Social norms and the tragedy of the commons

Inkee Jang*

April 15, 2016

Abstract

I study an environment in which selfish and normative agents share a common resource. Using Gul and Pesendorfer's temptation setup, I assume that normative individuals have a preference for reciprocity as well as a temptation to be selfish. I discuss the strategic interaction among players in this setting when types are public information. I show under which conditions there exist equilibria in which selfish players cooperate, avoiding the tragedy of the commons. I also show that in some circumstances, the social planner can Pareto-improve the social outcome by hiding or manipulating information.

Keywords: Reciprocity, Self control and temptation, The tragedy of the commons

1 Introduction

Traditional economic research has been established premising the assumption that human beings are self-interested. Many studies, however, recently question whether selfishness effectively explains certain social phenomenon to suggest alternative preferences. Preference of reciprocity is of increasing interest for many economists to complement what selfishness cannot explain. Rabin [1993] and Levine [1998] argue theoretically that being nice to others

*Department of Economics, Washington University in St. Louis, MO, USA.
E-mail: inkeejang@wustl.edu.

leads reciprocal behaviors. The power of reciprocity may also have an impact on social policy (Bowles and Gintis [2000]).

Reciprocal behavior is applicable to the problem of how to provide a public good to the society. Public good indeed presents the difficulty in a society with self-interested agents due to the classical prisoners' dilemma problem. Positive reciprocity in the context of public good indicates that individuals would like to cooperate keeping the public good in quality of others are also willing to cooperate, given that a contribution to the common resource represents a positive behavior, which incentivizes reciprocal agents to cooperate as well (Sugden [1984], Keser and Winden [2000]).

I adopt the self-control and temptation model of Gul and Pesendorfer [2001] to body out the preference of reciprocity in the model that every agent shares a common resource. To be more precise, I define a normative agent that has a normative utility function as a weighted average of every agent's outward payoff, and also has a temptation to behave as selfish agent, where selfish agent is defined as a classical self-interested agent. I extend Gul and Pesendorfer's model in a finite-stage extensive form game framework so that each agent's menu choice constitutes a history that leads to a different subgame.

I also take the notion of renegotiation-proofness (Pearce [1989], Ray [1994]) to effectively achieve a robust equilibrium. I call an equilibrium is renegotiation-proof if for each stage it is not dominated by any other subgame with given history and agents' previous menu choice. Pearce [1989] shows that renegotiation-proof equilibrium outcome lies on the Pareto optimal frontier. Ray [1994] extended the result by showing that such equilibrium outcome may exist as singleton or continuous sets on the Pareto optimal frontier. I demonstrate that under some condition, renegotiation-proof equilibrium in this paper's framework also lies on the Pareto optimal frontier.

Fehr and Schmidt [1999] show that even a minority of reciprocal agents

can induce a majority of selfish agents fully cooperate based on a punishment device. I first show that this result holds in the menu choice framework. I also prove that there exist renegotiation-proof equilibria that a considerable number of selfish agents cooperate, without imposing a worst equilibrium as a punishment. The existence of renegotiation-proof equilibrium inducing selfish agents' cooperation implies that preference of reciprocity still has a driving force even when all the agents are assumed to find the best outcome at each stage with given history, instead of some agents being ready to punish others' deviation. It is possible since existence of normative agents' temptation threats some selfish agents to keep cooperate. I also reveal in a model where agents are allowed to partially cooperate or defect that there exist symmetric renegotiation-proof equilibria where all the normative agents identically behave by holding a temptation partially and all the selfish agents also identically behave to cooperate partially.

2 Preliminaries

There is a finite number of agents in the society, each of whom utilizes a resource good to consume. Individual outputs are interpersonally comparable. There is a single common good from which every agent gets individual resource good. The common good is non-excludable and rival as usual, and the value varies depending on all the agents' behaviors.

Let $N = \{1, 2, \cdots, n\}$ be the set of agents and assume $n \geq 2$. Each agent decides to cooperate or defect toward the common good. Assume that the common good has a initial value of $w > 0$. I call 'c (cooperate)' as using the common good as much as any agent can without harming it at all, 'd (defect)' means overusing the common good to maximize personal payoff and to harm the common good. That is, if every agent plays c then the common good maintains its value as w and everyone consumes w. Moreover,

I assume that if K out of n agents play c and $(n-K)$ agents play d, then agents who chose c primarily get $w-(n-K)m_1$ and those who chose d get $w-(n-K)m_1+m_2$ where $m_2 > m_1 > 0$. In words, each agent who defects not only enjoys an additional m_2 as a personal payoff but also hurts the common good by m_1 so that the value of the common good after $n-K$ agents defect be $w-(n-K)m_1$. Every agent receives a resource good equal to the value of the common good first, and those who defect enjoy an additional value of m_2 where $m_1 < m_2 < n \cdot m_1$.

