Ostracism and Forgiveness
Online Appendix

S. Nageeb Ali and David A. Miller

February 26, 2016

This Supplementary Appendix contains several results for both permanent and temporary ostracism. Appendix B contains all results for permanent ostracism not in the main paper, namely:

- Section B.1 proves that permanent ostracism is efficient when communication is mechanical. Corollary 1 describes the best bilateral enforcement equilibrium.
- Section B.2 continues the proof of Theorem 1, accounting for mixed strategies at the stake-proposal stage, and behavior that is contigent on who communicates first.
- Section B.3 contains all our results pertaining to discrete time, including an example that illustrates how permanent ostracism can outperform bilateral enforcement, a general bound that applies across permanent ostracism equilibria in this setting, and a result indicating that the gain from permanent ostracism over bilateral enforcement vanishes with the period length.
- Section B.4 describes the possibilities that arise when communication is simultaneous and for which selection of equilibrium our negative result applies.
- Section B.5 describes our extension to general games and the extent to which "rewards for whistleblowers" can improve the outlook for permanent ostracism.

Appendix C contains all of our results for temporary ostracism not in the main paper, namely:

- Section C.1 proves that if forgiveness of all players were synchronized, then temporary ostracism could do no better than bilateral enforcement.
- Section C.2 considers the optimal rate of forgiveness for 4 players.

- Section C.3 constructs a temporary ostracism equilibrium for the case of 3 players that maximizes cooperation among all mutual effort equilibrium in which the distribution of stakes on the equilibrium path is stationary.

B Permanent ostracism

B.1 Permanent ostracism with mechanical communication

Under mechanical communication, consider a permanent ostracism equilibrium such that when partners i and j meet they first mechanically exchange information (each can send only that message which receives her entire history). A player is deemed guilty if he has ever deviated in any way. If given their pooled information they know that both of them are innocent and $n-\ell$ other players are guilty, then they both propose stakes $\overline{\phi}_\ell$ that solve

$$T(\phi) = \phi + (\ell - 1)\int_0^\infty e^{-rt}\lambda\phi\, dt; \tag{B.1}$$

and then they work at those stakes. Innocent players announce zero stakes and shirk with guilty players.

Proposition 1. *In any mutual effort equilibrium, no player earns an expected payoff greater than $\frac{\lambda}{r}\overline{\phi}_n$ in any relationship, regardless of whether communication is mechanical, evidentiary, or cheap. Under mechanical communication, there exists a permanent ostracism equilibrium that attains this bound for all players in all relationships.*

Proof. First we establish that a strategy profile satisfying the description above is an equilibrium when communication is mechanical. By construction innocent partners are indifferent between working and shirking. Since they always face stakes of zero, guilty players and their partners are also indifferent between working and shirking. No player can ever do strictly better by announcing any other stakes, since doing so would incur guilt. Thus, this strategy profile is an equilibrium.

To establish that this equilibrium is strongly efficient among the class of mutual effort equilibria requires comparison to those in which punishments are not permanent ostracism, and stakes depend on history in other ways.

Step 1 First we argue that for any mutual effort equilibrium, there exists an equilibrium with the same on path behavior in which once any player deviates from the equilibrium path, the off path behavior is that of permanent ostracism. Since permanent ostracism attains a deviating player's minmax payoff, if incentive conditions are satisfied with any other punishment, then they

remain satisfied when a player is punished by being permanently ostracized. So it suffices to restrict attention to equilibria in which the off path behavior coincides with the equilibrium defined above.

