Bad Apples in Symmetric Repeated Games

Takuo Sugaya and Alexander Wolitzky*

Stanford GSB and MIT

September 26, 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 in large populations is impossible when commitment types never contribute, even if monetary rewards can be targeted to contributors; however, provision is possible if non-contributors can be subjected to involuntary fines. A folk theorem under incomplete information provides a partial converse to our result. Along the way, we develop some general results on symmetric games with incomplete information and/or repeated play.

Keywords: repeated games, symmetric games, incomplete information, commitment types, large populations, population dominant action, free-rider problem

JEL codes: C72, C73, D82

*We thank Marina Halac and three anonymous referees for helpful comments. 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. The current paper shows that such robustness requires that it is possible to punish defectors without also punishing the rest of the population as severely: that is, that it is possible to hold defectors to payoffs below the population average. Intuitively, given that a large society is very likely to contain some defectors, if punishing defectors is too costly then rational players always prefer to pool with defectors, so in equilibrium everyone defects. If instead punishing defectors is cheaper, rational players can be induced to separate from defectors, and cooperation can prevail among rational players.

This paper builds closely on our earlier work (Sugaya and Wolitzky, 2020, henceforth SW20). There, 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." An action a_0 is pairwise dominant if whenever some player takes a_0 and another player takes a different action, the first player gets a strictly higher payoff than the second. The smoothness condition holds if, for example, each player is a bad apple with independent probability z, for any fixed $z \in (0,1)$ as the population size $N \to \infty$.

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

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

[1] This condition is equivalent to the stage game satisfying standard definitions of symmetry and anonymity

Under these conditions, as $N \to \infty$ social welfare in every Nash equilibrium converges to that where everyone takes a_0, regardless of how the players' actions are monitored. The logic is that if rational players frequently took actions other than a_0, 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_0, 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:

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.

These examples notwithstanding, 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.

(e.g., Plan, 2017, Theorem 1).

(It is convenient to consider a version of this game with a small transaction cost. For concreteness, assume that for each dollar player i sends to player j, player j receives only 99 cents, the remaining penny being wasted.)

Example 3: helping with externalities. The same as Example 1', but taking *Cooperate* generates a positive externality for all other players, in addition to benefiting one's partner.

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 pairwise dominant in Example 1' or 3, 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', 2', and 3, we generalize the notion of a pairwise dominant action to that of a "population dominant action." This is an action a_0 such that the payoff of any player who takes a_0 exceeds the average payoff in the population by an amount proportional to the fraction of the population who take actions other than a_0. 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*.[3] In contrast, *Defect*

[2] That is, their automorphism groups are player-transitive. We will explain this condition.

[3] The .01 comes from the assumed transaction cost. Without transaction costs, our arguments would

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. Finally, we will see that *Defect Against Everyone* is population dominant in Example 3 if and only if the externality is sufficiently large. Intuitively, a larger externality makes it more difficult to hold a free-rider's payoff below the population average.

In general, the existence of a population dominant action is tied to the impossibility of targeting punishment toward a specific player. If a population dominant action exists, a player who takes this action always obtains a payoff greater than the population average. It is therefore impossible to punish a player who takes such an action without also punishing the rest of the population just as severely. If instead no action is population dominant, then a player can be punished for taking any action without punishing the rest of the population as much. This distinction turns out to be crucial for supporting cooperative outcomes in large-population games with commitment types.

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. This result implies that the existence of a population dominant action is a major obstacle to cooperation in large populations. To see the intuition, suppose that the committed players in the population take population dominant actions, while the rational players may take different actions. By the definition of population dominance, on average the committed players obtain higher payoffs than the rational players. Moreover, if the distribution of the number of committed players is smooth, then if one rational player deviates to always taking her population dominant action, this has only a small effect on the population distribution of actions. Hence, the payoff of a rational player who deviates to always taking her population dominant action is close to the equilibrium payoff of a truly committed player. Since this deviation must be unprofitable in equilibrium, and since committed players obtain higher equilibrium payoffs than rational players, it follows that the

still show that no one can *Work*, but they would allow the possibility that some players might *Shirk* while transferring money back and forth.

equilibrium payoffs of committed players and rational players must be very similar. Finally, again by the definition of population dominance, this implies that rational players must also almost always take population dominant actions in equilibrium.

We also present a folk theorem for repeated games with incomplete information and perfect monitoring. When applied to symmetric games, this result implies that our anti-folk theorem is reasonably tight, and hence that the notion of population dominance cannot be greatly generalized. Together, our results imply that, for example, cooperation in large populations with commitment types is possible in Example 1' and in Example 3 with small externalities (for perfect monitoring) but not in Example 2' or in Example 3 with large externalities (for any monitoring structure).

A step in the proof of our anti-folk theorem is that, in symmetric games with public randomization, the average payoff across players from any equilibrium can be attained in an equilibrium where all players obtain the same payoff. This result is very natural but it appears to be novel, and it may be useful beyond our particular problem.

This paper connects to several strands of literature. First, a literature following Green (1980) and Sabourian (1990) studies large-population repeated games with complete information, focusing on the difficulty of monitoring a large number of players through a coarse "aggregate signal."[4] Second, the "reputation" literature studies how introducing a small amount of incomplete information can yield sharp anti-folk theorems in repeated games with patient players (Fudenberg and Levine, 1989; Mailath and Samuelson, 2006). Third, in incomplete information settings, several papers develop measures of the pivotality or influence of a player's type on an aggregate outcome, and give conditions under which most players' influence must be small in large populations (al-Najjar and Smorodinsky, 2000; McLean and Postlewaite, 2002). See SW20 for a more extensive discussion of related literature.

[4]For some recent results and further references on such models, see Sugaya and Wolitzky (2022).

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 product set of actions $A = \times_{i \in I} A_i$, and a payoff function $u_i : A \to \mathbb{R}$ for each $i \in I$. 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 $\pi : I \to I$ together with a bijection $\phi_i : A_i \to A_{\pi(i)}$ for each player i such that

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

where $\phi(a) \in A$ is the action profile defined by $\phi(a)_j = \phi_{\pi^{-1}(j)}\left(a_{\pi^{-1}(j)}\right)$ for all $j \in I$. This says that payoffs are invariant to simultaneously relabeling players according to π and relabeling actions according to ϕ. The game G is *symmetric* if its automorphism group is player-transitive: for all $i, j \in I$, there exists an automorphism (π, ϕ) on G such that $\pi(i) = j$.[5]

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. That is, an action is a mapping from the partner's identity to *Cooperate* or *Defect*. For $(x, y) \in \{C, D\}^2$, let $v(x, y)$ denote player 1's payoff in the two-player PD at action profile (x, y). 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 $\pi : I \to I$, the pair (π, ϕ) is

[5] This 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).

an automorphism, where ϕ is defined as $\phi_i\left(\left(a_{i,j}\right)_{j\in I\setminus\{i\}}\right) = \left(a_{\pi(i),\pi(j)}\right)_{j\in I\setminus\{i\}}$ 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 of $\bar{m}$ 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 ζ, but benefits each other player by $\beta/(N-1)$, where $\zeta,\beta > 0$. Recall also our assumption that one penny out of every dollar transferred is wasted. Then payoffs are given by

$$u_i(a) = \sum_{j\neq i}\frac{\beta\mathbf{1}\{w_j(a_j)=W\}}{N-1} + v\left(\sum_{j\neq i}\left(0.99m_{j,i}(a_j) - m_{i,j}(a_i)\right)\right) - \zeta\mathbf{1}\{w_i(a_i)=W\},$$

where v is a utility function for money, which is assumed to be strictly increasing, strictly concave, and bounded above.[6] Note that for any bijection $\pi : I \to I$, the pair (π,ϕ) 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 $m_{\pi(i),j}\left(a_{\pi(i)}\right) = m_{i,\pi^{-1}(j)}(a_i)$ 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 π on I, there exists a bijection ϕ such that (π,ϕ) 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^*. Let $a^* = (a_i^*)_{i\in I}$. 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

[6] We introduce a bounded utility function for money rather than just assuming quasi-linear utility because some of our results will require that payoff are bounded independently of N, which would not be the case with quasi-linear utility.

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 (π, ϕ) 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-transitively) symmetric, meaning that for any type profile $\theta \in \{R, B\}^N$ and any permutation $\pi : I \to I$, we have $p(\theta_1, \ldots, \theta_N) = p\left(\theta_{\pi(1)}, \ldots, \theta_{\pi(N)}\right)$. With a symmetric prior, we denote the probability that a given player is a commitment type by $z = \sum_{\theta:\theta_i=B} p(\theta)$.

Monitoring structures. Given a stage game G, a *monitoring structure* (Y, χ) consists of a finite product set of signals $Y = \times_{i \in I} Y_i$ 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 $\chi(y|a) = \mathbf{1}\{y_i = a \ \forall i \in I\}$.