I call an assignment of types made by nature at stage 0 a type assignment. Let $T = \{\text{normative}, \text{selfish}\}^n$ be the set of type assignments. For any $t = (t_1, t_2, \cdots, t_n) \in T$, t_i is agent i's type. Let $K(t)$ the number of normative agents in $t \in T$. Let's assume that c and d, as described in the previous paragraph, are the only two alternatives agents can play, and menu is defined as a nonempty subset of $\{c, d\}$. There are two types of agents: normative or selfish. Agents' actions take place according to the following timeline. Nature assigns each agent their type at stage 0, each agent $i \in N$ at stage 0 learns t_i and $K(t)$, and then picks a menu or delay its menu choice. Each agent at stage 1 learns t, and then holds its own menu if they already chose at the previous stage or choose its menu if its menu choice was delayed, and choose an alternative in its own menu at stage 2. That is, each agent can choose its menu when information is incomplete, or when more information is revealed. Consumption occurs only in stage 2. Empty set is not allowed for a choice of menu at stage 0 and 1, and one and only one alternative must be chosen at stage 2. Let $S_1 = \{\{c\}, \{d\}, \{c, d\}\}$ be the set of menus when potential alternatives are only c or d, and denote D as 'delay.' Then each agent chooses a menu or delay option in $\{\{c\}, \{d\}, \{c, d\}, D\}$ at stage 0, and chooses a menu in $\{\{c\}, \{d\}, \{c, d\}\}$ only if its stage-0 menu choice was D.

Denote the 3-stage game with a set of menus S as $G(S)$. Let (A_1, A_2, z) for $z \in A_2 = A_1 \subseteq S$ or $z \in A_2 \subseteq S, A_1 = D$ denote an agent's 3-stage menu

and alternative choices. I also define a subgame of $G(S)$ which starts with a given type assignment disregarding nature's role also as a 3-stage game, and denote it as $G(S,t)$ with a set of menus S and a given type assignment $t \in T$.

For the fixed set of agents N and set of menus S, subgame perfect equilibrium is defined as a solution concept from $G(S)$. I also define equilibrium component as a solution concept from $G(S)$ which starts after nature assigns a $t \in T$ as given, and denote this game as $G(S,t)$. That is, $G(S,t)$ for any $t \in T$ is a subgame of $G(S)$ where only nature made the decision. Let $E = \{e_\alpha\}_{\alpha \in I_{SPE}}$ be the set of subgame perfect equilibria where I is an index set. For each equilibrium $e_\alpha = \{e_\alpha^t\}_{t \in T} \in E$, e_α^t for each $t \in T$ is an equilibrium component.

Let $P := \{(A,a) : A \subseteq Z \text{ nonempty }, a \in A\}$. We define $\widehat{K} : P^n \to \mathbb{N}_+$ that $\widehat{K}(q)$ represents the number of alternatives at stage 2 that are c. We also define $\widehat{K}_i : P^n \to \mathbb{N}_+$ for each $i \in N$ to represent the number of alternatives that all the agents but i choose which are c.

For any collective choice vector $q = (q_i)_{i \in N} \in P^n$ where $q_i = (q_i^A, q_i^a) \in P$ is each agent i's choice pair of menu and alternative, each selfish agent i's utility function is $v_i : P^n \to \mathbb{N}_+$ where

$$v_i(q) = \begin{cases} -(n - \widehat{K}(q))m_1, & \text{if } a_i = c \\ -(n - \widehat{K}(q))m_1 + m_2, & \text{if } a_i = d \end{cases}.$$

Normative agents wish to maximize the collective utilities. They have a desire to follow the social norm: maximize the weighted average of everyone's payoff where the weight $(\varphi + 1) > 1$ is taken over agents who eventually choose 'c', but they are also tempted to be selfish to maximize their own payoff. With given $\hat{K}$ (number of agents who cooperate rather than me),

Utility function for normative agent i is $U_i : P^n \to \mathbb{N}_+$

$$U_i(q) = u(q) + v_i(q) - \max_{y \in A_i} v_i((A_i, y); q_{-i})$$

where $u(q) = \dfrac{(\varphi+1)\widehat{K}(q)(-(n-\widehat{K}(q))m_1) + (n-\widehat{K}(q))(-(n-\widehat{K}(q))m_1 + m_2)}{n+\varphi\widehat{K}(q)}$

for some $\varphi > 0$. As expressed above, all the selfish agents' utility functions are symmetric, as well as all the normative agents; utility functions. Since v_i and u_i for every agent $i \in N$ actually depend on i's choice of alternative and collective choices or alternatives of others, for any collective choive q we occationally abuse $v_i(q_i^a; \sum_{j\in N\setminus\{i\}} q_j^a)$ and $u(q_i^a; \sum_{j\in N\setminus\{i\}} q_j^a)$, or simply $v_i(z;\hat{K})$ and $u(z;\hat{K})$ for some $z \in \{c,d\}$ and some real number $\hat{K} \in [0, n-1]$, instead of $v_i(q)$ and $u(q)$.