Step 2 By Step 1, it suffices to establish that no permanent ostracism equilibrium supports cooperation at higher stakes than $\overline{\phi}_n$. In principle, stakes may be asymmetric across partnerships and history-dependent on the equilibrium path. Take any such equilibrium, and let $\phi_{ij}(h)$ denote the stakes that partners i and j would set if they meet at equilibrium-path history h. Notice that the payoff from working at history h is increasing in $\phi_{ik}(h')$ for every equilibrium path history h' that follows h. Let $\phi = \sup_{ij,h} \phi_{ij}(h)$: because stakes are uniformly bounded, $\phi < \infty$. For every equilibrium path history h^t and every player i, the continuation payoff after working is at most $n\frac{\lambda}{r}\phi$. Since there is some history along which $\phi_{ij}(h)$ is arbitrarily close to ϕ, it follows that

$$\frac{T(\phi)}{\phi} \leq 1 + (n-1)\frac{\lambda}{r} = \frac{T(\overline{\phi}_n)}{\overline{\phi}_n}. \tag{B.2}$$

Our assumptions on T imply that $\phi \leq \overline{\phi}_n$, so no mutual effort equilibrium supports stakes greater than $\overline{\phi}_n$ at any history. □

Corollary 1. *Since $\underline{\phi} = \overline{\phi}_2$, $\frac{\lambda}{r}\underline{\phi}$ is the highest payoff attainable from each relationship in any bilateral mutual effort equilibrium.*

B.2 Permanent ostracism

Proof of Theorem 1, continued. Here we prove that in every permanent ostracism equilibrium, each player's expected equilibrium payoff never exceeds that of bilateral enforcement, even when the equilibrium strategy profile may call for the players to randomize their stakes proposals and condition on who spoke first in the communication stage.

Let $\mathbb{E}\left[\phi_{ij}^t \middle| m_i^t, m_j^t, i\right]$ denote the expected stakes that are selected when player i sends message m_i^t first and then player j sends message m_j^t. Consider a pair of private histories (h_i^t, h_j^t) such that when players i and j meet at time t, they are both innocent and in equilibrium they expect to work at stakes greater than bilateral at least when player j speaks first: $\mathbb{E}\left[\phi_{ij}^t \middle| h_i^t, h_j^t, j\right] > \underline{\phi}$. Consider a private history $\hat{h}_i^t$ that coincides with h_i^t except that every other player has shirked on player i after the last interaction in h_i^t. Suppose that player j communicates first and sends the message h_j^t. In a permanent ostracism equilibrium, player i deems player j innocent, and so is supposed to report $\hat{h}_i^t$ truthfully; then both partners should propose stakes no greater than $\underline{\phi}$ so they can cooperate with each other while permanently ostracizing all the other players. Consider a deviation for player i in which he reports h_i^t rather than $\hat{h}_i^t$, he makes his stakes proposal strategy as if his true

private history had been h_i^t, and chooses to shirk regardless of what stakes are selected. This deviation is strictly profitable if

$$\mathbb{E}\left[T(\phi_{ij}) \mid h_i^t, h_j^t, j\right] > T\left(\mathbb{E}\left[\phi_{ij} \mid h_i^t, h_j^t, j\right]\right) > T(\underline{\phi}) = \underline{\phi} + \frac{\lambda}{r}\underline{\phi},$$

where the first inequality is implied by Jensen's Inequality and the strict convexity of T, the second inequality is implied by T being strictly increasing and the assumption that $\mathbb{E}\left[\phi_{ij}^t \middle| h_i^t, h_j^t, j\right] > \underline{\phi}$, and the equality is by definition of $\underline{\phi}$. Because this deviation is strictly profitable, it must be that in equilibrium $\mathbb{E}\left[\phi_{ij} \middle| h_i^t, h_j^t, j\right] \leq \underline{\phi}$ at all history pairs (h_i^t, h_j^t), and therefore, the maximum payoff player i expects from interacting with player j is $\frac{\lambda}{r}\underline{\phi}$. □

B.3 Permanent ostracism in discrete time

In the discrete time game, players may interact at times $0, \xi, 2\xi, ...$, where $\xi > 0$ specifies the *period length*, and $\lambda > 0$ is a parameter that specifies the frequency of interaction. Let $G = \frac{n(n-1)}{2}$ be the number of links in society. In each period, society is either *inactive* with probability $e^{-G\lambda\xi}$, in which case no link is selected; or it is *active* with probability $1 - e^{-G\lambda\xi}$, in which case a single link is selected. Conditional on society being active, each link is selected with equal probability. Let $p_\xi \equiv \frac{1}{G}\left(1 - e^{-G\lambda\xi}\right)$ be the probability that a particular link is selected in a period, and let $\delta \equiv e^{-r\xi}$ be the per-period discount factor. The continuous time game is the limit of this discrete time game as $\xi \to 0$. A key feature common to both settings is that there is zero probability that any player will ever meet multiple partners simultaneously.

