regulation compared to unconditional altruism when citizens' moral preferences are heterogeneous.

For matches with $y \in (y^r(0;\theta), y^r(1;\theta)]$, the game is a coordination game with three best-response equilibria. In the Pareto dominant equilibrium both citizens play C, and neither has an incentive to deviate. In this equilibrium the scope of self-regulation is the same as with unconditional altruism. In a second equilibrium, both citizens play N, since if the partner will play N, by playing C a citizen can gain only $b - c + \theta x e^{-\eta y}$, which is negative for $y \in (y^r(0;\theta), y^r(1;\theta)]$. The third equilibrium is in mixed strategies.[14]

The mixed strategy equilibrium and the (N, N) equilibrium identify a role for culture to the extent that it fosters unconditional altruism (as well as generalized morality) or it selects an equilibrium with a greater scope of self-regulation. Culture changes slowly, however, and citizens have the alternative of forming organizations to increase the scope of self-regulation, as considered in Section V. As shown in that section an organization can form with heterogeneous types even when in the absence of the organization all citizens would play the Pareto dominant equilibrium.

With heterogeneous preferences the equilibria are analogous to those with one type. Let the type 1s play C with probability μ and the type 2s play C with probability ρ. Then, $\delta = \beta\mu + (1-\beta)\rho$ is the probability that a partner in a match at y plays C. Define the boundaries $y_i^r(\delta;\theta),\ i = 1, 2,$ by

$$y_i^r(\delta;\theta) = \begin{cases} 0 & \text{if } (\delta + \theta(1-\delta))x \leq c - b \\ \dfrac{1}{\eta_i} \ln\Big(\dfrac{(\delta + \theta(1-\delta))x}{c-b}\Big) & \text{if } (\delta + \theta(1-\delta))x > c - b. \end{cases} \tag{10}$$

[14] In the mixed strategy equilibrium both citizens play C with probability $\delta(y;\theta)$ given by

$$\delta(y;\theta) \equiv \frac{c - b - \theta x e^{-\eta y}}{(1-\theta)x e^{-\eta y}}.$$

That is, given that the partner plays $\delta(y;\theta)$, a citizen is indifferent between playing C or N and hence is willing to play $\delta(y;\theta)$. The probability $\delta(y;\theta)$ is strictly increasing and strictly convex in y, i.e., because of limited morality a higher probability of a partner contributing is required to induce contributions for more distant matches. For matches $y > y^r(1;\theta)$ even a partner contributing with probability 1 is insufficient to induce the citizen to contribute. Note that $\delta(y^r(0;\theta)) = 0$ and $\delta(y^r(1;\theta)) = 1$. If $y^r(0;\theta) = 0$, the minimum probability δ^o of contributing even for a match $y = 0$ is

$$\delta^o = \frac{c - b - \theta x}{(1-\theta)x}.$$

The probability $\delta(y;\theta)$ of playing C is strictly increasing in θ, since

$$\frac{d\delta(y;\theta)}{d\theta} = \frac{c - b - x e^{-\eta y}}{(1-\theta)^2 x e^{-\eta y}} > 0,\ y \in (y^r(0;\theta), y^r(1;\theta)].$$

Consequently, the scope of self-regulation is increasing in the extent to which morality is unconditional rather than reciprocal. In the limit as $\theta \to 1$ contributions from all citizens are induced for matches $y \in [0, y^r(1;\theta)]$, as characterized for the case of unconditional altruism in Section II.

As above if $y_2^r(0;\theta)>0$, the dominant strategy for $y\leq y_2^r(0;\theta)$ is for all citizens to play C. For $y>y_2^r(1;\theta) = y_2^o$, the type 2's have a best response of playing N. If $y_2^r(0;\theta) = 0$, contributing is also a best-response for the type 2s and the type 1s for matches $y \in [0, y_2^o]$.

For type 1s, playing C is a best response if $y \in (y_2^o; y_1^r(0;\theta)]$, which is nonempty if $\theta x>c-b$ and

$$\Big(\frac{c-b}{x}\Big)^{1-\frac{\eta_1}{\eta_2}} \leq \theta. \tag{11}$$

For matches $y \in (\max\{y_2^o, y_1^r(0;\theta)\}, y_1^r(\beta;\theta)]$, which is nonempty for

$$\Big(\frac{c-b}{x}\Big)^{1-\frac{\eta_1}{\eta_2}} \leq \beta + \theta(1-\beta), \tag{12}$$

there are three best-response equilibria. In one, all type 1s play C, and in another all type 1s play N. The third is a mixed strategy equilibrium analogous to that for one type.[15] When (12) is satisfied there are sufficient type 1s that their reciprocal altruism induces them to contribute for $y \in (\max\{y_2^o, y_1^r(0;\theta)\}, y_1^r(\beta;\theta)]$. The second free-rider problem is then present, and the type 2s free ride on the contributions of the type 1s for those matches. The boundary $y_1^r(\beta;\theta)$ is strictly increasing in β and θ.

The inequality in (12) is satisfied, for example, if β or θ is large and η_2 is large, in which case the left side is approximately $\frac{c-b}{x}$. It is also satisfied if $c-b$ is small relative to x. If (12) is not satisfied, there are too few type 1s to induce contributions for $y>y_2^o$. There is then a single equilibrium for each match distance y. For $y\leq y_2^o$ all citizens play C, and for $y>y_2^o$, all citizens play N. In this case the type 2s cannot free ride on the type 1s because the free riding for a match

[15] In one mixed strategy equilibrium, if the type 1s play C, the type 2s have a mixed strategy

$$\rho(y;\theta) = \frac{c-b-(\beta+\theta(1-\beta))xe^{-\eta_2 y}}{(1-\beta)(1-\theta)xe^{-\eta_2 y}},\ y \in (y_2(\beta;\theta), y_2^o],$$

which satisfies $\rho(y_2^r(\beta;\theta)) = 0$ and $\rho(y_2^r(1;\theta)) = 1$. The type 1s play C when the type 2s play $\rho(y)$ for $y\leq y_2^o$ provided that $y_1^r(\beta;\theta)\geq y_2^o$. In this equilibrium, the type 1s play C on $[0, y_2^o]$. The probability $\rho(y;\theta)$ is strictly increasing in θ, so the probability of a contribution by a type 2 must be greater the more unconditional (higher η_2) is the altruism of citizens. There is also an equilibrium in which the type 1s play a mixed strategy $\mu(y;\theta)$ given by, for the case in which $y_2^o<y_1^r(\beta;\theta)$,

$$\mu(y;\theta) = \frac{c-b-\theta xe^{-\eta_1 y}}{\beta(1-\theta)xe^{-\eta_1 y}},\ y \in [y_2^o, y_1^r(\beta;\theta)].$$

The probability $\mu(y;\theta)$ is increasing in y, so a greater likelihood of contributions by the partner is needed to induce contributions for more distant matches. The probability is increasing in θ, so the less conditional is the altruism of citizens the higher is the probability of contributing for a given y.

$y>y_2^o$ would be sufficient to lead the type 1s not to contribute. The scope of self-regulation is then limited by the second free-rider problem when moral preferences exhibit reciprocal altruism. In all three equilibria no type 1 contributes for $y \in (\max\{y_2^o, y_1^r(\beta;\theta)\}, L]$, since a type 1 citizen can only count on a contribution from the other type 1s with whom she might be matched.

Free riding by the type 2s thus limits the self-regulation by the type 1s when altruism is reciprocal. With unconditional altruism the free riding by the 2s has no effect on the strategy of the 1s, but with reciprocal altruism a type 1 receives utility $xe^{-\eta y}$ only when matched with another type 1. This occurs only with probability β, so for matches $y \in (\max\{y_2^o, y_1^r(\beta;\theta)\}, y_1^o]$ the best response is not to contribute. Consequently, when the citizenry is heterogeneous, the scope of self-regulation is smaller with reciprocal than with unilateral altruism. The scope of self-regulation is increasing in β, since then there are fewer type 2s to free ride and more type 1s to reciprocate. Figure 6 illustrates the equilibrium with reciprocal altruism and a heterogeneous citizenry. The curve labeled η_1 is the expected reciprocated utility when only the type 1s contribute. As the figure illustates, the second free-rider problem limits the scope of self-regulation (i.e., $\max\{y_2^o, y_1^r(\beta;\theta)\}<y_1^o$) when moral preferences exhibit reciprocal altruism.

The expected utility $EU_1^r(\beta)$ for a type 1 in the Pareto dominant equilibrium with the greatest scope of self-regulation for the case in which (12) is satisfied is

$$\begin{aligned} EU_1^r(\beta) &= \Big[\int_0^{y_2^o}\Big(2b-c+xe^{-\eta_1 y}\Big)+\int_{y_2^o}^{y_1^r(\beta;\theta)}\Big((1+\beta)b-c+(\beta+\theta(1-\beta))xe^{-\eta_1 y}\Big)\Big]\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}dy \\ &= \frac{1}{1-e^{-\alpha L}}\Big[(2b-c)\Big(1-e^{-\alpha y_1^r(\beta;\theta)}\Big)-(1-\beta)b\Big(e^{-\alpha y_2^o}-e^{-\alpha y_1^r(\beta;\theta)}\Big) \\ &+ \frac{\alpha x}{\alpha+\eta_1}\Big((1-\theta)(1-\beta)e^{-(\alpha+\eta_1)y_1^r(\beta;\theta)}-(1-\beta)(1-\theta)e^{-(\alpha+\eta_1)y_2^o}\Big)\Big], \end{aligned} \tag{13}$$

where $-(1-\beta)b\Big(e^{-\alpha y_2^o}-e^{-\alpha y_1^r(\beta;\theta)}\Big)$ is the effect of the free-riding by the type 2s on a type 1. The expected utility $EU_2^r(\beta)$ for a type 2 for the case in which (12) is satisfied is

$$\begin{aligned} EU_2^r(\beta) &= \Big[\int_0^{y_2^o}\Big(2b-c+xe^{-\eta_2 y}\Big)+\int_{y_2^o}^{y_1^r(\beta;\theta)}\beta b\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)\Big]dy \\ &= \frac{1}{1-e^{-\alpha L}}\Big[(2b-c)\Big(1-e^{-\alpha y_2^o}\Big)+\beta b\Big(e^{-\alpha y_2^o}-e^{-\alpha y_1^r(\beta;\theta)}\Big)+\frac{\alpha x}{\alpha+\eta_2}\Big(1-e^{-(\alpha+\eta_2)y_2^o}\Big)\Big], \end{aligned} \tag{14}$$

where the term $\beta b\Big(e^{-\alpha y_2^o}-e^{-\alpha y_1^r(\beta;\theta)}\Big)$ is the gain to a type 2 from free-riding on the type 1s.

The results of this section are characterized in the following proposition.

Proposition 2: Reciprocal altruism transforms the dominant strategy game into a coordination game. With a homogeneous citizenry the scope of self-regulation in the Pareto dominant equilib-

rium is the same for all $\theta \in [0,1)$ as with unconditional altruism, but the scope is smaller in the other equilibria. In a heterogeneous citizenry the scope of self-regulation in the Pareto dominant, best-response equilibrium is smaller with reciprocal than unconditional altruism because the second free-rider problem reduces the incentive of citizens with stronger preferences to contribute. The expected scope of self-regulation is increasing in θ and β when (12) is satisfied. If (12) is (not) satisfied, citizens with stronger moral preferences contribute for a larger (the same) set of matches than do citizens with weaker moral preferences.

