# *A Few Bad Apples Spoil the Barrel: An Anti-Folk Theorem for Anonymous Repeated Games with Incomplete Information*

*American Economic Review 2020, 110(12): 3817–3835*
*https://doi.org/10.1257/aer.20200068*

# A Few Bad Apples Spoil the Barrel: An Anti-Folk Theorem for Anonymous Repeated Games with Incomplete Information†

*By* Takuo Sugaya and Alexander Wolitzky*

*We study anonymous repeated games where players may be "commitment types" who always take the same action. We establish a stark anti-folk theorem: if the distribution of the number of commitment types satisfies a smoothness condition and the game has a "pairwise dominant" action, this action is almost always taken. This implies that cooperation is impossible in the repeated prisoner's dilemma with anonymous random matching. We also bound equilibrium payoffs for general games. Our bound implies that industry profits converge to zero in linear-demand Cournot oligopoly as the number of firms increases.* (*JEL* C72, C73, D83)

The folk theorem of repeated games asserts that a group of rational players, however large, can cooperate if they are sufficiently patient and have enough information about each other's past behavior.[1] But it seems more realistic to assume that, if a group is large enough, it probably contains some irrational agents. For example, Kandori (1992) and Ellison (1994) show that rational players can support cooperation in the prisoner's dilemma with anonymous random matching by relying on *contagion strategies*: whenever a player sees anyone defect, she starts defecting against everyone. But Ellison (p. 578) also notes: "If one player were 'crazy' and always played $D$ [defect] … contagious strategies would not support cooperation. In large populations, the assumption that all players are rational and know their opponents' strategies may be both very important to the conclusions and fairly implausible."

In this paper, we show that the folk theorem fails for large groups when players are anonymous (so a player's payoff depends only on her own action and the number of opponents taking each action) and may be "commitment types" who always take the same action. For example, in the prisoner's dilemma with anonymous random matching, population size $N$, and discount factor $\delta$, cooperation is impossible

* Sugaya: Graduate School of Business, Stanford University (email: tsugaya@stanford.edu); Wolitzky: Department of Economics, MIT (email: wolitzky@mit.edu). Jeffrey Ely was the coeditor for this article. For helpful comments, we thank Daron Acemoglu, Glenn Ellison, Drew Fudenberg, Yuval Heller, Matt Jackson, George Mailath, Stephen Morris, Satoru Takahashi, Omer Tamuz, the anonymous referees, and seminar participants at Princeton, Stanford, and Wharton. We thank Eitan Sapiro-Gheiler for careful proofreading. Wolitzky acknowledges financial support from the NSF and the Sloan Foundation.

† Go to https://doi.org/10.1257/aer.20200068 to visit the article page for additional materials and author disclosure statements.

[1] The literature on the folk theorem is enormous. For a textbook treatment, see Mailath and Samuelson (2006).

when $N$ is large, even if $(1-\delta)N$ is small. Similarly, in linear-demand Cournot oligopoly, industry profits converge to zero as $N \to \infty$, even if $(1-\delta)N \to 0$.

The key assumption behind our results is that the distribution of the number of commitment types is "smooth": roughly speaking, for every number $n < N$, the probability that $n$ players are commitment types is close to the probability that $n+1$ players are commitment types. For instance, this assumption is satisfied if each player is a commitment type with independent probability $z$, for any fixed $z \in (0,1)$, and $N \to \infty$.

To see why the folk theorem fails with a smooth distribution of commitment types, observe that, if a rational player deviates from her equilibrium strategy by instead following the strategy of a commitment type, and if the number of "true" commitment types is $n$, then the population distribution of actions is exactly what it would have been if the rational player had not deviated and the number of commitment types had been $n+1$. Smoothness thus implies that a single deviation from the rational-type strategy to the commitment-type strategy has a small impact on the population distribution of actions. Therefore, the commitment type's action cannot perform much better than the rational type's equilibrium strategy against the equilibrium action distribution (a fact we formalize in Lemma 1). Finally, in many games this fact implies that the rational type's equilibrium strategy almost always prescribes the commitment type action, which yields an anti-folk theorem. For example, if the game is the prisoner's dilemma and the commitment type's action is defect, the rational type's equilibrium strategy must almost always defect.

More precisely, we consider anonymous repeated games with one rational type and one commitment type.[2] We first establish Lemma 1: the commitment-type strategy cannot yield a much higher payoff than the rational-type equilibrium strategy, where the size of the gap depends on the smoothness of the distribution of the number of commitment types. We then consider games with a "pairwise dominant" action $a^*$, meaning that, whenever one player takes action $a^*$ and another player takes a different action $a$, the player taking $a^*$ obtains a strictly higher payoff than the player taking $a$, and assume that commitment types take this action. For instance, defection is pairwise dominant in the prisoner's dilemma. Our main result (Theorem 1) shows that, as $N \to \infty$, the pairwise dominant action is almost always taken in every Nash equilibrium. We then briefly consider implications of Lemma 1 for games without a pairwise dominant action, showing in particular that industry profits converge to zero in linear-demand Cournot oligopoly as the number of firms increases.

The paper concludes by discussing implications of our approach beyond anonymous repeated games. One such implication is an elementary proof of a version of Mailath and Postlewaite's (1990) impossibility theorem for public good provision in large populations, which unlike existing proofs allows types to be correlated.

*Related Literature*.—This paper relates to several branches of literature. Most directly, we contribute to the literature on repeated games with anonymous random

[2] Our results can be extended to allow multiple commitment types at the cost of additional notation. We discuss this extension in Section V.

matching by showing that the existence of cooperative equilibria in such models is not robust to introducing a smooth distribution of commitment types.[3]

There are two related strands of literature on anti-folk theorems. First, several papers following Green (1980) and Sabourian (1990) consider large-population, complete-information games where the impact of each player's action on the aggregate signal distribution is small.[4] These papers give conditions, such that, for fixed $\delta$, all Nash equilibria of the repeated game converge to static Nash equilibria as $N \to \infty$; significantly, they do not give anti-folk theorems in the sense of persistent inefficiency as $\delta \to 1$. In contrast, with incomplete information we allow arbitrarily informative signals of actions (e.g., perfect monitoring) and give conditions, such that convergence to static Nash equilibrium obtains uniformly in $\delta$.

Second, like our paper, the "reputation" literature shows that introducing a small amount of incomplete information in repeated games can lead to anti-folk theorems (Mailath and Samuelson 2006). Reputation models typically consider a small number of long-run players (often only one), and are thus far from the large anonymous games we consider. In particular, the key argument that bounds a rational player's payoff in this literature (due to Fudenberg and Levine 1989) is that, if a rational player follows the strategy of a commitment type, this eventually causes her opponents to start taking favorable actions in response. In contrast, the key argument that bounds a rational player's payoff in our large anonymous games is that, if a rational player follows the strategy of a commitment type, this has only a small effect on the distribution of her opponents' actions.

Finally, a literature closer to mechanism design considers measures of the pivotality or influence of a player's type on an aggregate outcome, and gives conditions under which most players' influence must be small in large populations. For instance, al-Najjar and Smorodinsky (2000) show that, with independent types, players' average influence on a bounded, real-valued aggregate outcome goes to zero as $N \to \infty$. This result is distinct from our condition that the distribution of the number of agents with a specific type does not vary much with a particular player's type. And this distinction makes a difference: al-Najjar and Smorodinsky (2001) apply their notion of influence to continuation payoffs in repeated games to derive a Green-Sabourian-type result that depends on the order of limits between $N$ and $\delta$, while our results are uniform in $\delta$.[5]

## I. Model

A symmetric $N$-player stage game with action set $A$ and payoff function $u : A^N \to \mathbb{R}$ is *anonymous* if, for any $i \in I = \{1, \ldots, N\}$, any permutation $\pi$ on $I \backslash \{i\}$, and any action profile $\mathbf{a} = (a_j)_{j \in I} \in A^N$, we have $u_i(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_N) = u_i(a_{\pi(1)}, \ldots, a_{\pi(i-1)}, a_i, a_{\pi(i+1)}, \ldots, a_{\pi(N)})$: that is, a player's payoff depends only on her own action and the number of opponents taking each action. Fix a finite, anonymous stage game, and normalize the range of $u$ to lie in $[0, 1]$. Throughout the

[3] Without allowing commitment types, Kandori (1992) and Ellison (1994) showed that mutual cooperation is supportable in the prisoner's dilemma, and Deb, Sugaya, and Wolitzky (2020) established a general folk theorem.