Let $\underline{\phi}(\xi)$ be the maximum stakes in a mutual effort equilibrium under bilateral enforcement; then $\underline{\phi}(\xi)$ is the solution that binds

$$T(\phi) \leq \phi + \frac{\delta p_\xi}{1-\delta}\phi.$$

Similarly (cf. B.1), let $\overline{\phi}_n(\xi)$ be the maximum stakes in a mutual effort equilibrium under mechanical communication; i.e., $\overline{\phi}_n(\xi)$ is the solution that binds

$$T(\phi) \leq \phi + (n-1)\frac{\delta p_\xi}{1-\delta}\phi. \tag{B.3}$$

First we show by example that in discrete time players can cooperate at levels higher than $\underline{\phi}(\xi)$ using history-contingent strategies. Afterward, we show that, nonetheless, cooperation converges to bilateral enforcement levels as $\xi \to 0$.

Example 1. Consider the triangle depicted in Figure 1 and a history-dependent stakes profile in

which, at their meeting on the path of play at time t, Ann and Bob work at stakes $\phi > \underline{\phi}(\xi)$ if one of them reveals an interaction with Carol at $t - \xi$ that exhibits no deviation; otherwise Ann and Bob work at stakes $\underline{\phi}(\xi)$. If she did in fact work with Carol at time $t - \xi$, Ann is willing to reveal truthfully and work with Bob at stakes ϕ if

$$T(\phi) + \frac{\delta p_\xi}{1 - \delta\left(1 - 2p_\xi\right)} T(\underline{\phi}(\xi)) \leq \phi + \frac{2\delta p_\xi}{1-\delta} \begin{pmatrix} \left(1 - \delta(1 - 3p_\xi)\right)\phi \\ + \delta(1 - 3\delta p_\xi)\underline{\phi}(\xi) \end{pmatrix}.$$

The left-hand side includes Ann's payoff from shirking on Bob today and shirking on Carol in the future if she meets Carol before Bob does. Notice that when Ann shirks on Carol in the future, she does so at stakes $\underline{\phi}(\xi)$ because she cannot reveal an "on-path" interaction in the previous period. The right-hand side describes his payoff from working today and his discounted payoff from working in the future, where he is averaging between the payoffs he gains from sample paths where there are interactions in consecutive periods and sample paths where interactions occur without an interaction in the preceding period.

For every $\xi > 0$, this inequality is slack at $\phi = \underline{\phi}(\xi)$, so Ann is willing to work at stakes strictly greater than $\underline{\phi}$. Off path communication incentives are also satisfied: if Ann shirks on Bob, and Bob subsequently meets Carol, Bob is indifferent between revealing and concealing the truth, since in either case he and Carol shall set stakes $\underline{\phi}(\xi)$. This permanent ostracism equilibrium can support cooperation at levels higher than bilateral enforcement.

Yet, as $\xi \to 0$, these gains disappear since equilibrium path stakes exceed $\underline{\phi}(\xi)$ only when there was cooperation in the preceding period. Because the likelihood of interactions in two successive periods vanishes, the payoffs from such an equilibrium collapse to bilateral enforcement.[1]

Lemma 2. *In every permanent ostracism equilibrium,* $\mathbb{E}[\phi_{ij}|h_i^t, h_j^t] \leq \underline{\phi}(\xi)$ *for any pair of reported histories* (h_i^t, h_j^t) *in which there is no interaction at or after* $t - (n-2)\xi$.