Let $i \in N$ be any normative agent for a while. It is easy to check that both $v_i(\cdot,\widehat{K})$ and $u(\cdot,\widehat{K})$ increase in $\widehat{K}$. Moreover, if we define $g(\widehat{K}) := u(c;\widehat{K}) - u(d;\widehat{K})$ and $h(\widehat{K}) := u(c;\widehat{K}) + a\cdot v_i(c;\widehat{K}) - u(d;\widehat{K}) - a\cdot v_i(d;\widehat{K})$, then both $g(\widehat{K})$ and $h(\widehat{K})$ strictly increase in $\widehat{K}$. With given $\widehat{K}$, $g(\widehat{K}) \geq 0$ represents that normative agents prefer menu $\{c\}$ to $\{c,d\}$ at stage 1, and $h(\widehat{K}) \geq 0$ represents that when normative agents already chose $\{c,d\}$ as their menu, they prefer the alternative c to d.

I also assume

$$g(0) < 0, h(0) < 0, g(n-1) > 0, h(n-1) > 0.$$

In words, every normative agent always chooses $\{c,d\}$ as his menu when he knows that everyone else will defect, and always chooses $\{c\}$ as his menu when he knows that everyone else will cooperate. Moreover, every normative agent who already chose $\{c,d\}$ as his menu will eventually choose c as his alternative when he realizes that everyone else will cooperate, and vice versa.

From the fact that $g(\cdot)$ and $h(\cdot)$ strictly increase, there are a unique thresholds K_g, K_h that satisfy $g(K_g - 1) = h(K_h - 1) = 0$. It is easy to show that $K_g < K_h$. As K_g and K_h are not necessarily integers, we define $K_1 := \lceil K_g \rceil, K_2 := \lceil K_h \rceil$. That is, if there are at least K_1 agents in the society who are willing to play c eventually then every normative agent would

like to hole $\{c\}$ as the menu and play c eventually, and if there are at least K_2 agents in the society who are willing to play c eventually then every normative agent who already holds $\{c, d\}$ would like to choose c as the alternative. Note that $K_2 \geq K_1 \geq 2$ from the fact that $K_h > K_g > 0$. I assume that the degree of temptation a is large enough so that $K_1 < K_2$, and also assume that $K_1 > 1$ so that any normative agent will not cooperate when all the other agents are supposed to defect.

3 Equilibrium

3.1 Subgame perfect equilibrim: Imperfect full cooperation equilibrium

I first show in this section that there exists an equilibrium where all the agents cooperate, provided the society has minimum number of normative agents. For this purpose, we will see that the 'worst equilibrium' is used as a tool for a punishment by normative agents whenever any selfish agent tries to deviate from cooperation.

I start by defining three basic actions before introducing strategies. *Full defection* means playing $(\{d\}, \{d\}, d)$ and *full cooperation* means playing $(\{c\}, \{c\}, c)$. *Judgment reservation* means to choose D at stage 0, play $(D, \{c\}, c)$ if every menu already chosen at stage 0 is $\{c\}$, and play $(D, \{d\}, d)$ otherwise. Judgment reservation is an action which may be plausible only for normative agents. Note that the 'threshold' in judgment reservation is $\underline{K}$, which is the minimum level of normative agents required to effectively punish any selfish agent's deviation from cooperation.

Let's first describe two basic strategies using the basic actions. The *defecting strategy* in a game $G(S_1)$ for each $i \in N$ and each type assignment $t \in T$ is to fully defect no matter what t_i is and other agents' strategies are. The *cooperating strategy* in a game $G(S_1)$ for each $i \in N$ and each type

assignment $t \in T$ is to fully cooperate no matter what t_i is and what other agents' strategies are.

Definition 1. The *worst equilibrium* is an equilibrium where everyone plays the defecting strategy.

Any selfish agent has an individual incentive to fully defect in the worst equilibrium. Note that even any normative agent who knows that everyone else will eventually defect at stage 2 has incentive to defect because of his temptation. As a result, everyone gets the same utility level as $-nm_1 + m_2$ for any type assignment in the worst equilibrium.