[4] See also Levine and Pesendorfer (1995); Fudenberg, Levine, and Pesendorfer (1998); al-Najjar and Smorodinsky (2001); Pai, Roth, and Ullman (2016); and Awaya and Krishna (2016, 2019).

[5] Another way to appreciate the difference is to note that we allow correlated types.

paper, whenever we write a player's payoff as $u(\mathbf{a})$ (without an $i$-subscript on $u$), the first element of $\mathbf{a}$ refers to the player's own action and the remaining elements refer to the opponents' actions, which by anonymity can be ordered arbitrarily: thus, $u(a_i, a_{-i})$ is player $i$'s payoff when she takes action $a_i \in A$ and her opponents take actions $a_{-i} \in A^{(N-1)}$. In contrast, $u_i(\mathbf{a}) = u_i(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_N)$.[6]

The stage game is played repeatedly in periods $t = 1, 2, \ldots$. After taking an action in period $t$, each player $i$ observes a signal $y_{i,t}$ drawn from a probability distribution that depends on the history of past actions and signals $\left((\mathbf{a}_\tau, \mathbf{y}_\tau)_{\tau=1}^{t-1}, \mathbf{a}_t\right)$, where $\mathbf{y}_\tau = (y_{1,\tau}, \ldots, y_{N,\tau})$ and $\mathbf{a}_\tau = (a_{1,\tau}, \ldots, a_{N,\tau})$. A *history* for player $i$ at the beginning of period $t$ is thus $h_i^t = (a_{i,\tau}, y_{i,\tau})_{\tau=1}^{t-1}$, with $h_i^1 = \varnothing$. A *strategy* $\sigma_i$ for player $i$ maps histories $h_i^t$ to $\Delta(A)$ for each $t$.

Each player $i$ has a type $\theta_i \in \{R, B\}$, where $R$ is the *rational* type and $B$ is the *bad* (commitment) type.[7] Rational types maximize expected discounted payoffs with discount factor $\delta \in [0, 1)$. Bad types always play a particular action $a^* \in A$; we call this strategy *Always* $a^*$. A *strategy profile* $\sigma = (\sigma_i)_i$ specifies the strategy $\sigma_i$ that each player $i$ follows *when she is rational*; of course, when she is bad, she plays *Always* $a^*$.

There is a common prior $p$ on the set of players' types $\{R, B\}^N$, which we assume is symmetric: for every permutation $\pi$ on $I$ and every type profile $(\theta_1, \ldots, \theta_N)$, $p(\theta_1, \ldots, \theta_N) = p(\theta_{\pi(1)}, \ldots, \theta_{\pi(N)})$. The repeated game is thus parameterized by the tuple $\Gamma = (N, A, u, \delta, a^*, p)$. We denote the probability that a player is bad by $z = \sum_{\theta : \theta_i = B} p(\theta)$.

Given a strategy profile $\sigma$, denote player $i$'s expected discounted per-period payoff conditional on type profile $\theta$ by $U_i(\theta) \in [0, 1]$, and denote player $i$'s expected payoff by $U_i = \sum_\theta p(\theta) U_i(\theta) \in [0, 1]$. Since this expectation includes the possibility that $\theta_i = B$, we implicitly assume that bad types have the same utility function as rational types.[8]

Our results concern the set of equilibrium values of per-capita utilitarian social welfare, $\sum_i U_i / N$. This set is only expanded by letting the players access a public randomization device. Since the stage game and prior are symmetric, when public randomization is available any social welfare level attainable by an asymmetric strategy profile is also attained by the symmetric profile where public randomization is first used to randomly permute the players' strategies. We therefore allow public randomization and restrict attention to symmetric strategy profiles (and often drop the $i$ subscript from $U_i(\theta)$ and $U_i$).

Note that we allow $\delta = 0$, so our results apply equally to one-shot games.

[6] Note that symmetry and anonymity do not constrain the dimensionality of the set of feasible and individually rational payoffs. Indeed, this set is full-dimensional in all examples considered in this paper.

[7] We discuss the case with multiple commitment types in Section V.

[8] One could alternatively consider player $i$'s expected utility conditional on the event $\theta_i = R$. This would involve a little more notation while giving essentially the same results.

## II. Preliminaries

### A. *Lower Bound for Rational Players' Payoffs*

We first show that the bad type's action cannot perform much better than the rational type's equilibrium strategy against the equilibrium action distribution (Lemma 1). This *lower* bound for rational players' payoffs will later drive our anti-folk theorem, which in particular implies an *upper* bound for payoffs.

For $n \in \{0,1,\dots,N\}$, let $\mathcal{B}_n$ denote the event that the realized number of bad types is $n$, and let $p_n$ denote the probability of this event. Conditional on the event that a given player $i$ is rational ($\theta_i = R$), let $q_n$ denote the probability that $n$ out of the remaining $N-1$ players are bad. Since symmetry implies that

$$\Pr(\mathcal{B}_n \wedge \theta_i = R) = \frac{N-n}{N} p_n,$$

we see that $q_n$ is given by

$$q_n = \Pr(\mathcal{B}_n | \theta_i = R) = \frac{\Pr(\mathcal{B}_n \wedge \theta_i = R)}{\Pr(\theta_i = R)} = \frac{N-n}{N} \frac{p_n}{1-z}.$$

We also let $q_N = 0$ by convention. Next, 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

$$q_n^- = q_{n-1} \quad \text{for } n \in \{1,\dots,N\},$$

with $q_0^- = 0$ by convention. Note that, given the convention that $q_N = q_0^- = 0$, $q = (q_n)_{n=0}^N$ and $q^- = (q_n^-)_{n=0}^N$ are both probability distributions on $\{0,\dots,N\}$. Denote the total variation distance between these probability distributions by

$$\Delta_{q,q^-} = \max_{\mathcal{N} \subset \{0,\dots,N\}} \left| \sum_{n \in \mathcal{N}} (q_n - q_n^-) \right|. \tag{1}$$

Note that if a rational player deviates by playing *Always* $a^*$ instead of her equilibrium strategy and the realized number of bad types is $n-1$, then the population distribution of actions is the same as it would be if this player had not deviated and the realized number of bad types were $n$. Thus, from the perspective of a rational player (and assuming that the equilibrium strategy of rational players is something other than *Always* $a^*$), $q_n$ is the probability that $n$ players in the population play *Always* $a^*$ when she follows her equilibrium strategy, and $q_n^-$ is the probability that $n$ players in the population play *Always* $a^*$ when she deviates to *Always* $a^*$. The distance between the distributions $q$ and $q^-$, $\Delta_{q,q^-}$, is therefore a measure of the detectability of a deviation by a rational player from her equilibrium strategy to *Always* $a^*$.

Fix a symmetric Nash equilibrium $\sigma$. For $n \in \{0,1,\dots,N\}$, let $u_n$ denote the expected payoff of a random player in the population when there are $n$ bad types, given by

$$u_n = E\big[U_i(\theta) | \mathcal{B}_n\big].$$

Let $u_n^R$ denote a rational player's expected payoff when there are $n$ bad types, given by

$$u_n^R \;=\; E\big[U_i(\theta)|\theta_i \;=\; R, \mathcal{B}_n\big].$$

Let $u_n^B$ denote a bad player's expected payoff when there are $n$ bad types, given by

$$u_n^B \;=\; E\big[U_i(\theta)|\theta_i \;=\; B, \mathcal{B}_n\big].$$

Note that, for each $n$, we have

$$u_n \;=\; \frac{N-n}{N}u_n^R + \frac{n}{N}u_n^B.$$

Moreover,

$$U \;=\; \sum_{n=0}^{N} p_n u_n.$$

We let $u_0^B \;=\; 1$ by convention. This convention makes the following lemma, which gives the desired lower bound on rational players' payoffs, as strong as possible.