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, χ) is an admissible automorphism (π, ϕ) on (G, a^*, p) (defined above) together with a bijection $\psi_i : Y_i \to Y_{\pi(i)}$ for each player i such that

$$\chi(y|a) = \chi(\psi(y)|\phi(a)) \qquad \text{for all } i \in I, y \in Y, a \in A,$$

where $\phi(a)$ is defined above and $\psi(y) \in Y$ is the signal defined by $\psi(y)_j = \psi_{\pi^{-1}(j)}\left(y_{\pi^{-1}(j)}\right)$ for all $j \in I$. The tuple (G, a^*, p, Y, χ) 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, χ), 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 $h_i^t = (a_{i,\tau}, y_{i,\tau})_{\tau=1}^{t-1}$, with $h_i^1 = \emptyset$. A *strategy* σ_i for player i maps histories

h_i^t 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 σ_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 the normal form of a repeated game with commitment types Γ is itself a (static) game with commitment types, where a player's "action" is her repeated game strategy. In particular, player i's commitment "action" is the repeated game strategy *Always* a_i^*. A preliminary observation is that, when viewed in this way, Γ is symmetric if the tuple (G, a^*, p, Y, χ) is symmetric. The proof is straightforward and is deferred to the appendix.

Lemma 1 *If the tuple (G, a^*, p, Y, χ) is symmetric, then the normal form of the repeated game with commitment types $\Gamma = (G, a^*, p, Y, \chi, \delta)$ is a symmetric game with commitment types (with an infinite strategy set).*

We call such a game Γ a *symmetric repeated game with commitment types.*

2.2 Payoff-Symmetric Equilibria

We now show that, in any normal form 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). In this section only, we denote strategy profiles in this game by s, to emphasize the case where (G, a^*, p) is the normal form of a repeated game. We also allow the strategy set S to be infinite. Given a strategy profile

s and a type profile θ, 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 θ. 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

$$\begin{aligned} u_i^{n,R}(s) &= \sum_{\theta:|\theta|=n,\theta_i=R} \frac{\Pr(\theta)}{\Pr(|\theta| = n, \theta_i = R)} u_i(\rho(s,\theta)) \quad \text{and} \\ u_i^{n,B}(s) &= \sum_{\theta:|\theta|=n,\theta_i=B} \frac{\Pr(\theta)}{\Pr(|\theta| = n, \theta_i = B)} u_i(\rho(s,\theta)), \end{aligned}$$

where $u_i^{n,R}$ is well-defined for $n \in \{0,\ldots,N-1\}$ and $u_i^{n,B}$ is well-defined for $n \in \{1,\ldots,N\} = I$. Denote the corresponding population average payoffs by

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

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

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

Let S^* denote the set of NE in (G,p). Let $\Delta(S^*)$ denote the set of (Borel) 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. Linearly extending payoff functions to distributions over strategy profiles as usual, we call a distribution $\bar{s} \in \Delta(S^*)$ *payoff*

symmetric if

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

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

$$u^{n-1,R}(\bar{s}) = u^{n-1,R}(s^*) \quad \textit{and} \quad u^{n,B}(\bar{s}) = u^{n,B}(s^*) \qquad \textit{for all } n \in I.$$

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.[7] By symmetry, there is an admissible automorphism (π, ϕ) 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 $\left(u_k^{n,R}(s)\right)_{k \in I}$.[8] 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.

As an aside, note that Lemma 2 also applies 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 that the game is anonymous (i.e., (1) holds) and the commitment type actions are "pairwise dominant," which is a stronger condition than population dominance.

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

[8] A similar argument appears in Plan (2017).

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 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,

$$u_i\left(a_i^*; a_{-i}\right) - u_j\left(a_i^*; a_{-i}\right) \geq c\mathbf{1}\left\{a_j \neq a_j^*\right\} \quad \text{for all } i, j \in I, a_{-i} \in A_{-i}.$$

For instance, *Defect* is pairwise dominant in the PD with anonymous random matching, and *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 $\bar{m}$ is large).

Next, denote the population average payoff ("social welfare") at action profile a by

$$U\left(a\right) = \frac{1}{N}\sum_i u_i\left(a\right).$$

A profile of actions $a^* \in A$ is *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,

$$u_i\left(a_i^*; a_{-i}\right) - U\left(a_i^*; a_{-i}\right) \geq c\frac{\left|\left\{j \in I : a_j \neq a_j^*\right\}\right|}{N} \quad \text{for all } i \in I, a_{-i} \in A_{-i}.$$

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 *Shirk and Stiff* is population dominant in the public goods game with transfers, with c equal to the minimum of $.01v'\left(0\right)$ and the private cost of taking *Work*.[9]

Pairwise and population dominance are non-nested with the usual notion of dominance (i.e., $u_i\left(a_i^*; a_{-i}\right) \geq u_i\left(a_i; a_{-i}\right)$ for all $a_i \in A_i, a_{-i} \in A_{-i}$). For example, in the PD with non-anonymous random matching, *Defect Against Everyone* is dominant but not pairwise or

[9] This follows because if n players other than i each transfer \$1, the average money holdings of players $-i$ is at most $-.01n/\left(N-1\right)$, and hence, since v is concave, the average money utility of players $-i$ is at most $v\left(-.01n/\left(N-1\right)\right)$, which in turn is less than $v\left(0\right) - \left(.01n/\left(N-1\right)\right)v'\left(0\right)$. Hence, $u_i\left(a_i^*; a_{-i}\right) - U\left(a_i^*; a_{-i}\right) \geq .01v'\left(0\right)n/\left(N-1\right)$.

population dominant. One can also give examples where a pairwise or population dominant action is not dominant; SW20 gives such an example for pairwise dominance.

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."

It is interesting to consider the above definitions in the context of Example 3.

Helping with externalities. Consider the PD with non-anonymous random matching, where if a player cooperates she incurs a cost of $C > 0$, her partner incurs a benefit of $B > C$, and the other $n-2$ players in society each incur an externality of X. Player i's payoff is thus

$$-C\mathbf{1}\{i \text{ cooperates}\}+B\mathbf{1}\{i\text{'s partner cooperates}\}+X\,(\text{number of other players who cooperate}).$$

Observe that if $X = 0$ then this game reduces to the usual PD with non-anonymous matching (Example 1'), while if $X = B$ then, since the partner's identity becomes irrelevant, it reduces to the PD with anonymous matching (Example 1). We have thus already seen that *Defect Against Everyone* is population dominant if $X = B$, but not if $X = 0$. In general, it is easy to see that *Defect Against Everyone* is population dominant iff $X > (B-C)/2$.[10] This follows because, whenever a player takes *Cooperate* rather than *Defect*, the effect on social welfare is

$$\frac{1}{N}\left(B - C + (N-2)\,X\right),$$

while the effect on the utility of any third player is X, and

$$X > \frac{1}{N}\left(B - C + (N-2)\,X\right) \iff X > \frac{B-C}{2}.$$

Intuitively, when $X > (B-C)/2$, the free-rider problem in this game is relatively severe. Our results will imply that, if each player is committed to *Defect Against Everyone* with a small independent probability, then all players almost always defect in every equilibrium when $X > (B-C)/2$ and N is large (for any monitoring structure, uniformly in δ), but

[10] In contrast, the assumption that $B > C$ implies that shirking is not pairwise dominant.

there is an equilibrium where all rational players always cooperate when $X < (B-C)/2$ and δ is sufficiently high (for perfect monitoring, uniformly in N).

Returning to the general model, we let $b \geq 0$ denote the greatest impact on total population payoffs that can result from a player switching from a^* to another action. This is given by

$$b = \max_{a_i \in A_i, a_{-i} \in A_{-i}} N \left| U\left(a_i; a_{-i}\right) - U\left(a_i^*; a_{-i}\right) \right|.$$

Some of our result will require that, as $N \to \infty$, c is bounded away from 0 and b is bounded away from ∞. These conditions ensure that the advantage of a pairwise dominant action does not vanish, and that a significant fraction of players must take non-pairwise dominant actions in order to generate a level of social welfare significantly different from $U(a^*)$.