The qualitative results of this section are robust to a change in the model in which the action of a citizen is observable by the other citizen and citizens have an opportunity to more first. Unless the type 1s are numerous, moving first mitigates, but does not eliminate, the second free-rider problem and has no effect on the first free-rider problem. A type i has an incentive to move first for matches $y \in (y_2^o, y_i^+(\beta;\theta)]$, where $y_i^+(\beta;\theta) = L$ if $c \leq (1+\beta)b$ and otherwise

$$y_i^+(\beta;\theta) = \begin{cases} 0 & \text{if } (\beta+\theta(1-\beta))x \leq c-(1+\beta)b > 0 \\ \frac{1}{\eta_i}\ln\Big(\frac{(\beta+\theta(1-\beta))x}{c-(1+\beta)b}\Big) & \text{if } (\beta+\theta(1-\beta))x > c-(1+\beta)b > 0. \end{cases}$$

If the type 1s are not numerous ($\beta < \frac{c-b}{b}$), the boundary $y_1^+(\beta;\theta)$ is strictly less than L, so neither free-rider problem is eliminated. If the type 1s are numerous, both free-rider problem are eliminated, but in that case the free-rider problems were limited in their effect.

IV. Public Regulation

To increase their self-regulation, citizens could demand public regulation to induce or compel themselves to self-regulate. For example, the government could require citizens to purchase a carbon offset when flying or firms to pay a living wage in overseas factories. Regulation and its enforcement, however, is costly, and the cost could exceed the benefits. The regulation considered here does not involve the public provision of the public good or public redistribution but instead requires citizens to contribute. Regulation thus does not affect the private cost c of contributing.[16] In addition, the availability of regulation is assumed not to affect the altruism of citizens.

The first-best is to contribute for matches such that $2b - c + xe^{-\alpha\eta} \geq 0$, so the boundary y^{fb} on contributions is

$$y^{fb} = \begin{cases} \frac{1}{\eta}\ln\Big(\frac{x}{c-2b}\Big) & \text{if } c > 2b \\ L & \text{if } c \leq 2b. \end{cases} \tag{15}$$

[16] Tabellini considers the effect of government enforcement on the educational choices of parents for their children in a model in which players experience guilt if they do not cooperate.

If $c \leq 2b$, a regulator can command self-regulation by all citizens. If $c > 2b$, the regulator would have to observe the distance for each match to implement the first-best, which would be prohibitively costly or impossible. The regulation considered here is thus of two types. One is uniform ex ante regulation where all citizens are required to contribute. The other is regulation on demand where citizens choose to avail themselves of the regulatory powers based on their match distance.

To give ex ante uniform public regulation its best chance, assume that all citizens comply with the regulation.[17] The ex ante expected utility EU^R of a citizen subject to uniform regulation is

$$EU^R = \int_0^L \Big(2b - c + xe^{-\eta y}\Big) \frac{\alpha e^{-\alpha y}}{1 - e^{-\alpha L}} dy - t,$$

where t is the cost per citizen of regulation. The expected gain from regulation for both unconditional and reciprocal altruism is, using (3),

$$\begin{aligned} EU^R - EU^* &= \int_{y^o}^L \Big(2b - c + xe^{-\eta y}\Big) \frac{\alpha e^{-\alpha y}}{1 - e^{-\alpha L}} dy - t \\ &= (2b - c)\Big(\frac{e^{-\alpha y^o} - e^{-\alpha L}}{1 - e^{-\alpha L}}\Big) + \Big(\frac{\alpha x}{\alpha + \eta}\Big)\Big(\frac{e^{-(\alpha+\eta)y^o} - e^{-(\alpha+\eta)L}}{1 - e^{-\alpha L}}\Big) - t. \end{aligned} \tag{16}$$

Letting t^R equate to 0 the expression in (16), regulation is unanimously preferred if the cost is less than t^R, and otherwise self-regulation is preferred. The cut point t^R is increasing in b and x and decreasing in c, η, and L. Higher benefits b from self-regulation has two effects. First, it increases the gain $2b - c + xe^{-\eta y}$ from regulation for matches $y \in (y^o, L]$. Second, it increases the scope of self-regulation which reduces the set on which regulation is beneficial. The net effect, however, is to increase the gain from regulation in (16), so regulation is a normal good. The gain in (16) is also increasing in the strength of moral preferences (lower η), so regulation is a moral normal good. The demand for uniform public regulation thus is not crowded out by self-regulation.

Dixit (2003) considers whether self-enforcement or external enforcement, which is analogous to public regulation, is preferred in a society. He finds that for small L self-enforcement is better from a welfare perspective, whereas for large L external enforcement is better. This is equivalent to t^R in (16) being decreasing in L. That is, for a given $t > 0$ the difference in (16) is negative if the scope of self-regulation is high (y^o sufficiently close to L) in the absence of regulation. This provides an explanation for why a society would not support public regulation yet deal effectively with the externality. Conversely, when self-regulation is less extensive, as in a disparate society (high L) with a low scope of self-regulation, the demand for uniform regulation could be high. Even if all

[17] Alternatively, regulation could involve enforcement, as considered in Section V.C.

citizens ex ante prefer uniform regulation, however, those citizens who would have self-regulated in the absence of regulation are worse off ex post.

The gain from regulation in (16) is the same for both unconditional and reciprocal altruism when citizens are homogeneous and coordinate on the Pareto dominant equilibrium. When the citizenry is heterogeneous, however, the gain for public regulation is greater with reciprocal than unconditional altruism because with the former the second free-rider problem causes the type 1 citizens to restrict their self-regulation, as shown in Section III. Consequently, in a heterogeneous society the demand for uniform public regulation is greater when altruistic preferences are reciprocal than unconditional.

From a positive perspective regulation must be adopted by the citizenry. If a vote were taken ex ante, all citizens would vote for regulation if the cost were less than t^R, and otherwise would vote against it. Citizens, however, can control when they vote, so suppose that the vote takes place after the matches but before citizens play the self-regulation game. To give regulation its best chance, suppose it is selective rather than uniform; i.e., regulation is supplied only when demanded by the citizens in a match. Regulation on demand, however, must be approved by a majority of citizens before it can be made available.

Citizens with matches $y \in [0, y^o]$ would not support regulation, whereas some citizens with matches $y \in (y^o, L]$ would vote in favor of it depending on the cost t per citizen.[18] The cost may be viewed as a user fee, as in the case of verifying that all citizens purchased a carbon offset for their household emissions. The gain to a citizen in such a match is $2b-c+xe^{-\eta y}-t$, where t is incurred only when the matched citizens demand regulation. For $t<b$ citizens with matches $y \in (y^o, y^t(1;\theta)]$ benefit from regulation, where

$$y^t(1;\theta) = \begin{cases} \frac{1}{\eta}\ln\left(\frac{x}{c-2b+t}\right) & \text{if } c-2b+t>0 \\ L & \text{if } c-2b+t\leq 0. \end{cases}$$

Demanding regulation is individually rational only if $t\leq b$, so if $t>b$ no citizen demands regulation.

The citizens adopt regulation in a majority vote only if

$$\frac{y^t(1;\theta)-y^o}{L} > \frac{1}{2},$$

so regulation is never adopted if a majority of citizens would self-regulate; i.e., when the scope of

[18] Only those citizens who strictly benefit from regulation are assumed to vote for it, since regulation could also involve fixed costs covered by taxes rather than fees on those who use the regulation.

self-regulation is greater than $\frac{1}{2}$. Self-regulation then crowds out the political support for public regulation.[19]

The political support $y^t(1;\theta) - y^o$ for regulation on demand is independent of θ and hence is the same for unconditional and reciprocal altruism in a homogeneous society. If $y^t(1;\theta) \in (y^o, L)$, the support for regulation is increasing in the strength of moral preferences and in the quality of self-regulation. Regulation on demand is then a normal good when $y^t(1;\theta)<L$. If $y^t(1;\theta) = L$, the support for regulation is decreasing in x and increasing in η, so the stronger are moral preferences, the lower is the political support for regulation. Similarly, the support for regulation is decreasing in b and increasing in c when $y^t(1;\theta) = L$. The support for regulation on demand then is decreasing in the quality of the public good and in the strength of moral preferences, so self-regulation crowds out public regulation. As with uniform regulation the demand for regulation is greater in a heterogeneous society with reciprocal than with unconditional altruism.

The results of this section are summarized in the following proposition.

Proposition 3: The demand for ex ante uniform regulation is increasing in the quality of self-regulation and in the strength of moral preferences, but the number of matches for which citizens gain from regulation is decreasing in both. Regulation on demand is ex post individually rational, and when regulation benefits every match $y^t(1;\theta) = L$, its political support is decreasing in both the quality of self-regulation and the strength of moral preferences. Regulation on demand is then crowded out by self-regulation. If $y^t(1;\theta)<L$ and regulation is not too costly ($b>t$), the political support for regulation on demand is increasing in the quality of self-regulation and the strength of moral preferences, but that support can be less than a majority. The political support for regulation is the same with unconditional and reciprocal altruism in a homogeneous society, but in a heterogeneous society the demand for regulation is greater with reciprocal than with unconditional altruism.

V. Privately-Organized Self-Regulation

This section considers private alternatives to public regulation where citizens utilize a voluntary, self-regulation organization to affect their behavior. In the context of the model an organization can affect the scope of self-regulation in three ways. First, it could allow citizens to reveal their type, allowing them to coordinate their behavior. Second, an organization could certify that a citizen contributed to the public good in a prior period and make that information available to

[19] Public regulation could also be supported by the citizenry if it lowered the cost c of self-regulating.

citizens in the current period. Third, an organization could provide enforcement that raises the cost of not contributing. Sections V.C and V.D consider enforcement provided by non-profit organizations and by for-profit firms, respectively, and compare the strengths of their enforcement. In Section VI social pressure from NGOs funded by voluntary donations from citizens increases self-regulation. The analysis in these sections does not explain the process by which an organization is formed but instead explains whether citizens avail themselves of the services of the organization.

A. A Social Label Organization

A social label organization allows its members to interact only with other members. For example, fast food chains can restrict their purchases to suppliers that practice humane treatment of food animals, and farmers employing those practices can supply only those chains. Similarly, retailers can buy only from overseas suppliers that meet certain standards for working conditions in their factories, and suppliers meeting those standards can concentrate their sales on retailers that sell products produced under those standards. With public regulation in Section IV, the equilibria are the same with both unconditional and reciprocal altruism, but the equilibria with a social label organization depend importantly on the nature of preferences. A social label organization cannot exist with unconditional altruism, whereas it can exist with reciprocal altruism.

Consider two types of citizens with $\eta_1<\eta_2$. Citizens cannot credibly reveal their types to others, but they can join a social label organization that attracts particular types of members. The organization is open to all citizens, and the social label received, i.e., by joining the organization, is publicly observable. Both types of citizen are assumed to be distributed uniformly on the circle, and assume that a match selects a distance y and places the citizen before citizens of both types. Citizens who join the organization can interact among themselves, and those who do not join interact with other citizens not in the organization. If only type 1s join the organization, they can avoid the free riding by the type 2s, which then increases the scope of their self-regulation. The screening instrument is the membership fee g of the organization.[20]

To determine if a social label organization can attract type 1s but not type 2s, first consider unconditional altruism. Citizens with stronger moral preferences (η_1) can gain because their partner would contribute for matches with distance up to y_1^o rather than y_2^o. Their expected gain relative to having no organization is $\Delta EU_1^u = EU_1^* - EU_1^\beta$, where the superscript u denotes unconditional altruism, EU_1^* is given in (3) with η_1 replacing η and y_1^o replacing y^o, and EU_1^β is given in (6).

[20] The organization can be thought of as formed by an entrepreneur, who receives the membership fees.

This can be evaluated as

$$\Delta EU_1^u = \frac{(1-\beta)b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^o}\Big)}{1-e^{-\alpha L}},$$

which results from avoiding the loss ($(1-\beta)b$) due to free riding by the type 2s.