LEMMA 1: *For any anonymous game and any symmetric Nash equilibrium*, *the following bounds apply:*

(*i*) *Rational player payoff bound*: $\sum_{n=0}^{N-1} q_n u_n^R \;\geq\; \sum_{n=0}^{N-1} q_n u_n^B - \Delta_{q,q^-}$.

(*ii*) *Social welfare bound*: $U \;\geq\; \sum_{n=0}^{N}\; p_n u_n^B - \big(1-z\big)\Delta_{q,q^-}$.

PROOF:

In any Nash equilibrium, a rational player must prefer her equilibrium strategy to deviating to *Always* $a^*$. A rational player's equilibrium payoff is $\sum_{n=0}^{N-1} q_n u_n^R$. If a rational player instead plays *Always* $a^*$, then for each realized number of "true" bad types $n$, she receives the same payoff as that received in equilibrium by a bad type when the true number of bad types is $n+1$. Thus, her expected payoff from such a deviation is $\sum_{n=0}^{N-1} q_n u_{n+1}^B$. Now note that

$$\begin{aligned}
(2)\qquad \sum_{n=0}^{N-1} q_n u_{n+1}^B \;&=\; \sum_{n=0}^{N-1} q_n u_n^B + \sum_{n=0}^{N-1} q_n u_{n+1}^B - \sum_{n=0}^{N-1} q_n u_n^B \\
&=\; \sum_{n=0}^{N-1} q_n u_n^B + \sum_{n=1}^{N} q_n^- u_n^B - \sum_{n=0}^{N-1} q_n u_n^B \\
&=\; \sum_{n=0}^{N-1} q_n u_n^B + \sum_{n=0}^{N} q_n^- u_n^B - \sum_{n=0}^{N} q_n u_n^B \\
&=\; \sum_{n=0}^{N-1} q_n u_n^B - \sum_{n=0}^{N} \big(q_n - q_n^-\big) u_n^B \\
&\geq\; \sum_{n=0}^{N-1} q_n u_n^B - \Delta_{q,q^-}.
\end{aligned}$$

Here, the third equality follows because $q_0^- = q_N = 0$, and the final inequality follows because (recalling that $u_n^B \in [0,1]$)

$$\sum_{n=0}^{N}(q_n - q_n^-)\,u_n^B \;\leq\; \sum_{n:q_n \geq q_n^-}(q_n - q_n^-)\,u_n^B \;\leq\; \sum_{n:q_n \geq q_n^-}(q_n - q_n^-) \;=\; \Delta_{q,q^-}.$$

This establishes the rational player payoff bound.

To derive the social welfare bound, let $r_n = \Pr(\mathcal{B}_n|\theta_1 = B)$, and note that

$$p_n \;=\; \begin{cases} (1-z)\,q_0 & \text{if } n = 0 \\ (1-z)\,q_n + z r_n & \text{if } 1 \leq n \leq N-1 \\ z r_N & \text{if } n = N. \end{cases}$$

Putting this together with the definition of $U$ and the rational player payoff bound, we obtain

$$\begin{aligned} U \;&=\; (1-z)\sum_{n=0}^{N-1} q_n u_n^R + z\sum_{n=1}^{N} r_n u_n^B \\ &\geq\; (1-z)\left(\sum_{n=0}^{N-1} q_n u_n^B - \Delta_{q,q^-}\right) + z\sum_{n=0}^{N} r_n u_n^B \\ &=\; \left((1-z)\,q_0 + \left(\sum_{n=1}^{N-1}(1-z)\,q_n + z\,r_n\right) + z\,r_N\right) u_n^B - (1-z)\,\Delta_{q,q^-} \\ &=\; \sum_{n=0}^{N} p_n u_n^B - (1-z)\,\Delta_{q,q^-}. \blacksquare \end{aligned}$$

## B. *Smooth Type Distributions*

The payoff bounds established in Lemma 1 are most significant when $\Delta_{q,q^-}$ is small. We say that a sequence $(N,p)_N$ with $N \to \infty$ (where, for each $N$, $p$ is a symmetric prior on $\{R,B\}^N$) has a *smooth distribution of bad types* if

$$\lim_{N\to\infty} \Delta_{q,q^-} \;=\; 0.$$

Similarly, a sequence of games indexed by $N$, $(\Gamma)_N$, has a *smooth distribution of bad types* if this true of $(N,p)_N$. We now discuss when a sequence $(N,p)_N$ has a smooth distribution of bad types.

Suppose $p$ is log-concave: $p_n/p_{n-1} \geq p_{n+1}/p_n$ for all $n \in \{1,\ldots,N-1\}$.[9] Then the maximum in (1) is attained by a set $\mathcal{N}$ that takes a "threshold" form $\mathcal{N} = \{n^*,\ldots,N\}$ for some threshold $n^* \in \{0,\ldots,N\}$. This yields

$$\Delta_{q,q^-} \;=\; q_{n^*-1} \;=\; \frac{N-n^*+1}{N}\,\frac{p_{n^*-1}}{1-z}.$$

[9] See Bagnoli and Bergstrom (2005) for a survey of log-concave probability distributions, with many examples.

Therefore, when $p$ is log-concave, the sequence $(N,p)_N$ has an smooth distribution of bad types if and only if $\max_{n\leq N-1}((N-n)/N)\,p_n \to 0$.

This is a mild condition. For example, when the players' types $(\theta_i)_{i\in I}$ are independent, $p$ is log-concave,[10] and $\max_{n\leq N-1}((N-n)/N)\,p_n \to 0$ whenever $z$ remains bounded away from 0 and 1 as $N \to \infty$. More generally, a log-concave distribution of bad types is smooth if the probability that the number of bad types takes on any particular value $n$ converges to zero as $N \to \infty$.

For an example where the distribution of bad types is *not* smooth, consider independent types with $zN$ held constant at some $\bar{n} \in \mathbb{N}$ as $N \to \infty$, so the distribution of the number of bad types converges to a Poisson distribution with parameter $\bar{n}$. Then

$$\Delta_{q,q^-} = q_{\bar{n}} = \frac{N-\bar{n}}{N}\binom{N}{\bar{n}}\left(\frac{\bar{n}}{N}\right)^{\bar{n}}\left(\frac{N-\bar{n}}{N}\right)^{N-\bar{n}-1}.$$

For instance, if $\bar{n} = 1$ (i.e., on average, there is exactly one bad player in the population) then

$$\Delta_{q,q^-} = \left(\frac{N-1}{N}\right)^{N-1} \sim \frac{1}{e}.$$

Thus, Lemma 1 can provide a meaningful bound even if on average there is only a single bad player. If instead $\bar{n}$ is sufficiently large, then Stirling's approximation gives

$$\Delta_{q,q^-} = \frac{N-\bar{n}}{N}\binom{N}{\bar{n}}\left(\frac{\bar{n}}{N}\right)^{\bar{n}}\left(\frac{N-\bar{n}}{N}\right)^{N-\bar{n}-1} \sim \frac{1}{\sqrt{2\pi\bar{n}}}.$$

Thus, Lemma 1 can provide a tight bound even if the expected number of bad types stays finite as $N \to \infty$.

If $p$ is not log-concave, then $\Delta_{q,q^-}$ need not converge to 0 even if $\max_{n\leq N-1}((N-n)/N)\,p_n \to 0$. For example, $\Delta_{q,q^-} = 1$ if the number of bad types is known in advance to be even. In this (rather artificial) case, the conclusion of Lemma 1 is vacuous.

