Bad Apples in Symmetric Repeated Games

Takuo Sugaya and Alexander Wolitzky*
Stanford GSB and MIT
March 21, 2022

Abstract

We study large-population repeated games where players are symmetric but not anonymous, so player-specific rewards and punishments are feasible. Players may be commitment types who always take the same action. Even though players are not anonymous, we show that an anti-folk theorem holds when the commitment action is “population dominant,” meaning that it secures a payoff greater than the population average payoff. For example, voluntary public goods provision is impossible in large populations when commitment types never contribute, even if monetary rewards can be directed to specific players. However, provision is possible if non-contributors can be subjected to involuntary fines.

Keywords: repeated games, symmetric games, incomplete information, commitment types, large populations, population dominant action

JEL codes: C72, C73, D82

*Wolitzky acknowledges financial support from the NSF.
1 Introduction

Large-population repeated games model social cooperation in settings including community resource management (Ostrom, 1990), voluntary public goods provision (Miguel and Gugerty, 2005), informal risk-sharing (Ligon, Thomas, and Worrall, 2002), and interactions in online marketplaces (Friedman and Resnick, 2001). In reality, large groups inevitably contain a few agents who do not behave cooperatively, so it is natural to investigate when social cooperation is robust to introducing a small share of uncooperative agents. In recent work (Sugaya and Wolitzky, 2020, henceforth SW20), we showed that cooperation is impossible in large-population repeated games under two conditions:

1. The game has a “pairwise dominant” action, each player may be a commitment type who always takes this action—what we call a bad apple—and the distribution of the number of bad apples in the population is “smooth.” For example, the latter condition holds if each player is a bad apple with independent probability $z$, for any fixed $z \in (0, 1)$ as the population size $N \to \infty$. An action $a^*$ is pairwise dominant if whenever some player takes $a^*$ and another player takes a different action, the first player gets a strictly higher payoff than the second.

2. Players are symmetric and anonymous: for any action profile $(a_1, \ldots, a_N)$, any permutation $\pi$ on the set of player-names $I = \{1, \ldots, N\}$, and any player $i \in I$, we have

$$u_i(a_1, \ldots, a_N) = u_{\pi(i)}(a_{\pi^{-1}(1)}, \ldots, a_{\pi^{-1}(N)}).$$

Under these conditions, as $N \to \infty$ social welfare in every Nash equilibrium converges to that where everyone takes $a^*$, regardless of how the players’ actions are monitored. The logic is that if rational players frequently took actions other $a^*$, bad apples would get substantially higher payoffs than rational players. A rational player would therefore deviate by following the bad-apple strategy of always taking $a^*$, if this deviation were undetectable. Finally, the smoothness assumption implies that this deviation is almost undetectable when $N$ is large (even if actions are perfectly monitored).

Granting that anyone could turn out to be a bad apple with some small (independent) probability, this “anti-folk theorem” precludes cooperation in a range of environments, including the following two:

\footnote{This condition is equivalent to the stage game satisfying standard definitions of symmetry and anonymity (e.g., Plan, 2017, Theorem 1).}
Example 1: prisoner’s dilemma (PD) with anonymous random matching. Each period, players match in pairs, uniformly at random, to play a standard one-shot, two-player PD. Players do not observe their partner’s identity before choosing actions (Cooperate or Defect).

Example 2: public goods game. Each period, players decide whether to Work or Shirk, where working is privately costly but benefits everyone else.

These two examples are actually one and the same, because playing Cooperate without knowing the partner’s identity is a kind of public good provision. Note that Defect is pairwise dominant in Example 1; so is Shirk in Example 2.

While these examples are canonical, anonymity is a very restrictive assumption, because it rules out player-specific rewards and punishments. For instance, the following games (described formally later on) are symmetric but not anonymous:

Example 1’: PD with non-anonymous random matching. The same as Example 1, but players observe their partner’s identity before choosing actions.

Example 2’: public goods game with transfers. The same as Example 2, but each player also has the option of sending money to any other player, simultaneously with the Work/Shirk decisions.

(It is convenient to consider a version of this game where sending money is slightly wasteful. For concreteness, assume that for each dollar player \(i\) sends to player \(j\), player \(j\) receives only 99 cents, the remaining penny being wasted.)

These games violate condition (1) because actions have different payoff consequences for different players, but they are still symmetric.\(^2\) So, when the population is large and likely contains a few bad apples, does the folk theorem hold in these games or not?

The current paper shows that our earlier anti-folk theorem extends to all symmetric games. Unlike in anonymous games, in symmetric games a player may care about which of her opponents are bad apples, not just how many bad apples there are in the population. Nonetheless, a player’s expected payoff conditional on the event that there are \(n\) bad apples remains well-defined, and we can reproduce our earlier arguments working with these expected payoffs.

However, while our anti-folk theorem extends to symmetric, non-anonymous games, these games rarely have pairwise dominant actions. For instance, the action Defect Against Everyone is not

\(^2\)That is, their automorphism groups are player-transitive. We will explain this condition.
pairwise dominant in Example 1’, and the action *Shirk and Don’t Send Anyone Money* (or, for short, *Shirk and Stiff*) is not pairwise dominant in Example 2’. This follows because, for example, a player who takes *Shirk and Stiff* can get a lower payoff than another player who takes a more generous action, if the latter player receives enough money from third parties.

To address games like Examples 1’ and 2’, we generalize the notion of a pairwise dominant action to that of a “population dominant action.” This is an action $a^*$ such that the payoff of any player who takes $a^*$ exceeds the average payoff in the population by an amount proportional to the fraction of the population who take actions other than $a^*$. For example, *Shirk and Stiff* is population dominant in Example 2’, because the payoff of a player who takes *Shirk and Stiff* exceeds the average payoff in the population by at least .01 times the fraction of players who take actions other than *Shirk and Stiff*. In contrast, *Defect Against Everyone* is not population dominant in Example 1’, because the payoff of a player who takes *Defect Against Everyone* can be lower than the payoffs of the other players in the population if they cooperate with each other while defecting against the player taking *Defect Against Everyone*. In this case, deviating from a more cooperative strategy to *Defect Against Everyone* is unprofitable.

The main result of the current paper is that our earlier anti-folk theorem extends not only to symmetric (non-anonymous) games, but also to games with a population dominant (non-pairwise dominant) action. We also present a folk theorem, which shows that the anti-folk theorem is reasonably tight when actions are perfectly monitored. In particular, cooperation in a large population with a few bad types is possible in Example 1’ (for perfect monitoring) but not in Example 2’ (for any monitoring structure).

Roughly speaking, our results show that a key determinant of cooperation in large populations is whether a deviator can be held to a payoff that is lower than that of the rest of the population. This observation highlights the importance of coercion in supporting large-group cooperation. If player-specific incentives can be provided only through rewards (as in Example 2’), a player who acts selfishly and never rewards others will outperform the rest of the population, and the presence of a few such players destroys cooperation. If instead player-specific incentives can be provided through punishments (as in Example 1’, or a variant of Example 2’ where players can impose fines on each other), selfish players can be held below the population payoff, and cooperation can prevail.

---

3 The .01 comes from the assumed inefficiency of transfers. If transfers were perfectly efficient, our arguments would still show that no one can *Work*, but they would allow the possibility that some players might *Shirk* while transferring money back and forth.

4 This theorem is a variant of prior results on repeated games with incomplete information (e.g., Hörner and Lovo, 2009).
2 Preliminaries

2.1 Model

We consider symmetric repeated games with commitment types. These are repeated games where the stage game, the prior over players’ types (rational or committed), and the monitoring structure are all symmetric. This section introduces the model and the relevant symmetry notions. This material is relatively standard but somewhat notation-heavy; it can be skimmed on a first reading.

Stage games. An $N$-player stage game $G = (I, A, u)$ consists of a finite set of players $I = \{1, \ldots, N\}$, a finite action set $A_i$ for each player $i \in I$, and a payoff function $u_i : A \rightarrow \mathbb{R}$ for each $i \in I$ (where $A = \prod_{i \in I} A_i$ and $u(a) = (u_i(a))_{i \in I}$ for $a \in A$). Throughout the paper, we normalize the range of each $u_i$ to lie in $[0, 1]$. An automorphism on $G$ (Nash, 1951) is a bijection $\pi : I \rightarrow I$ together with a bijection $\phi_i : A_i \rightarrow A_{\pi(i)}$ for each player $i$ such that

$$u_i(a) = u_{\pi(i)}(\phi(a)) \text{ for all } i \in I, a \in A,$$

(2)

where $\phi(a) \in A$ is the action profile defined by $\phi(a)_j = \phi_{\pi^{-1}(j)}(a_{\pi^{-1}(j)})$ for all $j \in I$. This says that payoffs are invariant to simultaneously relabeling players according to $\pi$ and relabeling actions according to $\phi$. The game $G$ is symmetric if its automorphism group is player transitive: for all $i, j \in I$, there exists an automorphism $(\pi, \phi)$ on $G$ such that $\pi(i) = j$.