Next, following SW20, let $\mathcal{B}_n$ denote the event that $|\theta| = n$, and let q_n denote the probability of $\mathcal{B}_n$ conditional on the event that a given player is rational: $q_n = \Pr\left(\mathcal{B}_n | \theta_i = R\right)$ for $n \in \{0, \ldots, N-1\}$. 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, $q = (q_n)_{n=0}^N$ 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

$$\Delta_{q,q^-} = \max_{\mathcal{N} \subseteq \{0,\ldots,N\}} \left| \sum_{n \in \mathcal{N}} \left(q_n - q_n^-\right) \right|.$$

As discussed in SW20, Δ_{q,q^-} is a measure of the detectability of a deviation by the rational type of player i to the strategy *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 $q \in \Delta\left(\{0, \ldots, N\}\right)$ 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 the commitment probability 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 Γ, this is

just a normalization; but when we consider a sequence of repeated games $(\Gamma)_N$, it requires that payoffs are bounded independent of N.[11]

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

$$|U - U(a^*)| \leq (1-z)\, b \frac{1+c}{c} \Delta_{q,q^-}. \tag{2}$$

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

$$\lim_{N\to\infty} |U_N - U_N(a^*)| = 0. \tag{3}$$

Theorem 1 extends the main result in SW20 by generalizing anonymity to symmetry, and pairwise dominance to population dominance. 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.[12]

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 $\mathcal{B}_n$ is equal to the expected payoff of a bad type conditional on $\mathcal{B}_{n+1}$.

[11] However, Theorem 1 goes through if payoffs are bounded by a function $\bar{u}(N)$, and the smoothness condition is strengthened to $\lim_{N\to\infty} \bar{u}(N)\, \Delta_{q,q^-} = 0$. (In this case, the right-hand side of (2) must be multiplied by $\bar{u}(N)$.) For example, this condition holds if $\bar{u}(N)$ is linear in N and types are independent with a fixed commitment probability z, as in this case Δ_{q,q^-} converges to 0 exponentially fast in N. Note that, since payoffs are bounded by a linear function of N in Example 3, the conclusion of Theorem 1 applies in this example.

[12] Applied to the public goods game with transfers, Theorem 1 is reminiscent of the impossibility theorem of Mailath and Postlewaite (1990). However, their theorem concerns a static game with two levels of public good provision and independent types. See SW20 for a more detailed comparison with Mailath and Postlewaite.

That is, for any payoff-symmetric strategy profile σ, we have

$$\sum_{\theta:|\theta|=n,\theta_i=R} \Pr\left(\theta|\mathcal{B}_n,\theta_i=R\right)\mathbb{E}\left[u_i\left(Always\ a_i^*;\sigma_{-i}\right)|\theta\right]=u^{n+1,B} \quad \text{for all } i\in I, n\in\{0,\ldots,N-1\}. \tag{4}$$

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 θ_{-i} such that $|\theta_{-i}|=n$, we have

$$\begin{aligned}\Pr\left(\theta_{-i}|\mathcal{B}_n,\theta_i=R\right) &= \frac{\Pr\left(\theta_{-i}|\theta_i=R\right)}{\sum_{\tilde{\theta}_{-i}:|\tilde{\theta}_{-i}|=n}\Pr\left(\tilde{\theta}_{-i}|\theta_i=R\right)} \\ &= \frac{1}{\binom{N-1}{n}}=\frac{\Pr\left(\theta_{-i}|\theta_i=B\right)}{\sum_{\tilde{\theta}_{-i}:|\tilde{\theta}_{-i}|=n}\Pr\left(\tilde{\theta}_{-i}|\theta_i=B\right)}=\Pr\left(\theta_{-i}|\mathcal{B}_{n+1},\theta_i=B\right)\end{aligned} \tag{5}$$

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

$$\begin{aligned}&\sum_{\theta:|\theta|=n,\theta_i=R}\Pr\left(\theta|\mathcal{B}_n,\theta_i=R\right)\mathbb{E}\left[u_i\left(Always\ a_i^*;\sigma_{-i}\right)|\theta\right] \\ =&\sum_{\theta_{-i}:|\theta_{-i}|=n}\Pr\left(\theta_{-i}\left||\theta_{-i}|=n,\theta_i=R\right.\right)\mathbb{E}\left[u_i\left(Always\ a_i^*;\sigma_{-i}\right)|\theta\right] \\ =&\sum_{\theta_{-i}:|\theta_{-i}|=n}\Pr\left(\theta_{-i}\left||\theta_{-i}|=n,\theta_i=B\right.\right)\mathbb{E}\left[u_i\left(Always\ a_i^*;\sigma_{-i}\right)|\theta\right]=u^{n+1,B},\end{aligned}$$

which yields (4).[13]

Equation (4) lets us generalize the key lemma of SW20 as follows.

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

Proof. The equilibrium payoff of the rational type of player i equals $\sum_{n=0}^{N-1}q_nu^{n,R}$. If instead

[13]Note that (5) requires our assumption that the prior is N-transitively symmetric. If we imposed only a weaker form of symmetry that allowed certain players' types to be especially strongly correlated, (5) would be violated and the conclusion of Theorem 1 would typically fail, because society could detect a deviation by player i to *Always* a_i^* by checking specific other players' types.

the rational type of player i deviates to *Always* a_i^*, her expected payoff equals

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

where the first equation is by definition of q_n, and the second is by (4). Hence, in any payoff-symmetric 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 (cf. equation (2) in that paper), 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, $\sum_{n=0}^{N-1} q_n u^{n,R} \geq \sum_{n=0}^{N-1} q_n u^{n,B} - \Delta_{q,q^-}$. ■

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 σ, 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 θ 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

$$u^B \geq U + \frac{m}{N} c = \frac{n}{N} u^B + \frac{N-n}{N} u^R + \frac{m}{N} c, \quad \text{or equivalently } u^B \geq u^R + \frac{m}{N-n} c. \tag{6}$$

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

$(1-\delta)\sum_t \delta^{t-1} u_{i,t}(\theta)$. Since (6) holds for every θ and a that arise with positive probability conditional on $\mathcal{B}_n$, we have, for all t and θ,

$$\frac{1}{n}\sum_{i:\theta_i=B} u_{i,t}(\theta) \geq \frac{1}{N-n}\sum_{i:\theta_i=R} u_{i,t}(\theta) + \frac{\sum_{i\in I}\Pr\left(a_{i,t}\neq a_i^*|\theta\right)}{N-n}c.$$

Taking a discounted sum over periods, and taking the expectation over $\theta : |\theta| = n$, we have

$$\frac{1}{n}\mathbb{E}\left[\sum_{i:\theta_i=B} u_i(\theta)|\mathcal{B}_n\right] \geq \frac{1}{N-n}\mathbb{E}\left[\sum_{i:\theta_i=R} u_i(\theta)|\mathcal{B}_n\right] + \frac{(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\sum_{i\in I}\Pr\left(a_{i,t}\neq a_i^*|\mathcal{B}_n\right)}{N-n}c.$$

Next, note that

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\sum_{i:\theta_i=B} u_i(\theta)|\mathcal{B}_n\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i\in I}\mathbf{1}\{\theta_i=B\}u_i(\rho(\sigma,\theta))|\mathcal{B}_n\right] \\
&= \frac{1}{n}\sum_{i\in I}\Pr(\theta_i=B|\mathcal{B}_n)\,\mathbb{E}\left[u_i(\rho(\sigma,\theta))|\mathcal{B}_n,\theta_i=B\right] \\
&= \frac{1}{n}\sum_{i\in I}\frac{n}{N}u_i^{n,B}(\sigma) = \frac{1}{N}\sum_{i\in I}u_i^{n,B}(\sigma) = u^{n,B}(\sigma),
\end{aligned}$$

and similarly $\frac{1}{N-n}\mathbb{E}\left[\sum_{i:\theta_i=R} u_i(\theta)|\mathcal{B}_n\right] = u^{n,R}(\sigma)$. So we have

$$u^{n,B} \geq u^{n,R} + \gamma_n c, \tag{7}$$

where

$$\gamma_n = \frac{(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\sum_{i\in I}\Pr\left(a_{i,t}\neq a_i^*|\mathcal{B}_n\right)}{N-n} = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\frac{1}{N}\sum_{i\in I}\Pr\left(a_{i,t}\neq a_i^*|\mathcal{B}_n,\theta_i=R\right).$$

With Lemma 3 and inequality (7) 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

$$\Delta_{q,q^-} \geq \sum_{n=0}^{N-1} q_n\left(u^{n,B}-u^{n,R}\right) \geq \sum_{n=1}^{N-1} q_n\left(u^{n,B}-u^{n,R}\right) \geq \sum_{n=1}^{N-1} q_n\gamma_n c,$$

where the first inequality is by Lemma 3, the second is by $q_0\left(u^{0,B}-u^{0,R}\right)\geq 0$, and the third is by (7). Now define $\gamma=\sum_{n=0}^{N-1}q_n\gamma_n$. Since $q_0=q_0-q_0^-\leq\Delta_{q,q'}$ and $\gamma_0\in[0,1]$, 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

$$\begin{aligned}&(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\frac{1}{N}\sum_{i\in I}\sum_{n=0}^{N-1}\Pr\left(\mathcal{B}_n\wedge\theta_i=R\right)\Pr\left(a_{i,t}\neq a_i^*|\mathcal{B}_n,\theta_i=R\right)\\&=(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\frac{1}{N}\sum_{i\in I}\sum_{n=0}^{N-1}(1-z)\,q_n\Pr\left(a_{i,t}\neq a_i^*|\mathcal{B}_n,\theta_i=R\right)=(1-z)\sum_{n=0}^{N-1}q_n\gamma_n=(1-z)\,\gamma.\end{aligned}$$

Therefore, expected social welfare differs from $U\left(a^*\right)$ by at most $(1-z)\,b\gamma\leq(1-z)\,b\frac{1+c}{c}\Delta_{q,q^-}$. This yields (2), and taking $\Delta_{q,q^-}\to 0$ yields (3). ■

It is straightforward to extend Theorem 1 to games with multiple populations, where the players within 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* (where working is privately costly but benefits all other players), 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 how a deep-pocketed principal (or a group of known-rational principals) can induce effort by a large group of agents, while the agents would be unable to support effort by transferring money among themselves.