## III. Anti-Folk Theorem for Games with a Pairwise Dominant Action

We say that an action $a^* \in A$ is *pairwise dominant* if there exists a positive number $c > 0$ such that, for any action $a \neq a^*$, if player $i$ takes $a^*$, player $j$ takes $a$, and the remaining players take any actions $a_{-ij} \in A^{(N-2)}$, then player $i$'s payoff exceeds player $j$'s by at least $c$: that is,

$$u(a^*,a,a_{-ij}) - u(a,a^*,a_{-ij}) > c \qquad \text{for all } a\ (\neq a^*) \in A,\quad a_{-ij} \in A^{(N-2)}.$$

To interpret this definition, note that if the impact of a single opponent's action on a player's payoff is small, then $u(a,a^*,a_{-ij}) \approx u(a,a,a_{-ij})$, so the definition of

[10] Here $p_n = \binom{N}{n}\varepsilon^n(1-\varepsilon)^{N-n}$, and hence $p_n/p_{n-1} = ((N-n+1)/n)(\varepsilon/(1-\varepsilon))$, which is decreasing in $n$.

a pairwise dominant action reduces to that of a dominant action, with $c$ equal to the minimum payoff gain from taking $a^*$ rather than another action. For example, this equivalence holds in large-population anonymous random matching games, where $u(a_i, a_{-i}) = \big(1/(N-1)\big)\sum_{j \neq i} \hat{u}(a_i, a_j)$ for some function $\hat{u} : A^2 \to \mathbb{R}$ fixed independent of $N$. More generally, a dominant action is also pairwise dominant if it imposes a negative externality on other players, in that $u\big(a, a^*, a_{-ij}\big) \leq u\big(a, a, a_{-ij}\big)$ for all $a, a_{-ij}$; and a non-dominant action can be pairwise dominant only if it imposes a sufficiently large negative externality.

In this section, we assume that the action $a^*$ played by bad types is pairwise dominant.[11] Denote social welfare when everyone takes the pairwise dominant action by $U^* = u(a^*, \ldots, a^*) \in [0,1]$. We also let $b > 0$ denote the greatest impact on social welfare that can result from a player switching from $a^*$ to another action, given by

$$b = \sup_{a_i \in A, a_{-i} \in A^{(N-1)}} \left| \sum_{j=1}^{N} \big( u_j(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_N) - u_j(a_1, \ldots, a_{i-1}, a_i^*, a_{i+1}, \ldots, a_N) \big) \right|.$$

**Example 1** (**Prisoner's Dilemma with Anonymous Random Matching**)**:** Suppose that in each period players match in pairs (uniformly at random and independently across periods) to play the prisoner's dilemma stage game

(3)

| | $C$ | $D$ |
|---|---|---|
| $C$ | $\frac{1+L}{1+G+L}, \frac{1+L}{1+G+L}$ | $0, 1$ |
| $D$ | $1, 0$ | $\frac{L}{1+G+L}, \frac{L}{1+G+L}$ |

where $G, L > 0$. This is an anonymous $N$-player game, where a player's stage-game payoff when she takes action $a \in \{C, D\}$, $m$ out of her $(N-1)$ opponents take action $D$, and her remaining $(N-m-1)$ opponents take action $C$, is given by

$$\frac{N-m-1}{N-1} u(a, C) + \frac{m}{N-1} u(a, D).$$

In this game, it can be easily checked that action $D$ is pairwise dominant, with

$$c = \frac{1}{1+G+L}\Big( \min\{G, L\} + \frac{1}{N-1}\big(1 + \max\{G, L\}\big) \Big),$$

$$b = \frac{1 + |G - L|}{1+G+L}.$$

Our main result is the following anti-folk theorem for anonymous repeated games with a pairwise dominant action.

[11] We thus consider games that have a pairwise dominant action. Clearly, any game has at most one such action.

THEOREM 1: *For any anonymous repeated game* $\Gamma$ *with a pairwise dominant action, in any Nash equilibrium social welfare* $U$ *satisfies*

$$|U - U^*| \leq (1-z)b\frac{1+c}{c}\Delta_{q,q^-}. \tag{4}$$

*In particular, for any sequence* $(\Gamma)_N$ *of anonymous repeated games with a pairwise dominant action satisfying* $\liminf_{N\to\infty} c_N > 0$ *and* $\limsup_{N\to\infty} b_N < \infty$ *and a smooth distribution of bad types, and any corresponding sequence of Nash equilibrium social welfare levels* $(U)_N$*, we have*

$$\lim_{N\to\infty} |U_N - U_N^*| = 0. \tag{5}$$

For example, consider the repeated prisoner's dilemma with anonymous random matching where bad types always defect. If we fix the payoff parameters $G$ and $L$ and vary $N$ and $\delta$, then along any sequence with a smooth distribution of bad types, social welfare converges to the payoff from mutual defection, $L/(1+G+L)$. Crucially, this conclusion does not depend on how $\delta$ varies along the sequence: for instance, it applies even if $(1-\delta)N \to 0$.

To see the intuition for Theorem 1, note that as in Lemma 1, bad players' expected discounted equilibrium payoffs cannot be much greater than rational players'. However, since $a^*$ is pairwise dominant, bad players' payoffs exceed rational players' by at least $c$ multiplied by the expected discounted frequency with which rational players take actions other than $a^*$. Therefore, this frequency must be small: that is, rational players must almost always take $a^*$. Finally, when bad players always take $a^*$ and rational players almost always take $a^*$, social welfare is close to $U^*$.[12]

PROOF:

Fix a symmetric Nash equilibrium $\sigma$. For $n \in \{1,\ldots,N-1\}$, let $\gamma_n$ denote the "expected discounted frequency" with which a rational player takes an action other than $a^*$ when there are $n$ bad types, given by

$$\gamma_n = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\sum_{h_i^t}\Pr^{\sigma}(h_i^t|\theta_i = R, \mathcal{B}_n)\left(1-\sigma_{i,t}(h_i^t)[a^*]\right).$$

We note that

$$u_n^B \geq u_n^R + \gamma_n c \quad \text{for all } n \in \{1,\ldots,N-1\}. \tag{6}$$

This inequality holds because, in any period where a rational player takes an action other than $a^*$, a bad player's payoff exceeds her payoff by at least $c$. So if a rational player takes an action other than $a^*$ in period $t$ with probability

[12] More precisely, we show that rational players take actions other than $a^*$ with frequency at most $((1+c)/c)\Delta_{q,q^-}$. Since the probability that a given player is rational is $1-z$ and the payoff impact of switching a single player's action from $a^*$ to another action is at most $b$, we see that $|U-U^*|$ is at most $(1-z)b((1+c)/c)\Delta_{q,q^-}$.

$$\gamma_{n,t} = \sum_{h_i^t} \Pr^{\sigma}(h_i^t|\theta_i = R, \mathcal{B}_n)\big(1 - \sigma_{i,t}(h_i^t)[a^*]\big),$$

a bad player's expected period-$t$ payoff exceeds hers by at least $\gamma_{n,t}c$, and taking a discounted sum over periods implies that a bad player's repeated game payoff exceeds a rational player's by at least $(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\gamma_{n,t}c = \gamma_n c$.