Let us formalize Examples 1’ and 2’, and check that they symmetric.

PD with non-anonymous random matching. For each player $i$, $A_i = \{C, D\}^I \setminus \{i\}$, with the interpretation that the $j \neq i$-coordinate of $a_i$ (which we denote as $a_{i,j}$) is $i$’s action upon meeting $j$. For $(x, y) \in \{C, D\}^2$, let $v(x, y)$ denote player 1’s payoff in the two-player PD payoff at action profile $(x, y)$. Then payoffs in the PD with non-anonymous random matching are given by $u_i(a) = \sum_{j \neq i} v(a_{i,j}, a_{j,i}) / (N - 1)$. Note that for any bijection $\pi : I \rightarrow I$, the pair $(\pi, \phi)$ is an automorphism, where $\phi$ is defined as $\phi_i \left( (a_{i,j})_{j \in I \setminus \{i\}} \right) = (a_{\pi(i), \pi(j)})_{j \in I \setminus \{i\}}$ for all $i \in I$ and $a_i \in A_i$. This implies that the game is symmetric.

Public goods game with transfers. Let $M_i$ denote the set of vectors $m_i \in \{0, \ldots, \bar{m}\}^{I \setminus \{i\}}$ whose components sum to at most $\bar{m}$. For each player $i$, $A_i = \{W, S\} \times M_i$: player $i$ chooses Work or Shirk, along with a non-negative integer amount of money to send to each opponent, up to a total

---

5This is a standard, general notion of symmetry. For much more on symmetry in $N$-player games, see, e.g., Stein (2011), Hefti (2017), Plan (2017), Ham (2021).
of \( m \) dollars. Let \( w_i(a_i) \in \{ W, S \} \) denote the first component of \( a_i \), and let \( m_{i,j}(a_i) \) denote the amount of money \( i \) sends to \( j \) under \( a_i \). Suppose that taking Work entails a private cost of \( c \), but benefits each other player by \( b/(N-1) \). Recall also our assumption that one penny out of every dollar transferred is wasted. Then payoffs are given by

\[
 u_i(a) = \sum_{j \neq i} \frac{b1 \{ w_j(a_j) = W \}}{N-1} + \nu \left( \sum_{j \neq i} (0.99m_{j,i}(a_j) - m_{i,j}(a_i)) \right) - c1 \{ w_i(a_i) = W \},
\]

where \( \nu \) is a utility function for money, which is assumed to be strictly increasing, strictly concave, and bounded above.\(^6\) Note that for any bijection \( \pi : I \rightarrow I \), the pair \((\pi, \phi)\) is an automorphism, where, for all \( i \in I \) and \( a_i \in A_i \), the action \( a_\pi(i) = \phi_i(a_i) \in A_\pi(i) \) is defined as \( w_\pi(i) \left( a_\pi(i) \right) = w_i(a_i) \) and \( m_{\pi(i),j}(a_\pi(i)) = m_{i,\pi^{-1}(j)}(a_i) \) for all \( j \neq \pi(i) \). So the game is symmetric.

Note that both of these examples are not only symmetric but also \( N \)-transitively symmetric, meaning that for any permutation \( \pi \) on \( I \), there exists a bijection \( \phi \) such that \((\pi, \phi)\) is an automorphism. An example of a game that is symmetric but not \( N \)-transitively symmetric is a game “played on a circle,” where each player cares only about her neighbors’ actions.

**Commitment types.** We consider games where each player \( i \) has a type \( \theta_i \in \{ R, B \} \), where \( R \) is the rational type and \( B \) is the bad (or “commitment”) type. For each player \( i \), there is a commitment action \( a^*_i \in A_i \) such that if \( \theta_i = B \) then player \( i \) is constrained to take \( a^*_i \). There is a common prior \( p \) on the set of players’ types \( \{R, B\}^N \). It will be convenient to adopt the accounting convention that the rational and commitment type of each player have the same payoff function, despite having different strategy sets.

We call a triple \((G, a^*, p)\) a game with commitment types. In such a game, we say that an automorphism \((\pi, \phi)\) of \( G \) is admissible if it maps each player \( i \)'s commitment action to player \( \pi(i) \)'s, so that \( \phi_i(a^*_i) = a^*_{\pi(i)} \) for all \( i \). We say that a game with commitment types \((G, a^*, p)\) is symmetric if the group of admissible automorphisms of \( G \) is player-transitive, and in addition the prior \( p \) is \( (N\text{-transitively}) \) symmetric, meaning that for any type profile \( \theta \in \{R, B\}^N \) and any permutation \( \pi : I \rightarrow I \), we have \( p(\theta_1, \ldots, \theta_N) = p(\theta_{\pi(1)}, \ldots, \theta_{\pi(N)}) \). With a symmetric prior, we denote the probability that a given player is a commitment type by \( z = \sum_{\theta_i = B} p(\theta) \).

**Monitoring structures.** Given a stage game \( G \), a monitoring structure \((Y, \chi)\) consists of a

\(^6\)We introduce a bounded utility function for money rather than just assuming quasi-linear utility because our results will require that payoff are bounded independently of \( N \), which would not be the case with quasi-linear utility.
finite set of signals of the form \( y = (y_i)_{i \in I} \) and a family of conditional probability distributions \( \chi(y | a) \), one for each action profile \( a \in A \). For example, perfect monitoring describes the case where \( Y_i = A \) for each player \( i \), and \( \chi(y | a) = 1 \{ y_i = a \ \forall \ i \in I \} \).

We will need a notion of symmetry that jointly applies to stage games (including games with commitment types) and monitoring structures. We say that an admissible automorphism for the tuple \( (G, a^*, p, Y, \chi) \) is an admissible automorphism \( (\pi, \phi) \) on \( G \) (defined above) together with a bijection \( \psi_i : Y_i \rightarrow Y_{\pi(i)} \) for each player \( i \) such that

\[
\chi(y | a) = \chi(\psi(y) | \phi(a)) \quad \text{for all } i, y \in Y, a \in A,
\]

where \( \phi(a) \) is defined above and \( \psi(y) \in Y \) is the signal defined by \( \psi(y)_j = \psi_{\pi^{-1}(j)}(y_{\pi^{-1}(j)}) \) for all \( j \in I \). The tuple \( (G, a^*, p, Y, \chi) \) is symmetric if the group of its admissible automorphisms is player transitive and the prior \( p \) is symmetric.

**Repeated games.** A repeated game with commitment types \( \Gamma = (G, a^*, p, Y, \chi, \delta) \) consists of a stage game \( G \), a profile of commitment actions \( a^* \), a prior \( p \in \Delta \left( \{ R, B \}^N \right) \), a monitoring structure \( (Y, \chi) \), and a discount factor \( \delta \in [0, 1) \). In each period \( t = 1, 2, \ldots \), the players take actions \( a_t \), the period-\( t \) signal \( y_t \) is drawn according to \( \chi(y_t | a_t) \), and each player \( i \) observes \( y_{i,t} \), the \( i \) component of \( y_t \). A history for player \( i \) at the beginning of period \( t \) takes the form \( h^t_i = (a_{i,\tau}, y_{i,\tau})_{\tau=1}^{t-1} \), with \( h^1_i = \emptyset \). A strategy \( \sigma_i \) for player \( i \) maps histories \( h^t_i \) to elements of \( \Delta(A_i) \), for each \( t \). For each player \( i \), the commitment type of player \( i \) is constrained to play \( a_i^* \) in every period—that is, to play the strategy Always \( a_i^* \)—while the rational type of player \( i \) chooses a strategy \( \sigma_i \) to maximize her expected discounted payoff. We can also let the players observe the outcome of a public randomizing device in each period, but this is not essential: our folk and anti-folk theorems both hold irrespective of the availability of public randomization.

Note that a repeated game with commitment types \( \Gamma \) is an example of a game with commitment types, where a player’s “action” is her repeated game strategy. A preliminary observation is that, when viewed as such a game, \( \Gamma \) is symmetric if the tuple \( (G, a^*, p, Y, \chi) \) is symmetric. The proof is straightforward and is deferred to the appendix.

**Lemma 1** If the tuple \( (G, a^*, p, Y, \chi) \) is symmetric, the repeated game with commitment types \( \Gamma = (G, a^*, p, Y, \chi, \delta) \) is a symmetric game with commitment types.

We call such a game \( \Gamma \) a symmetric repeated game with commitment types.
2.2 Payoff-Symmetric Equilibria

We now show that, in any symmetric game with commitment types where players observe the outcome of a public randomizing device at the beginning of the game, it is without loss to focus on equilibria where a player’s expected payoff conditional on her own type and the event that the number of bad types in the population is \(n\) is the same across players. By Lemma 1, the same conclusion applies to symmetric repeated games with commitment types, when public randomization is available. Since public randomization only expands the equilibrium payoff set, our main result (the anti-folk theorem given in the next section) applies a fortiori without public randomization.