4 Folk Theorem

We now present a folk theorem for repeated games with incomplete information and perfect monitoring. This theorem covers asymmetric games, but, when specialized to symmetric games, it implies a partial converse to Theorem 1. This shows that the population dominance concept 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 θ. 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 $\prod_{i \in R(\theta)} \Delta(A_i) \times \prod_{i \in B(\theta)} \mathbf{1}\{a_i = a_i^*\}$. Denote player i's minmax payoff in the game $\Gamma(\theta)$ by $\underline{v}_i^{\theta_{-i}} = \min_{\alpha \in \Delta^*(A_{-i}(\theta_{-i}))} \max_{a_i \in A_i} u_i(a_i; \alpha_{-i})$. 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

$$F^*(\theta) = \left\{ v \in F(\theta) : v_i > \underline{v}_i^{\theta_{-i}} \ \forall i \text{ s.t. } \theta_i = R \right\}.$$

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

$$\mathbb{E}\left[v_i(\theta) \,|\theta_i = R\right] > \mathbb{E}\left[\max\left\{v_i(\theta_i = B; \theta_{-i}), \min_{v \in \mathrm{cl}(F^{**}(\theta))} v_i\right\} |\theta_i = R\right] \quad \text{for all } i \in I.^{14} \quad (8)$$

Note that this definition imposes strict versions of both individual rationality (rational player i's payoff exceeds her smallest payoff in $\mathrm{cl}(F^{**}(\theta))$) and incentive compatibility (rational

[14] Here $\mathrm{cl}(\cdot)$ denotes closure in the relative topology.

player i's payoff exceeds her payoff when following the bad-type strategy). Finally, we say that an expected payoff vector $v \in [0,1]^N$ is *consistent with FIRIC* if there exists a FIRIC family of 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, Γ. For any payoff vector $v \in [0,1]^N$ consistent with FIRIC and any $\varepsilon > 0$, there exists $\bar{\delta} < 1$ such that, for every $\delta > \bar{\delta}$, there exists a sequential equilibrium in Γ 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 all 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.)

In contrast, for any symmetric game where the commitment action profile a^* is population dominant, any payoff vector consistent with FIRIC is close to $u(a^*)$ when Δ_{q,q^-} is small. This follows because, by the same argument as in the proof of Theorem 1, incentive compatibility implies that the expected discounted fraction of periods in which players take actions other than a^* (i.e., the variable γ defined in the proof of Theorem 1), is bounded by $\frac{1+c}{c}\Delta_{q,q^-}$, where c is the parameter in the definition of population dominance.

We sketch the proof, deferring the details to the appendix.[15] 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 $\theta(h^t)$ 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 $v(\theta(h^t))$, and that have

[15] The proof is a variation of existing arguments (e.g., Fudenberg and Maskin, 1986; Hörner and Lovo, 2009; Hörner, Lovo, and Tomala, 2011). As compared to the latter two papers, our construction is simpler because we do not require that the equilibrium is "belief-free." Indeed, non-trivial belief-free equilibria typically do not exist in our setting, because a player who is certain that all of her opponents are commitment types can only take a static best response.

the further property that continuation payoffs under the action sequence are always close to $v(\theta(h^t))$.[16] Off the equilibrium path, the players take a sequence of actions that achieve a payoff vector close to $\text{argmin}_{v' \in \text{cl}(F^{**}(\theta(h^t)))} v'_{i(h^t)}$. 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 $\Gamma(\theta(h^t))$. Therefore, since $v(\theta(h^t)) \in F^{**}(\theta(h^t))$, the payoff vector $v(\theta(h^t))$ is attainable in a continuation equilibrium as in Fudenberg and Maskin (1986). Moreover, 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 $\theta(h^t)$ 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 θ_{-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 $\min_{v \in \text{cl}(F^{**}(\theta))} v_i$ (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.

To conclude this section, we show how, when applied to symmetric games, Theorem 2 implies a partial converse to Theorem 1. Fix a symmetric game with a commitment action profile a^*. For each number of bad types n, fix a mixed action profile

$$\alpha^n \in \underset{\alpha \in \Delta(A)}{\text{argmax}}\, u^{n,R}(\alpha) - u^{n,B}(\alpha).$$

That is, α^n maximizes the payoff difference between rational players and bad ones. Next, for any type profile $\theta \in \{R, B\}^N$, let $v^*(\theta)$ denote the payoff vector where rational players take $\alpha^{|\theta|}$. We will show that, if the commitment action profile a^* does not satisfy a slightly generalized version of population dominance, then the family of payoff vectors $(v^*(\theta))_{\theta \in \{R,B\}^N}$ is incentive compatible, and hence, by Theorem 2, can be obtained in equilibrium by pa-

[16] Sorin (1986) and Fudenberg and Maskin (1991) showed that such a sequence exists.

tient players, if these payoffs vectors are also individually rational. In other words, for any symmetric game where our anti-folk theorem does not apply, there exists a strategy profile where rational players outperform bad players, and this strategy profile can be supported as an equilibrium if it is individually rational and the players are patient. This observation implies that Theorem 1 cannot be extended much further.

We say that the commitment action profile a^* satisfies *generalized population dominance* if there exists a positive number $c > 0$ such that

$$\sum_{n=0}^{N-1} \frac{N}{N-n} q_n c_n \geq 0,$$

where

$$c_n = \min_{\alpha \in \Delta(A)} \left(u^{n,B}(\alpha) - u^n(\alpha) - c \frac{\mathbb{E}\left[|\{i \in I : a_i \neq a_i^*\}| \, \alpha, \mathcal{B}_n \right]}{N} \right) \quad \text{for all } n \in \{0, \ldots, N-1\}.$$

We note that population dominance can be replaced with generalized population dominance in Theorem 1.[17]

Now suppose that a^* does not satisfy generalized population dominance for any $c > 0$. Then, for $c_n^* = \max_{\alpha \in \Delta(A)} u^n(\alpha) - u^{n,B}(\alpha)$, we have $\sum_n \frac{N}{N-n} q_n c_n^* \geq 0$. We claim that if this inequality holds with Δ_{q,q^-} slack, so that $\sum_n \frac{N}{N-n} q_n c_n^* - \Delta_{q,q^-} > 0$, then the family of payoff vectors $(v^*(\theta))_{\theta \in \{R,B\}^N}$ is incentive compatible: that is,

$$\mathbb{E}\left[v_i^*(\theta) \, | \theta_i = R \right] > \mathbb{E}\left[v_i^*(\theta_i = B; \theta_{-i}) \, | \theta_i = R \right] \quad \text{for all } i \in I.^{18} \tag{9}$$

To see why this is true, note that, by symmetry, (9) 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 α^n for each realized number of bad types n, we have

$$u^{n,B} = u^n - c_n^* = \frac{n}{N} u^{n,B} + \frac{N-n}{N} u^{n,R} - c_n^*, \quad \text{and hence} \quad u^{n,B} \leq u^{u,R} - \frac{N}{N-n} c_n^*.$$

[17] To see why, by the same proof as (7), we have $u^{n,B} \geq u^{n,R} + (N/(N-n)) c_n + \gamma_n c$. Taking an expectation and using $\sum_{n=1}^{N-1} (N/(N-n)) q_n c_n \geq 0$, this implies that $\sum_{n=1}^{N-1} q_n \left(u^{n,B} - u^{n,R} \right) \geq \sum_{n=1}^{N-1} q_n \gamma_n c$. The rest of the proof is unchanged.

[18] Note that (9) is the same as (8), but without individual rationality.