Combining the rational player payoff bound from Lemma 1 with (6), and recalling that $u_0^B = 1$ by convention, we obtain

$$\Delta_{q,q^-} \geq \sum_{n=0}^{N-1} q_n\big(u_n^B - u_n^R\big) \geq \sum_{n=1}^{N-1} q_n\big(u_n^B - u_n^R\big) \geq \sum_{n=1}^{N-1} q_n\gamma_n c.$$

Now define $\gamma = \sum_{n=0}^{N-1} q_n\gamma_n$. Since $q_0 = q_0 - q_0^- \leq \Delta_{q,q'}$, we have

$$\gamma = q_0\gamma_0 + \sum_{n=1}^{N-1} q_n\gamma_n \leq \Delta_{q,q^-} + \frac{1}{c}\Delta_{q,q^-} = \frac{1+c}{c}\Delta_{q,q^-}.$$

Finally, the ex ante probability that a given player takes an action other than $a^*$ in period $t$ equals $(1-z)\sum_{n=0}^{N-1} q_n\gamma_{n,t}$. Hence, (per-capita) expected social welfare in period $t$ differs from $U^*$ by at most $(1-z)b\sum_{n=0}^{N-1} q_n\gamma_{n,t}$. Therefore, ex ante expected welfare differs from $U^*$ at most by

$$(1-z)b(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\sum_{n=0}^{N-1} q_n\gamma_{n,t} = (1-z)b\gamma \leq (1-z)b\frac{1+c}{c}\Delta_{q,q^-}.$$

This yields (4), and taking $\Delta_{q,q^-} \to 0$ yields (5). ∎

Let us clarify the difference between dominant and pairwise dominant actions under the conditions on payoffs required by Theorem 1. If an action $a^*$ is strictly dominant, the minimum payoff gain from taking $a^*$ rather than another action is bounded away from 0 as $N \to \infty$, and the total externality $b$ imposed by switching from $a^*$ to another action is bounded as $N \to \infty$, then $a^*$ is pairwise dominant for sufficiently large $N$, and thus Theorem 1 implies that $a^*$ is almost always played. But the converse is false: even when $N \to \infty$ while $c$ and $b$ remain bounded (so the payoff conditions of Theorem 1 are satisfied), a pairwise dominant action does not need to be dominant. Therefore, Theorem 1 applies in some games without a strictly dominant action.

For example, suppose that $A = \{a^0, a^1, a^*\}$ and payoffs are given by

$$u(a_i, a_{-i}) = \begin{cases} 1/2 & \text{if } a_i = a^* \\ 1 & \text{if } a_i = a^1 \text{ and } a_j = a^0 \text{ for all } j \neq i \\ 0 & \text{otherwise.} \end{cases}$$

Note that $a^*$ is pairwise dominant with $c = 1/2$, since whenever some player takes $a^*$ and another player takes a different action, the first player's payoff is $1/2$ and the second player's payoff is 0; however, $a^*$ is not dominant, because $a_i = a^1$ is the unique best response when $a_j = a^0$ for all $j \neq i$. Moreover, $b$ is also equal to $1/2$: switching a player's action from $a^*$ to $a^1$ changes her own payoff by $1/2$ without affecting anyone else's payoff; and switching her action from $a^*$ to $a^0$ decreases her own payoff by $1/2$, while either affecting no one else's payoff or

increasing a single other player's payoff by 1. Thus, assuming a smooth distribution of bad types, Theorem 1 implies that $a^*$ is almost always played.

However, when $N \to \infty$ a pairwise dominant action $a^*$ does satisfy the weaker condition that, for any other action $a$, $a^*$ is a strictly better-response than $a$ against any mixture of $a$ and $a^*$ actions.

PROPOSITION 1: *Fix a sequence of stage games* $(N,A,u)_N$ *with a pairwise dominant action* $a^*$ *satisfying* $\liminf_{N\to\infty} c_N > 0$ *and* $\limsup_{N\to\infty} b_N < \infty$. *There exists* $\bar{N}$ *such that, for all* $N > \bar{N}$, *all* $a \in A$, *and all* $M \in \{0,\ldots,N-1\}$, *we have*

$$u\left(a^*,(a)^M,(a^*)^{N-M-1}\right) > u\left(a,(a)^M,(a^*)^{N-M-1}\right),$$

*where* $(a)^M$ *is the vector of M a terms and* $(a^*)^{N-M-1}$ *is the vector of* $(N-M-1)$ $a^*$ *terms.*

PROOF:

The proof follows easily from the definitions of $c$ and $b$ and is thus omitted.

We also note that Theorem 1 generalizes to stochastic games where the payoff function $u$ depends on the profile of players' histories $h^t = \left(h_i^t\right)_{i\in I}$. Specifically, letting $u^t(a;h^t)$ denote the period-$t$ stage-game payoff at action profile $a$ and history profile $h^t$, assume that

$$(7)\quad u^t\left(a^*,a,a_{-ij};h^t\right) - u^t\left(a,a^*,a_{-ij};h^t\right) > c, \quad \forall a \left(\neq a^*\right) \in A,\ a_{-ij} \in A^{(N-2)},\ t,\ h^t,$$

and let

$$b = \sup_{a_i\in A, a_{-i}\in A^{(N-1)}, t, h^t} \left| \sum_{j=1}^{N} \left( u_j^t(a_1,\ldots,a_{i-1},a_i,a_{i+1},\ldots,a_N;h^t) - u_j^t(a_1,\ldots,a_{i-1},a_i^*,a_{i+1},\ldots,a_N;h^t) \right) \right|.$$

With these values for $c$ and $b$, Theorem 1 holds verbatim for stochastic games, by the same proof.

**Example 2** (**Non-Uniform Matching**)**:** Consider again the prisoner's dilemma with anonymous random matching and payoff matrix (*3*), but now allow the matching process to be non-uniform, non-stationary, and history-dependent. Specifically, assume that, given history profile $h^t = \left(h_i^t\right)_i$ at the beginning of period $t$, players $i$ and $j$ meet in period $t$ with probability $\psi_{ij}(h^t)$. Continue to assume that the matching process is symmetric across players ex ante: for any permutation $\pi$ on $I$, we have $\psi_{ij}\left(\left(h_k^t\right)_k\right) = \psi_{\pi(i)\pi(j)}\left(\left(h_{\pi(k)}^t\right)_k\right)$ for all $(i,j,t,h^t)$. In order for the action $D$ to remain pairwise dominant, we must assume that the meeting probabilities between any two players cannot be too unequal: otherwise, some player who takes $D$ could receive

a lower payoff than another player who takes $C$, if the latter player is much more likely to meet a partner who takes $C$. In particular, letting

$$R := \frac{\sup_{i,j,t,h^t} \psi_{ij}(h^t)}{\inf_{i,j,t,h^t} \psi_{ij}(h^t)},$$

we assume that

$$R^2 < 1 + G.$$

This assumption implies that action $D$ remains pairwise dominant, so Theorem 1 holds with the appropriate choice of $c > 0$.[13]

(To see why $D$ is pairwise dominant whenever $R^2 < 1 + G$, suppose player $i$ takes $D$, player $j$ takes $C$, and fraction $\alpha_{-ij}$ of the remaining players take $C$. Note that, for any set of players $\mathcal{S} \subset I$, any $h^t$-measurable events $\mathcal{E}$ and $\mathcal{E}'$, and any $k \in I$, we have

$$1/R \leq \frac{\Pr(\mu_k(t) \in \mathcal{S} | \mathcal{E})}{\Pr(\mu_k(t) \in \mathcal{S} | \mathcal{E}')} \leq R.$$

Hence, by Bayes' rule, the probability that player $i$ meets an opponent who takes $C$ is at least $\alpha_{-ij}/R$, and the probability that player $j$ meets an opponent who takes $C$ is at most $\min\{R\alpha_{-ij}, 1\}$. So the difference between player $i$'s payoff and player $j$'s payoff is at least

$$\frac{\alpha_{-ij}}{R} + \left(1 - \frac{\alpha_{-ij}}{R}\right)\frac{L}{1+G+L} - \min\{R\alpha_{-ij}, 1\}\frac{1+L}{1+G+L}.$$

This expression is strictly positive whenever $R^2 < 1 + G$.[14])

## IV. General Games

The payoff bounds established in Lemma 1 can also be useful in games without a pairwise dominant action. As an example of such an analysis, in this section we derive an implication of Lemma 1 involving the concave closure of the payoff function, and use it to show that industry profits in linear Cournot oligopoly converge to zero as the number of firms increases.