Consider any game with commitment types \((G, a^*, p)\). Given a strategy profile \(s\) and a type profile \(\theta\), let

\[
\rho (s, \theta)_i = \begin{cases} 
  s_i & \text{if } \theta_i = R \\
  a_i^* & \text{if } \theta_i = B
\end{cases}, \quad \text{for each } i \in I.
\]

If each player \(i\) takes strategy \(s_i\) when she is rational, \(\rho (s, \theta)\) is the strategy profile in game \(G\) that is actually played at type profile \(\theta\). Let \(|\theta| = |i \in I : \theta_i = B|\), and denote player \(i\)'s expected payoff under strategy profile \(s\) conditional on the event that \(|\theta| = n\) and \(\theta_i = R\) (resp., \(\theta_i = B\)) by

\[
u^{n,R}_i (s) = \sum_{\theta:|\theta|=n, \theta_i = R} \Pr (\theta \mid \theta_i = R) u_i (\rho (s, \theta)) \quad \text{and} \quad u^{n,B}_i (s) = \sum_{\theta:|\theta|=n, \theta_i = B} \Pr (\theta \mid \theta_i = B) u_i (\rho (s, \theta)),
\]

where \(u^{n,R}_i\) is well-defined for \(n \in \{0, \ldots, N - 1\}\) and \(u^{n,B}_i\) is well-defined for \(n \in \{1, \ldots, N\}\).

Denote the corresponding population average payoffs by

\[
u^{n,R} (s) = \frac{1}{N} \sum_{i \in I} u^{n,R}_i (s), \quad \nu^{n,B} (s) = \frac{1}{N} \sum_{i \in I} u^{n,B}_i (s), \quad \text{and} \quad \nu (s) = \frac{N - n}{N} \nu^{n,R} (s) + \frac{n}{N} \nu^{n,B} (s).
\]

To ease notation, we let \((s'_i; s_{-i}) := (s_1, \ldots, s_{i-1}, s'_i, s_{i+1}, \ldots, s_N)\), the strategy profile where \(i\) takes \(s'_i\) and her opponents take \(s_{-i}\). A strategy profile \(s \in S\) is a Bayes Nash equilibrium (NE) in the game \((G, s^*, p)\) if

\[
\sum_{\theta} \Pr (\theta) u_i (\rho (s, \theta)) \geq \sum_{\theta} \Pr (\theta) u_i (\rho ((s'_i; s_{-i}) , \theta)) \quad \text{for all } i \in I, s'_i \in S_i.
\]

\footnote{In this subsection we denote a generic strategy profile by \(s\) to indicate that the results apply equally to one-shot or repeated games. We also allow the strategy sets to be infinite.}
Let $S^*$ denote the set of NE in $(G, p)$. Let $\Delta (S^*)$ denote the set of probability distributions over $S^*$. Note that any distribution in $\Delta (S^*)$ can be attained in an equilibrium with public randomization at the beginning of the game.

**Lemma 2** Let $(G, a^*, p)$ be a symmetric game with commitment types. For any $s^* \in S^*$, there exists $s \in \Delta (S^*)$ such that

$$u_i^{n-1,R}(s) = u_i^{n-1,R}(s^*) \text{ and } u_i^{n,B}(s) = u_i^{n,B}(s^*) \text{ for all } i \in I, n \in I. \quad (4)$$

The proof is somewhat lengthy and is deferred to the appendix, but the main idea is simple. Fix a NE $s$, and suppose that $u_i^{n,R}(s) < u_j^{n,R}(s)$ for some $i, j, n$.

By symmetry, there is an admissible automorphism $(\pi, \phi)$ such that $\pi (i) = j$. Since $s$ is a NE, a simple argument implies that the strategy profile $s' = \phi (s)$ is also a NE, and moreover that the vector \( \left( u_k^{n,R}(s') \right)_{k \in I} \) is a permutation of the vector \( \left( u_k^{n,R}(s) \right)_{k \in I} \). Therefore, the distribution $s'' = .5s + .5s'$ is in $\Delta (S^*)$. Furthermore, payoffs under $s''$ are the average of those under $s$ and $s'$, so since payoffs under $s$ and $s'$ are permutations of each other, payoffs under $s''$ are more equal across players than those under $s$. Thus, for any NE with unequal payoffs, we can construct a NE with more equal payoffs, which yields the conclusion of the lemma.

We call a NE that satisfies (4) payoff symmetric, and henceforth restrict attention to such equilibria. Note that Lemma 2 applies a fortiori to symmetric games without commitment types. While it is very natural that payoff symmetric equilibria are without loss in symmetric games with public randomization, we are not aware of a reference for this result.

### 3 Anti-Folk Theorem

We now present our main result: in symmetric repeated games where the commitment type actions $a^*$ are “population dominant” and the prior $p$ is “smooth,” as $N \to \infty$ social welfare in every NE converges to that where $a^*$ is always played. This generalizes the main result of SW20, which assumed the game is anonymous (i.e., (1) holds) and the commitment type actions are “pairwise dominant,” which is a stronger condition than population dominance.

We first introduce the relevant definitions. A profile of actions $a^* \in A$ is *pairwise dominant* if there exists a positive number $c > 0$ such that the payoff of any player $i$ who takes $a^*_i$ is no less

---

8 The case where $u_i^{n,B}(s) < u_j^{n,B}(s)$ for some $i, j, n$ is analogous.

9 A similar argument appears in Plan (2017).
than any other player’s payoff, and exceeds the payoff of any player \( j \) who takes \( a_j \neq a_j^* \) by at least \( c \): that is,
\[
u_i (a_i^*; a_{-i}) - u_j (a_j^*; a_{-i}) \geq c 1 \{ a_j \neq a_j^* \} \quad \text{for all } i, j \in I, a_{-i} \in A_{-i}.
\]

For instance, \textit{Defect} is pairwise dominant in the PD with anonymous random matching, and \textit{Shirk} is pairwise dominant in the public goods game, but no action is pairwise dominant in the PD with non-anonymous random matching or the public goods game with transfers (when \( m \) is large).

Next, denote the population average payoff (“social welfare”) at action profile \( a \) by
\[
U (a) = \frac{1}{N} \sum_i u_i (a).
\]

A profile of actions \( a^* \in A \) is \textit{population dominant} if there exists a positive number \( c > 0 \) such that the payoff of any player \( i \) who takes \( a_i^* \) exceeds the population average payoff by at least \( c \) times the fraction of the population whose actions differ from \( a^* \): that is,
\[
u_i (a_i^*; a_{-i}) - U (a_i^*; a_{-i}) \geq c \left| \{ j \in I : a_j \neq a_j^* \} \right| \quad \text{for all } i \in I, a_{-i} \in A_{-i}.
\]

Clearly, a pairwise dominant action is also population dominant. Note that no action is population dominant in the PD with non-anonymous random matching, but \textit{Shirk} and \textit{Stiff} is population dominant in the public goods game with transfers, with \( c \) equal to the minimum of \(.01v' (0)\) and the private cost of taking \textit{Work}.

We assume that, whenever a pairwise or population dominant action profile \( a^* \) exists, it is also the commitment action profile. The interpretation is that we are focusing on situations where the commitment types are “selfish.”

We also let \( b \geq 0 \) denote the greatest impact on total population payoffs that can result from a player switching from \( a^* \) to another action, which is given by
\[
b = \max_{a_i \in A_i, a_{-i} \in A_{-i}} N \left| U (a_i; a_{-i}) - U (a_i^*; a_{-i}) \right|.
\]

Next, following SW20, let \( B_n \) denote the event that \( \lvert \theta \rvert = n \), and let \( q_n \) denote the probability of \( B_n \) conditional on the event that a given player is rational: \( q_n = \Pr (B_n | \theta_i = R) \) for

\[\text{This follows because if } n \text{ players other than } i \text{ each transfer } $1, \text{ the average money holdings of players } -i \text{ is at most } -0.01n/(N-1), \text{ and hence, since } v \text{ is concave, the average money utility of players } -i \text{ is at most } v(-0.01n/(N-1)), \text{ which in turn is less than } v(0) - (.01n/(N-1)) v'(0). \text{ Hence, } u_i (a_i^*; a_{-i}) - U (a_i^*; a_{-i}) \geq .01v'(0) n/(N-1).\]
n \in \{0, \ldots, N - 1\}. Similarly, conditional on the event that a given player is rational, denote the probability that \( n - 1 \) out of the remaining \( N - 1 \) players are bad by \( q_n^- = q_{n-1}^- \) for \( n \in \{1, \ldots, N\} \).