Taking the expectation over n gives

$$\sum_{n=0}^{N-1} q_n u^{n+1,B} = \sum_{n=0}^{N-1} q_n u^{n,B} + \sum_{n=0}^{N} \left(q_n^- - q_n\right) u^{n,B} \leq \sum_{n=0}^{N-1} q_n u^{n,R} - \sum_{n=0}^{N-1} \frac{N}{N-n} q_n c_n^* + \Delta_{q,q^-} < \sum_{n=0}^{N-1} q_n u^{n,R},$$

as desired.

5 Conclusion

This paper has investigated when a large group of symmetric players can support cooperation when each of them might be committed to defection. Our main result is that cooperation in this environment requires that it is possible to punish defectors without simultaneously punishing the rest of the population as severely. For example, voluntary public goods provision is impossible when the only available incentive instruments are the withdrawal of provision and monetary rewards targeted to contributors; however, involuntary fines targeted to non-contributors restore the possibility of provision. In addition, in the PD with non-anonymous random matching, cooperation is possible if and only if it does not provide large positive externalities to third parties.

We have presented our results in a simple model with one rational type, one commitment type, and (for the folk theorem) perfect monitoring. Extensions to multiple rational or commitment types are straightforward; see SW20 for a discussion of these extensions in the anonymous case. In particular, our anti-folk theorem extends whenever players are committed to population dominant actions with positive (independent) probability, even if there is also a positive probability that they may be committed to different strategies. The simple commitment types considered here are thus "canonical," in the same sense as in the reputation literature (e.g., Fudenberg and Levine, 1989).

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 between the discount factor, the population size, and the precision of monitoring (Sugaya and Wolitzky, 2022). 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, 2021). The interaction between incomplete information and private monitoring more generally is a fairly open area.[19]

6 Appendix: Omitted Proofs

6.1 Proof of Lemma 1

Fix distinct $i, j \in I$. Since (G, a^*, p, Y, χ) is symmetric, there exists an admissible automorphism (π, ϕ, ψ) on G such that $\pi(i) = j$. We construct an admissible automorphism $\left(\tilde{\pi}, \tilde{\phi}\right)$ on Γ such that $\tilde{\pi}(i) = j$, where here admissibility means that $\tilde{\phi}_k(\textit{Always } a_k^*) = \textit{Always } a^*_{\tilde{\pi}(k)}$ for all $k \in I$. First, for each player i and period t, let H_i^t denote the set of player i's period t histories, and define a bijection $\eta_i^t : H_i^t \to H_{\pi(i)}^t$ by

$$\eta_i^t\left((a_{i,\tau}, y_{i,\tau})_{\tau=1}^{t-1}\right) = \left(\phi_i(a_{i,\tau}), \psi_i(y_{i,\tau})\right)_{\tau=1}^{t-1} \qquad \text{for all } h_i^t \in H_i^t.$$

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

$$\tilde{\sigma}_{\pi(i)}\left(h^t_{\pi(i)}\right)\left[a_{\pi(i)}\right] = \sigma_i\left(\left(\eta_i^t\right)^{-1}\left(h^t_{\pi(i)}\right)\right)\left[\phi_i^{-1}\left(a_{\pi(i)}\right)\right] \qquad \text{for all } h^t_{\pi(i)} \in H^t_{\pi(i)}, a_{\pi(i)} \in A_{\pi(i)}. \tag{10}$$

Since η_i^t and ϕ_i are bijections, $\tilde{\phi}_i$ is also a bijection. Also, since (π, ϕ) is admissible, $\phi_i(a_i^*) = a^*_{\pi(i)}$, and hence $\tilde{\phi}_i(\textit{Always } a_i^*) = \textit{Always } a^*_{\tilde{\pi}(i)}$.

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

[19] A couple exceptions are Yamamoto (2014) and Sugaya and Yamamoto (2020).

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

$$\begin{aligned}
\Pr^\sigma\left(a_t, y_t | h^t\right) &= \prod_i \sigma_i\left(h^t_i\right)[a_{i,t}]\,\chi(y_t|a_t) \\
&= \prod_i \tilde{\sigma}_{\pi(i)}\left(\eta^t_i\left(h^t_i\right)\right)[\phi_i(a_{i,t})]\,\chi(y_t|a_t) \\
&= \prod_i \tilde{\sigma}_{\pi(i)}\left(\eta^t_i\left(h^t_i\right)\right)[\phi_i(a_{i,t})]\,\chi(\psi(y_t)|\phi(a_t)) \\
&= \prod_{\pi^{-1}(i)} \tilde{\sigma}_i\left(\eta^t_{\pi^{-1}(i)}\left(h^t_{\pi^{-1}(i)}\right)\right)\left[\phi_{\pi^{-1}(i)}\left(a_{\pi^{-1}(i),t}\right)\right]\chi(\psi(y_t)|\phi(a_t)) \\
&= \Pr^{\tilde{\sigma}}\left(\phi(a_t), \psi(y_t)|\eta^t\left(h^t\right)\right)
\end{aligned}$$

where the second line follows from (10), the third line follows from admissibility of (π, ϕ), the fourth line changes the index from i to $\pi^{-1}(i)$, and the fifth line is by definition of η^t. Given this, by induction on t, for each t and $a_t \in A$, we have

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

Since (π, ϕ) is an automorphism on G, we have

$$\begin{aligned}
u_i(\sigma) &= (1-\delta)\sum_t \delta^{t-1}\sum_{a_t}\Pr^\sigma(a_t)\,u_i(a_t) = (1-\delta)\sum_t \delta^{t-1}\sum_{a_t}\Pr^\sigma(a_t)\,u_{\pi(i)}(\phi(a_t)) \\
&= (1-\delta)\sum_t \delta^{t-1}\sum_{a_t}\Pr^{\tilde{\sigma}}(\phi(a_t))\,u_{\pi(i)}(\phi(a_t)) = u_{\pi(i)}(\tilde{\sigma}),
\end{aligned}$$

as desired.

6.2 Proof of Lemma 2

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

$$u_i(\rho(s,\theta)) = u_{\pi(i)}((\phi \circ \rho)(s,\theta)) = u_{\pi(i)}(\rho(\phi(s), \pi(\theta))), \tag{11}$$

where $\pi(\theta)$ is the type profile defined by $\pi(\theta)_j = \theta_{\pi^{-1}(j)}$ for all $j \in I$. Here, the first equality holds because (π, ϕ) is an automorphism, and the second holds because, since (π, ϕ) is admissible, for each $i \in I$ we have

$$\phi_{\pi^{-1}(i)}\left(\rho(s,\theta)_{\pi^{-1}(i)}\right) = \begin{cases} \phi_{\pi^{-1}(i)}\left(s_{\pi^{-1}(i)}\right) & \text{if } \theta_{\pi^{-1}(i)} = R \\ s_i^* & \text{if } \theta_{\pi^{-1}(i)} = B \end{cases} = \rho\left(\phi(s), \pi(\theta)\right)_i.$$

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

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

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

$$v^{n-1,R} = \left(v_{(n-1)N+i}\right)_{i=1}^N \quad \text{and} \quad v^{n,B} = \left(v_{N^2+(n-1)N+i}\right)_{i=1}^N.$$

Note that v is given by the concatenation of the vectors $v^{n-1,R}$ for $n \in I$, followed by the concatenation of the vectors $v^{n-1,B}$ 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 $v^{n-1,R}$, 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 $v^{n,B}$, for each $i \in I$ and $n \in I$. That is, for each $n \in I$, the $(n-1)N+1^{st}$ through $(n-1)(N+1)^{st}$ coordinates 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 $N^2+(n-1)(N+1)^{st}$ coordinates 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 $\succsim$ denote the lexicographic order on $\mathbb{R}^{2N^2}$. Let $\hat{w}$ denote a maximal element of F in the lexicographic order.[20] Define the vector $\bar{w} \in \mathbb{R}^{2N^2}$ by letting $\bar{w}_{(n-1)N+i} = u^{n-1,R}$ and $\bar{w}_{N^2+(n-1)N+i} = u^{n,B}$ for all $i \in I$ and $n \in I$. Note that $\bar{w} \succsim \hat{w}$; otherwise, there would exist $n \in I$ such that $\hat{w}_{(n-1)N+i} \geq u^{n-1,R}$ for all $i \in \{1,\ldots,N\}$, with strict inequality for some i, which implies that $(1/N)\sum_{i=1}^{N} \hat{w}_{(n-1)N+i} > u^{n-1,R}$ (or, symmetrically, $n \in I$ such that $\hat{w}_{N^2+(n-1)N+i} \geq u^{n,B}$ for all $i \in \{1,\ldots,N\}$ with strict inequality for some i, implying that $(1/N)\sum_{i=1}^{N} \hat{w}_{N^2+(n-1)N+i} > u^{n,B}$), a contradiction.