The social label organization can exist if there is a fee $g \leq \Delta EU_1^u$ that no type 2 would be willing to pay. If a type 2 does not join the organization, he interacts only with other type 2s and his utility is given in (3) with η_2 and y_2^o replacing η and y^o, respectively. If he joins the organization, he interacts only with type 1s and hence free rides with probability one for all matches $y \in (y_2^o, y_1^o]$. The gain ΔEU_2^u for a type 2 is then

$$\Delta EU_2^u = b\Big(\frac{e^{-\alpha y_2^o} - e^{-\alpha y_1^o}}{1-e^{-\alpha L}}\Big), \tag{17}$$

which is greater than the gain to a type 1 citizen, so there is no fee that can separate the types.[21] When altruism is unconditional, adverse selection precludes a social label organization that includes type 1s but not type 2s.[22]

If citizens have reciprocal altruism, a social label organization can result in separation. With reciprocal altruism the type 1s gain not only from the contribution provided by the other type 1s but also from the reciprocation of their altruism. The type 2s have no such gain, since they do not contribute for $y > y_2^o$. To provide the toughest test for an organization, assume that in the absence of an organization citizens play the best of the self-regulation equilibria characterized in Section III. That is, in the absence of an organization the type 1s contribute for $y \in [0, y_1^m(\beta;\theta) \equiv \max\{y_2^o, y_1^r(\beta;\theta)\}]$, whereas the type 2s contribute for $y \in [0, y_2^o]$. If only type 1s join the organization, they interact

[21] A process by which the organization could be formed is as follows. Any citizen can move first and join by paying g, but the type 2s have no incentive to do so. If the types 1s joined and membership in the organization were observable, the type 2s would understand that those who joined were type 1s and hence they would join. The type 1s then could not obtain the gain in (16) and would not join.

[22] An example of a failed social label organization is Responsible Care formed by firms in the chemical industry to improve safety and environmental performance in factories in the aftermath of the Bhopal tragedy. The firms joining Responsible Care included those with good and bad safety and environmental records, and the subsequent performance of the firms that joined the organization was no better than those firms that did not join (King and Lenox, 2000, 2002). The performance of Responsible Care participants subsequently improved after enforcement mechanisms were put in place.

with each other for matches $y \leq y_1^o$. Their expected gain ΔEU_1^r, where ^{r} denotes reciprocal, then is

$$\begin{aligned}\Delta EU_1^r = &\frac{1}{1-e^{-\alpha L}}\Big[(2b-c)\Big(e^{-\alpha y_1^r(\beta;\theta)} - e^{-\alpha y_1^o}\Big) + (1-\beta)b\Big(e^{-\alpha y_2^o} - e^{-\alpha y_1^m(\beta;\theta)}\Big) \\ &+ \frac{\alpha x}{\alpha+\eta_1}\Big((1-\theta)(1-\beta)\Big(e^{-(\alpha+\eta_1)y_2^o} - e^{-(\alpha+\eta_1)y_1^o}\Big) \\ &+ (\theta+\beta(1-\theta))\Big(e^{-(\alpha+\eta_1)y_1^m(\beta;\theta)} - e^{-(\alpha+\eta_1)y_1^o}\Big)\Big)\Big].\end{aligned}$$

The expected gain to a type 2 who joins the organization is given by (17).

If there exists a membership fee g satisfying

$$\begin{aligned}\Delta EU_2^u < g \leq \Delta EU_1^r = \Delta EU_2^u + \frac{1}{1-e^{-\alpha L}}\Big[&\int_{y_1^m(\beta;\theta)}^{y_1^o}\Big(b-c+xe^{-\eta_1 y}\Big)\alpha e^{-\alpha y}dy \\ + &\int_{y_2^o}^{y_1^m(\beta;\theta)}\Big(-\beta b + (1-\theta)(1-\beta)xe^{-\eta_1 y}\Big)\alpha e^{-\alpha y}dy\Big].\end{aligned} \tag{18}$$

self-selection results in an organization with only type 1s as members. To show that such a g exists, consider first the case in which (12) is not satisfied, in which case $y_1^m(\beta;\theta) = y_2^o$ and the second integral in (18) is 0. Then since $b - c + xe^{-\eta_1 y} > 0$ for $y \in (y_2^o, y_1^o)$, there exists a membership fee such that only type 1s join. If $y_1^m(\beta;\theta) > y_2^o$, a sufficient but not necessary condition for the second integral to be non-negative is $c \geq b\Big(1 + \frac{\beta(\beta+\theta(1-\beta))}{(1-\theta)(1-\beta)}\Big)$. To show this, note that $y_1^r(\beta;\theta)$ is defined by $b - c + (\beta + \theta(1-\beta))xe^{-\eta_1 y_1^r(\beta;\theta)} = 0$. Substituting this into the second integrand in (18) yields

$$\frac{(1-\theta)(1-\beta)}{\beta+\theta(1-\beta)}\Big(c - b\Big(1 + \frac{\beta(\beta+\theta(1-\beta))}{(1-\theta)(1-\beta)}\Big)\Big).$$

In the case of pure ($\theta = 0$) reciprocal altruism this reduces to $\frac{1-\beta}{\beta}\Big(c - b\frac{\beta^2}{1-\beta}\Big)$.

For parameter values such that $y_1^m(\beta;\theta) = y_1^r(\beta;\theta) > y_2^o$, the second integral can be negative, whereas the first integral is positive. When $y_1^r(\beta;\theta)$ is low, a membership fee exists satisfying (18), but as $y_1^r(\beta;\theta)$ approaches y_1^o, such a membership fee may not exist. This, however, requires that β or θ be high. For pure reciprocal altruism β must be high, but in that case citizens contribute with high probability for matches $y \in (y_2^o, y_1^o]$, so a social label organization can accomplish little. That is, the type 1s contribute on a sufficiently large set of matches in the absence of an organization that their willingness to pay for an organization is lower than any fee that would not attract type 2s.

As argued in Section IV the political support for public regulation can be crowded out by unorganized self-regulation. To determine if self-regulation crowds out the value of a social label organization, first note that the value due to the organization results from the additional contributions of the type 1s for matches $y \in (y_1^m(\beta;\theta), y_1^o]$. The set on which additional contributions

occur can be shown to be increasing in the quality of self-regulation and the strength of moral preferences.[23] In this sense unorganized self-regulation does not crowd out a social label organization.

Reciprocal rather than unconditional altruism thus can give rise to organized self-regulation. Moreover, the social label organization expands the scope of self-regulation when altruism is reciprocal. That is, the organization enables type 1s to contribute for matches up to y_1^o, whereas with unorganized self-regulation they self-regulate only up to $y_1^m(\beta;\theta)$. A social label organization expands the scope of self-regulation, however, only by eliminating the effect of free-riding by those with weaker moral preferences on the willingness of those with stronger moral preferences to self-regulate. That is, the organization mitigates the second free-rider problem but not the first. A social label organization thus allows citizens to expand the scope of self-regulation but not beyond the scope with unconditional altruism. These results are summarized in the following proposition.

Proposition 4: A social label organization that separates the types can exist in a heterogeneous citizenry when preferences reflect reciprocal altruism but not when altruism is unconditional. The organization expands the scope of self-regulation, but the scope is bounded above by that with unconditional altruism. The increase in self-regulation due to a social label organization is increasing in the quality of self-regulation and the strength of moral preferences, so the demand for a social label organization is not crowded out by unorganized self-regulation.

B. A Certification Organization

A social label organization expands the scope of self-regulation when altruism is reciprocal by separating the types of citizens, which avoids free riding by those with weaker moral preferences. In contrast, a certification organization expands the scope of self-regulation when altruism is reciprocal by inducing citizens with weaker moral preferences to contribute for a larger set of matches so that in the next period they can free ride on citizens with stronger moral preferences. Similarly, citizens with stronger moral preferences have an incentive to separate from those with weaker moral preferences for another set of matches, which also expands the scope of self-regulation.

[23] To show this consider the case in which $y_1^m(\beta;\theta) = y_2^o$. Then,

$$\frac{d(y_1^o - y_2^o)}{db} = \Big(\frac{1}{\eta_1} - \frac{1}{\eta_2}\Big)\Big(\frac{1}{c-b}\Big) > 0,$$

and

$$\frac{d(\tilde{y}_1^o - \tilde{y}_2^o)}{d\Delta\eta} \mid_{\Delta y=0} = \Big(-\frac{1}{\eta_1^2} + \frac{1}{\eta_2^2}\Big)\ln\Big(\frac{x}{c-b}\Big) < 0,$$

where $\tilde{y}_i^o = \frac{1}{\eta_i + \Delta\eta}\ln(\frac{x}{c-b}), i = 1, 2$. When $y_1^m(\beta;\theta) = y_1^r(\beta;\theta)$, the same results obtain.

A certification organization thus can expand self-regulation, but the scope of self-regulation is bounded above by that with unconditional altruism.

This section considers a two-period extension of the model with two types of citizens and reciprocal altruism and shows that a certification organization induces pooling and separation that expands the scope of self-regulation in the first period. To affect the scope of self-regulation, information must be provided to future match partners about a citizen's play in the first period. The information system that accomplishes this is not modeled here. One simple type of information system is for a citizen in the first period match to give a certificate to her partner if and only if he contributed in a match of distance y. The certificate then can be shown to her matched partner next period or posted on a secure Internet site that can be checked by future trading partners. This system, however, has opportunities for fraud, counterfeiting the certificate, or corruption, paying the first-period partner to give a certificate when N is played. A more elaborate information system, as in the model of the law merchant by Milgrom, North, and Weingast (1990) and in Dixit (2003a), however, could resolve the issue of the credibility of the certificate. Alternatively, an independent NGO could grant the certificate. Hence, a citizen who contributes in the first period will be assumed to receive a certificate from an organization that identifies her action along with the socioeconomic distance of the match. With such an information system in place, type 2 citizens can have an incentive to contribute in the first period for some matches in which they would not contribute in a single-period model. In addition, for more matches the type 1s contribute to separate from the type 2s.

For type 2 citizens to contribute in the first period they must be able to free-ride on the contributions of the type 1s in the second period. This requires that (12) be satisfied, which is assumed here, and in addition (11) is assumed not to be satisfied to simplify the exposition. The analysis proceeds by conjecturing an equilibrium with pooling on an interval $(y_2^o, y_2^c]$ in the first period and separation on an interval $(y_1^r(\beta;\theta), y_1^c]$ and determines the set of matches such that no citizen prefers to deviate. The intuition is developed here, and the equilibrium is verified and a proof of Proposition 5 below are presented in the Appendix.

For citizens who pool in the first period for a set of match distances, all their potential period-two partners have the same beliefs about their type at the beginning of period two as at the beginning of period one. The period-two equilibrium for a match between such citizens then is the same as the single-period equilibrium characterized in Section III. Again the Pareto dominant equilibrium is considered, and the expected period-two utility for a type 2 with pooling is $EU_2^r(\beta)$

given in (14). If a type 2 chooses N in the first period for matches in an interval $(y_2^o, y_2^c]$, his type is revealed, and in the second period no partner will contribute for matches $y{>}y_2^o$, as shown in the Appendix. The expected period-two utility then is EU_2^o given in (3) with η_2 and y_2^o replacing η and y^o, respectively. The utility difference ΔEU_2 for a type 2 citizen from playing C in period one versus playing N is then

$$\Delta EU_2 = b - c + (\delta + \theta(1-\delta))xe^{-\eta_2 y} + \tau(EU_2^r(\beta) - EU_2^o), \tag{19}$$

where $\tau \in (0, 1]$ is the discount factor.