Fixing a strategy profile $\sigma$, let $\boldsymbol{\alpha}_t(\sigma) \in \Delta(A^N)$ denote the resulting distribution over action profiles in period $t$. Let $\boldsymbol{\alpha}(\sigma) = (1-\delta)\sum_{t=1}^{\infty} \delta^{t-1} \boldsymbol{\alpha}_t(\sigma) \in \Delta(A^N)$. That is, for each action profile $\mathbf{a} \in A^N$, $\boldsymbol{\alpha}(\sigma)$ is the "discounted frequency" with which $\mathbf{a}$ is played under $\sigma$.

Next, let $U(\mathbf{a}) = (1/N)\sum_{i=1}^{N} u_i(\mathbf{a})$ denote social welfare at action profile $\mathbf{a} \in A^N$, and let $U(\boldsymbol{\alpha}) = \sum_{\mathbf{a} \in A^N} \boldsymbol{\alpha}[\mathbf{a}] U(\mathbf{a})$ denote expected social welfare at action profile distribution $\boldsymbol{\alpha} \in \Delta(A^N)$. Let $\bar{U} : \Delta(A^N) \to [0,1]$ denote the *concavification* of $U$: that

[13] And with the same value of $b$ as in the uniform random matching case: $b = (1 + |G - L|)/(1 + G + L)$.

[14] This follows from straightforward algebra, considering separately the cases where $R\alpha_{-ij} < 1$ and $R\alpha_{-ij} \geq 1$.

is, the smallest concave function $\bar{U}$ that satisfies $\bar{U}(\boldsymbol{\alpha}) \geq U(\boldsymbol{\alpha})$ for all $\boldsymbol{\alpha} \in \Delta(A^N)$. Finally, let $\underline{u}: \Delta(A^{N-1}) \to [0,1]$ denote the *convexification* of the function $u(a^*, \cdot): \Delta(A^{N-1}) \to [0,1]$ given by $u(a^*, \boldsymbol{\alpha}_{-i}) = \sum_{a_{-i} \in A^{(N-1)}} \boldsymbol{\alpha}_{-i}[a_{-i}] u(a^*, a_{-i})$: that is, the greatest convex function $\underline{u}$ that satisfies $\underline{u}(a^*, \boldsymbol{\alpha}_{-i}) \leq u(a^*, \boldsymbol{\alpha}_{-i})$ for all $\boldsymbol{\alpha}_{-i} \in \Delta(A^{N-1})$.

PROPOSITION 2: *For any anonymous game and any symmetric Nash equilibrium $\sigma$, we have*

$$\bar{U}\big(\boldsymbol{\alpha}(\sigma)\big) \geq \underline{u}\big(\boldsymbol{\alpha}_{-i}(\sigma)\big) - (1-z)\,\Delta_{q,q^-}.$$

Proposition 2 follows from the social welfare bound of Lemma 1, because (as the proof shows) $\bar{U}\big(\boldsymbol{\alpha}(\sigma)\big) \geq U$ and $\underline{u}\big(\boldsymbol{\alpha}(\sigma)\big) \leq \sum_{n=0}^{N} p_n u_n^B$.

PROOF:

Let $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t(\sigma)$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\sigma)$. We have

$$\begin{aligned}\bar{U}(\boldsymbol{\alpha}) = \bar{U}\Big((1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\boldsymbol{\alpha}_t\Big) &\geq (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\bar{U}(\boldsymbol{\alpha}_t)\\ &\geq (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}U(\boldsymbol{\alpha}_t) = U,\end{aligned}$$

where the first inequality follows because $\bar{U}$ is concave and the second follows because $\bar{U}$ is everywhere greater than $U$. Similarly, we have

$$\begin{aligned}\underline{u}(\boldsymbol{\alpha}_{-i}) = \underline{u}\Big(a^*, (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\boldsymbol{\alpha}_{-i,t}\Big) &\leq (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\underline{u}(a^*, \boldsymbol{\alpha}_{-i,t})\\ &\leq (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u(a^*, \boldsymbol{\alpha}_{-i,t}) \leq \sum_{n=0}^{N} p_n u_n^B,\end{aligned}$$

where the first inequality follows because $\underline{u}$ is convex, the second follows because $\underline{u}$ is everywhere less than $u(a^*, \cdot)$, and the third follows because of our convention that $u_0^B = 1$. The result follows from combining these inequalities with the social welfare bound of Lemma 1. ∎

**EXAMPLE 3** (**Linear Cournot Oligopoly**): Suppose that in every period each of $N$ firms produces quantity $a_i \geq 0$ and payoffs are given by $u_i(a_i, a_{-i}) = \max\{1 - \sum_{j=1}^{N} a_j, 0\}\, a_i$.[15] We assume that bad types always take the static Nash equilibrium action, so $a^* = 1/(N+1)$. (An interpretation is that bad types are not aware that the other firms are trying to collude, and thus expect the static equilib-

[15] We thus normalize marginal costs to zero, and assume that prices are bounded by zero to keep payoffs bounded. The assumption that the kink in the demand curve is located at marginal cost is not essential. Also, as our proofs have covered only finite games (although they extend to the case where $A$ is a compact metric space and $u$ is continuous), the set of feasible quantity levels may be taken to be finite.

rium to be played.) We also assume that the distribution of the number of bad types is smooth. Under these assumptions, expected industry profits $\sum_{i=1}^{N} U_i$ converge to zero along any sequence of Nash equilibria as $N \to \infty$ (regardless of how $\delta$ may vary with $N$, including the case where $(1-\delta)N \to 0$).[16]

To see this, first note that $u_n^B \in \left[0, 1/(N+1)\right]$ for each $n$. This implies that $\Delta_{q,q^-}$ can be replaced by $\Delta_{q,q^-}/(N+1)$ in the statements of Lemma 1 and Proposition 2.[17] Moreover, it is without loss to restrict attention to symmetric equilibria in which industry output $\sum_i a_i$ is bounded by 1 with probability 1.[18] Note that the proof of Proposition 2 involves only on-path action profiles, so the resulting payoff bound continues to apply if we restrict attention to output profiles where $\sum_i a_i \leq 1$ when defining $\bar{U}$ and $\underline{u}$. With this restriction, industry profits are concave in $\sum_i a_i$, and hence in $\mathbf{a}$. Therefore, $\bar{U}(\boldsymbol{\alpha}) = U(\boldsymbol{\alpha})$ for all $\boldsymbol{\alpha} \in \Delta(A^N)$. Similarly, $u(a^*, a_{-i}) = \left(1 - a^* - \sum_{j \neq i} a_j\right) a^*$, which is linear in $\sum_{j \neq i} a_j$, and hence in $a_{-i}$. Therefore, $\underline{u}(\boldsymbol{\alpha}_{-i}) = u(a^*, \boldsymbol{\alpha}_{-i})$ for all $\boldsymbol{\alpha}_{-i} \in \Delta(A^{N-1})$. Letting $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\sigma) \in \Delta(A^N)$, Proposition 2 now implies that

$$U(\boldsymbol{\alpha}) \geq u(a^*, \boldsymbol{\alpha}_{-i}) - \Delta_{q,q^-}/(N+1).$$

To complete the proof, let $\bar{a} = E^{\boldsymbol{\alpha}}\left[\sum_{i=1}^{N} a_i\right]$. Since $\left(1 - \sum_{i=1}^{N} a_i\right)\sum_{i=1}^{N} a_i$ is concave in $\sum_{i=1}^{N} a_i$, we have

$$U(\boldsymbol{\alpha}) \leq \frac{1}{N}(1-\bar{a})\bar{a}.$$

Similarly, since $E^{\boldsymbol{\alpha}}\left[\sum_{j \neq i} a_j\right] = ((N-1)/N)\bar{a}$ by symmetry, we have

$$u(a^*, \boldsymbol{\alpha}_{-i}) = \left(1 - \frac{N-1}{N}\bar{a} - \frac{1}{N+1}\right)\frac{1}{N+1}.$$

We conclude that

$$\left(1 - \frac{N-1}{N}\bar{a} - \frac{1}{N+1}\right)\frac{1}{N+1} - \frac{1}{N}(1-\bar{a})\bar{a} \leq \frac{1}{N+1}\Delta_{q,q^-}, \quad \text{or}$$

$$1 - \frac{N-1}{N}\bar{a} - \frac{1}{N+1} - \frac{N+1}{N}(1-\bar{a})\bar{a} \leq \Delta_{q,q^-}.$$

[16] In this example we consider total payoffs rather than per-capita payoffs, since with a fixed demand curve the maximum feasible per-firm profits go to zero as $N \to \infty$.