We now argue that $\hat{w} \succsim \bar{w}$. Suppose toward a contradiction that $\bar{w} \succ \hat{w}$. Let $m \in \{1,\ldots,2N^2\}$ denote the smallest index such that $\hat{w}_m > \bar{w}_m$. 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 $\hat{w}_{(n-1)N+i} > \bar{w}_{(n-1)N+i}$, $(1/N)\sum_{i=1}^{N} \hat{w}_{(n-1)N+i} = u^{n-1,R}$, and the vector $\left(\hat{w}_{(n-1)N+i}\right)_{i=1}^{N}$ 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

$$i \in \operatorname*{argmin}_{k \in I} v_k^{n-1,R} \quad \text{and} \quad j \in \operatorname*{argmax}_{k \in I} v_k^{n-1,R},$$

and note that $v_i^{n-1,R} < v_j^{n-1,R}$. Moreover, note that for all $n' < n$ and $i \in I$, we have $v_i^{n'-1,R} = u^{n'-1,R}$ 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,p) is symmetric, there exists an admissible automorphism (π,ϕ) such that $\pi(i) = j$ and $u_k(\rho(\tilde{s},\theta)) = u_{\pi(k)}(\rho(\phi(\tilde{s}),\pi(\theta)))$ for each $k \in I$, $\theta \in \Theta$, and $\tilde{s} \in S$. 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

[20] Note that the lexicographic order admits a maximum on a compact subset of $\mathbb{R}^{2N^2}$.

$\hat{s}_{k'} = \phi_{k'}^{-1}(\hat{s}_k)$. For every strategy profile $\tilde{s}$, we have

$$\begin{aligned}\sum_\theta \Pr(\theta)\, u_{k'}(\rho(\tilde{s},\theta)) &= \sum_\theta \Pr(\theta)\, u_k(\rho(\phi(\tilde{s}),\pi(\theta))) \\ &= \sum_\theta \Pr(\pi(\theta))\, u_k(\rho(\phi(\tilde{s}),\pi(\theta))) \\ &= \sum_\theta \Pr(\theta)\, u_k(\rho(\phi(\tilde{s}),\theta)), \end{aligned} \tag{12}$$

where the first line follows because (π,ϕ) is an admissible automorphism (so (11) holds), the second line follows because π is symmetric, and the third line follows by rearranging the sum. Hence, we have

$$\begin{aligned}\sum_\theta \Pr(\theta)\, u_k(\rho(s',\theta)) &= \sum_\theta \Pr(\theta)\, u_k(\rho(\phi(s),\theta)) = \sum_\theta \Pr(\theta)\, u_{k'}(\rho(s,\theta)) \\ &\geq \sum_\theta \Pr(\theta)\, u_{k'}(\rho((\hat{s}_{k'};s_{-k'}),\theta)) \\ &= \sum_\theta \Pr(\theta)\, u_k\left(\rho\left(\left(\hat{s}_k;\phi(s)_{-k}\right),\theta\right)\right) = \sum_\theta \Pr(\theta)\, u_k\left(\rho\left(\left(\hat{s}_k,s'_{-k}\right),\theta\right)\right), \end{aligned}$$

where the first and last equalities follow because $s' = \phi(s)$, the second and third equalities follow by (12) and $\hat{s}_{k'} = \phi_{k'}^{-1}(\hat{s}_k)$, 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 $v'_{(n'-1)N+k} = u_k^{n'-1,R}(s')$ and $v'_{N^2+(n'-1)N+k} = u_k^{n',B}(s')$. Since $s' \in S^*$, the resulting vector v' lies in U. Moreover, by (11) and symmetry of π, for each $k \in I$ and $n' \in I$ we have

$$\begin{aligned}u_k^{n'-1,R}(s) &= \sum_{\theta:|\theta|=n'-1,\theta_k=R} \frac{\Pr(\theta)}{\Pr(|\theta|=n'-1,\theta_k=R)} u_k(\rho(s,\theta)) \\ &= \sum_{\theta:|\theta|=n'-1,\theta_k=R} \frac{\Pr(\theta)}{\Pr(|\theta|=n'-1,\theta_k=R)} u_{\pi(k)}(\rho(\phi(s),\pi(\theta))) \\ &= \sum_{\theta:|\theta|=n'-1,\theta_{\pi(k)}=R} \frac{\Pr(\theta)}{\Pr\left(|\theta|=n'-1,\theta_{\pi(k)}=R\right)} u_{\pi(k)}(\rho(\phi(s),\pi(\theta))) \\ &= u_{\pi(k)}^{n-1,R}(s'). \end{aligned}$$

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 $\left(v'_{kN+i}\right)_{i=1}^{N}$ 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, $\bar{s}\in\Delta\left(S^{**}\right)$. Let

$$\bar{v}=u\left(\bar{s}\right)=\frac{1}{2}\left(v+v'\right),$$

and note that $f\left(\bar{v}\right)\in F$. Since, for all $n'<n$ and $k\in I$, we have $v_k^{n'-1,R}=u^{n'-1,R}$, it follows that

$$\bar{v}_k^{n'-1,R}=u^{n'-1,R}\quad\text{for all } k\in I,\ n'<n. \tag{13}$$

In addition, since $\pi\left(i\right)=j$, $i\in\operatorname{argmin}_{k\in I}v_k^{n-1}$, and $j\in\arg\max_{k\in I}v_k^{n-1}$, we have

$$\bar{v}_i^{n-1,R}=\frac{1}{2}\left(v_i^{n-1,R}+v_j^{n-1,R}\right)>v_i^{n-1,R}. \tag{14}$$

Moreover, since $\left(v'\right)^{n-1,R}$ is a permutation of $v^{n-1,R}$, and $\bar{v}=\frac{1}{2}\left(v+v'\right)$, we also have

$$\bar{v}_k^{n-1,R}\ \geq\ v_i^{n-1,R}\quad\text{for all } k\in I,\quad\text{and} \tag{15}$$

$$\bar{v}_k^{n-1,R}\ >\ v_i^{n-1,R}\quad\text{for all } k\in I\backslash\operatorname*{argmin}_{k'\in I}v_{k'}^{n-1,R}. \tag{16}$$

Since $i\in\operatorname{argmin}_{k\in I}v_{k'}^{n-1,R}$, (13), (14), (15), and (16) together imply that $f\left(\bar{v}\right)\succ f\left(v\right)$. But this contradicts the maximality of $\hat{w}=f\left(v\right)$ in F. We can thus conclude that $\hat{w}\succsim\bar{w}$.

Since $\bar{w}\succsim\hat{w}$ and $\hat{w}\succsim\bar{w}$, we conclude that $\hat{w}=\bar{w}$, proving the lemma.

6.3 Proof of Theorem 2

Fix such a payoff vector v and $\varepsilon>0$. Let $\left(v\left(\theta\right)\right)_{\theta\in\{R,B\}^N}$ satisfy (8) and $v=\mathbb{E}\left[v\left(\theta\right)\right]$.

Let $\Theta^{\text{int}}\subseteq\{R,B\}^N$ denote the set of type profiles such that the projection of $F^*\left(\theta\right)$ on the set of rational-player payoff vectors has non-empty relative interior, and hence $F^{**}\left(\theta\right)=F^*\left(\theta\right)$. By (8), there exists a constant $\eta>0$ such that

$$\begin{gathered}v_i\left(\theta\right)\geq\min_{v'\in\text{cl}(F^{**}(\theta))}v'_i+\eta\quad\text{for all }\theta\in\Theta^{\text{int}},\\ \mathbb{E}\left[v_i\left(\theta\right)|\theta_i=R\right]>\mathbb{E}\left[\max\left\{v_i\left(\left(\theta_i=B;\theta_{-i}\right)\right),\min_{v'\in\text{cl}(F^{**}(\theta))}v'_i\right\}|\theta_i=R\right]+\eta\quad\text{for all } i\in I.\end{gathered} \tag{17}$$

Lemma 4 *There exists $\bar{\delta} < 1$ such that, for every $\delta > \bar{\delta}$, the following conditions hold:*

1. *For each $\theta \in \Theta^{\text{int}}$, there exists a sequence of pure action profiles $\{\alpha_t(\theta)\}_{t=1}^{\infty}$ such that $v(\theta) = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u(\alpha_t(\theta))$ and, for each t, $|(1-\delta)\sum_{t'=t}^{\infty}\delta^{t'-1}u(\alpha_{t'}(\theta)) - v(\theta)| < \eta/4$.*

2. *For each $\theta \notin \Theta^{\text{int}}$, there exists sequence of mixed action profiles $\{\alpha_t(\theta)\}_{t=1}^{\infty}$ such that $\alpha_t(\theta)$ is a static Nash equilibrium for every θ, $v(\theta) = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u(\alpha_t(\theta))$, and, for each t, $|(1-\delta)\sum_{t'=t}^{\infty}\delta^{t'-1}u(\alpha_{t'}(\theta)) - v(\theta)| < \eta/4$.*

3. *For each $\theta \in \Theta$ and $i \in I$, there exists a subgame-perfect equilibrium $\sigma^i(\theta)$ in the game $\Gamma(\theta)$ with equilibrium payoff v satisfying $v_i \leq \min_{v' \in \text{cl}(F^{**}(\theta))} v'_i + \eta/4$ for all $i \in I$.*

Proof. Follows from Fudenberg and Maskin (1991), observing that the set of feasible and individually rational payoffs has non-empty interior when $\theta \in \Theta^{\text{int}}$, and that $F^{**}(\theta)$ is defined as the convex hull of the set of static Nash equilibrium payoffs when $\theta \notin \Theta^{\text{int}}$. ■

Without loss, we take $\bar{\delta} \geq 1 - \eta/8$.