We can construct an equilibrium that induces selfish agents to cooperate using the worst equilibrium as a threat. It is easy to think as the following: no matter what $t \in T$ is, all the normative agents delay at stage 0 and be ready to play $(D, \{d\}, d)$ unless every selfish chooses $\{c\}$ at stage 0. There are, however, two issues that disrupt this plan work as an equilibrium. First, at least $\underline{K}$ normative agents are required to let the punishment as strong as selfish agents' incentive to defect. Second, if there are more than or equal to K_1 selfish agents, then each selfish agent has incentive to defect from full cooperation to full defect given that all the other selfish play full cooperation, since there are enough cooperation from the selfish so that normative agents play $(D, \{c\}, c)$ anyway. When $K(t) < \underline{K}$ regarding the first issue, there is no way for normative agents to strategically keep selfish agents cooperate so that everyone plays as in the worst equilibrium. However, when $\underline{K} \leq K_1$ and $K(t) \in [\underline{K}, n - K_1]$, at most $K_1 - 1$ selfish agents fully cooperate at stage 0 is a necessary condition to construct an equilibrium component, and and normative agents play $(D, \{c\}, c)$ if at least $K_1 - 1$ selfish agents hold $\{c\}$ at this equilibrium component.

One question arises in the discussion of the previous paragraph: how can we determine who will fully cooperate when $K(t) \in [\underline{K}, n-K_1]$? Note that no

agent can learn anyone's type including its own type at the beginning of the game, which means that selfish agents cannot prearrange who will fully cooperate and who will fully defect before the game starts since they don't know whether they will be selfish or not before the nature announces $t \in T$. I propose a 'numbering' process for this commitment. That is, number every agent from 1 to n, and when $t \in T$ is announced at the beginning of stage 1, it becomes a common knowledge. There are $n!$ different ways to reorder n agents. Define a permutation $\pi : N \to N$ be a mapping that assigns everyone a unique number. Denote $\pi_1, \pi_2, \cdots, \pi_{n!}$ be all distinct permutations, and Π be the set of permutations. Let's assume that nature not only randomly assigns everyone's type, but also a permutation so that every agent is 'numbered' at stage 0. Moreover, assume that all the agents learn their own type and their position among agents with same type. To specify, if $N = \{1, 2, 3, 4, 5\}, t =$ (normative, normative, selfish, normative, selfish), $\pi = (3, 1, 2, 5, 4)$, then agent 2 is the first normative, agent 3 is the first selfish, agent 1 is the second normative, agent 5 is the second selfish, and agent 4 is the third normative. After nature picks π, every agent immediately learn its own type and its own position as well as $K(t)$.

To come back to the discussion how to decide which selfish agents are supposed to fully cooperate, now we have an 'order' of group of selfish agent and normative agents respectively. Therefore, we can indicate that the 'first' $K_1 - 1$ selfish agents fully cooperate at stage 0 when $K(t) \in [\underline{K}, n - K_1]$. Let's define a *cooperative strategy* reflecting the discussion above. *Cooperative strategy* in a game $G(S_1)$ for each $i \in N$ and each type assignment $t \in T$ is as follows: i) If I am selfish and $K(t) < \underline{K}$ then fully defect, ii) if I am selfish and $K(t) \geq \max\{\underline{K}, n - K_1 + 1\}$ then fully cooperate, iii) if I am selfish and $K(t) \in [\underline{K}, n - K_1]$, then fully cooperate if I am the first $K_1 - 1$ selfish and fully defect otherwise, iv) if I am normative then reserve judgment.

Definition 2. The *imperfect full cooperation equilibrium* is an equilibrium

where everyone plays the cooperative strategy.

In imperfect full cooperation equilibrium, all the normative agents delay their decision to see how selfish agents behave, and perform a punishment by playing as in the worst equilibrium. Note that every normative agent indeed has incentive to play $(D, \{d\}, d)$ if they know that ay selfish agent deviates to fully defect so that strictly less than $K_1 - 1$ selfish agents fully cooperate and all the other normative agents play $(D, \{d\}, d)$. In an imperfect full cooperation equilibrium, as a result, every agent holds $\{c\}$ as its menu and play c when $K(t) \geq \max\{\underline{K}, n - K_1 + 1\}$, everyone defects when $K(t) < \underline{K}$, and partial cooperation occurs when $\underline{K} + K_1 \leq n$ and $K(t) \in [\underline{K}, n - K_1]$. The following proposition reflects this result.