The term $b - c + xe^{-\eta_2 y}$ in (19) (for $\delta = 1$) is negative for $y \in (y_2^o, y_1^o]$, so to free ride in period two the type 2 citizen must incur a loss in period one. The gain $EU_2^r(\beta) - EU_2^o$ from free-riding in the second period is positive and independent of the match distance in the first period, whereas the first period loss is increasing in y. Consequently, a type 2 has an incentive to contribute in period one for some y close to y_2^o. The strongest incentive for a type 2 to play N in the first period is for a match $y = y_2^c$, which has the largest period-one loss. For that match the type 2 will not deviate if

$$b - c + xe^{-\eta_2 y_2^c} + \tau(EU_2^r(\beta) - EU_2^o) \geq 0, \tag{20}$$

where $EU_2^r(\beta) - EU_2^o = b\Big(\frac{e^{-\alpha y_2^o} - e^{-\alpha y_1^r(\beta;\theta)}}{1-e^{-\alpha L}}\Big)$. By definition of y_2^o, $c - b = xe^{-\eta_2 y_2^o}$, and substituting this into (20) yields

$$x\Big(e^{-\eta_2 y_2^o} - e^{-\eta_2 y_2^c}\Big) \leq \tau b\Big(\frac{e^{-\alpha y_2^o} - e^{-\alpha y_1^r(\beta;\theta)}}{1 - e^{-\alpha L}}\Big). \tag{21}$$

The right side is positive and independent of y_2^c, whereas the left side is 0 for $y_2^c = y_2^o$ and increasing in y_2^c. Consequently, for a y_2^c sufficiently close to y_2^o the inequality is satisfied. Let $\bar{y}_2^c$ be defined by (21) as an equality. Then, for matches $y \in [0, \bar{y}_2^c]$ the type 2s contribute in the first period, which induces the type 1s to contribute on that interval. The type 2s thus pool with the type 1s over a larger set of matches. The boundary $\bar{y}_2^c$ is strictly increasing in τ, so the greater is the gain from free-riding in the second period the larger is the first-period pooling interval.

As shown in the Appendix the type 1s gain in period two from having their type revealed in period one for matches $y \in (\bar{y}_2^c, y_1^r(\beta;\theta)]$ where they contribute and the type 2s do not contribute. This gain results because a revealed type 1 could be matched with a revealed type 1 in period one, in which case they both contribute for $y \in [0, y_1^o]$. Type 1s thus have an incentive to separate from the type 2s for an additional set of matches $(y \in y_1^r(\beta;\theta), \bar{y}_1^c(\beta;\theta)]$, where $\bar{y}_1^c(\beta;\theta)$ is defined and characterized in the Appendix. A certification organization thus expands the scope of self-regulation for both types 1 and 2 in period one. The organization, however, cannot expand

the scope of self-regulation beyond that with unconditional altruism. That is, like a social label organization a certification organization mitigates the second but not the first free-rider problem.

The results for a certification organization are summarized in the following proposition.

Proposition 5: With reciprocal altruism and a certification organization, assume that (11) is not satisfied and (12) is satisfied.[24] (A) An equilibrium exists in which all citizens contribute in the first period for matches $y \in [0, \bar{y}_2^c]$, $\bar{y}_2^c > y_2^o$. (B) Type 1s contribute for matches $y \in [0, \bar{y}_1^c(\beta;\theta)]$ in period one and thus separate from the 2s, where $\bar{y}_1^c(\beta;\theta) > y_1^r(\beta;\theta)$. (C) There is no equilibrium in which the types separate for all $y \in (y_2^o, y_1^o]$. (D) A certification organization has no effect on behavior if preferences reflect unconditional altruism.

With reciprocal altruism a certification organization increases the scope of self-regulation in the first-period by the type 2s because they are in a dilemma. If they do not contribute for matches $y \in (y_2^o, \bar{y}_2^c]$ they reveal their type and will not be able to free ride on their partner in the second period for matches $y \in [y_2^o, y_1^r(\beta;\theta)]$. So the threat of being excluded by the type 1s in the second period expands the scope of self-regulation by inducing the 2s to pool with the 1s on a larger set of matches. This requires that there is an opportunity for free-riding in period two, which requires that there are sufficient type 1s for (12) to be satisfied. Similarly, the type 1s expand their scope of self-regulation in the first period by separating from the 2s for an additional set of matches. This allows them to expand the scope of their contributions in the second period when matched with a type 1 whose type was revealed by a certificate. In contrast, with unconditional altruism citizens have dominant strategies, so the type 1s have no incentive to separate and the type 2s have no incentive to pool, since they can free ride in the second period without a certificate. A certification organization thus cannot exist with unconditional altruism.

In period two when (12) is satisfied the type 2s contribute only for $y \in [0, y_2^o]$, and the type 1s contribute for $y \in [0, y_2^r(\beta;\theta)]$ if either their or their partner's type was not revealed in period one. If a type 1 is revealed by contributing for matches $(\bar{y}_2^c, \bar{y}_1^c(\beta;\theta)]$, she contributes for $y \in [0, y_1^o]$ if she is matched with another type 1 whose type was revealed. If matched with a type 2 whose type was revealed, she contributes for $y \in [0, y_2^o]$. Letting $z = \frac{e^{-\alpha \bar{y}_2^c} - e^{-\alpha \bar{y}_1^c(\beta;\theta)}}{1 - e^{-\alpha L}}$ denote the probability of a first-period match $y \in (\bar{y}_2^c, \bar{y}_1^c(\beta;\theta)]$, the expected scope of self-regulation for the type 1s is given in $(A3)$ in the Appendix. The difference between that expected scope and the single-period scope

[24] If (12) is not satisfied, the equilibrium is that both types contribute for $y \leq y_2^o$, and neither type contributes for $y > \tilde{y}_1^o$, where $\tilde{y}_1^o < y_1^o$. For a set $(\tilde{y}_2, \tilde{y}_1^o]$ the type 1s contribute and the type 2s do not contribute. For matches $y \in (y_2^o, \tilde{y}_2]$, both types play mixed strategies. A certification organization thus also expands the scope of self-regulation when (12) is not satisfied.

of self-regulation $y_1^r(\beta;\theta)$ is

$$\beta z^2(y_1^o - y_1^r(\beta;\theta)) + (1-\beta)z^2(y_2^o - y_1^r(\beta;\theta)). \tag{22}$$

A certification organization thus expands the scope of self-regulation in both periods if

$$\beta(y_1^o - y_1^r(\beta;\theta)) > (1-\beta)(y_1^r(\beta;\theta) - y_2^o).$$

This condition can be satisfied, for example, if η_2 is close to η_1 in which case the right side is close to 0.

C. Enforcement Organizations

1. Non-Profit Self-Imposed Enforcement

A social label organization screens the types of citizens, and it can expand the scope of self-regulation when citizens have reciprocal altruism. A certification organization expands the scope of self-regulation by providing information to future trading partners about a citizen's play in the first period. An organization could also have an enforcement capability. Citizens may be thought of as pledging to contribute in the face of incentives to free-ride, and enforcement raises the cost of breaking that pledge. The firms participating in the FLA have an incentive to shirk on meeting working condition standards, so independent inspections are used and enforcement takes the form of internal reporting within the FLA and with board approval the release of the inspection reports to the public. Participation is assumed to be voluntary, so citizens subject themselves to enforcement; i.e., enforcement is self-imposed. Citizens self-impose enforcement because it disciplines them to contribute for a larger set of matches. Enforcement provided by a non-profit organization is considered in this section, and enforcement by a for-profit firm is considered in the next section.

Enforcement is assumed to take the form of punishment or harm h in the event that a citizen violates her pledge to contribute.[25] The harm could be reputation damage from public exposure in the case of the members of the FLA. Enforcement is assumed to be available everywhere on the circle, and the strength of enforcement is taken to be h.[26] To make the threat of harm credible, the organization must develop the capability of (i) determining whether a citizen contributed and

[25] In this sense enforcement is analogous to a contract with a penalty for breech. The situations in which self-regulation occurs, however, are generally those in which a contract would be costly to enforce in a court. Moreover, the participants in an organization such as the FLA would be reluctant to turn jurisdiction over to a court.

[26] In a repeated game with random matching Kandori (1992) showed that cooperation is sustained if there is enough local punishment. See also Ellison (1994). Here punishment is administered by the organization, and enforcement is self-imposed.

(ii) delivering harm. Developing the capability has a cost $f(h)$ per citizen that is assumed to be strictly increasing and strictly convex in h. The non-profit organization then charges $f(h)$ to each citizen who demands enforcement. The decision to impose enforcement is assumed to be made by citizens after the match has been drawn but before the play of the self-regulation game.

Consider the case of a homogeneous citizenry and reciprocal altruism.[27] To provide the toughest test for non-profit enforcement, assume that in the absence of the organization the Pareto dominant equilibrium is played. Citizens with matches $y \in [0, y^o]$ have a best response of contributing, so the only citizens with a demand for enforcement are those with more distant matches. A citizen who pledges to contribute and demands enforcement still faces the first free-rider problem and will contribute if

$$(1+\delta)b - c + (\delta + \theta(1-\delta))xe^{-\eta y} - f(h) \geq \delta b - h,$$

so enforcement is demanded only if $f(h)<h$, and attention is restricted to that case. Citizens with matches $y \in (y^o, \hat{y}(1;\theta)]$ benefit from enforcement when $f(h)<h$, where

$$\hat{y}(\delta;\theta) = \begin{cases} 0 & \text{if } (\delta+\theta(1-\theta))x \leq c + f(h) - b - h \\ \dfrac{1}{\eta}\ln\Big(\dfrac{(\delta+\theta(1-\delta))x}{c+f(h)-b-h}\Big) & \text{if } (\delta+\theta(1-\theta))x > c + f(h) - b - h. \end{cases} \tag{23}$$

The organization expands ($\hat{y}(\delta;\theta)>y^o$) the scope of self-regulation for $h>f(h)$, and $\hat{y}(\delta;\theta)$ is decreasing in $f(h)$, so the scope of self-regulation is limited by the cost of organization and enforcement.[28] Demanding enforcement is individually rational for citizens with matches $y>y^o$ if $2b - c + (\delta + \theta(1-\delta))xe^{-\eta y} - f(h) \geq 0$, which requires $f(h)<b$ for matches $y \geq y^o$. That is, the fee must be less than the benefits from the partner's contribution.

The non-profit organization can choose the strength h of its enforcement, and it is assumed to maximize the aggregate utility of those using its enforcement services. The aggregate utility EU^n in the Pareto dominant equilibrium ($\delta = 1$) is

$$EU^n = \int_{y^o}^{\hat{y}(1;\theta)} \Big(2b - c + xe^{-\eta y} - f(h)\Big) \frac{\alpha e^{-\alpha y}}{(1-e^{-\alpha L})} dy.$$

[27] The results in this section also hold for unconditional altruism by letting $\theta = 1$.

[28] If $f(h)$ is paid ex ante as in the case of insurance, it is sunk and does not affect the ex post enforcement choice. For the Pareto dominant equilibrium enforcement then takes place on an interval $(y^o, \tilde{y}(1;\theta)]$, where $\tilde{y}(1;\theta) = L$ if $c - b - h \leq 0$ and otherwise

$$\tilde{y}(1;\theta) \equiv \begin{cases} 0 & \text{if } x \leq c - b - h \\ \dfrac{1}{\eta}\ln\Big(\dfrac{x}{c-b-h}\Big) & \text{if } x > c - b - h. \end{cases}$$

Since $\tilde{y}(1;\theta) > \hat{y}(1;\theta)$, when $f(h)$ is sunk the scope of self-regulation is greater than when $f(h)$ is paid ex post.

The optimal strength h^* of enforcement satisfies the first-order condition, for $\hat{y}(1;\theta)<L$,

$$(b-h^*)\alpha e^{-\alpha\hat{y}(1;\theta)}\Big(\frac{1-f'(h^*)}{\eta(c+f(h^*)-b-h^*)}\Big)-f'(h^*)\Big(e^{-\alpha y^o}-e^{-\alpha\hat{y}(1;\theta)}\Big)=0. \tag{24}$$

The second-order condition is assumed to be satisfied, a sufficient condition for which is $\alpha\geq\eta$. The second term in (24) is the marginal cost of enforcement, and the first term is the marginal gain from enforcement. The marginal effect of enforcement on the scope of self-regulation is proportional to $1-f'(h^*)$, and $b-h^*$ is the incentive to free-ride of the citizen in the most distant match $y=\hat{y}(1;\theta)$ for which enforcement is demanded.