Proof. Suppose otherwise: consider a pair of messages (h_i^t, h_j^t) such that $\mathbb{E}[\phi_{ij}|h_i^t, h_j^t] > \underline{\phi}(\xi)$, and there is no interaction at or before $t-(n-2)\xi$. Let $\hat{h}_i^t$ be a history that is identical to h_i^t except that in the previous $(n-2)\xi$ periods, player i has met every player other than j, who has proceeded to shirk on player i. Suppose player j communicates h_j^t first. Once player i reveals history $\hat{h}_i^t$, the maximal stakes that the two can work at is $\underline{\phi}(\xi)$ resulting in an expected payoff of

$$\underline{\phi}(\xi) + \frac{\delta p_\xi}{1-\delta}\underline{\phi}(\xi).$$

[1] The payoff difference between this equilibrium and bilateral enforcement is $\frac{2\delta p_\xi}{1-\delta}\left(1 - \delta(1-3\delta p_\xi)\right)\left(\phi - \underline{\phi}(\xi)\right)$, which converges to zero as $\xi \to 0$.

Consider the expected payoff from a deviation in which player i reveals only h_i^t, chooses a proposal using the equilibrium strategy after histories (h_i^t, h_j^t), and chooses to shirk whatever stakes are selected:

$$\mathbb{E}\left[T(\phi_{ij}) \mid h_i^t, h_j^t\right] > T\big(\mathbb{E}[\phi_{ij}|h_i^t, h_j^t]\big) > T(\underline{\phi}(\xi)) = \underline{\phi}(\xi) + \frac{\delta p_\xi}{1-\delta}\underline{\phi}(\xi),$$

where the first two inequalities are implied by our assumptions on T and Jensen's Inequality, and the equality is by definition of $\underline{\phi}(\xi)$. Since the payoff from deviation exceeds that from truthful communication, the strategy profile is not an equilibrium. □

Theorem 3. *For every $\varepsilon > 0$, there exists $\overline{\xi} > 0$ such that for all discrete time games with period length $\xi < \overline{\xi}$, in every permanent ostracism equilibrium each player's expected continuation payoff for every on-path history (including that at time 0) is at most $\varepsilon + \frac{(n-1)p_\xi}{1-\delta}\underline{\phi}(\xi)$, where the latter is her payoff from private bilateral enforcement.*

Proof. We proceed by constructing a strategy profile $\hat{\sigma}$ whose payoffs bound those of any permanent ostracism equilibrium with strategic communication. We suppose that whenever an interaction happens, its timing (though not its outcome) is publicly observed by all players. We break time into blocks of length $(n-2)\xi$. In this profile, along the path of play, players cooperate at stakes $\underline{\phi}(\xi)$ when no interaction is observed in the previous or current block; and at stakes $\overline{\phi}_n(\xi)$ otherwise. Recall that the probability of there being no interaction in a block of length $(n-2)\xi$ can be written as $(1 - Gp_\xi)^{n-2}$.

We first argue that every permanent ostracism equilibrium with strategic communication has equilibrium path payoffs that are less than those of $\hat{\sigma}$. Since any stakes that satisfy the incentives for permanent ostracism also satisfy the effort incentive for mechanical communication (B.3) with slack, it follows that in any permanent ostracism equilibrium with strategic communication, $\mathbb{E}[\phi_{ij}|h_i^t, h_j^t] < \overline{\phi}_n(\xi)$ for every ij and every pair of messages (h_i^t, h_j^t). By Lemma 2, no permanent ostracism equilibrium with strategic communication can do better than $\hat{\sigma}$.