Proposition 1. *If $\underline{K} + K_1 > n$ and $K(t) \geq \underline{K}$, then there exists an equilibrium component where every agent chooses its menu as $\{c\}$ and cooperates at stage 2 in $G(S_1, t)$.*

Proposition 1 shows that if K_1 is large enough, $\underline{K}$ can be small with the following holds: with given conditions, even $\underline{K}$ normative agents can have a driving force to make all the selfish agents fully cooperate. This result indicates that this paper is in line with Fehr and Schmidt [1999]'s result: even a small number of normative agents can induce a majority of selfish agents to cooperate using punishment device given some circumstances satisfied.

3.2 Subgame perfect equilibrium with social planner

I extend the model including a social planner who has a power to control the type information at stage 0. Social planner makes a decision right after the nature picks $t \in T$ and $\pi \in \Pi$ whether to let every agent learn $K(t)$ or not. I call such 3-stage game with a set of menus S and with a social planner $G'(S)$ and such game with given t as a type assigement $G'(S, t)$. Let's redefine the process of the game $G'(S_1)$.

Stage 0: Nature picks $t \in T$ and $\pi \in \Pi$. Social planner decides whether to reveal $K(t)$ or not. Each agent i learns its own type and position, learn $K(t)$ only when the social planner allows, and then choose their menu in $\{\{c\}, \{d\}, \{c, d\}, D\}$.
Stage 1: Agents learn t, all the agents' stage 0 menu choice, and then those who delayed the menu choice choose their menu in $\{\{c\}, \{d\}, \{c, d\}\}$.
Stage 2: Agents learn all the others' menu choice at stage 1, and then choose an alternative from their own menu.

I define *fully cooperative strategy* in $G'(S_1)$ that the social planner always conceal $K(t)$, every selfish agent fully cooperates no matter what $t \in T$ and $\pi \in \Pi$ is, and every normative agent reserves judgment. In a state that every agent plays fully cooperative strategy, every selfish plays $(\{c\}, \{c\}, c)$ and every normative plays $(D, \{c\}, c)$ at any $t \in T$ and $\pi \in \Pi$. I call

Definition 3. The *full cooperation equilibrium* is an equilibrium where everyone including the social planner plays the fully cooperative strategy.

Even though the fully cooperation equilibrium is not guaranteed to exist, it is likely to exist when n is large enough or the personal benefit from defect is not too large compared to hurting the common good.

Lemma 1. *If* $\dfrac{m_2}{m_1} < \dfrac{n+1}{2}$ *then full cooperation equilibrium exists.*

Sketch. If a selfish agent must deviate at stage 0, deviating to $(\{d\}, \{d\}, d)$ maximizes its payoff. Let such selfish agent $i \in N$. When $K(t) = K$, a selfish agent who deviates from $(\{c\}, \{c\}, c)$ to $(\{d\}, \{d\}, d)$ has additional payoff of $(m_2 - m_1) - K \cdot m_1$, where $(m_2 - m_1)$ comes from its own deviation, and $K \cdot m_1$ comes from K normative agents' defects. Moreover, given that $t \in T$ is announced and i is selfish, $\mathbb{P}(K(t) = K) = \binom{n-1}{K}$. Therefore i keeps

playing $(\{c\}, \{c\}, c)$ if

$$\sum_{i=0}^{n-1} \binom{n-1}{i} (m_2 - m_1 - i \cdot m_1) < 0.$$

That is, $\frac{m_2}{m_1} < M = \frac{n+1}{2}$. □

According to lemma 1, with given any finite m_1 and m_2, there always exists the full cooperation equilibrium whenever n is large enough. The required condition for lemma 1 is more likely to be satisfied than proposition 1 in a sense that it only requires a large economy, not even requiring anything for $\underline{K}$ nor K_1 which is related to the characteristics of agents' type. To summarize, subgame perfect equilibrium has a plausible feature in this model.

3.3 Renegotiation-proof equilibrium

We saw how threat effectively works in imperfect full cooperation equilibrium in the previous section. However, when any selfish agent deviates from cooperating strategy, and if there are still enough selfish agents fully cooperating or there are enough normative agents in the society, then there is actually no reason for normative agents not to cooperate beyond their initial plan. The notion of renegotiation-proofness is brought to this paper with this motivation. Renegotiation-proof equilibrium is defined that all the agents can 'renegotiate' at each subgame regardless of their initial commmitment to find the best solution given type assigement, permutation, and history of agents' previous menu choices.