By convention, let \( q_N = q_0^- = 0 \). With this convention, \( q^n = (q_n)_{n=0}^N \) and \( q^- = (q_n^-)_{n=0}^N \) are both probability distributions on \( \{0, \ldots, N\} \). Denote the total variation distance between these probability distributions by

\[
\Delta_{q, q^-} = \max_{N \in \{0, \ldots, N\}} \left| \sum_{n \in N} (q_n - q_n^-) \right|. \tag{5}
\]

As discussed in SW20, \( \Delta_{q, q^-} \) is a measure of the detectability of a deviation by the rational type of player \( i \) to the strategy \( \text{Always } a_i^* \).

We say that a sequence of repeated games indexed by \( N, (\Gamma)_N \), has a smooth distribution of bad types if \( \lim_{N \to \infty} \Delta_{q, q^-} = 0 \). For example, this condition holds if the distribution \( q \in \Delta(\{0, \ldots, N\}) \) is log-concave for all \( N \) and \( \lim_{N \to \infty} q_n = 0 \) for all \( n \). In particular, this is the case if types are independent and \( z \) is fixed independent of \( N \). See SW20 for further examples and discussion of the smoothness condition.

We are ready to state our main result. Note that the formulas in the theorem rely on our assumption that \( u_i(a) \in [0, 1] \) for all \( i \in I \) and \( a \in A \). For a fixed repeated game \( \Gamma \), this is just a normalization; but when we consider a sequence of repeated games \( (\Gamma)_N \), it entails assuming that payoffs are bounded independent of \( N \).

**Theorem 1** For any symmetric repeated game with commitment types \( \Gamma \) with a population dominant action profile \( a^* \), in any Nash equilibrium social welfare \( U \) satisfies

\[
|U - U(a^*)| \leq (1 - z) b \frac{1 + c}{c} \Delta_{q, q^-}. \tag{6}
\]

In particular, for any sequence \( (\Gamma)_N \) of such games that satisfies \( \lim \inf_{N \to \infty} c_N > 0 \) and \( \lim \sup_{N \to \infty} b_N < \infty \) and has a smooth distribution of bad types, and any corresponding sequence of Nash equilibrium social welfare levels \( (U)_N \), we have

\[
\lim_{N \to \infty} |U_N - U_N(a^*)| = 0. \tag{7}
\]

For example, Theorem 1 implies that for large \( N \), social welfare in any NE in the public goods game with transfers is close to \( \sum_i v_i(0) / N \)—the welfare level that results when everyone plays Shirk.
and Stiff—whenever commitment types play Shirk and Stiff and the distribution of commitment types is smooth. We emphasize that this conclusion holds even though this game is not anonymous and does not have a pairwise dominant action.\textsuperscript{11}

The proof of Theorem 1 follows the proof in SW20, with two new ideas. First, a key point in SW20 is that if the rational type of player $i$ deviates to Always $a_i^*$, then her expected payoff conditional on $B_n$ is equal to the expected payoff of a bad type conditional on $B_{n+1}$. That is, for any payoff-symmetric strategy profile $\sigma$, we have

$$\sum_{\theta:|\theta|=n, \theta_i=R} \Pr(\theta|B_n, \theta_i=R) \mathbb{E}[u_i(\text{Always } a_i^*; \sigma_{-i})|\theta] = u^{n+1,B} \text{ for all } i, n \in \{0, \ldots, N-1\}.$$ \hspace{1cm} (8)

The first step in proving Theorem 1 is showing that this equation remains valid in symmetric games. To see this, note that for all $i \in I$, $n \in \{0, \ldots, N-1\}$, and $\theta_{-i}$ such that $|\theta_{-i}|=n$, we have

$$\Pr(\theta_{-i}|B_n, \theta_i=R) = \frac{\Pr(\theta_{-i}|\theta_i=R)}{\sum_{\tilde{\theta}_{-i}:|\tilde{\theta}_{-i}|=n} \Pr(\tilde{\theta}_{-i}|\theta_i=R)} = \frac{1}{\binom{N-1}{n}} \frac{\Pr(\theta_{-i}|\theta_i=B)}{\sum_{\tilde{\theta}_{-i}:|\tilde{\theta}_{-i}|=n} \Pr(\tilde{\theta}_{-i}|\theta_i=B)} = \Pr(\theta_{-i}|B_{n+1}, \theta_i=B),$$ \hspace{1cm} (9)

where the middle equalities hold because the prior is symmetric. This in turn implies that

$$\sum_{\theta:|\theta|=n, \theta_i=R} \Pr(\theta|B_n, \theta_i=R) \mathbb{E}[u_i(\text{Always } a_i^*; \sigma_{-i})|\theta] = \sum_{\theta_{-i}:|\theta_{-i}|=n, \theta_i=R} \Pr(\theta_{-i}|\theta_i=R) \mathbb{E}[u_i(\text{Always } a_i^*; \sigma_{-i})|\theta] = \sum_{\theta_{-i}:|\theta_{-i}|=n, \theta_i=B} \Pr(\theta_{-i}|\theta_i=B) \mathbb{E}[u_i(\text{Always } a_i^*; \sigma_{-i})|\theta] = u^{n+1,B},$$

which yields (8).\textsuperscript{12}

Equation (8) lets us generalize the key lemma of SW20 as follows.

**Lemma 3** For any symmetric game with commitment types and any payoff symmetric NE,
\[ \sum_{n=0}^{N-1} q_n u^{n,R} \geq \sum_{n=0}^{N-1} q_n u^{n,B} - \Delta_{q,q^-}, \text{ with the convention that } u^{0,B} = 1. \]

**Proof.** The equilibrium payoff of the rational type of player \( i \) equals \( \sum_{n=0}^{N-1} q_n u^{n,R} \). If instead the rational type of player \( i \) deviates to *Always* \( a_i^* \), her expected payoff equals

\[
\sum_{\theta} \Pr(\theta|\theta_i = R) \mathbb{E}[u_i (\text{Always } a_i^*; \sigma_{-i}) | \theta] = \sum_{n=0}^{N-1} q_n \sum_{\theta:|\theta| = n, \theta_i = R} \Pr(\theta|\theta_n, \theta_i = R) \mathbb{E}[u_i (\text{Always } a_i^*; \sigma_{-i}) | \theta] = \sum_{n=0}^{N-1} q_n u^{n+1,B},
\]

where the first equation is by definition of \( q_n \) and the second is by (8). Hence, in any NE, we must have \( \sum_{n=0}^{N-1} q_n u^{n,R} \geq \sum_{n=0}^{N-1} q_n u^{n+1,B} \). However, by the same argument as in SW20, we have

\[ \sum_{n=0}^{N-1} q_n u^{n+1,B} \geq \sum_{n=0}^{N-1} q_n u^{n,B} - \Delta_{q,q^-}. \]

Therefore, \( \sum_{n=0}^{N-1} q_n u^{n,R} \geq \sum_{n=0}^{N-1} q_n u^{n,B} - \Delta_{q,q^-} \). \( \square \)

Now we can prove Theorem 1. Here, the novelty relative to SW20 involves comparing bad types’ payoffs to the average payoff among players in the population who do not take their population dominant actions, and showing that this comparison implies that the population dominant actions must almost always be taken.

**Proof of Theorem 1.** We restrict attention to payoff symmetric equilibria \( \sigma \), which is without loss by Lemma 2. We first show that, whenever \( |\theta| \in \{1, \ldots, N - 1\} \), in every period the average payoff among bad types exceeds the average payoffs among rational types by at least \( c \) times the fraction of rational types who take actions other than \( a^* \). To see this, for any type profile \( \theta \) with \( |\theta| = n \in \{1, \ldots, N - 1\} \) and any action profile \( a \) with \( a_i = a_i^* \) for all \( i \) with \( \theta_i = B \), let \( m(a) = |\{i \in I : a_i \neq a_i^*\}| \), the number of players who take actions other than \( a^* \). Denote the average payoffs among bad types, rational types, and all players by

\[ u^B = \frac{1}{n} \sum_{i: \theta_i = B} u_i(a), \quad u^R = \frac{1}{N - n} \sum_{i: \theta_i = R} u_i(a), \quad \text{and} \quad U = \frac{n}{N} u^B + \frac{N - n}{N} u^R. \]

Since \( a^* \) is population dominant, we have

\[ u^B \geq U + \frac{m}{N} c = \frac{n}{N} u^B + \frac{N - n}{N} u^R + \frac{m}{N} c, \text{ or equivalently } u^B \geq u^R + \frac{m}{N - n} c. \quad (10) \]

Now denote player \( i \)'s expected payoff in period \( t \) conditional on type profile \( \theta \) by \( u_{i,t}(\theta) = \mathbb{E}[u_i(a_t) | \theta] \), and denote her overall expected payoff conditional on \( \theta \) by \( u_i(\theta) = (1 - \delta) \sum_t \delta^{t-1} u_{i,t}(\theta) \).