[17] In particular, the last step of the derivation of (2) uses the fact that $u_n^B \in [0,1]$ to conclude that $\sum_{n=0}^{N-1} q_n u_n^R \geq \sum_{n=0}^{N-1} q_n u_n^B - \Delta_{q,q^-}$; if the upper bound for $u_n^B$ is replaced by $1/(N+1)$, the corresponding conclusion is $\sum_{n=0}^{N-1} q_n u_n^R \geq \sum_{n=0}^{N-1} q_n u_n^B - \Delta_{q,q^-}/(N+1)$.

[18] This follows by considering the relaxed problem where the only available strategies are the rational-type equilibrium strategy and the bad-type strategy. In this relaxed problem, if industry output ever exceeds 1, the output of rational types can be reduced so that industry output exactly equals 1 without affecting anyone's payoff from following either strategy.

This inequality implies that if $\lim_{N\to\infty}\Delta_{q,q^-}=0$ then $\lim_{N\to\infty}\bar{a}=1$ as well. (Otherwise, the left-hand side would not converge to 0.) Finally, industry profits equal $NU(\boldsymbol{\alpha})\leq(1-\bar{a})\bar{a}$, which converges to zero when $\lim_{N\to\infty}\bar{a}=1$.

## V. Discussion

We conclude by assessing the prospects for extending our results to settings with multiple types, to mechanism design problems, and to non-anonymous games.

### A. *Multiple Commitment Types*

Our results extend straightforwardly to the case with one rational type and $K$ commitment types, each of whom is committed to an arbitrary repeated game strategy $\sigma^k$. In this case, let the vector $\mathbf{n}\in\{0,\ldots,N\}^K$ count the realized number of players of each commitment type, and let $q_{\mathbf{n}}$ be the probability of $\mathbf{n}$ conditional of the event that a given player is rational. For each commitment strategy $\sigma^k$, let $q_{\mathbf{n}}^{\sigma^k}$ denote the probability that the realized number of players of each commitment type differs from $\mathbf{n}$ in that one fewer player is committed to $\sigma^k$. Note that, if a single rational player deviates by playing $\sigma^k$, then $q_{\mathbf{n}}^{\sigma^k}$ is the probability that number of players who play each commitment strategy is given by $\mathbf{n}$. Let $\Delta_{q,q^{\sigma^k}}=\max_{\mathcal{N}\subset\{0,\ldots,N\}^K}\left|\sum_{\mathbf{n}\in\mathcal{N}}\left(q_{\mathbf{n}}-q_{\mathbf{n}}^{\sigma^k}\right)\right|$. A straightforward extension of Lemma 1 then implies that $\sum_{\mathbf{n}}q_{\mathbf{n}}u_{\mathbf{n}}^R\geq\sum_{\mathbf{n}}q_{\mathbf{n}}u_{\mathbf{n}}^{\sigma^k}-\Delta_{q,q^{\sigma^k}}$ for every commitment strategy $\sigma^k$, where the sum is taken over all vectors $\mathbf{n}$, and $u_{\mathbf{n}}^{\sigma^k}$ is the expected utility of the $\sigma^k$ commitment type conditional on vector $\mathbf{n}$. Finally, in games with a pairwise dominant action $a^*$, Theorem 1 holds with $\Delta_{q,q^{Always\,a^*}}$ in place of $\Delta_{q,q^-}$ (by the same proof), with the slight modification that an additional $+bz$ term must be added to the right-hand sides of equations (4) and (5) to reflect the fact that commitment types other than the *Always* $a^*$ type can take actions besides $a^*$.

Although our results extend to the case with multiple commitment types, we have focused on the case with a single commitment type because calculating $\Delta_{q,q^{\sigma^k}}$ (the total variation distance between two $K$-dimensional probability distributions) is much simpler when $K=1$. However, in two leading special cases computing $\Delta_{q,q^{\sigma^k}}$ for arbitrary $K$ is not much harder than it is when $K=1$. The first is when types are independent across players: in this case, to compute $\Delta_{q,q^{\sigma^k}}$ we need only keep track of the number of $\sigma^k$ commitment types, as in the $K=1$ case. The second is when types can be correlated but, conditional on the event that a given set of players are not rational, their specific commitment types are determined independently: in this case, to compute $\Delta_{q,q^{\sigma^k}}$ we need only keep track of the number of $\sigma^k$ commitment types and the number of rational types, and thus calculate the distance between two 2-dimensional distributions.

### B. *Multiple Rational Types*, *Incentive Compatibility*, *and Mechanism Design*

Lemma 1 can also be extended to settings with multiple rational types. Suppose there are $K$ rational types and (for simplicity) no commitment types. Let $\mathbf{n}\in\{0,\ldots,N\}^K$ count the realized number of players of each type. Fixing

a pair of types $\left(\theta_i, \hat{\theta}_i\right)$ and conditioning on the event that a given player has type $\theta_i$, let $q_{\mathbf{n}}^{\theta_i}$ be the probability of $\mathbf{n}$, let $q_{\mathbf{n}}^{\theta_i,\hat{\theta}_i}$ denote the probability that the realized number of players of each type differs from $\mathbf{n}$ in that one fewer player has type $\hat{\theta}_i$ and one more player has type $\theta_i$, and let $\Delta_{q^{\theta_i},q^{\theta_i,\hat{\theta}_i}} = \max_{\mathcal{N}\subset\{0,\ldots,N\}^K}\left|\sum_{\mathbf{n}\in\mathcal{N}}\left(q_{\mathbf{n}}^{\theta_i} - q_{\mathbf{n}}^{\theta_i,\hat{\theta}_i}\right)\right|$. Next, fixing a symmetric Nash equilibrium, let $u_{\mathbf{n}}^{\theta_i}$ denote the equilibrium expected utility of a type $\theta_i$ player conditional on the vector $\mathbf{n}$. Let $u_{\mathbf{n}}^{\theta_i,\hat{\theta}_i}$ denote the expected utility that a type $\theta_i$ player receives from following the equilibrium strategy of a type $\hat{\theta}_i$ player, conditional on $\mathbf{n}$. (Put differently, $u_{\mathbf{n}}^{\theta_i,\hat{\theta}_i}$ is the expected utility according to the type $\theta_i$ utility function of the equilibrium outcome obtained by type $\hat{\theta}_i$ conditional on $\mathbf{n}$.) Lemma 1 then implies that $\sum_{\mathbf{n}} q_{\mathbf{n}}^{\theta_i} u_{\mathbf{n}}^{\theta_i} \geq \sum_{\mathbf{n}} q_{\mathbf{n}}^{\theta_i} u_{\mathbf{n}}^{\theta_i,\hat{\theta}_i} - \Delta_{q^{\theta_i},q^{\theta_i,\hat{\theta}_i}}$.