The optimal enforcement by the non-profit organization when $\hat{y}(1;\theta)<L$ is summarized in the following proposition, which is proven in conjunction with the proof of Proposition 7 in Section V.C.2.[29]

Proposition 6: When $\hat{y}(1;\theta)<L$, the optimal strength h^* of enforcement for a non-profit organization has the following properties:

(i) $f'(h^*)<1$; (ii) $f(h^*)<h^*<b$; (iii) $b-h^*>(\leq)\ 2b-c-f(h^*) \iff \hat{y}(1;\theta)<(=)\ L$; (iv) h^* is independent of x and θ; (v) $\hat{y}(1;\theta)>y^o$, so the scope of self-regulation is increased by non-profit enforcement and the increased scope $\hat{y}(1;\theta)-y^o$ is constant in x and θ; (vi) A sufficient condition for h^* to be increasing in c is $\alpha\geq\eta$.

Enforcement extends to the point at which its marginal cost is less than its marginal effect on self-regulation; i.e., $f'(h^*)<1$, and the scope of self-regulation is increased by enforcement since $h^*>f(h^*)$. The optimal strength of enforcement is independent of θ, since contributions occur for matches $y\in[0,\hat{y}(1;\theta)]$. Enforcement is also independent of x, since that parameter affects y^o and $\hat{y}(1;\theta)$ in the same proportion. Non-profit enforcement addresses the first free-rider problem and expands the scope of self-regulation beyond that with unconditional altruism. The individual rationality condition is satisfied since $b>h^*>f(h^*)$, so all citizens with matches $y\in(y^o,\hat{y}(1;\theta)]$ demand enforcement.

Sufficient but not necessary conditions for the demand $\hat{y}(1;\theta)-y^o$ for enforcement to be increasing in the quality of self-regulation and the strength of moral preferences is that h^* is

[29] If the cost of enforcement is sufficiently low, the optimal enforcement could result in contributions by all citizens ($\hat{y}(1;\theta)=L$). The non-profit organization then chooses the strength h^* of enforcement to minimize $f(h)$ and satisfies

$$c+f(h^*)-b-h^*=xe^{-\eta L}.$$

increasing in b and decreasing in η. The effect of b and η on the optimal enforcement, however, is ambiguous in sign. The demand may also be increasing in the quality of self-regulation and the strength of moral preferences if h^* is decreasing in b and increasing in η, particularly if $1 - f'(h^*)$ is small. In these cases, self-selection does not crowd out the demand for non-profit enforcement in the sense that the demand $(\hat{y}(1;\theta) - y^o)$ is increasing in the quality of self-regulation and the strength of moral preferences.

2. For-Profit Self-Imposed Enforcement

A society could rely on the for-profit sector rather than non-profit organizations for enforcement. For-profit enforcement is common in providing security, and organizations such as the FLA require that independent organizations conduct the inspections of overseas factories. This section explores whether enforcement of the form in the preceding section will be supplied in the marketplace by a profit-maximizing firm and how that enforcement compares with that by a non-profit organization. Enforcement on demand and reciprocal altruism are again considered.

The firm chooses the strength h of enforcement and the price p for enforcement. Enforcement overcomes the free-rider problem for matches $y \in (y^o, \hat{y}^\pi(1;\theta)]$, where $\hat{y}^\pi(1;\theta)$ is defined as in (23) with p replacing $f(h)$. The firm has demand only if its enforcement is individually rational ($p<b$) for citizens and overcomes the free-rider problem ($h>p$). The conditions $p<b$ and $h<b$ are ignored in the following analysis and then in Proposition 7 shown to be satisfied in the optimal enforcement policy.

If the firm incurs the cost $f(h)$ of enforcement only for those citizens who demand enforcement, its profit Π is given by[30]

$$\Pi = (p - f(h))\Big(\frac{e^{-\alpha y^o} - e^{-\alpha \hat{y}^\pi(1;\theta)}}{1 - e^{-\alpha L}}\Big),$$

which is assumed to be strictly concave in p and h. In contrast to non-profit enforcement, the firm does not take into account the citizens' utility from altruism other than though its effect on demand.

The optimal price is characterized first, and then the optimal enforcement is characterized. The first-order condition for the optimal price $\hat{p}(h)$ is

$$e^{-\alpha y^o} - e^{-\alpha \hat{y}^\pi(1;\theta)} + (\hat{p}(h) - f(h))\alpha e^{-\alpha \hat{y}^\pi(1;\theta)} \frac{d\hat{y}^\pi(1;\theta)}{dp} = 0, \tag{25}$$

[30] This assumes that the firm cannot observe the match distance and price discriminate based on y.

where

$$\frac{d\hat{y}^{\pi}(1;\theta)}{dp} = -\frac{1}{\eta(c+\hat{p}(h)-b-h)} < 0.$$

A sufficient but not necessary condition for the second-order condition to be satisfied is $\alpha \geq \eta$. The condition in (25) implies that $\hat{p}(h) > f(h)$, so the price charged is greater than the cost of enforcement. A sufficient condition for the price to be strictly increasing in h is $\alpha \geq \eta$. The profit-maximizing enforcement $\hat{h}$ satisfies the first-order condition

$$(\hat{p}(\hat{h}) - f(\hat{h}))\alpha e^{-\alpha\hat{y}^{\pi}(1;\theta)}\frac{\partial\hat{y}^{\pi}(1;\theta)}{\partial h} - f'(\hat{h})\Big(e^{-\alpha y^{o}} - e^{-\alpha\hat{y}^{\pi}(1;\theta)}\Big) = 0. \tag{26}$$

If $h > b$, demand is determined by citizen's individual rationality condition, the price satisfies (25) with b replacing h in $\hat{y}^{\pi}(1;\theta)$, in which case $\frac{\partial\hat{y}^{\pi}(1;\theta)}{\partial h} = 0$. Then, the first term in (26) equals 0, so (26) cannot be satisfied. This implies that the optimal strength of enforcement satisfies $\hat{h} \leq b$, as characterized in (26).

The properties of the equilibrium are summarized in the following proposition and related to those with non-profit enforcement, and the proofs are in the Appendix.

Proposition 7: (A) The optimal enforcement policy of a profit-maximizing firm satisfies:

(i) $f'(\hat{h}) = 1$, (ii) $\hat{p}'(\hat{h}) = 1$; (iii) $b \geq \hat{h} > \hat{p}(\hat{h}) > f(\hat{h})$; (iv) $\hat{y}^{\pi}(1;\theta) > y^{o}$; (v) $\frac{d\hat{p}(\hat{h})}{db} - \frac{d\hat{h}}{db} > 0$; $\frac{d\hat{p}(\hat{h})}{dc} - \frac{d\hat{h}}{dc} < 0$; (vi) $\frac{d\hat{y}^{\pi}(1;\theta)}{db} > 0$; $\frac{d\hat{y}^{\pi}(1;\theta)}{dc} < 0$; (vii) $\hat{p}(h)$ and $\hat{h}$ are independent of x and θ; (viii) $\hat{y}^{\pi}(1;\theta) - y^{o}$ is strictly increasing in the quality of self-regulation and the strength of moral preferences.

(B) The enforcement policies of the non-profit organization and the profit-maximizing firm have the following relations:

(i) $\hat{h} > h^{*}$; (ii) $f'(\hat{h}) > f'(h^{*})$; (iii) $\hat{p}(\hat{h}) > f(h^{*})$; (iv) Both non-profit and for-profit enforcement expand the scope of self-regulation beyond that with unconditional altruism.

At the optimal for-profit enforcement $\hat{h}$, the marginal price equals the marginal cost, since p and h are perfect substitutes in the boundary $\hat{y}^{\pi}(1;\theta)$ of enforcement; i.e., only the difference between $\hat{p} = \hat{p}(\hat{h})$ and $\hat{h}$ affects demand. An increase in the net benefits from self-regulation increases the difference between $\hat{p}$ and $\hat{h}$, which decreases the demand for enforcement. That is, the firm responds to a higher quality opportunity for self-regulation by increasing price by more than it increases the strength of enforcement. The demand $(\hat{y}^{\pi}(1;\theta) - y^{o})$ for for-profit enforcement is strictly increasing in the quality of self-regulation and the strength of moral preferences. The demand for for-profit enforcement thus is not crowded out by unorganized self-regulation.

Enforcement by a for-profit firm is more aggressive ($\hat{h}>h^*$) than that by a non-profit organization. This results because the firm has a first-order incentive to increase its price, which reduces demand. Demand can be increased, however, by more stringent enforcement, which is carried to the point at which the marginal cost equals the marginal revenue product of enforcement. Both for-profit and non-profit enforcement increase the scope of self-regulation beyond the scope with unconditional altruism by addressing the first free-rider problem, but which provides the greater increase cannot be determined without more specific assumptions.

VI. Social Pressure

An alternative to public regulation and organized self-regulation by a non-profit organization or for-profit firm is reliance on social pressure to strengthen the incentives to self-regulate. Environmental NGOs pressure firms to reduce the harmful environmental impacts of their activities. Other NGOs pressure firms to improve the working conditions in overseas factories and to pay a living wage. The basic instrument of activist NGOs is "naming and shaming." That is, identifying a citizen who has failed to contribute and informing others of that failure, resulting in shame. This social pressure typically is funded by voluntary donations by citizens, and those donations face their own free-rider problem. Nevertheless, many NGOs are well-funded by membership dues and donations, which are encouraged by tax-deductibility.

This section introduces an activist NGO funded by voluntary donations by citizens. Those donations support the naming and shaming of citizens who fail to self-regulate in their matches. Social pressure and naming and shaming mitigate the first free-rider problem by imposing harm in the form of shame. The harm to the citizen could vary depending on the nature of the self-regulation problem. For example, an environmental issue could result in more harm than a working conditions issue in overseas factories. Epstein and Schnietz (2002) found that the protests at the failed 1999 Seattle WTO talks resulted in a statistically significant decrease in the market values of firms targeted as environmentally abusive and but had no significant effect on the market value of firms targeted as having abusive labor condition in overseas factories. Similarly, the harm from not addressing an environmental issue could vary by industry. In contrast to the enforcement models in the previous section in which participation by a citizen was voluntary, all citizens are potentially subject to naming and shaming by the activist. Citizen-funded self-regulation thus serves as an enforcement mechanism to mitigate the free-rider problem. Since the capacity of NGOs to monitor self-regulation games is limited, monitoring and enforcement are assumed to be probabilistic and a function of the donations received from citizens.

Assume that each citizen can donate an amount a to the activist, and the total donations A received are used to detect the play in a match with probability $q = Q(A)$, where $Q(0) = 0$ and $Q'(A)>0$. When play is detected, the activist can impose harm h on a citizen who played N, where h is assumed to be exogenous and hence not chosen by the activist but could depend on the action of other citizens. For example, a firm may not incur as much harm if it is revealed to have played N when all other firms also played N. In contrast, if the other firms played C, the firm could incur significant harm. In the latter case the harm results from the public shame of having been disclosed as having played N when others played C. Shame is considered here.

The donations to the activist are made ex ante before the public goods game takes place. With reciprocal altruism and the threat of incurring shame, a citizen contributes if[31]

$$(1+\delta)b - c + (\delta + \theta(1-\delta))xe^{-\eta y} \geq \delta(b - qh). \tag{27}$$

If $qh>c-b$, the expected social pressure is sufficiently great that citizens contribute to the public good for all L. The focus is on the case in which the detection probability and social pressure are not that strong. The equilibrium strategy of a citizen then is to contribute if and only if $y \leq y^q(\delta;\theta)$, where

$$y^q(\delta;\theta) \equiv \begin{cases} 0 & \text{if } (\delta+\theta(1-\delta))x \leq c-b-\delta qh \\ \dfrac{1}{\eta}\ln\Big(\dfrac{(\delta+\theta(1-\delta))x}{c-b-\delta qh}\Big) & \text{if } (\delta+\theta(1-\delta))x > c-b-\delta qh. \end{cases}$$

The boundary $y^q(\delta;\theta)$ is strictly increasing and strictly convex in q, and for $q = 0$, the boundary is $y^0 = y^r(\delta;\theta)$. Consequently, greater donations to the activist expand the scope of self-regulation and expand it beyond the scope with unconditional altruism.