Since (10) holds for every \( \theta \) and \( a \) that arise with positive probability conditional on \( B_n \), we have,
for all $t$ and $\theta$,
\[
\frac{1}{n} \sum_{i: \theta_i = B} u_{i,t}(\theta) \geq \frac{1}{N-n} \sum_{i: \theta_i = R} u_{i,t}(\theta) + \frac{\sum_{i \in I} \Pr(a_{i,t} \neq a_{i}^{*}\mid \theta)}{N-n} c.
\]

Taking a discounted sum over periods, and taking the expectation over $\theta : |\theta| = n$, we have
\[
\frac{1}{n} \mathbb{E} \left[ \sum_{i: \theta_i = B} u_i(\theta) \mid \mathcal{B}_n \right] \geq \frac{1}{N-n} \mathbb{E} \left[ \sum_{i: \theta_i = R} u_i(\theta) \mid \mathcal{B}_n \right] + \frac{(1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} \sum_{i \in I} \Pr(a_{i,t} \neq a_{i}^{*}\mid \mathcal{B}_n)}{N-n} c.
\]

Next, note that
\[
\frac{1}{n} \mathbb{E} \left[ \sum_{i: \theta_i = B} u_i(\theta) \mid \mathcal{B}_n \right] = \frac{1}{n} \mathbb{E} \left[ \sum_{i \in I} \mathbb{I}\{\theta_i = B\} u_i(\rho(\sigma, \theta)) \mid \mathcal{B}_n \right] = \frac{1}{n} \sum_{i \in I} \Pr(\theta_i = B\mid \mathcal{B}_n) \mathbb{E}[u_i(\rho(\sigma, \theta)) \mid \mathcal{B}_n, \theta_i = B] = \frac{1}{n} \sum_{i \in I} u_i^B(\sigma) = \frac{1}{N} \sum_{i \in I} u_i^{n,B}(\sigma) = u^B(\sigma),
\]
and similarly $\frac{1}{N-n} \mathbb{E} \left[ \sum_{i: \theta_i = R} u_i(\theta) \mid \mathcal{B}_n \right] = u^{n,R}(\sigma)$. So we have
\[
u^{n,B} \geq u^{n,R} + \gamma_n c, \tag{12}
\]
where
\[
\gamma_n = (1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} \sum_{i \in I} \Pr(a_{i,t} \neq a_{i}^{*}\mid \mathcal{B}_n) = (1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} \frac{1}{N} \sum_{i \in I} \Pr(a_{i,t} \neq a_{i}^{*}\mid \mathcal{B}_n, \theta_i = R).
\]

With Lemma 3 and inequality (12) in hand, the rest of the proof follows SW20; we include the remaining steps for completeness. Recalling that $u^{0,B} = 1$ by convention and $u^{0,R} \in [0,1]$ by assumption, we obtain
\[
\Delta_{q, q'} \geq \sum_{n=0}^{N-1} q_n (u^{n,B} - u^{n,R}) \geq \sum_{n=1}^{N-1} q_n (u^{n,B} - u^{n,R}) \geq \sum_{n=1}^{N-1} q_n \gamma_n c.
\]
Define $\gamma = \sum_{n=0}^{N-1} q_n \gamma_n$. Since $q_0 = q_0 - q_0^\top \leq \Delta_{q,q}$, we have

$$\gamma = q_0 \gamma^0 + \sum_{n=1}^{N} q_n \gamma_n \leq \Delta_{q,q} + \frac{1}{c} \Delta_{q,q} = \frac{1 + c}{c} \Delta_{q,q}.$$  

Finally, the discounted sum of the expected fraction of players who take actions other than $a^*$ equals

$$(1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} \frac{1}{N} \sum_{i \in I} \sum_{n=0}^{N-1} \Pr (B_n \land \theta_i = R) \Pr (a_{i,t} \neq a^*_i | B_n, \theta_i = R)$$

$$= (1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} \frac{1}{N} \sum_{i \in I} \sum_{n=0}^{N-1} (1 - z) q_n \Pr (a_{i,t} \neq a^*_i | B_n, \theta_i = R) = (1 - z) \sum_{n=0}^{N-1} q_n \gamma_n = (1 - z) \gamma.$$  

Therefore, expected social welfare differs from $U(a^*)$ by at most $(1 - z) b \gamma \leq (1 - z) b \frac{1 + c}{c} \Delta_{q,q}$. This yields (6), and taking $\Delta_{q,q} \to 0$ yields (7).  

It is straightforward to extend Theorem 1 to games with multiple populations, where the players in each population are symmetric. For example, consider a variant of the public goods game with transfers where there are two populations, agents and principals. In every period, each agent chooses Work or Shirk, and each principal chooses an amount of money to send to each agent. (These choices can be simultaneous or sequential.) Suppose that each agent is committed to Shirk with independent probability $z_A > 0$, and each principal is committed to Stiff (i.e., send no money) with independent probability $z_P > 0$. Then the above arguments can be modified to show that, as $N \to \infty$, all principals almost always Stiff; and, given this, all agents almost always Shirk. In contrast, if $z_P = 0$, so the principals are known to be rational (or, alternatively, if there is a single principal with sufficiently deep pockets), then there is an equilibrium where rational agents always Work, and the principal(s) send money to each agent if and only if she works. This example illustrates the importance of having a deep-pocketed principal (or a group of known-rational principals) for incentivizing effort by a large group of agents.

4 Folk Theorem

We now present a folk theorem for repeated games with incomplete information, which serves as a partial converse to Theorem 1. This shows that our notion of population dominance cannot be greatly generalized.
For each type profile \( \theta \in \{R, B\}^N \), let \( \Gamma (\theta) \) denote the complete information repeated game where it is common knowledge that the players’ types are described by \( \theta \). Let \( R (\theta) = \{ i : \theta_i = R \} \) and \( B (\theta) = \{ i : \theta_i = B \} \). Let \( A (\theta) = \prod_{i \in R (\theta)} A_i \times \prod_{i \in B (\theta)} \{ a_i^* \} \), let \( \Delta (A (\theta)) \) denote the set of probability distributions on \( A (\theta) \), and let \( \Delta^* (A (\theta)) \) denote the set of independent mixtures on \( A (\theta) \), given by \( \prod_{i \in R (\theta)} \Delta (A_i) \times \prod_{i \in B (\theta)} 1 \{ a_i = a_i^* \} \). Denote player \( i \)’s minmax payoff in the game \( \Gamma (\theta) \) by \( v_i^\theta = \min_{\alpha \in \Delta^* (A_{-i}(\theta_{-i}))} \max_{a_i \in A_i} u_i (a_i; \alpha_{-i}) \). Denote the set of feasible payoffs in \( \Gamma (\theta) \) by

\[
F (\theta) = \left\{ v \in [0, 1]^N : \exists \alpha \in \Delta^* (A (\theta)) \; \text{s.t.} \; u (\alpha) = v (\theta) \right\},
\]

and denote the set of feasible and individually rational payoffs in \( \Gamma (\theta) \) by

\[
F^* (\theta) = \left\{ v \in F (\theta) : v_i \geq v_i^\theta \; \forall i \; \text{s.t.} \; \theta_i = R \right\}.
\]

Define the set \( F^{**} (\theta) \) to be equal to \( F^* (\theta) \) on the set of rational-player payoff vectors \([0, 1]^{\mid i: \theta_i = R \mid}\) has non-empty relative interior, and to equal the convex hull of the set of static NE payoffs in \( G (\theta) \) otherwise. Next, say that a family of payoff vectors \( (v (\theta))_{\theta \in \{R, B\}^N} \), with \( v (\theta) \in F^{**} (\theta) \) for each \( \theta \), is feasible, individually rational, and incentive compatible (or FIRIC) if

\[
\mathbb{E} [v_i (\theta) \mid \theta_i = R] > \mathbb{E} \left[ \max \left\{ v_i ((\theta_i = B; \theta_{-i})) , \min_{v \in F^{**} (\theta)} v_i \right\} \mid \theta_i = R \right] \text{ for all } i \in I \quad (13)
\]

Finally, say that an expected payoff vector \( v \in [0, 1]^N \) is consistent with FIRIC if there exists a family of FIRIC payoff vectors \( (v (\theta))_{\theta \in \{R, B\}^N} \) such that \( v = \mathbb{E} [v (\theta)] \).

**Theorem 2** Fix any repeated game with commitment types and perfect monitoring, \( \Gamma \). For any payoff vector \( v \in [0, 1]^N \) consistent with FIRIC and any \( \varepsilon > 0 \), there exists \( \delta < 1 \) such that, for every \( \delta > \delta \), there exists a sequential equilibrium in \( \Gamma \) with an expected payoff vector \( v' \) satisfying \( |v_i - v'_i| \leq \varepsilon \) for all \( i \in I \).