Recall that θ^* is the realized set of rational players. For any history h^t, let $\theta(h^t) \subseteq \theta^*$ denote the set of players who have ever taken any action other than a^* at h^t: we call this the set of players who have *revealed rationality at* h^t.

We now fix an arbitrary static NE in the one-shot game where θ^* is distributed according to p, each player $i \in \theta^*$ is restricted to take actions in $A \setminus \{a^*\}$, and each player $i \notin \theta^*$ is restricted to take $\{a^*\}$. Let $\alpha_0 \in \Delta^*(A(\theta^*))$ be the equilibrium strategy. Let $\bar{\Sigma}_i$ be the set of strategies for player i where she takes $\alpha_{i,0}$ in period 1.

We will prove that each $v(\theta^*)$ is attainable without fully constructing the equilibrium strategy profile: in particular, we will not construct the continuation strategy of a rational player who does not reveal rationality in the first period, leaving this defined implicitly. Formally, we will construct a *quasi-equilibrium*, which is a strategy profile in $\bar{\Sigma}$ that satisfies the following conditions for each player $i \in \theta^*$:

1. Player i takes $\alpha_{i,0}$ in period 1.

2. For each period $t \geq 2$ and history h^t such that $i \in \theta(h^t)$, it is optimal for player i to follow the equilibrium strategy at history h^t, conditional on the event that $\theta^* = \theta(h^t)$.

3. It is optimal for player i to follow the equilibrium strategy in period 1.

We first establish that this suffices to deliver the theorem.

Lemma 5 *Fix a strategy profile $\sigma \in \bar{\Sigma}$. For any belief system μ such that (σ, μ) satisfies Kreps-Wilson consistency, and for any $t \geq 2$, after each (possibly off-path) history h^t, every rational player i believes with probability 1 that $\theta(h^t) \cup \{i\} = \theta^*$.*

Proof. It is immediate that every rational player believes with probability 1 that $\theta(h^t) \cup \{i\} \subseteq \theta^*$. Thus, we prove that every rational player believes with probability 1 that, for any $j \notin \theta(h^t) \cup \{i\}$, player j is a commitment type. Since $j \notin \theta(h^t) \cup \{i\}$, player j takes a_j^* for all periods $1, ..., t-1$. For any completely mixed sequence of strategy profiles converging to σ, conditional on any sequence of the other players' actions, this action sequence for player j is played with non-vanishing probability in the (positive probability) event that player j is bad, but is played with vanishing probability when player j is rational (as $\alpha_{j,0}$ puts zero probability on a_j^*). Hence, the corresponding limit beliefs put probability 1 on the event that player j is rational. ■

Lemma 6 *For any quasi-equilibrium, there exists an outcome-equivalent sequential equilibrium.*

Proof. Given a quasi-equilibrium σ^*, we can construct an outcome-equivalent strategy profile σ^{**} by specifying that, for each period t, (i) if either $t = 1$ or h^t satisfies $i \in \theta(h^t)$, player i follows σ_i^*, and (ii) if $t > 1$ and $i \in \theta^* \backslash \theta(h^t)$, player i takes a (dynamic) best response given the belief that $\theta^* = \theta(h^t) \cup \{i\}$ (and hence, by (i), given the belief that players $-i$ follow σ_{-i}^*). Since $\sigma^{**} \in \bar{\Sigma}$, by Lemma 5, at any history h^t with $t \geq 2$, every rational player i believes that $\theta^* = \theta(h^t) \cup \{i\}$. Thus, σ_i^{**} is sequentially rational given Conditions 1–3 in the definition of a quasi-equilibrium. ■

We now construct a quasi-equilibrium that attains $v(\theta)$ whenever $\theta^* = \theta$, for each θ. In period 1, each player $i \in \theta^*$ takes a_i according to $\alpha_{0,i}$ in period 1. In period $t \geq 2$, players follow an automaton strategy profile with state (θ, ω), where $\theta \subseteq I$ and $\omega \in \{0\} \cup (\{3, ..., t\} \times I)$. The initial state is $\theta = \theta(h^2)$ and $\omega = 0$.

Given the current state (θ, ω) and calendar time t, actions are determined as follows: (i) If $\omega = 0$, then player i takes $\alpha_{i,t-1}(\theta)$ specified in Lemma 4.[21] (ii) If $\omega = (n, \tau)$ for some $n \in I$ and $\tau \leq t$, then player i takes $\sigma_i^n(\theta)(h^{\tau:t})$, where $h^{\tau:t} = (a_\tau, ..., a_{t-1})$ is the history following period τ (with $h^{\tau:\tau} = \{\emptyset\}$).

Given the current state (θ, ω) and realized action profile a_t, the next-period state (θ', ω') is determined as follows: (i) If $\omega = 0$ and each player i takes $a_{i,t} \in \operatorname{supp} \alpha_{i,t-1}(\theta)$, then $(\theta', \omega') = (\theta, 0)$. (ii) If $\omega = 0$ and there is a unique player i who takes $a_{i,t} \notin \operatorname{supp} \alpha_{i,t-1}(\theta)$, then $\theta' = \theta \cup \{i\}$ and $\omega' = (i, t+1)$. (iii) If $\omega = (n, \tau)$ and each player i takes $a_{i,t} \in \operatorname{supp} \sigma_i^n(\theta)(h^{\tau:t})$, then $\theta' = \theta$ and $\omega' = \omega$. (iv) If $\omega = (n, \tau)$ and there is a unique player $i \in \theta$ who takes $a_{i,t} \notin \operatorname{supp} \sigma_i^n(\theta)(h^{\tau:t})$, then $\theta' = \theta$ and $\omega' = \omega$. (v) If $\omega = (n, \tau)$ and there is a unique player $i \notin \theta$ who takes $a_i \notin \sigma_i^n(\theta)(h^{\tau:t}) = \{a_i^*\}$, then $\theta' = \theta \cup \{i\}$ and $\omega' = (i, t+1)$.

Conditional on each realization of the set of rational players θ^*, this strategy profile delivers payoff $v(\theta^*)$ from period 2 onward. Since $v = \mathbb{E}[v(\theta)]$, for sufficiently large δ the ex ante expected payoffs are within ε of v.

It remains to verify that the strategy profile is a quasi-equilibrium. Condition 1 holds by construction. For Condition 2, note that if $\omega = 0$ and $\theta(h^t) \in \Theta^{\text{int}}$, then, by Lemma 4, the on-path continuation payoff is at least $v(\theta(h^t)) - \eta/4$, while the deviation payoff is at most $(1-\delta)(1) + \delta\left(\min_{v' \in \text{cl}(F^{**}(\theta(h^t)))} v_i' + \eta/4\right)$. By (17) and $\delta \geq 1 - \eta/8$, the former quantity is greater, so the prescribed strategy is optimal. If instead $\omega = 0$ and $\theta(h^t) \notin \Theta^{\text{int}}$, then the prescribed strategy is a sequence of static Nash equilibria. If instead $\omega \neq 0$, then the prescribed strategy $\sigma^i(\theta)$ is a subgame-perfect equilibrium.

Finally, for Condition 3, note that

$$\begin{aligned} & \mathbb{E}\left[(1-\delta)\sum_{t=1}^{\infty} u_i(a_t) \,|\sigma^*, i \in \theta^*\right] \\ = \;& \mathbb{E}\left[(1-\delta)\, u_i(\alpha) + \delta v_i(\theta^*) \,|i \in \theta^*\right] \\ \geq \;& \mathbb{E}\left[v_i(\theta^*) \,|i \in \theta^*\right] - (1-\delta)(1) \geq \mathbb{E}\left[v_i(\theta^*) \,|i \in \theta^*\right] - \frac{\eta}{8}. \end{aligned} \tag{18}$$

[21] Note that the time index is shifted by one since the game $\Gamma(\theta)$ in Lemma 4 does not have the revelation stage.