Before we define renegotiation-proof equilibrium, we need to formally establish the information that agents observe at the beginning of each stage. I call history be the collective information that agents share at the beginning of each stage. I denote h^ℓ for $\ell = 0, 1, 2$ be a history at stage ℓ for game

$G(\cdot)$ or $G'(\cdot)$. $h^0 = K(t)$ after $t \in T$ is announced by nature in game $G(\cdot)$, and $h^0 = K(t)$ or $\emptyset$ depending on the social planner's choice. A history at stage 1 contains the type assignment, permutation, and all the agents' menu choice at stage 0. Formally, H^1 is the set of histories at stage 1 where $H^1 = \{(A_1, \cdots, A_n, t, \pi) : A_i \in \{\{c\}, \{d\}, \{c, d\}, D\}\ \forall i \in N, \text{ and } t \in T, \pi \in \Pi\}$. For any $h^1 = (A_1, \cdots, A_n, t, \pi) \in H^1$, denote $h_i = A_i$ for $i \in N$, $h^1_{n+1} = t$, and $h^1_{n+2} = \pi$. A history at stage 2 is based on the previous history at stage 1. For any $h^1 \in H^1$, $H^2(h^1)$ is the set of histories at stage 2 given that the previous history at stage 1 is h^1, formally, $H^2(h^1) = \{(A_1, \cdots, A_n, t, \pi) : A_i \in \{\{c\}, \{d\}, \{c, d\}\}\ \forall i \in N$ where $A_i = h^1_i$ if $h^1 \neq D$, and $t = h^1_{n+1}, \pi = h^1_{n+2}\}$. Let $H^2 = \bigcup_{h^1 \in H^1} H^2(h^1)$.

For each $\ell \in 1, 2$ and any $h^\ell \in H^\ell$, denote $G_\ell(\cdot, h^\ell)$ the subgame of $G(\cdot)$ which starts at stage ℓ with given history h^ℓ, denote $E_\ell(h^\ell)$ the set of subgame perect equilibria starts at stage ℓ with given history h^ℓ, and let $E_\ell = \bigcup_{h^\ell \in H^\ell} E_\ell(h^\ell)$. Note that there is no concept of equilibrium component for subgames since type assignment is already a common knowledge at the beginning of stage 1.

$F^t = (F^t_1, \cdots, F^t_n) : E \to \mathbb{R}^n$ is a mapping from an equilibrium to a vector of all the agents' utility when type assignment is t. Let $F = \{F^t\}_{t \in T}$ be the utility function for any type assignment and for every agent.

Here is some useful notation. For any $e_\alpha, e_\beta \in E$ and any $t \in T$, we write $F^t(e_\alpha) \geq F^t(e_\beta)$ if $F^t_i(e_\alpha) \geq F^t_i(e_\beta)$ for all $i \in N$, $F^t(e_\alpha) > F^t(e_\beta)$ if $F^t_i(e_\alpha) \geq F^t_i(e_\beta)$ for all $i \in N$ and $F^t(e_\alpha) \neq F^t(e_\beta)$.

I also define Pareto improvement and Pareto optimality in a classical way.

Definition 4. For any $e_\alpha, e_\beta \in E$, e_α Pareto improves e_β if

$$F(e_\alpha) > F(e_\beta).$$

e_α is Pareto optimal if there is no equilibrium in E that Pareto improves e_α.

I define Pareto improvement and Pareto optimality of $E_1(h^1)$ and $E_2(h^2)$ for some $h^1 \in H^1$ and $h^2 \in H^2(h^1)$ in line with Definition 4.

Definition 5. For each $\ell = 1, 2$, any $h^\ell \in H^\ell$, any $e_{\alpha_\ell}, e_{\beta_\ell} \in E_\ell(h^\ell)$, e_{α_ℓ} Pareto improves e_{β_ℓ} if

$$F(e_{\alpha_\ell}) > F(e_{\beta_\ell}).$$

$e_{\alpha_\ell} \in E_\ell(h^\ell)$ is Pareto optimal if there is no equilibrium in $E_\ell(h^\ell)$ that Pareto improves e_{α_ℓ}.

Pareto improvement and optimality for subgame is defined based on the same history.

Let's abbreviate the concept of Pareto improvement and Pareto optimality for equilibrium component as follows:

Definition 6. For any $e_\alpha, e_\beta \in E$ and any $t \in T$, e_α^t Pareto improves e_β^t if

$$F^t(e_\alpha) > F^t(e_\beta).$$

e_α^t is Pareto optimal if there is no equilibrium component with type t that Pareto improves e_α^t.