For example, in the PD with non-anonymous random matching, let \( v (\theta) \) be the payoff vector that results when pairs of rational players cooperate with each other, while everyone defects against commitment types. It is easy to see that the family \( (v (\theta))_{\theta \in \{R, B\}^N} \) is FIRIC. Hence, Theorem 2 implies that the corresponding ex ante payoff vector can be approximated in sequential equilibrium when the players are sufficiently patient. (In this example, the payoff vector can actually be exactly attained.)
We only sketch the proof, which is a variation of existing arguments (e.g., Fudenberg and Maskin, 1986; Koren, 1992; Hörner and Lovo, 2009). Fix a family of FIRIC payoff vectors \((v(\theta))_{\theta \in \{R,B\}^N}\) such that \(v = \mathbb{E}[v(\theta)]\). For any history \(h^t\), let \(\hat{\theta}(h^t)\) denote the set of players that have “revealed rationality” at history \(h^t\) by previously taking some action \(a_i \neq a_i^*\), and let \(i(h^t)\) denote the identity of the most recent player (if any) to have deviated from equilibrium play at history \(h^t\). All rational players are supposed to reveal rationality in the first period of the game. Subsequently, on the equilibrium path the players take a sequence of actions that achieve the payoff vector \(v(\hat{\theta}(h^t))\), and that have the further property that continuation payoffs under the action sequence are always close to \(v(\hat{\theta}(h^t))\).\(^{13}\) Off the equilibrium path, the players take a sequence of actions that achieve a payoff vector \(v \in \arg\min_{v' \in F^{**}(\hat{\theta}(h^t))} v'_i(h^t)\). Since rational players are supposed to reveal rationality immediately, at any history \(h^t\) all players who have revealed rationality believe that the continuation game is the complete information game \(\Gamma(\hat{\theta}(h^t))\), so since \(v(\hat{\theta}(h^t)) \in F^{**}(\hat{\theta}(h^t))\), the payoff vector \(v(\hat{\theta}(h^t))\) is attainable in a continuation equilibrium as in Fudenberg and Maskin (1986). Since the family of payoff vectors \((v(\theta))_{\theta \in \{R,B\}^N}\) is incentive compatible, and continuation payoffs conditional on each set of revealed-rational players \(\hat{\theta}(h^t)\) are approximately constant, it is optimal for a rational player to reveal rationality in the first period (rather than never revealing rationality, or waiting to reveal rationality until a later period). In particular, if player \(i\) does not reveal rationality in period 1, then conditional on each opposing type profile \(\theta_{-i}\), her continuation payoff cannot exceed the maximum of \(v_i((\theta_i = B; \theta_{-i}))\) (her continuation payoff if she never reveals rationality) and \(\min_{v \in F^{**}(\theta)} v_i\) (her continuation payoff subsequent to revealing rationality after period 1) by more than an arbitrarily small amount. Finally, at off-path histories a rational player who has not yet revealed rationality (contrary to equilibrium play) may or may not prefer to do so, but her play at these histories is irrelevant for the other players’ incentives, so she can be prescribed an arbitrary best response. Overall, the equilibrium construction is similar to that in Hörner and Lovo (2009) or Hörner, Lovo, and Tomala (2011), but it is simpler because we do not require that the equilibrium is “belief-free.”

We now show that if the commitment action profile \(a^*\) does not satisfy a slightly generalized version of population dominance, then there is a family of feasible payoff vectors \((v(\theta))_{\theta \in \{R,B\}^N}\) that satisfies a version of incentive compatibility. Together with Theorem 2, this implies that Theorem 1 cannot be extended much further.

We say that \(a^*\) satisfies \textit{generalized population dominance} if there exists a positive number \(c > 0\)

\(^{13}\)Sorin (1986) and Fudenberg and Maskin (1991) showed that such a sequence exists.
such that
\[
\sum_{n=0}^{N-1} \frac{N}{N-n} q_n c_n \geq 0,
\]
where
\[
c_n = \min_{\alpha \in \Delta(A)} \left( u^{n,B}(\alpha) - u^n(\alpha) - c \frac{\mathbb{E}[\{i \in I : a_i \neq a^*_i\}]}{N} \right) \text{ for all } n \in \{0, \ldots, N-1\}.
\]

We note that population dominance can be replaced with generalized population dominance in Theorem 1.

Now suppose that \( a^* \) does not satisfy generalized population dominance for any \( c > 0 \). Then, for \( \hat{c}_n = \min_{\alpha \in \Delta(A)} u^{n,B}(\alpha) - u^n(\alpha) \), we have \( \sum_n \frac{N}{N-n} q_n \hat{c}_n \leq 0 \). We claim that if this inequality holds with \( q;\gamma \) slack, so that \( \sum_n \frac{N}{N-n} q_n \hat{c}_n + \Delta q;\gamma \leq 0 \), then there exists a family of feasible payoffs \( \{v(\theta)\}_{\theta \in \{R,B\}^N} \) such that
\[
\mathbb{E}[v_i(\theta) | \theta_i = R] > \mathbb{E}[v_i((\theta_i = B; \theta_{-i}) | \theta_i = R) \text{ for all } i \in I.
\]

This inequality is the same as (13), except that we are neglecting individual rationality.

To see why this is true, note that, by symmetry, (14) is equivalent to \( \sum_{n=0}^{N-1} q_n u^{n,R} > \sum_{n=0}^{N-1} q_n u^{n+1,B} \).

When players take \( \hat{s}^n \in \arg \min_s u^{n,B}(s) - u^n(s) \) for each realized number of bad types \( n \), we have
\[
u^{n,B} = u^n + \hat{c}_n = \frac{n}{N} u^{n,B} + \frac{N-n}{N} u^{n,R} + \hat{c}_n, \text{ and hence } u^{n,B} \leq u^{n,R} + \frac{N}{N-n} \hat{c}_n.
\]

Taking the expectation over \( n \), we obtain
\[
\sum_{n=0}^{N-1} q_n u^{n+1,B} = \sum_{n=0}^{N-1} q_n u^{n,B} + \sum_{n=0}^{N} (q^- - q_n) u^{n,B} \leq \sum_{n=0}^{N-1} q_n u^{n,R} + \sum_{n=0}^{N-1} \frac{N}{N-n} q_n \hat{c}_n + \Delta q;\gamma < \sum_{n=0}^{N-1} q_n u^{n,R}.
\]

5 Conclusion

To induce members of a large population to take cooperative actions rather than behaving like “bad apples,” players must be identifiable (i.e., non-anonymous) and must be subject to player-specific rewards and punishments. This paper has investigated what forms these rewards and punishments must take. The main result is that it is necessary (and almost sufficient) that the available incentive
instruments can reduce a bad apple’s payoff to that of the rest of the population. Voluntary monetary transfers among players do not satisfy this condition, but “coercive” instruments—like fines or non-cooperation directed at deviators—do.

We have presented our results in a simple model with one rational type, one commitment type, and perfect monitoring. Extensions to multiple rational or commitment types are straightforward; see SW20 for a discussion of these extensions in the anonymous case. Imperfect monitoring raises interesting issues, some of which we have pursued in other work. In large-population repeated games with imperfect public monitoring, the prospects for cooperation depend on the interaction among the discount factor, the population size, and the information content of the monitoring structure (Sugaya and Wolitzky, 2021a). As for private monitoring, in the PD with non-anonymous random matching where players only observe their partner’s actions, cooperation is possible only if players are sufficiently patient relative to the population size, or if the game is augmented with cheap talk (Sugaya and Wolitzky, 2021b). The interaction between incomplete information and private monitoring more generally is a fairly open area.
References


6 Appendix: Proofs of Lemmas 1 and 2

6.1 Proof of Lemma 1

Fix distinct $i, j \in I$. Since $(G, a^*, p, Y, \chi)$ is symmetric, there exists an admissible automorphism $(\pi, \phi, \psi)$ on $G$ such that $\pi(i) = j$. We construct an admissible automorphism $((\hat{\pi}, \hat{\phi})$ on $G$ such that $\hat{\pi}(i) = j$, where this admissibility means that $\hat{\phi}_k (\text{Always } a^*_k) = \text{Always } a^*_k$ for all $k \in I$. First, for each player $i$ and period $t$, let $H^t_i$ denote the set of player $i$’s period $t$ histories, and define a bijection $\eta^i_t : H^t_i \rightarrow H^t_{\pi(i)}$ by

$$\eta^i_t((a_{i, \tau}, y_{i, \tau})_{\tau=1}^{t-1}) = (\phi_i(a_{i, \tau}), \psi_i(y_{i, \tau}))_{\tau=1}^{t-1} \text{ for all } h^t_i \in H^t_i.$$  

Next, let $\hat{\pi} = \pi$ and define $\hat{\phi}$ as follows: for each player $i$ and strategy $\sigma_i$, define $\hat{\phi}_i(\sigma_i)$ to be the strategy $\hat{\sigma}_{\pi(i)}$ that satisfies

$$\hat{\sigma}_{\pi(i)}(h^t_{\pi(i)})[a_{\pi(i)}] = \sigma_i\left((\eta^i_t)^{-1}(h^t_{\pi(i)})\right)\left[\phi^1_i(a_{\pi(i)})\right] \text{ for all } h^t_{\pi(i)} \in H^t_{\pi(i)}, a_{\pi(i)} \in \pi_{\pi(i)}. \quad (15)$$

Since $\eta^i_t$ and $\phi_i$ are bijections, $\hat{\phi}_i$ is also a bijection. Also, since $(\pi, \phi)$ is admissible, $\phi_i(a^*_i) = a^*_{\pi(i)}$, and hence $\hat{\phi}_i(\text{Always } a^*_i) = \text{Always } a^*_{\pi(i)}$.