With social pressure all citizens contribute for matches $y \in [0, y^q(1;\theta)]$ in the Pareto dominant equilibrium. A citizen recognizes that her donation to the activist induces other citizens as well as herself to contribute for a larger set of matches. As discussed in Section II a citizen's altruism pertains to the increase in utility her actions provide to others. Her donation increases social pressure which induces self-regulation by citizens for a larger set of matches. The gain she provides through her donation to the two citizens in a match is $4b - 2c + 2xe^{-\eta y}$. This gain is provided for matches $y \in (y^{q'}(\delta;\theta), y^q(\delta;\theta)]$, where q' is the detection probability without her donation and q is the probability with her donation. As an approximation view each citizen as an atom, and assume that there are $2M$ citizens. Thus, $q' = Q(\sum_{j\neq i} a_j)$ and $q = Q(\sum_j a_j)$, where a_j is the donation of

[31] If the activist harms the citizen whenever N is played, the right side of (27) is replaced by $\delta b - qh$. The Pareto dominant equilibrium is the same as with shame.

citizen j. The expected utility EU_i^A of a citizen i in the Pareto dominant equilibrium is then

$$\begin{aligned} EU_i^A = &\int_0^{y^{q'}(1;\theta)} \Big(2b - c + xe^{-\eta y}\Big)\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)dy \\ &+ \int_{y^{q'}(1;\theta)}^{y^{q}(1;\theta)} M\Big(4b - 2c + 2xe^{-\eta y}\Big)\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)dy - a_i. \end{aligned} \tag{28}$$

Donations thus are motivated in (28) by the gains to citizens in the additional matches in which they contribute in self-regulation game.

The optimal donation by a citizen satisfies the first-order condition which in a symmetric equilibrium is

$$2M(b - Q(A^*)h)\Big(\frac{\alpha e^{-\alpha y^q(1;\theta)}}{1-e^{-\alpha L}}\Big)\Big(\frac{Q'(A^*)h}{\eta(c-b-Q(A^*)h)}\Big) - 1 = 0, \tag{29}$$

where $A^* = \sum_i a_i^*$ is the equilibrium aggregate contributions. A sufficient condition for the second-order condition to be satisfied is $Q(A)$ concave and $\alpha{>}\eta$.[32] The effect of the donation on the scope of self-regulation is valued at the marginal incentive $(b - Q(A^*)h)$ to free ride multiplied by the marginal probability $\frac{\alpha e^{-\alpha y^q(1;\theta)}}{1-e^{-\alpha L}}$ of a match in the expanded set of matches on which contributions take place.[33]

The form of the first-order condition in (29) is directly analogous to that in (24) for non-profit enforcement and in (26) for for-profit enforcement. All three first-order conditions equate the marginal cost to the organization or donor to the marginal benefit composed of the incentive to free-ride of the citizen at the boundary of contributions, the marginal probability of being at the boundary scaled by the interval on which enforcement is applied, and the marginal effect of enforcement on the boundary of contributions.

Since each citizen takes the utility of all citizens into account at the margin, the social pressure is collectively optimal. Although social pressure expands the scope of self-regulation by addressing the first free-rider problem, a free-rider problem remains for some matches unless the detection probability is high. The results of this section are summarized as:

[32] The second derivative is

$$\frac{d^2EU_i^A}{da_i^2} = 2M(b-Q(A^*)h)\Big(\frac{\alpha e^{-\alpha y^q(1;\theta)}}{1-e^{-\alpha L}}\Big)\Big[(-\alpha+\eta)\Big(\frac{Q'(A^*)h}{\eta(c-b-Q(A^*)h)}\Big)^2 + \frac{hQ''(A^*)}{\eta(c-b-Q(A^*)h)^2}\Big],$$

which is negative for $\alpha{>}\eta$.

[33] If the detection probability is sufficiently high that $b{\leq}Q(A^*)h$, all citizens contribute to the public good. Then, (29) is not satisfied, and citizens do not donate to the activist at the margin. Citizens then have a participation game.

Proposition 8: Social pressure funded by voluntary donations increases the scope of self-regulation by addressing the first free-rider problem. With reciprocal (or unconditional) altruism donations to the activist in the Pareto dominant equilibrium are collectively optimal given the technology $Q(A)$, but social pressure results in the second-best self-regulation unless $Q(A^*)h \geq c-b$. The equilibrium contribution a^* is strictly decreasing in x, and $\frac{da^*}{dc} > (=)(<) 0$ as $\alpha > (=)(<) \eta$ and $\frac{da^*}{db} > 0$ if $\alpha \leq \eta$.

The effect on the donations a^* of the quality of self-regulation and the strength of moral preferences cannot be identified unambiguously. If a^* is nondecreasing in b and nonincreasing in η, the impact $y^q(1;\theta) - y^o$ on self-regulation of social pressure is increasing in the quality of self-regulation and the strength of moral preferences. The same is true if a^* is decreasing in b and increasing in η but not "too large."[34] If these conditions are satisfied, the demand for social pressure is not crowded out by unorganized self-regulation.

VIII. Conclusions

Self-regulation includes the private provision of public goods and private redistribution and can result from a variety of motivations, including self-interest, forestalling public or private politics, and moral concerns. Two forms of moral preferences that seem natural are limited morality and reciprocal altruism. Limited morality can overcome the incentive to free ride for interactions among citizens who are close on some dimension but not for those who are distant. With reciprocal altruism a second free-rider problem can further limit self-regulation.

Self-regulation with moral preferences thus faces two levels of a free-rider problem. The first occurs for matches in which both players have incentives not to self-regulate. The second occurs in a heterogeneous society for matches in which citizens with stronger moral preferences self-regulate and those with weaker moral preferences free ride.

With heterogeneous moral preferences reciprocal altruism results in a smaller scope of self-regulation than does unconditional altruism. This is due to the second free-rider problem – citizens with weaker moral preferences free ride on the self-regulation by those with stronger moral preferences, which reduces their self-regulation. Some voluntary organizations help citizens mitigate this second free-rider, whereas other organizations mitigate the first free-rider problem. Social label and certification organizations are limited in the sense that they deal only with the second free-

[34] That is,

$$\frac{d(y^q(1;\theta) - y^o)}{db} = \frac{h}{\eta}\Big(\frac{Q(A^*) + (c-b)Q'(A^*)\frac{da^*}{db}}{(c-b-Q(A^*)h)(c-b)}\Big),$$

which is positive if $\frac{da^*}{db} > -\frac{Q(A^*)}{(c-b)Q'(A^*)}$.

rider problem, whereas enforcement and social pressure organizations deal with the first free-rider problem.

When moral preferences reflect reciprocal altruism, social label and certification organizations increase the scope of self-regulation. A social label organization allows those with similar preferences to interact among themselves. This allows citizens with stronger moral preferences to avoid the second free-rider problem, which then elicits expanded self-regulation by those citizens. A social label organization expands the scope of self-regulation by sorting citizens according to the strength of their moral preferences. A certification organization expands self-regulation by inducing pooling for some matches by those citizens with weaker moral preferences and for other matches separation by those with stronger moral preferences. Pooling results because the opportunity to free ride in the second period outweighs the loss from contributing in the first period, which with limited altruism is small for some matches. Citizens with stronger moral preferences have an incentive to separate from those with weaker moral preferences by self-regulating for other matches so that they can increase the mutually beneficial self-regulation in the second period. Social label and certification organizations address the second free-rider problem and hence expand the scope of self-regulation with reciprocal altruism but not beyond that which would result with unconditional altruism. Neither social label nor certification organizations can exist with unconditional altruism.

A voluntary enforcement organization expands the scope of self-regulation beyond that with unconditional altruism by addressing the first free-rider problem and does so by imposing harm on a citizen who fails to contribute. Enforcement is self-imposed by citizens and can be provided by both non-profit and for-profit organizations. A profit-maximizing firm provides stronger enforcement than does a non-profit organization, but it charges a price higher than the fee required by the non-profit organization. The demand for enforcement by a for-profit firm is increasing in the quality of self-regulation and the strength of moral preferences.

An alternative to voluntary, private organization is to rely on social pressure by an activist NGO funded by voluntarily donations by citizens. Naming and shaming can impose harm on citizens by publicly disclosing their failure to self-regulate. Citizens face a collective action problem in their donations to the NGO, but reciprocal altruism is sufficient to overcome this problem. The scope of self-regulation is increased, since social pressure addresses the first free-rider problem, but unless the detection probability is high, the first free-rider problem is not eliminated.

Public regulation is an alternative to voluntary self-regulation and the organizations that support it, but the political support for public regulation can be crowded out by self-regulation.

From a normative perspective the value of public regulation is increasing in the quality of self-regulation and the strength of moral preferences. From a positive perspective the political support for regulation is reduced by self-regulation, and if the scope of self-regulation is greater than half the citizenry, regulation is not adopted. Since private organizations expand the scope of self-regulation, they also reduce the political support for public regulation. The demand for public regulation is greater with reciprocal than with unconditional altruism, but so is the incentive to form a private self-regulation organization.

Self-regulation organizations have two advantages and one disadvantage relative to public regulation. First, they can be formed quickly by a minority without majority approval and can respond to particular forms of the free-rider problem. Second, the demand or value of the private organization may not be crowded out by self-regulation as can be the political support for public regulation. The disadvantage of private organizations is that their enforcement activities may only mitigate the free-rider problem, whereas in principle, but perhaps not in practice, public regulation can eliminate the problem.

Appendix

Period-Two Equilibrium with a Certification Organization

As in Section V.B assume that (11) is not satisfied and (12) is satisfied, so a type 2 has an incentive to free ride in the second period for $y \in (y_2^o, y_1^r(\beta;\theta)]$. In the first period separation occurs for matches $y \in [y_1^r(\beta;\theta), \bar{y}_1^c(\beta;\theta)]$, where $\bar{y}_1^c(\beta;\theta)$ is defined below. The equilibrium in the second period depends on whether a citizen's type has been revealed. If a partner's type has not been revealed, the citizen's beliefs about the type of her partner are the same as her prior beliefs. The second-period equilibrium is then as characterized for the single-period model in Section III.

Consider a period-two match of a citizen whose type has been revealed and a citizen whose type has not been revealed. A revealed type 2 has a best response of contributing for $y \in [0, y_2^o]$, since both a type 1 and a type 2 partner have a best response of contributing for $y \in [0, y_2^o]$. A revealed type 1 has a best response of contributing for $y \in [0, y_1^r(\beta;\theta)]$, since a type 1 partner's best response is to contribute for $y \in [0, y_1^r(\beta;\theta)]$ and a type 2 partner's best response is to contribute for $y \in [0, y_2^o]$. A type 1 citizen whose type has not been revealed thus has a best response of contributing for $y \in [0, y_1^r(\beta;\theta)]$ with a revealed type 1 and for matches $y \in [0, y_2^o]$ with a revealed type 2. A type 2 citizen whose type has not been revealed contributes only for matches $y \in [0, y_2^o]$.

Next consider a match of a type i and type j, $i = 1, 2$, $j = 1, 2$, each of whose types has been revealed. Two type 1s have best responses to contribute for $y \in [0, y_1^o]$, and two type 2's have best responses of contributing for $y \in [0, y_2^o]$. A type 1 and a type 2 have best responses of contributing for $y \in [0, y_2^o]$, since (11) is not satisfied.

In period two a type 2 whose type has been revealed is in equilibria in which his partner contributes only for $y \in [0, y_2^o]$. His expected utility EU_2^o is thus given by (3) with η_2 and y_2^o replacing η and y^o, respectively.

A type 1 whose type has been revealed is in equilibria in which both players contribute for $y \in [0, y_2^o]$, and with probability βq she is in equilibria in which she and her partner contribute for $y \in [0, y_1^o]$. With probability $(1-\beta)z$ she is in equilibria in which she and her type 2 partner contribute for $y \in [0, y_2^o]$, where

$$z = \frac{e^{-\alpha y_1^r(\beta;\theta)} - e^{-\alpha \bar{y}_1^c(\beta;\theta)}}{1 - e^{-\alpha L}}$$

is the probability that a type 1 citizen receives a certificate for first-period matches $y \in (y_1^r(\beta;\theta), \bar{y}_1^c(\beta;\theta)]$.