Of course, we have simply shown that this inequality is one implication of *incentive compatibility*: the fact that a type $\theta_i$ player prefers to follow her own equilibrium strategy rather than the equilibrium strategy of type $\hat{\theta}_i$. But it may be a useful implication in some mechanism design problems. For example, this form of Lemma 1 can be used to give an elementary proof of a version of Mailath and Postlewaite's (1990) impossibility theorem for large-population public good provision, assuming that the type space is discrete and the prior is symmetric and satisfies $\lim_{N\to\infty}\Delta_{q^{\theta_i},q^{\theta_i,\hat{\theta}_i}} = 0$ for all $\left(\theta_i, \hat{\theta}_i\right)$ but is not necessarily independent. (In contrast, Mailath and Postlewaite's proof requires independence, as do all other proofs of their result that we aware of, such as that of al-Najjar and Smorodinsky 2000.[19])

The proof can be easily sketched: recall that a type $\theta_i$ player receives utility $\theta_i y - t_i$, where $y \in \{0,1\}$ is the public provision level and $t_i$ is the player's payment. Thus, $\sum_{\mathbf{n}} q_{\mathbf{n}}^{\theta_i}\left(u_{\mathbf{n}}^{\theta_i,\hat{\theta}_i} - u_{\mathbf{n}}^{\theta_i}\right)$ equals the difference between the expected payment of a type $\theta_i$ player and a type $\hat{\theta}_i$ player, according to the beliefs of a type $\theta_i$ player. Taking $\hat{\theta}_i$ to be the lowest type and normalizing this type to zero, individual rationality for type $\hat{\theta}_i$ implies that this type makes non-positive payments, so $\sum_{\mathbf{n}} q_{\mathbf{n}}^{\theta_i}\left(u_{\mathbf{n}}^{\theta_i,0} - u_{\mathbf{n}}^{\theta_i}\right)$ weakly exceeds the expected payment of a type $\theta_i$ player according to her own beliefs, $t(\theta_i)$. Thus, for each type $\theta_i$, Lemma 1 implies that $t(\theta_i) \leq \Delta_{q^{\theta_i},q^{\theta_i,0}}$.[20] Since $\lim_{N\to\infty}\Delta_{q^{\theta_i},q^{\theta_i,0}} = 0$ for each $\theta_i$, we have $\lim_{N\to\infty} t(\theta_i) = 0$ for each $\theta_i$. Taking an expectation over $\theta_i$ and applying the law of iterated expectations then implies that ex ante expected per-capita payments converge to zero as $N \to \infty$. Therefore, if the per-capita cost of providing the good is positive and constant in $N$ (as Mailath and Postlewaite assume), the probability that it is provided also converges to zero as $N \to \infty$.

[19] Independence (combined with full support) is much stronger than our condition that $\lim_{N\to\infty}\Delta_{q^{\theta_i},q^{\theta_i,\hat{\theta}_i}} = 0$ for all $\left(\theta_i, \hat{\theta}_i\right)$. For example, our condition is satisfied whenever types are conditionally independent with full support (given some common random variable). On the other hand, Mailath and Postlewaite allow continuous types and an asymmetric prior. We can allow continuous types if the prior satisfies an appropriate continuity condition, but symmetry is crucial for our approach. Also, while Mailath and Postlewaite's proof requires independence, in Appendix 2 of their paper they present an example with correlated types, the logic of which is similar to that of our result.

[20] Recall that Lemma 1 assumes bounded utilities, which in the current context implies bounded transfers. This explains why mechanisms like those of Crémer and McLean (1988) are not effective. Note that Mailath and Postlewaite's Appendix 2 example similarly assumes bounded transfers.

### C. *Non-Anonymous Games*

Our analysis depends critically on the anonymity assumption: a player's payoff is a function of her own action and the number of opponents taking each action. This assumption is what makes the distribution of the number of bad types so important. To apply our approach in more general games, one would need to find some statistic $\omega$ of the vector of players' types $(\theta_i)_{i\in I}$ with the properties that (1) a player's payoff depends only on her own action and type and $\omega$, and (2) the distribution of $\omega$ is not very responsive to a change in a single player's type. In anonymous games, $\omega$ is the number of players of each type. In games with multiple populations (e.g., buyers and sellers) where players are anonymous within each population (as in the "semi-anonymous" games studied by Kalai 2004), $\omega$ could be taken to be the number of players of each type in each population. Whether there are other interesting classes of games where such a statistic can be found is an open question.

An important example to which our results do *not* directly extend is the repeated prisoner's dilemma with non-anonymous random matching (where players observe their partners' identities before taking actions). In this game, for any $N$ and any distribution of the number of bad types, it is straightforward to support cooperation among rational types when $\delta$ is sufficiently high: simply prescribe "bilateral grim trigger strategies," where each player views herself as playing a separate 2-player repeated game with each opponent and plays grim trigger in all of them.

However, some trace of the negative conclusion of Theorem 1 does survive in non-anonymous random matching games. Bilateral grim trigger strategies are robust to allowing bad types but support cooperation only if $(1-\delta)N \to 0$, and are thus ineffective in very large populations. In contrast, Kandori (1992) and Ellison (1994) showed that contagion strategies support cooperation whenever $(1-\delta)\log N \to 0$; however, such strategies are not robust to bad types. In a companion paper (Sugaya and Wolitzky 2020), we generalize this observation to show that introducing bad types logarithmically decreases the maximum population size for which cooperation is sustainable in the repeated prisoner's dilemma with non-anonymous random matching.

## REFERENCES

**Al-Najjar, Nabil I., and Rann Smorodinsky.** 2000. "Pivotal Players and the Characterization of Influence." *Journal of Economic Theory* 92 (2): 318–42.

**Al-Najjar, Nabil I., and Rann Smorodinsky.** 2001. "Large Nonanonymous Repeated Games." *Games and Economic Behavior* 37 (1): 26–39.

**Awaya, Yu, and Vijay Krishna.** 2016. "On Communication and Collusion." *American Economic Review* 106 (2): 285–315.

**Awaya, Yu, and Vijay Krishna.** 2019. "Communication and Cooperation in Repeated Games." *Theoretical Economics* 14 (2): 513–53.

**Bagnoli, Mark, and Ted Bergstrom.** 2005. "Log-Concave Probability and Its Applications." *Economic Theory* 26 (2): 445–69.

**Crémer, Jacques, and Richard P. McLean.** 1988. "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions." *Econometrica* 56 (6): 1247–57.

**Deb, Joyee, Takuo Sugaya, and Alexander Wolitzky.** 2020. "The Folk Theorem in Repeated Games with Anonymous Random Matching." *Econometrica* 88 (3): 917–64.

**Ellison, Glenn.** 1994. "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *Review of Economic Studies* 61 (3): 567–88.

**Fudenberg, Drew, and David K. Levine.** 1989. "Reputation and Equilibrium Selection in Games with a Patient Player." *Econometrica* 57 (4): 759–78.

**Fudenberg, Drew, David Levine, and Wolfgang Pesendorfer.** 1998. "When Are Nonanonymous Players Negligible?" *Journal of Economic Theory* 79 (1): 46–71.

**Green, Edward J.** 1980. "Noncooperative Price Taking in Large Dynamic Markets." *Journal of Economic Theory* 22 (2): 155–82.

**Kalai, Ehud.** 2004. "Large Robust Games." *Econometrica* 72 (6): 1631–65.

**Kandori, Michihiro.** 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59 (1): 63–80.

**Levine, David K., and Wolfgang Pesendorfer.** 1995. "When Are Agents Negligible?" *American Economic Review* 85 (5): 1160–70.

**Mailath, George J., and Andrew Postlewaite.** 1990. "Asymmetric Information Bargaining Problems with Many Agents." *Review of Economic Studies* 57 (3): 351–67.

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

**Pai, Mallesh M., Aaron Roth, and Jonathan Ullman.** 2016. "An Antifolk Theorem for Large Repeated Games." *ACM Transactions on Economics and Computation* 5 (2): 10.

**Sabourian, Hamid.** 1990. "Anonymous Repeated Games with a Large Number of Players and Random Outcomes." *Journal of Economic Theory* 51 (1): 92–110.

**Sugaya, Takuo, and Alexander Wolitzky.** 2020. "Communication and Community Enforcement." Unpublished.