By contrast, we can bound $\max_{\sigma_i} \mathbb{E}\left[(1-\delta)\sum_{t=1}^{\infty} u_i(a_t)\,|\sigma_i, \sigma^*_{-i}\right]$ as follows. Taking any action outside $\operatorname{supp}\alpha_{0,i}$ other than a_i^* in period 1 is unprofitable, as it leads to a lower payoff in period 1 (as α_0 is a static equilibrium) and the same continuation payoff starting in period 2 (since Condition 2 has established that, for player $i \in \theta(h^t)$, it is optimal to follow the equilibrium strategy). We thus focus on strategies that take a_i^* in period 1. By Lemma 5, player i believes with probability 1 that $\theta(h^1) = \theta^*\backslash\{i\}$. By construction, for each θ^*, conditional on the event $\theta^*\backslash\{i\} = \theta$, taking an action $a_{i,t} \notin \operatorname{supp}\alpha_{i,t-1}(\theta)$ in period t at state $(\theta, 0)$ gives a continuation payoff of at most $\min_{v' \in \mathrm{cl}(F^{**}(\theta^*))} v_i' + \eta/4$. Since the sequence of equilibrium action distributions $\alpha_t = \alpha_{t-1}(\theta^*\backslash\{i\})$ is deterministic for $t \geq 2$, by optimally choosing the period T in which player i reveals rationality, we have

$$\begin{aligned}
&\max_{\sigma_i} \mathbb{E}\left[(1-\delta)\sum_{t=1}^{\infty} u_i(a_t)\,|\sigma_i, \sigma^*_{-i}, i \in \theta^*\right] \\
\leq\ & (1-\delta) + \mathbb{E}\left[\begin{array}{c} \max_{T\geq 2}(1-\delta)\sum_{t=2}^{T-1}\delta^{t-1}u_i(\alpha_{t-1}(\theta^*\backslash\{i\})) \\ +(1-\delta)\delta^{T-1}(1) + \delta^T\left(\min_{v' \in \mathrm{cl}(F^{**}(\theta^*))} v_i' + \frac{\eta}{4}\right) \end{array} \Big|\, i \in \theta^*\right].
\end{aligned}$$

For each θ^*, recalling that $v_i(\theta^*\backslash\{i\}) = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(\alpha_t(\theta^*\backslash\{i\}))$, we have

$$\begin{aligned}
&\left|(1-\delta)\sum_{t=1}^{T}\delta^{t-1}u_i(\alpha_t(\theta^*\backslash\{i\})) - \left(1-\delta^T\right)v_i(\theta^*\backslash\{i\})\right| \\
\leq\ & \delta^T\left|v_i(\theta^*\backslash\{i\}) - (1-\delta)\sum_{t=T+1}^{\infty}\delta^{t-T-1}u_i(\alpha_t(\theta^*\backslash\{i\}))\right| \leq \frac{\eta}{4}, \qquad (19)
\end{aligned}$$

where the latter inequality follows from Lemma 4. Hence, we have

$$\begin{aligned}
&\max_{\sigma_i} \mathbb{E}\left[(1-\delta)\sum_{t=1}^{\infty} u_i\left(a_t\right)|\sigma_i, \sigma_{-i}^*, i \in \theta^*\right] \\
\leq\ & (1-\delta) + \mathbb{E}\left[\max_{T\geq 2}(1-\delta)\sum_{t=2}^{T-1}\delta^{t-1}u_i\left(\alpha_{t-1}\left(\theta^*\backslash\{i\}\right)\right) + (1-\delta)\,\delta^{T-1} + \delta^T\left(\min_{v'\in \mathrm{cl}(F^{**}(\theta^*))} v_i' + \frac{\eta}{4}\right)|i\in\theta^*\right] \\
\leq\ & 3(1-\delta) + \mathbb{E}\left[\max_{T\geq 1}(1-\delta)\sum_{t=1}^{T}\delta^{t-1}u_i\left(\alpha_t\left(\theta^*\backslash\{i\}\right)\right) + \delta^{T+1}\min_{v'\in \mathrm{cl}(F^{**}(\theta^*))} v_i'|i\in\theta^*\right] + \frac{\eta}{4} \\
\leq\ & 3(1-\delta) + \mathbb{E}\left[\max_{T\geq 1}\left(1-\delta^T\right)v_i\left(\theta^*\backslash\{i\}\right) + \delta^{T+1}\min_{v'\in \mathrm{cl}(F^{**}(\theta^*))} v_i'|i\in\theta^*\right] + \frac{\eta}{2} \\
\leq\ & \mathbb{E}\left[\max\left\{v_i\left(\theta^*\backslash\{i\}\right), \min_{v'\in \mathrm{cl}(F^{**}(\theta^*))} v_i'\right\}|i\in\theta^*\right] + \frac{7\eta}{8},
\end{aligned}$$

where the second inequality changes the time index, the third uses (19), and the fourth uses $\delta > 1-\eta/8$. Finally, (17) and (18) imply Condition 3.

References

[1] Al-Najjar, Nabil I., and Rann Smorodinsky. "Pivotal Players and the Characterization of Influence." *Journal of Economic Theory* 92.2 (2000): 318-342.

[2] Friedman, Eric J., and Paul Resnick (2001), "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10.2: 173-199.

[3] Fudenberg, Drew, and David K. Levine. "Reputation and Equilibrium Selection in Games with a Patient Player." *Econometrica* 57.4 (1989): 759-778

[4] Fudenberg, Drew, and Eric Maskin (1986), "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54.3: 533-554.

[5] Fudenberg, Drew, and Eric Maskin (1991), "On the Dispensability of Public Randomization in Discounted Repeated Games." *Journal of Economic Theory* 53.2: 428-438.

[6] Green, Edward J. "Noncooperative Price Taking in Large Dynamic Markets." *Journal of Economic Theory* 22.2 (1980): 155-182.

[7] Ham, Nicholas (2021), "Notions of Anonymity, Fairness, and Symmetry for Finite Strategic-Form Games." *working paper.*

[8] Hefti, Andreas (2017), "Equilibria in Symmetric Games: Theory and Applications." *Theoretical Economics* 12.3: 979-1002.

[9] Hörner, Johannes, and Stefano Lovo (2009), "Belief-Free Equilibria in Games with Incomplete Information." *Econometrica* 77(2): 453-487.

[10] Hörner, Johannes, Stefano Lovo, and Tristan Tomala (2011), "Belief-Free Equilibria in Games with Incomplete Information: Characterization and Existence." *Journal of Economic Theory* 146.5: 1770-1795.

[11] Ligon, Ethan, Jonathan P. Thomas, and Tim Worrall (2002), "Informal Insurance Arrangements with Limited Commitment: Theory and Evidence from Village Economies." *Review of Economic Studies* 69.1: 209-244.

[12] Mailath, George J., and Andrew Postlewaite (1990), "Asymmetric Information Bargaining Problems with Many Agents." *Review of Economic Studies* 57.3: 351-367.

[13] Mailath, George J., and Larry Samuelson. *Repeated Games and Reputations: Long-Run Relationships.* Oxford University Press, 2006.

[14] McLean, Richard, and Andrew Postlewaite (2002), "Informational Size and Incentive Compatibility." *Econometrica* 70.6: 2421-2453.

[15] Miguel, Edward, and Gugerty, Mary K. (2005), "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics* 89: 2325-2368.

[16] Nash, John (1951), "Non-Cooperative Games." *Annals of Mathematics* 54.2: 286-295.

[17] Ostrom, Elinor (1990), *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.

[18] Plan, Asaf (2017), "Symmetric n-Player Games." *Working paper.*

[19] Sabourian, Hamid. "Anonymous Repeated Games with a Large Number of Players and Random Outcomes." *Journal of Economic Theory* 51.1 (1990): 92-110.

[20] Sorin, Sylvain (1986), "On Repeated Games with Complete Information." *Mathematics of Operations Research* 11.1: 147-160.

[21] Stein, Noah (2011), "Exchangeable Equilibria." *Doctoral thesis*, MIT.

[22] Sugaya, Takuo, and Alexander Wolitzky (2020), "A Few Bad Apples Spoil the Barrel: An Anti-Folk Theorem for Anonymous Repeated Games with Incomplete Information." *American Economic Review* 110: 3817-3835.

[23] Sugaya, Takuo, and Alexander Wolitzky (2022), "Repeated Games with Many Players." *Working paper.*

[24] Sugaya, Takuo, and Alexander Wolitzky (2021), "Communication and Community Enforcement." *Journal of Political Economy* 129: 2595-2628.

[25] Sugaya, Takuo, and Yuichi Yamamoto (2020), "Common Learning and Cooperation in Repeated Games." *Theoretical Economics* 15.3: 1175-1219.

[26] Yamamoto, Yuichi (2014), "Individual Learning and Cooperation in Noisy Repeated Games." *Review of Economic Studies* 81.1: 473-500.