The period-two expected utility $EU_1^c(\beta)$ of a revealed type 1 is thus

$$\begin{aligned} EU_1^c(\beta) &= \Big[(1-z)\int_0^{y_2^o}\Big(2b-c+xe^{-\eta_1 y}\Big)+(1-z)\int_{y_2^o}^{y_1^r(\beta;\theta)}\Big(b(1+\beta)-c+(\beta+\theta(1-\beta))xe^{-\eta_1 y}\Big) \\ &\quad +\beta z\int_0^{y_1^o}\Big(2b-c+xe^{-\eta_1 y}\Big)+(1-\beta)z\int_0^{y_2^o}\Big(2b-c+xe^{-\eta_1 y}\Big)\Big]\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)dy \\ &= EU_1^r(\beta)-z(1-\beta)\int_{y_2^o}^{y_1^r(\beta;\theta)}\Big(b-c+\theta xe^{-\eta_1 y}\Big)\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)dy \\ &\quad +z\beta\int_{y_1^r(\beta;\theta)}^{y_1^o}\Big(2b-c+xe^{-\eta_1 y}\Big)\Big(\frac{\alpha e^{-\alpha y}}{1-e^{-\alpha L}}\Big)dy, \end{aligned} \tag{A1}$$

which satisfies $EU_1^c(\beta)>EU_1^r(\beta)$.

The gain in period two identified in (A1) provides an incentive for a type 1 to expand the set for which she contributes. A type 1 then plays C in period one for matches such that

$$b-c+(\beta+\theta(1-\beta))xe^{-\eta_1 y}+\tau(EU_1^c(\beta)-EU_1^r(\beta))\geq 0. \tag{A2}$$

Defining $\bar{y}_1^c(\beta;\theta)$ by the equality in (A2) (where z is a function of $\bar{y}_1^c(\beta;\theta)$), a type 1 contributes for $y\in[0,\max\{\bar{y}_1^c(\beta;\theta),y_1^o\}]$ in period one. This does not affect the strategy of a type 2 in period one. The expected utility for a type 2 does not receive a certificate in period one is $EU_2^r(\beta)$.

The scope of self-regulation in period two: A type 2 contributes for $y\in[0,y_2^o]$ regardless of whether he received a certificate in period one. A type 1 contributes for $y\in[0,y_1^r(\beta;\theta)]$ if her type has not been revealed or her partner's type not been revealed. If the types of both partners have been revealed, a type 1 contributes for $y\in[0,y_1^o]$ when matched with a type 1. The probability that the types of both partners were revealed in period one is z^2, so the expected scope of self-regulation by type 1s is

$$(1-z^2)y_1^r(\beta;\theta)+\beta z^2 y_1^o+(1-\beta)q^2y_2^o, \tag{A3}$$

which is used to obtain (22).

Proof of Proposition 5: (A) Conjecture an equilibrium in which all citizens contribute for matches $y\leq\bar{y}_2^c$. The play in the first period thus provides no information about the type of a citizen with a match in that interval, so the beliefs of all citizens are the same as their prior beliefs. The equilibrium in the second period is then that characterized in Section III in the Pareto dominant equilibrium. The expected period-two utilities are then $EU_1^r(\beta)$ and $EU_2^r(\beta)$ given in (13) and (14), respectively.

A type 2 has a best-response strategy of playing C for $y \in [0, y_2^o]$ in period one, so consider a $y \in (y_2^o, \bar{y}_2^c]$. If a type 2 deviates by playing N, she avoids the loss $b - c + xe^{-\eta_2 y}$ in period one and is revealed as a type 2 under the standard off-the-equilibrium-path belief refinements. Consider the Pareto dominant period-two equilibrium in which the revealed type 2 plays C for $y \in [0, y_2^o]$. As shown above in the proof of the period-two equilibrium any period-two partner has a best response of contributing for $y \leq y_2^o$ and not contributing for $y > y_2^o$. As argued in the text for $y \in (y_2^o, \bar{y}_2^c)$, the gain in period one from playing N is exceeded by the loss in period two. A type 2 thus has no incentive to deviate by playing N on $y \in [0, \bar{y}_2^c]$.

Similarly, a type 2 has no incentive to deviate by playing C for $y > \bar{y}_2^c$. If he deviates and is believed to be a type 1, in the second period his expected utility will be $EU_2^r(\beta)$ rather than EU_2^o if he had played N. By definition of $\bar{y}_2^c$ a type 2 has no incentive to so deviate.

Consider a type 1 in period one. If the type 1 deviates and plays N for a match $y \in [0, \bar{y}_2^c]$ her period one utility is lower than if she plays C, since $b - c + xe^{-\eta_1 y} > 0$. Also, citizens believe that she is a type 2 under the standard off-the-equilibrium-path belief refinements, and her period two partners will contribute only for matches $y \in [0, y_2^o]$. A type 1 thus has no incentive to deviate by playing N for $y \in [0, \bar{y}_2^c]$.

Consider a deviation by a type 1 of playing C for a period one match $y > \bar{y}_1^c(\beta; \theta)$. This results in a lower utility in period 1 and reveals the citizen as a type 1. As shown in the proof of the period-two equilibrium, a type 1 cannot gain for $y > \bar{y}_1^c(\beta; \theta)$ by definition of $\bar{y}_1^c(\beta; \theta)$.

(B) This has been shown in the text.

(C) To show that there is no separating equilibrium on $(y_2^o, y_1^r(\beta; \theta)$, consider the incentives of a type 2 to contribute for some $y > y_2^o$. A type 2 gains $\tau b\left(\frac{e^{-\alpha y_2^o} - e^{-\alpha y_1^r(\beta;\theta)}}{1 - e^{-\alpha L}}\right)$ from free riding on the 1s in period two and loses $c - b - x^{-\eta_2 y}$ in the first period. As argued in the context of (21), for some $y > y_2^o$ the loss in period one is exceeded by the gain in period 2, so a type 2 will play C. Q.E.D.

(D) With unconditional altruism a citizen has a dominant strategy in both periods, and the argument presented in Section V.B shows that a certification organization cannot expand self-regulation.

Proof of Proposition 7: A(i) follows from substituting (25) into (26). Differentiating $\hat{p}(h)$ in (25) and using A(i) yields A(ii). A(iii) is implied by (26) and the requirement that the firm is profitable. A(iv) is implied by (26) given A(iii). A(v) follows from totally differentiating (25) and (26). A(vi) follows because $\hat{y}^\pi(1; \theta)$ is decreasing in $\hat{p}(h) - h$. A(vii) results because (25) and (26)

are independent of x and independent of θ when $\delta = 1$. A(viii) follows from totally differentiating (26) with respect to $\hat{h}$ and b, c, and η.

Proof of Proposition 6: The derivative $\frac{dEU^n}{dh}$ in (24) evaluated at $h = \hat{h}$ is negative, which given the strict concavity of EU^n at h^* implies that $h^* < \hat{h}$. Then, since $f(h)$ is strictly increasing, $f(h^*) < f(\hat{h}) = 1$. Then, $b > h^*$ is implied by (24). Property (iii) then follows directly from $\hat{y}(1;\theta) < L$. That h^* is independent of x and θ follows directly from (24), which is independent of x and of θ when $\delta = 1$. Property (vi) follows directly from differentiating (24) and simplifying.

Parts B(i) and (ii) of Proposition 7 have been established above, and (iii) is immediate from (26) and B(ii). B(iv) is immediate from $h^* > f(h^*)$ and $\hat{h} > \hat{p}(\hat{h})$. Q.E.D.

References

Andreoni, James. 1988. "Privately provided public goods in a large economy: the limits of altruism." *Journal of Public Economics*. 35: 57-73.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving." *Economic Journal*. 100: 464-477.

Andreoni, James and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*. 70: 737-753.

Banfield, E.C. 1958. *The Moral Basis of a Backward Society*. New York: The Free Press.

Baron, David P. 2007a. "Corporate Social Responsibility and Social Entrepreneurship." *Journal of Economics and Management Strategy*. 16: 683-717.

Baron, David P. 2007b. "Managerial Contracting and Corporate Social Responsibility." *Journal of Public Economics*. (forthcoming)

Baron, David P. 2007c. "A Positive Theory of Moral Management, Social Pressure, and Corporate Social Performance." *Journal of Economics and Management Strategy*. (forthcoming).

Baron, David P. and Daniel Diermeier. 2007. "Strategic Activism and Nonmarket Strategy." *Journal of Economics and Management Strategy*. 16: 599-634.

Bohnet, Iria and Bruno S. Frey. 1999. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *American Economic Review*. 89: 335-339.

Calveras, Aleix, Juan-Jose Ganuza, and Gerard Llobet. 2007. "Regulation, Corporate Social Responsibility and Activism." *Journal of Economics and Management Strategy*. 16: 719-740.

Dixit, Avinash. 2003a. "On Modes of Economic Governance." *Econometrica*. 71: 449-481.

Dixit, Avinash. 2003b. "Trade Expansion and Contract Enforcement." *Journal of Political Economy*. 111: 1293-1317.

Dixit, Avinash. 2004. *Lawlessness and Economics*. Princeton University Press: Princeton, NJ

Ellison, Glenn. 1993. "Learning, Local Interaction, and Coordination." *Econometrica*. 61: 1047-1072.

Ellison, Glenn. 1994. "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *Review of Economic Studies*. 61: 567-588.

Epstein, Marc J. and Katherine Schnietz. 2002. "Measuring the Cost of Environmental and Labor Protest to Globalization: An Event Study of the Failed 1999 Seattle WTO Talks." *International Trade Journal*. 16: 129-160.

Eshel, Ilan, Larry Samuelson, and Avner Shaked. 1998. "Altruists, Egoists, and Hooligans in a Local Interaction Model." *American Economic Review*. 88: 157-179.

Graff Zivin, Joshua and Arthur Small. 2005. "A Modigliani-Miller Theory of Altruistic Corporate Social Responsibility." *Topics in Economic Analysis & Policy*. 5: Article 10.

Kandori, Michihiro. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies*. 59: 63-80.

Keser, Claudia and Frans van Winden. 2000. "Conditional Cooperation and Voluntary Contributions to Public Goods." *Scandinavian Journal of Economics*. 102: 23-39.

King, Andrew A. and Michael J. Lenox. 2000. "Industry Self-Regulation with Sanctions: The Chemical Industry's Responsible Care Program." *Academy of Management Journal*. 43: 698-716.

King, Andrew A. and Michael J. Lenox. 2002. "Exploring the Locus of Profitable Pollution Reduction." *Management Science*. 48: 289-299.

Kotchen, Matthew J. 2006b. "Green Markets and Private Provision of Public Goods." *Journal of Political Economy*. 114: 816-834.

La Ferrara, Eliana. 2003. "Kin Groups and Reciprocity: A Model of Credit Transactions in Ghana." *American Economic Review*. 93: 1730-1759.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*. 1: 593-622.

Levy, Gilat and Ronny Razin. 2007. "A Theory of Religious Organizations." Working paper, London School of Economics, London, UK.

Lyon, Thomas P. and John W. Maxwell. 2004. *Corporate Environmentalism and Public Policy*. Cambridge, UK: Cambridge University Press.

Maxwell, John W., Thomas P. Lyon, and Steven C. Hackett. 2000. "Self-Regulation and Social Welfare: The Political Economy of Corporate Environmentalism." *Journal of Law and Economics*. 43: 583-618.

Milgrom, Paul, Douglas North, and Barry Weingast. 1990. "The Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics*. 2:1-20.

Prakash, Aseem and Matthew Potoski. 2006. *The Voluntary Environmentalists*. Cambridge, UK: Cambridge University Press.

Prakash, Aseem and Matthew Potoski, eds. 2007. *Voluntary Programs: A Club Theory Perspective*. MIT Press, Cambridge, MA (forthcoming).