For any subset $E_{\alpha_\ell} \subset E_\ell$, any equilibria $e_0 \in E$ and $e_1 \in E_1$, and for each $\ell = 1, 2$, say that an equilibrium $e_{\ell-1}$ is *supported by* E_{α_ℓ} if any subgame at stage ℓ in $e_{\ell-1}$ results in an equilibrium in E_{α_ℓ}. Let $\mathcal{E}_2(E_2(h^2)) = \{e_2 \in E_2(h^2) : e_2 \text{ is Pareto optimal}\}$ for any $h^2 \in H^2$. For any $h^1 \in H^1$, let $\Phi_1(E_1(h^1))$ be the set of equilibria in $E_1(h^1)$ supported by $\bigcup_{h^2 \in H^2(h^1)} \mathcal{E}_2(E_2(h^2))$. Define $\mathcal{E}_1(E_1(h^1))$ be the set of equilibria in $\Phi_1(E_1(h^1))$ not Pareto improved by any $e_1 \in \Phi_1(E_1(h^1))$.

I define a renegotiation-proof equilibrium as a refinement of subgame perfect equilibrium, adopting the basic notion from Pearce [1989]. Renegotiation-proof equilibrium is a subgame perfect equilibrium that every subgame is Pareto optimal.

Definition 7. For any set of menus S, A subgame perfect equilibrium $e \in E$ in $G(S)$ is renegotiation-proof if it is supported by $\mathcal{E}_1(E_1(h^1))$ where $h^1 \in H^1$ is generated by e. Denote $E_{RPE} = \{e \in E : \ e \text{ is renegotiation-proof}\}$.

The following lemma demonstrates that the analysis at the final stage in any renegotiation-proof equilibrium is simple.

Lemma 2. *At stage 2 in any renegotiation-proof equilibrium of $G(S_1)$ with any given collective choice of menus and any type assignment, choice of alternatives is deterministic.*

I introduce two main patterns of renegotiation-proof equilibria. The first is *selfish-move-first (SMF) equilibrium*, and the second is *normative-move-first (SMF) equilibrium*. Selfish-move-first (SMF) strategy is defined as follows: i) if I am selfish and $\underline{K} \leq K(t) < K_1$ and I am the first $K_1 - K(t)$ selfish then play $(\{c\}, \{c\}, c)$, ii) if I am selfish not in the case of i) then play $(\{d\}, \{d\}, d)$, iii) if I am normative then play the defecting strategy.

Definition 8. A *SMF equilibrium* is a renegotiation-proof equilibrium where everyone plays the Selfish-move-first strategy.

In SMF equilibrium all the normatives delay their menu choice at stage 0 and see how selfish agents 'well behave.' Selfish agents, however, know that normative agents will collectively act to achieve Pareto optimality at each stage so that minimum number of selfish agents required for normative agents to play $(D, \{c\}, c)$ cooperate.

Normative-move-first (NMF) strategy is defined as follows:
If I am selfish, then delay at stage 0, and at stage 1 if the number of normative holding $\{c, d\}$ is at least $\underline{K}$ and I am the first $K_2 - K(t)$ selfish then play $(D, \{c\}, c)$, otherwise play $(D, \{d\}, d)$.
If I am normative, then choose $\{c, d\}$ at stage 0 if I am the first $\underline{K}$ normative and $\underline{K} \leq K(t) < K_2$, and choose $\{c\}$ otherwise. At stage 2, if i am holding

$\{c, d\}$ then play c only if the number of selfish agents who holds $\{c\}$ and normative agents is at least K_2.

Definition 9. A *NMF equilibrium* is a renegotiation-proof equilibrium where everyone plays the Normative-move-first strategy.

In NMF equilibrium all the selfishs delay their menu choice at stage 0 to see whether enough number of normative agents are really choosing $\{c, d\}$ at stage 0 so that selfish agents are supposed to be punished if they do not 'well behave.'

I define a Separating equilibrium as a combination of a SMF equilibrium and NMF equilibrium 2.

Definition 10. A *Separating equilibrium* is a renegotiation-proof equilibrium where there exists $T_1 \subset T$ such that everyone plays the SMF strategy if $t \in T_1$ and everyone plays the NMF strategy if $t \in T \backslash T_1$.

Proposition 2. *If $K_2 - K_1 < \underline{K}$ then any separating equilibrium is Pareto optimal in E_{RPE}. If $K_2 - K_1 \geq \underline{K}$ then a separating equilibrium is Pareto optimal in E_{RPE} if any $t \in T_1$ satisfies $K(t) > K_2 - \underline{K}$.*

Proposition 2 says that a separating equilibrium is Pareto optimal in E_{RPE} under a circumstance. Separating equilibrim, obviously, is not the only renegotiation-proof equilibrium concept in $G(S_1)$. However, the following proposition shows that if we care of efficiency, or Pareto optimality, then the set of separating equilibria is enough to deal with. In other words, all the equilibrium outcome achieved by separating equilibria covers all the Pareto optimal allocations of renegotiation-proof equilibria