We approximate the payoffs for $\hat{\sigma}$ for small ξ by decomposing payoffs within each $(n-2)\xi$ block and ignoring errors from discounting that are $O(\xi)$. π_H denotes the continuation payoff at the start of a block when there was an interaction in the previous block, and π_L when there was

no interaction. Then

$$\pi_L = \underbrace{(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\pi_L}_{\text{No interaction in this block}} + \sum_{k=1}^{n-2}\underbrace{\binom{n-2}{k}(Gp_\xi)^k(1-Gp_\xi)^{n-2-k}\left(\frac{n-1}{G}\big(\underline{\phi}(\xi)+(k-1)\overline{\phi}(\xi)\big)+e^{-r\xi(n-2)}\pi_H\right)}_{k\text{ interactions in this block}}+O(\xi),$$

where the first term is the expected payoff from there being no interactions in this block, the second term is the payoff from there being k interactions in this block, and the third term are discounting errors. The particular term $\underline{\phi}(\xi)+(k-1)\overline{\phi}(\xi)$ is the average level of cooperation when there are k interactions in the block. Similarly, we derive

$$\pi_H = (1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\pi_L + \sum_{k=1}^{n-2}\binom{n-2}{k}(Gp_\xi)^k(1-Gp_\xi)^{n-2-k}\left(\frac{n-1}{G}k\overline{\phi}_n(\xi)+e^{-r\xi(n-2)}\pi_H\right)+O(\xi),$$

where the middle term is different because each interaction in this block has stakes $\overline{\phi}_n(\xi)$. Subtracting the first equation from the second yields

$$\pi_H-\pi_L=\sum_{k=1}^{n-2}\binom{n-2}{k}(Gp_\xi)^k(1-Gp_\xi)^{n-2-k}\frac{n-1}{G}\big(\overline{\phi}_n(\xi)-\underline{\phi}(\xi)\big)+O(\xi).$$

Substituting the above expression into that for π_H and re-arranging yields:

$$\pi_H=\sum_{k=1}^{n-2}\binom{n-2}{k}\frac{(Gp_\xi)^k(1-Gp_\xi)^{n-2-k}}{1-e^{-r\xi(n-2)}}\frac{(n-1)}{G}\left(\begin{array}{l}\underline{\phi}(\xi)(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\\ +\overline{\phi}(\xi)\big(k-(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\big)\end{array}\right)+O(\xi).$$

Notice that $(Gp_\xi)^k=(1-e^{-G\lambda\xi})^k$ is $O(\xi^k)$ as $\xi\to 0$. Therefore $\frac{(Gp_\xi)^k(1-Gp_\xi)^{n-2-k}}{1-e^{-r\xi(n-2)}}\to\frac{G\lambda}{r(n-2)}$ for $k=1$ as $\xi\to 0$, and for $k\geq 2$ is $O(\xi^{k-1})$. Since $\overline{\phi}(\xi)$ converges, now we can write, more simply,

$$\pi_H=\frac{(n-1)\lambda}{r}\left(\underline{\phi}(\xi)(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}+\overline{\phi}(\xi)\big(1-(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\big)\right)+O(\xi).$$

Since $(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\to 1$ while $1-(1-Gp_\xi)^{n-2}e^{-r\xi(n-2)}\to 0$ as $\xi\to 0$, we conclude that $\pi_H\to\frac{(n-1)\lambda}{r}\underline{\phi}(0)$ as $\xi\to 0$. Therefore, for every $\varepsilon>0$, there exists $\overline{\xi}$ such that if $\xi<\overline{\xi}$, π_H is not more than ε greater than $\frac{(n-1)p_\xi}{1-\delta}\underline{\phi}(\xi)$, the payoff from private bilateral enforcement. □

B.4 Simultaneous communication in each interaction

Our results thus far relied on a communication protocol in which partners speak sequentially in each interaction. That protocol allows us to study ex post incentive constraints for at least one partner in each interaction. If instead players communicate simultaneously, a player's belief about what his partner already knows affects his incentives to reveal his information. Given this uncertainty, the freedom under weak perfect Bayesian equilibrium to construct arbitrary beliefs off the equilibrium path can be exploited to generate incentives to communicate in permanent ostracism equilibria. We illustrate how using two examples:

1. In Figure 1, suppose that when Ann shirks on Bob, Bob assigns high probability to Ann having shirked on Carol in the past. Consider a strategy profile in which if both parties report simultaneously that Ann is guilty, they work perpetually at stakes $\underline{\phi}$; but if only one party reports on it, then they work at small stakes $\eta > 0$ thereafter.
2. Consider a larger population, and suppose as in the history used in the proofs of Theorem 1 and Lemma 2, every player has shirked on player i since the last time players i and j met. Suppose that player i believes with high probability that some of these players have shirked on player j. Consider a strategy profile in which if players i and j commonly know that they are the only ones who are innocent, they work and set stakes $\underline{\phi}$; but commonly know that someone is guilty without commonly knowing that everyone else is guilty, they set stakes $\eta > 0$.

Characterizing the set of equilibria generated by this potentially rich set of first-order and second-order off-path beliefs is beyond our scope here. Instead, by imposing two natural selection criteria we show that obtaining payoffs above bilateral enforcement levels requires equilibria with implausible properties. These selection criteria are an adaptation of bilateral rationality from Ghosh and Ray (1996) and a richness condition that rules out certain unreasonable off-path beliefs.

Definition 2. *A permanent ostracism equilibrium is* ***bilaterally rational*** *if for histories* (h_i^t, h_j^t)*, if players i and j are innocent then they work at stakes* $\phi_{ij}(h_i^t, h_j^t) \geq \underline{\phi}$*.*

Bilateral rationality precludes a pair of partners from working at stakes strictly below $\underline{\phi}$ when each deems the other to be innocent. In the context of ostracism, bilateral rationality is motivated by the idea that innocent players should not be punished, where any continuation payoff below $\underline{\phi}$ within a relationship is classified as a punishment.

The second condition restricts off-path beliefs. Let $\hat{H}_j$ be the set of private histories for player j in which she believes that all players are innocent. In contrast, let $\tilde{H}_i^{\varepsilon}(j)$ be the set of private histories for player i in which the only innocent players are i and j, in the past $\varepsilon > 0$ interval of

real time there are $n-1$ interactions in which each other player $k \neq j$ has shirked for the first time on player i, and player i does not know of any interaction in which player j would have learned of any player being guilty.

Definition 3. *Off-path beliefs in a permanent ostracism equilibrium are **rich** if there exists $\underline{p} > 0$ such that for every sufficiently small $\varepsilon > 0$, for every pair ij, for every private history in $\tilde{H}_i^{\varepsilon}(j)$, player i believes that*

- *with probability at least $\underline{p}$, player j's private history is in $\hat{H}_j$;*
- *with probability $O(\varepsilon)$, player j has learned that player i interacted with someone in the past $\varepsilon > 0$ interval of real time.*

Richness implies the following: suppose within the last ε length of time, all players other than j have just shirked for the first time on player i. Player i must then believe that it is somewhat likely that player j has neither seen any defections nor has learned that player i has had any interactions during the last ε length of time. We view richness to be a natural condition on off-path beliefs, particularly because it applies when ε is sufficiently small.

Bilateral rationality and rich beliefs ensure that permanent ostracism does no better than bilateral enforcement.

Proposition 2. *With simultaneous, evidentiary communication in every interaction, in every bilaterally rational permanent ostracism equilibrium with rich beliefs, each player's expected equilibrium payoff never exceeds that of bilateral enforcement.*

Proof. Consider a bilaterally rational permanent ostracism equilibrium with rich off-path beliefs. Let h_i^t be a private history for player i on the equilibrium path in which he meets j at time t, and there are no interactions in the most recent $\varepsilon > 0$ interval of real time. Suppose towards a contradiction that player i's expected stakes conditional h_i^t are strictly greater than $\underline{\phi}$. Without loss of generality, consider a history $\hat{h}_i^t \in \tilde{H}_i^{\varepsilon}(j)$ that coincides with h_i^t, except that in the previous $\varepsilon > 0$ interval of real time, every other neighbor $k \neq j$ has shirked on player i. If player i reveals history $\hat{h}_i^t$ then he at best expects to work at stakes $\underline{\phi}$ in perpetuity. If instead he conceals the fact that he has been shirked on, then, because his off-path beliefs are rich, he expects:

- with probability at least $\underline{p}$, player j's private history is in $\hat{H}_j$, in which case they will set stakes $\phi_{ij}(h_i^t, h_j^t) > \underline{\phi}$;
- with probability $O(\varepsilon)$, player j knows that player i has been shirked on within the last ε interval of real time, in which case player j will ostracize player i for reporting a deviant message;

- otherwise, player j will report that some other players are guilty but still consider player i to be innocent, in which case they will cooperate at stakes at least $\underline{\phi}$ (by bilateral rationality).

Since the second case vanishes while the first does not as $\varepsilon \to 0$, there exists $\varepsilon > 0$ sufficiently small that it is a profitable deviation for player i to conceal that he has been shirked on, and then himself to shirk if they set stakes strictly greater than $\underline{\phi}$ (as in the first and possibly some of the third cases above). □

B.5 General games (Rewarding whistleblowers with asymmetric play)

In this section we generalize the environment to allow the stage game to differ across partnerships and be asymmetric. When they meet, players i and j play stage game $G_{\{ij\}}$, in which they simultaneously choose actions from A_{ij} and A_{ji}, respectively, and player i's utility is $u_{ij} : \mathcal{A}_{ij} \times \mathcal{A}_{ji} \to \mathbb{R}$ (where $\mathcal{A}_{ij}$ is the mixed extension of A_{ij}). Player i's minmax payoff in $G_{\{ij\}}$ is $u_{ij}^{\min}$.

There are no payoff interdependencies across relationships, and each player's payoff is the sum of her payoffs from her relationships. We focus on a class of games in which it is straightforward to generalize what permanent ostracism means.

Assumption 1. *For each player i, and in every game $G_{\{ij\}}$, there is a Nash equilibrium $\left(\underline{\alpha}_{ij}, \underline{\alpha}_{ji}\right) \in \mathcal{A}_{ij} \times \mathcal{A}_{ji}$ that attains each player's minmax in that game.*

Assumption 1 guarantees that in each game, each player finds it incentive compatible to maximally punish the other in their bilateral relationship without requiring intertemporal incentives. Apart from being satisfied in several moral hazard settings, Assumption 1 typifies those environments in which each player has the power to unilaterally sever a relationship, since that is a Nash equilibrium that attains the minmax within those games. For games in which Assumption 1 fails, our results pertain to equilibria in which guilty players are punished by Nash reversion.

First, we describe bilateral enforcement: in relationship ij, this is the set of subgame perfect equilibrium payoffs in the repeated play of $G_{\{ij\}}$ at rate λ_{ij}. Let $\overline{U}_{ij}$ denote the highest payoffs that player i can attain in any subgame perfect equilibrium of this game, starting at an $\{ij\}$ interaction.

For this environment, we define a *generalized permanent ostracism* strategy profile as one in which an innocent player continues to communicate and "cooperate" with other innocent players, but suspends communication and shifts to minmaxing anyone who shirks on her. "Cooperation" in this context can involve non-stationary behavior, but we impose the constraint that the stage game action profile two innocent partners should play at a given history should not depend on the order in which they were recognized to communicate.[2] Our definition does not restrict how

[2] This constraint was not needed in Theorem 1 because both players' stage game payoffs were tied to the same

an innocent player should interact with guilty players who have not deviated on her; it allows for strategy profiles in which she ostracizes them as well as strategy profiles in which she does not.