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review.* 83: 1281-1302.

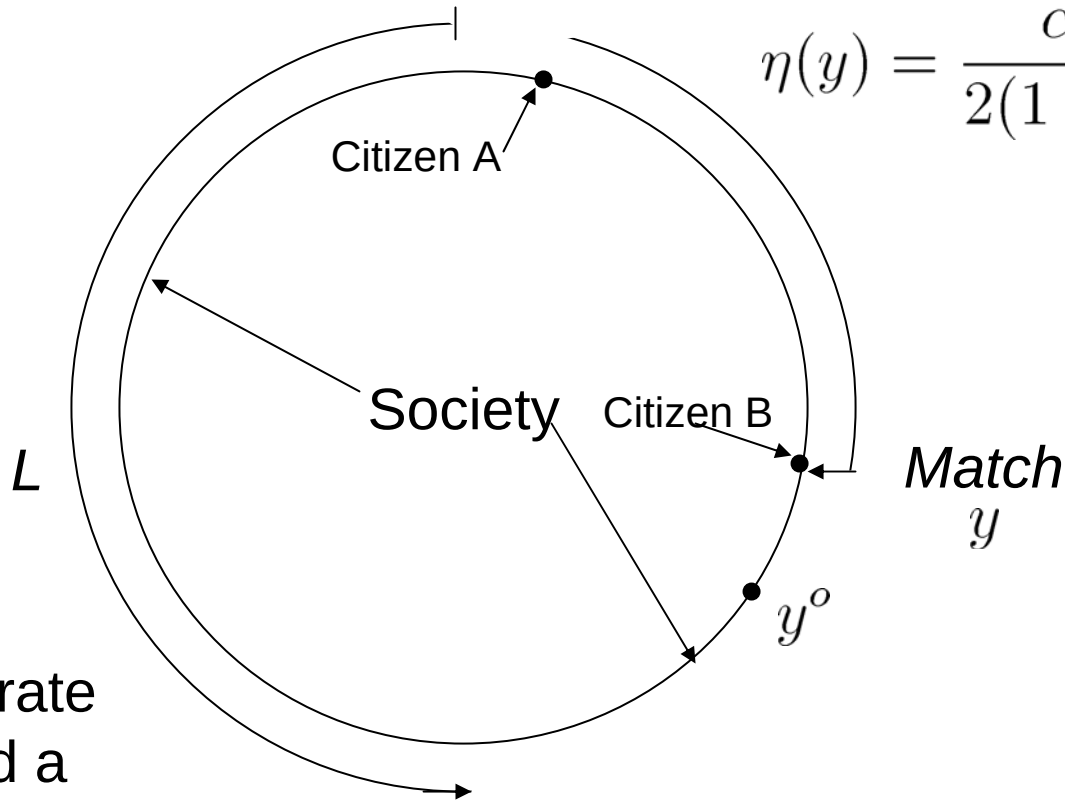
Rabin, Matthew. 1998. "Psychology and Economics." *Journal of Economic Literature.* 36: 11-46.

Siegel, Donald S. and Donald F. Vitaliano. 2007. "An Empirical Analysis of the Strategic Use of Corporate Social Responsibility." *Journal of Economics and Management Strategy.* 16: 773-792.

Tabellini, Guido. 2007. "The Scope of Cooperation: norms and incentives." Working paper, Bocconi University.

Figure 1

Society and Matching
(continuum of citizens located on a circle)



L is how disparate or factionalized a society is.
α parameterizes how close or distant matches are.

Figure 2

Representation of Moral (Altruistic/Warm Glow) Preferences

	Generalized	Limited
Unconditional	x	$xe^{-\eta y}$
Conditional (reciprocal)	θx	$\theta x e^{-\eta y}$

$$\theta \in [0, 1)$$

Figure 3

Basic Self-Regulation Game
Unconditional Altruism

Citizen B

Citizen A	Contribute	Not contribute
Contribute	$2b - c + xe^{-\eta y}$ $2b - c + xe^{-\eta y}$	$b - c + xe^{-\eta y}$ b
Not contribute	b $b - c + xe^{-\eta y}$	0 0

Strategic neutrality

Assumptions: $b - c + xe^{-\eta L} < 0$

$$b - c + x > 0$$

Figure 4

Basic Model
Unconditional Altruism

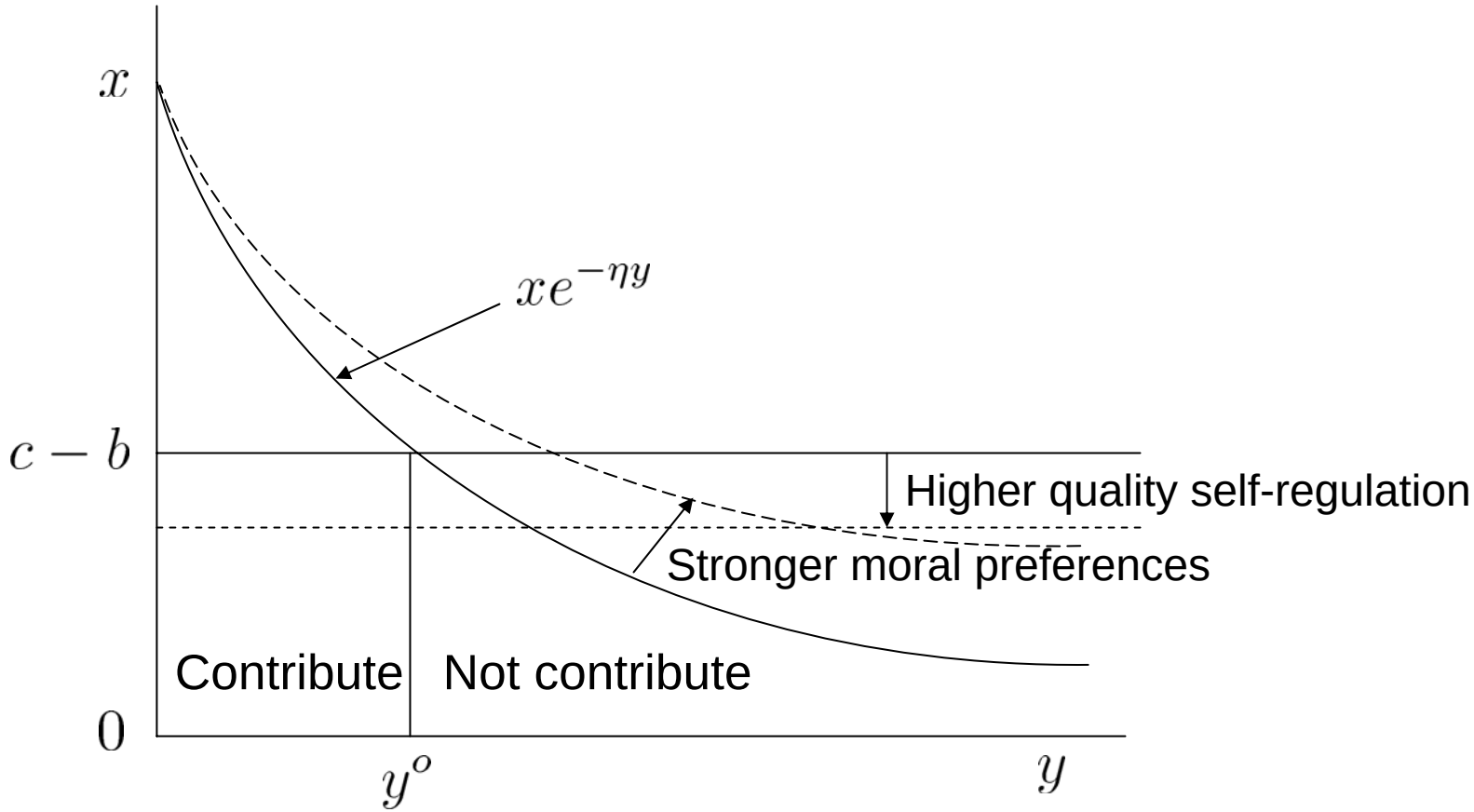


Figure 5
Basic Model
Unconditional Altruism
Heterogeneous Citizens

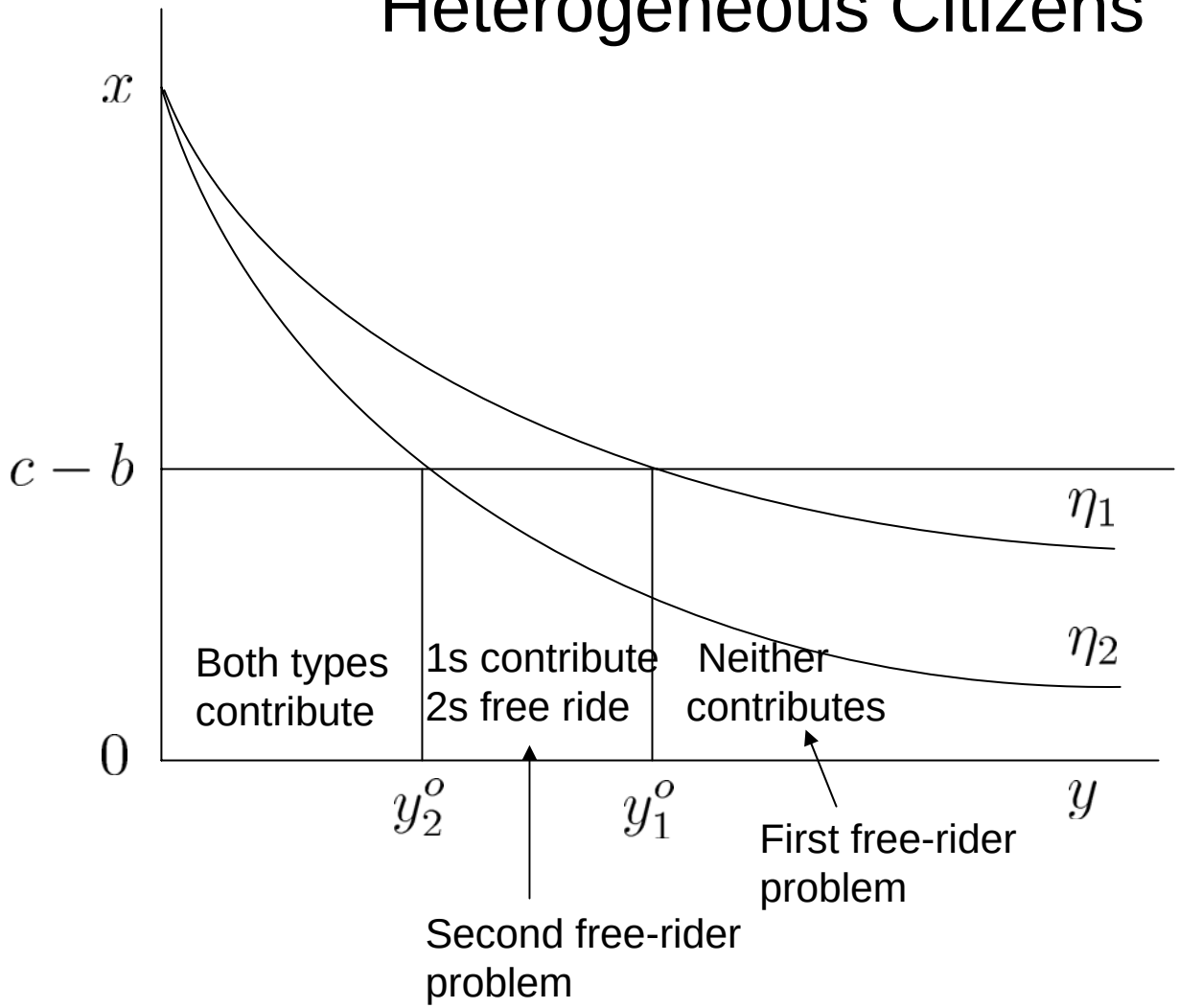


Figure 6
Reciprocal Altruism
Heterogeneous Citizenry