It remains to show that $u_i(\sigma) = u_{\pi(i)}(\hat{\sigma})$. For each $h^t = ((a_{i, \tau}, y_{i, \tau})_{\tau=1}^{t-1})_{i \in I}$, define $\eta^t(h^t)$ by

$$\eta^t(h^t)_j = \eta^i_{t-1}(j)(h^t_{\pi^{-1}(j)}) \text{ for all } j \in I. \text{ Then for all } h^t \text{ and } (a_t, y_t) \in A_t \times Y_t, \text{ we have}$$

$$\Pr^\sigma(a_t, y_t|h^t) = \prod_i \sigma_i(h^t_i)[a_{i, t}] \chi(y_t|a_t)$$

$$= \prod_i \hat{\sigma}_{\pi(i)}(\eta^i_t(h^t_i))\left[\phi_i(a_{i, t})\right] \chi(y_t|a_t)$$

$$= \prod_i \hat{\sigma}_{\pi(i)}(\eta^i_t(h^t_i))\left[\phi_i(a_{i, t})\right] \chi(\psi(y)\mid\phi(a_t))$$

$$= \prod_i \hat{\sigma}_i(\eta^i_{t-1}(h^t_{\pi^{-1}(i)})\left[\phi_{\pi^{-1}(i)}(a_{\pi^{-1}(i), t})\right] \chi(\psi(y)\mid\phi(a_t))$$

$$= \Pr^{\hat{\sigma}}(\phi(a_t), \psi(y_t)\mid\eta^t(h^t))$$

where the second line follows from (15), the third line follows from (3), the fourth line changes the index from $i$ to $\pi^{-1}(i)$, and the fifth line is by definition of $\eta^i_t$. Given this, by induction on $t$, for each $t$ and $a_t \in A$, we have

$$\Pr^\sigma(a_t) = \sum_{y_t, h^t} \Pr^\sigma(a_t, y_t|h^t) \Pr^\sigma(h^t) = \sum_{y_t, h^t} \Pr^{\hat{\sigma}}(\phi(a_t), \psi(y_t)\mid\eta^t(h^t)) \Pr^{\hat{\sigma}}(\eta^t(h^t)) = \Pr^{\hat{\sigma}}(\phi(a_t)).$$

Since $(\pi, \phi)$ is an automorphism on $G$, we have

$$u_i(\sigma) = (1 - \delta) \sum_t \delta^{t-1} \sum_{a_t} \Pr^\sigma(a_t) u_i(a_t) = (1 - \delta) \sum_t \delta^{t-1} \sum_{a_t} \Pr^\sigma(a_t) u_{\pi(i)}(\phi(a_t))$$

$$= (1 - \delta) \sum_t \delta^{t-1} \sum_{a_t} \Pr^{\hat{\sigma}}(\phi(a_t)) u_{\pi(i)}(\phi(a_t)) = u_{\pi(i)}(\hat{\sigma}),$$

21
as desired.

### 6.2 Proof of Lemma 2

We first note a preliminary fact used later in the proof: if \((\pi, \phi)\) is an admissible automorphism on \(G\), then

\[
u_i(\rho(s, \theta)) = u_{\pi(i)}((\phi \circ \rho)(s, \theta)) = u_{\pi(i)}(\rho(\phi(s), \pi(\theta))).
\]

(16)

Here, the first equality holds because \((\pi, \phi)\) is an automorphism, and the second holds because, since \((\pi, \phi)\) is admissible, for each \(i \in I\) we have

\[
\phi_{\pi^{-1}(i)}(\rho(s, \theta)s) = \begin{cases} 
\phi_{\pi^{-1}(i)}(s_{\pi^{-1}(i)}) & \text{if } \theta_{\pi^{-1}(i)} = R \\
\phi_{\pi^{-1}(i)}(s_{\pi^{-1}(i)}) & \text{if } \theta_{\pi^{-1}(i)} = B 
\end{cases} = \rho(\phi(s), \pi(\theta)).
\]

Now fix any \(s^* \in S^*\). To simplify notation, for each \(n \in I\), let \(u^{n-1}{_R} = u^{n-1}{_R}(s^*)\) and \(u^{n}{_B} = u^{n}{_B}(s^*)\). Define

\[S^{**} = \{s \in S^* : u^{n-1}{_R}(s) = u^{n-1}{_R} \text{ and } u^{n}{_B}(s) = u^{n}{_B} \forall n \in I\} \quad \text{and} \quad U = \left\{v \in \mathbb{R}^{2N^2} : \exists s \in S^{**} \text{ s.t. } u^{n-1}{_R}_i(s) = v_{(n-1)N+i} \text{ and } u^{n}{_B}_i(s) = v_{N^2+(n-1)N+i} \forall i \in I, n \in I \right\}.
\]

Thus, \(v \in U\) iff there is an equilibrium \(s\) such that \(v\) is the vector of conditional expected utilities under \(s\) for each player, where the vector \(v\) first lists, for each \(n \in \{0, \ldots, N-1\}\), each player’s expected payoff conditional on being rational when there are \(n\) bad players; and then lists, for each \(n \in \{1, \ldots, N\}\), each player’s expected payoff conditional on being bad when there are \(n\) bad players. Note that the set \(U\) is compact by standard arguments.

Given \(v \in \mathbb{R}^{2N^2}\), for each \(n \in I\), define the \(N\)-dimensional vectors

\[u^{n-1}{_R}_i = (v_{(n-1)N+i})^N_{i=1} \quad \text{and} \quad u^{n}{_B}_i = (v_{N^2+(n-1)N+i})^N_{i=1}.
\]

Note that \(v\) is given by the concatenation of the vectors \(u^{n-1}{_R}\) for \(n \in I\), followed by the concatenation of the vectors \(u^{n}{_B}\) for \(n \in I\). Now define a new vector \(f(v) \in \mathbb{R}^{2N^2}\) by letting \((f(v))_{(n-1)N+i}\) equal the \(i^{th}\)-lowest component of the vector \(u^{n-1}{_R}\), for each \(i \in I\) and \(n \in I\); and letting \((f(v))_{N^2+(n-1)N+i}\) equal the \(i^{th}\)-lowest component of the vector \(u^{n}{_B}\), for each \(i \in I\) and \(n \in I\). That is, for each \(n \in I\), the \((n-1)N + 1^{st}\) through \((n-1)(N+1)^{st}\) components of the vector \(f(v)\) are equal to the increasing rearrangement of the vector \(v^{n-1}{_R}\), and the \(N^2+(n-1)N+1^{st}\) through \(N^2+(n-1)(N+1)^{st}\) components of the vector \(f(v)\) are equal to the increasing rearrangement of the vector \(v^{n}{_B}\). Let

\[F = \left\{w \in \mathbb{R}^{2N^2} : \exists v \in U \text{ s.t. } f(v) = w \right\}.
\]

Note that \(F\) is compact, because \(U\) is compact and \(f\) is continuous. Note also that

\[(1/N) \sum_{i=1}^{N} w_{(n-1)N+i} = u^{n-1}{_R} \quad \text{and} \quad (1/N) \sum_{i=1}^{N} w_{N^2+(n-1)N+i} = u^{n}{_B} \quad \text{for all } w \in F \text{ and } n \in I.
\]

Let \(\succ\) denote the lexicographic order on \(\mathbb{R}^{2N^2}\). Let \(\bar{w}\) denote a maximal element of \(F\) in the lexicographic order.\(^{15}\) Define the vector \(\bar{w} \in \mathbb{R}^{2N^2}\) by letting \(\bar{w}_{(n-1)N+i} = u^{n-1}{_R}\) and

\[\ldots
\]

\[^{15}\text{Note that the lexicographic order admits a maximum on a compact subset of } \mathbb{R}^{2N^2}.
\]
\[ \bar{w}_{N^2 + (n-1)N + i} = u^{n,B} \] for all \( i \in I \) and \( n \in I \). Note that \( \bar{w} \succcurlyeq \hat{w} \); otherwise, there would exist \( n \in I \) such that \( \bar{w}_{(n-1)N + i} \geq u^{n-1,R} \) for all \( i \in \{1, \ldots, N\} \), with strict inequality for some \( i \), which implies that \( (1/N) \sum_{i=1}^{N} \bar{w}_{(n-1)N + i} > u^{n-1,R} \) (or, symmetrically, \( n \in I \) such that \( \bar{w}_{N^2 + (n-1)N + i} \geq u^{n,B} \) for all \( i \in \{1, \ldots, N\} \) with strict inequality for some \( i \), implying that \( (1/N) \sum_{i=1}^{N} \bar{w}_{N^2 + (n-1)N + i} > u^{n,B} \)), a contradiction.