Proposition 3. *Any Pareto optimal outcome of renegotiation-proof equilibrim can be achieved by a separating equilibrium.*

3.4 Partial cooperation or defect

I now assume that each agent can pick the degree of cooperation or defect. To be specific, letting 0 represent c and 1 represent d, $a \in (0,1)$ means partial cooperation and defect. If $a \in [0,1]$ is close to 0 it represents a cooperating behavior with a tiny amount of defect. Formally, define $S_2 = \{A \subseteq [0,1] : A \text{ is compact } \}$ a set of menus. The timeline is as follows:
Stage 0: Nature picks $t \in T$ and $\pi \in \Pi$. Each agent i learns t_1, its own position, and then chooses its menu $\in [0,1]$ or delays menu choice.
Stage 1: Agents learn all the agents' stage 0 menu choice, and then those who delayed the menu choice choose their menu $\subseteq [0,1]$.
Stage 2: Agents learn all the others' menu choice, and choose an alternative from their own menu.

Let's focus on 'symmetric equilibria': all selfish agents share the same strategy as well as all normative agents.

I start with simplifying the menu choice for normative agents.

Lemma 3. *For any compact $A \subset [0,1]$, any $z \in A$, and any K, $U((A,z);K) \leq U([0, max_{a \in A} a]; K)$.*

Lemma 4. *For any K, if $A \in S$ maximizes $\max_{z \in A} U((A,z);K)$, then $\max_{z \in A} U((A,z);K) = \max_{z \in A} U([0, max_{a \in A} a]; K)$*

By lemma 3 and 4, we can assume that normative agents choose their menu in a form of $[0,a]$ with $a \in [0,1]$.

I show that there exist symmetric equilibria in $G(S_2)$ corresponding to those in $G(S_1)$.

The symmetric defecting strategy:
If I am selfish, then play $\left(\{\frac{K_g - K(t)}{n - K(t)}\}, \{\frac{K_g - K(t)}{n - K(t)}\}, \frac{K_g - K(t)}{n - K(t)} \right)$ if $K(t) =$

$K_1 - 1$, play $(\{d\}, \{d\}, d)$ otherwise.
If I am normative, then delay at stage 0, and at stage 1 if the collective cooperation of selfish and the number of normative is at least K_1 then play $(D, \{0\}, 0)$, otherwise $(D, [0, 1], 1)$.

Definition 11. The *Symmetric worst equilibrium* is an equilibrium in $G(S_2)$ where everyone plays the symmetric defecting strategy.

Selfish-move-first (SMF) strategy:
If I am selfish and i) $m \leq K(t) < K_g$ then play $\left(\{\frac{K_g - K(t)}{n - K(t)}\}, \{\frac{K_g - K(t)}{n - K(t)}\}, \frac{K_g - K(t)}{n - K(t)}\right)$,
ii) otherwise play $(\{1\}, \{1\}, 1)$.
If I am normative then play the defecting strategy.

Definition 12. A *SMF equilibrium* is an equilibrium where everyone plays the Selfish-move-first strategy.

Normative-move-first (NMF) strategy:
If I am selfish, then delay at stage 0, and at stage 1 if the maximum of punishment normatives can collectively perform is at least m then play $\left(D, \{\frac{K_h - K(t)}{n - K(t)}\}, \frac{K_h - K(t)}{n - K(t)}\right)$, otherwise play $(D, \{1\}, 1)$.
If I am normative, then choose $\left[0, \frac{m}{K(t)}\right]$ at stage 0 if $m \leq K(t) < K_h$, and choose $\{0\}$ otherwise. At stage 2, play 0 only if the number of collective cooperation by selfish agents and normative agents is at least K_h, and play $\frac{m}{K_h}$ otherwise.

Definition 13. A *NMF equilibrium* is an equilibrium where everyone plays the Normative-move-first strategy.

References

S. Bowles and H. Gintis. The evolution of strong reciprocity. mimeo, University of Massachusetts, 2000.

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

F. Gul and W. Pesendorfer. Temptation and self-control. *Econometrica*, 69:6: 1403–1435, 2001.

C. Keser and F. Winden. Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1):23–39, 2000.

D. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1:593–622, 1998.

D. Pearce. Renegotiation-proof equilibria: Collective rationality and intertemporal cooperation (cowles foundation discussion paper no. 855). Cowles Foundation for Research in Economics, Yale University, 1989.

M. Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83:5:1281–1302, 1993.

D. Ray. Internally renegotiation-proof equilibrium sets: Limit behavior with low discounting. *Games and Economic Behavior*, 6:162177, 1994.

R. Sugden. Reciprocity: The supply of public goods through voluntary contributions. *Economic Journal*, 84:772–787, 1984.