We now argue that \( \bar{w} \succcurlyeq \hat{w} \). Suppose toward a contradiction that \( \bar{w} \succ \hat{w} \). Let \( m \in \{1, \ldots, 2N^2\} \) denote the smallest index such that \( \bar{w}_m > \hat{w}_m \). Suppose that \( m \leq N^2 \), so there exist \( n \in I \) and \( i \in I \) satisfying \( m = (n-1)N + i \). (The \( m > N^2 \) case is symmetric and omitted.) Let \( v \) denote an element of \( U \) such that \( f(v) = \hat{w} \). Since \( \bar{w}_{(n-1)N + i} > \bar{w}_{(n-1)N + i} \), \( (1/N) \sum_{i=1}^{N} \bar{w}_{(n-1)N + i} = u^{n-1,R} \), and the vector \( (\bar{w}_{(n-1)N + i})_i \) is a rearrangement of the vector \( v^{n-1,R} \), not all components of \( v^{n-1,R} \) are equal. Let \( i, j \in I \) satisfy
\[
i \in \arg\min_{k \in I} v^{n-1,R}_k \quad \text{and} \quad j \in \arg\max_{k \in I} v^{n-1,R}_k,
\]
and note that \( v^{n-1,R}_i < v^{n-1,R}_j \). Moreover, note that for all \( n' < n \) and \( i \in I \), we have \( v^{n'-1,R}_i = w^{n'-1,R} \) by minimality of \( m \).

Let \( s \in S^* \) satisfy \( u^{n-1,R}(s) = v^{n-1,R} \) for all \( n \in I \). Since \((G, \rho)\) is symmetric, there exists an admissible automorphism \((\pi, \phi)\) such that \( \pi(i) = j \) and \( u_k(\rho(\bar{s}, \theta)) = u_{\pi(k)}(\rho(\phi(\bar{s}), \pi(\theta))) \) for each \( k \in I \), \( \theta \in \Theta \), and \( \bar{s} \in S \). Let \( s' = \phi(s) \). We claim that \( s' \) is a NE. To see this, fix a player \( k \in I \) and a strategy \( \hat{s}_k \in S_k \). Let \( k' = \pi^{-1}(k) \), and let \( \hat{s}_{k'} = \phi^{-1}(\hat{s}_k) \). For every strategy profile \( \bar{s} \), we have
\[
\sum_{\theta} \Pr(\theta) u_{k'}(\rho(\phi(\bar{s}), \pi(\theta))) = \sum_{\theta} \Pr(\theta) u_k(\rho(\phi(\bar{s}), \pi(\theta))) = \sum_{\theta} \Pr(\pi(\theta)) u_k(\rho(\phi(\bar{s}), \pi(\theta))) = \sum_{\theta} \Pr(\theta) u_k(\rho(\phi(\bar{s}), \theta)) \tag{17}
\]
where the first line follows because \((\pi, \phi)\) is an admissible automorphism (so (16) holds), the second line follows because \( \pi \) is symmetric, and the third line follows by rearranging the sum. Hence, we have
\[
\sum_{\theta} \Pr(\theta) u_k(\rho(s', \theta)) = \sum_{\theta} \Pr(\theta) u_k(\rho(\phi(s), \theta)) = \sum_{\theta} \Pr(\theta) u_{k'}(\rho(s, \theta)) \geq \sum_{\theta} \Pr(\theta) u_{k'}(\rho(\hat{s}_{k'}(s_{-k'}), \theta)) = \sum_{\theta} \Pr(\theta) u_k(\rho(\hat{s}_{k'}(s_{-k'}), \theta)) = \sum_{\theta} \Pr(\theta) u_k(\rho((\hat{s}_k; s_{-k}), \theta)) = \sum_{\theta} \Pr(\theta) u_k(\rho((\hat{s}_k; s'_{-k}), \theta)) \tag{17}
\]
where the first and last equalities follow because \( s' = \phi(s) \), the second and third equalities follow by (17) and \( \hat{s}_{k'} = \phi^{-1}(\hat{s}_k) \), and the inequality follows because \( s \) is a NE. As this inequality holds for any \( k \in I \) and \( \hat{s}_k \in S_k \), we see that \( s' \) is a NE.

Next, for each \( k \in I \) and \( n' \in I \), let \( v'_{n'-1,N + k} = u^{n'-1,R}_k (s') \) and \( v'_{N^2 + (n'-1)N + k} = u^{n',B}_k (s') \). Since \( s' \in S^* \), the resulting vector \( v' \) lies in \( U \). Moreover, by (16) and symmetry of \( \pi \), for each
Since \( n' \in I \) we have
\[
\begin{align*}
u_k^{n'-1,R}(s) &= \sum_{\theta:|\theta|=n'-1, \theta_k=R} \frac{\Pr(\theta)}{\Pr(|\theta|=n'-1, \theta_k=R)} u_k(\rho(s, \theta)) \\
&= \sum_{\theta:|\theta|=n'-1, \theta_k=R} \frac{\Pr(\theta)}{\Pr(|\theta|=n'-1, \theta_k=R)} u_{\pi(k)}(\rho(\phi(s), \pi(\theta))) \\
&= \sum_{\theta:|\theta|=n'-1, \theta_{\pi(k)}=R} \frac{\Pr(\theta)}{\Pr(|\theta|=n'-1, \theta_{\pi(k)}=R)} u_{\pi(k)}(\rho(\phi(s), \pi(\theta))) \\
&= u_{\pi(k)}^{n'-1,R}(s').
\end{align*}
\]
Similarly, \( u_k^{n',B}(s) = u_{\pi(k)}^{n',B}(s') \) for each \( k \in I \) and \( n' \in I \). Therefore, for each \( k \in \{0, \ldots, 2N^2 - N\} \), the vector \( (v_{kN+i})_{i=1}^N \) is a permutation of \( (v_{kN+i})_{i=1}^N \), which in particular implies that \( s' \in S^* \).

Now define the distribution \( \bar{s} \) to be a 50:50 mixture over \( s \) and \( s' \). Clearly, \( \bar{s} \in \Delta(S^*) \). Let
\[
\bar{v} = u(\bar{s}) = \frac{1}{2} (v + v'),
\]
and note that \( f(\bar{v}) \in F \). Since, for all \( n' < n \) and \( k \in I \), we have \( v_k^{n'-1,R} = u^{n'-1,R} \), it follows that
\[
\bar{v}^{n'-1,R} = u^{n'-1,R} \quad \text{for all } k \in I, \ n' < n. \tag{18}
\]
In addition, since \( \pi(i) = j, i \in \arg\min_{k \in I} v_k^{n-1}, \) and \( j \in \arg\max_{k \in I} v_k^{n-1}, \) we have
\[
\bar{v}_{\pi(k)}^{n-1,R} = \frac{1}{2} \left( v_{\pi(k)}^{n-1,R} + v_{j}^{n-1,R} \right) > v_{\pi(k)}^{n-1,R}. \tag{19}
\]
Moreover, since \( (v')^{n-1,R} \) is a permutation of \( v^{n-1,R} \), and \( \bar{v} = \frac{1}{2} (v + v') \), we also have
\[
\bar{v}_k^{n-1,R} \geq v_i^{n-1,R} \quad \text{for all } k \in I, \quad \text{and} \tag{20}
\bar{v}_k^{n-1,R} > v_i^{n-1,R} \quad \text{for all } k \in I \setminus \arg\min_{k' \in I} v_{k'}^{n-1,R}. \tag{21}
\]
Since \( i \in \arg\min_{k \in I} v_k^{n-1,R} \), (18), (19), (20), and (21) together imply that \( f(\bar{v}) \succ f(v) \). But this contradicts the maximality of \( \hat{w} = f(v) \) in \( F \). We can thus conclude that \( \hat{w} \succeq \bar{w} \).

Since \( \hat{w} \succeq \bar{w} \) and \( \hat{w} \succeq \bar{w} \), we conclude that \( \hat{w} = \bar{w} \), proving the lemma.