A behavioral strategy for player i is a function $\sigma_i = (\sigma_i^{\mathrm{M}}, \sigma_i^{\mathrm{A}})$, where σ_i^{M} specifies her reporting strategy and σ_i^{A} specifies her (mixed) action choice. Let $A(j, t, h_i^t, m_i^t, m_j^t)$ be the support of player i's equilibrium actions in $G_{\{ij\}}$ when meeting partner j at history h_i^t, after exchanging messages $\left(m_i^t, m_j^t\right)$ (in either order). Player i deems player j *innocent* in history h_i^t—i.e., $j \in \mathcal{I}_i(h_i^t)$—if there is no evidence in $\mathcal{E}_i(h_i^t)$ that player j has deviated from σ_j. By contrast, player i deems player j *guilty*—i.e., $j \in \mathcal{G}_i(\hat{h})$—if there exists an interaction $z^\tau \in \mathcal{E}_i(\hat{h})$ that involves players i and j in which a_j^τ is not in $A(i, h_i^\tau, m_j^\tau, m_i^\tau)$.

Definition 4. *An assessment σ is a **generalized permanent ostracism** assessment, if for every player i, every private history h_i^t, and every partner $j \neq i$, if i meets j at h_i^t and $i \in \mathcal{I}_i(h_i^t)$, then:*

1. *If i speaks first and $j \in \mathcal{I}\left(h_i^t\right)$, or if j speaks first and j's message m_j^t satisfies $\mathcal{E}_i^j(h_i^t, t) \subseteq m_j^t$ and $j \in \mathcal{I}\left(h_i^t \cup m_j^t\right)$, then she sends the truthful message h_i^t.*
2. *If $j \in \mathcal{I}\left(h_i^t \cup m_j^t\right)$, then i believes with probability 1 that j has not deviated, and plays action $\sigma_i^{\mathrm{A}}(j, h_i^t, h_i^t, m_j^t)$.*
3. *If $j \in \mathcal{G}_i(h_i^t)$, then she plays action $\underline{\alpha}_{ij}$.*

A generalized permanent ostracism profile guarantees that a player continues to communicate and "cooperate" with those who are innocent, but requires that she shift to minmaxing anyone who shirks on her. Our definition does not restrict how she should interact with players she learns are considered personally guilty by others. The following result also applies to generalized permanent ostracism equilibria in our basic model—i.e., in which shirking may occur on the equilibrium path.

Theorem 4. *Consider a generalized permanent ostracism equilibrium, σ. For any partnership ij, consider a mixed action profile α^* in $G_{\{ij\}}$. If α^* is played on the equilibrium path, then:*

$$\max_{a \in A_{ij}} u_{ij}(a, \alpha_{-i}^*) + \frac{\lambda_{ij}}{r} u_{ij}^{\min} \leq \overline{U}_{ij}. \tag{B.4}$$

stakes; here the two players could face very different incentives in the stage game. The constraint is tantamount to imposing ex post incentive constraints in the communication stage: in any meeting between innocent players, each partner must be willing to reveal truthfully regardless of what the other partner reveals. Relaxing this constraint would allow whichever player speaks second to be provided ex post incentives; the player who speaks first, if he knew of anyone who was guilty, could believe that his current partner was very likely already aware of that fact, relaxing his incentive constraint. Weak perfect Bayesian equilibrium allows such beliefs, but we don't find it plausible to impose them at all off-path histories. Moreover, even if we did allow for such beliefs, the bound on payoffs we identify in Theorem 4 would still apply at any history to whichever player speaks second.

If $G_{\{ij\}}$ is symmetric for every pair ij, and σ prescribes symmetric behavior on the equilibrium path, then player i's expected equilibrium payoff is bounded above by $\sum_{j\neq i} \frac{\lambda_{ij}}{r+\lambda_{ij}} \overline{U}_{ij}$.

Proof. First, observe that any bilateral enforcement equilibrium on link *ij* must satisfy (B.4) for every stage game action profile α^* that may be played on the equilibrium path.

The argument is similar to Theorem 1. Consider a history h_i^t on the equilibrium path, at which α^* is the prescribed stage game action profile; and another history $\hat{h}_i^t$ identical to h_i^t except that after the last interaction in $h_i^t \cup h_j^t$, player i has met each player $k \in \mathcal{N}\backslash\{i,j\}$, and $k \in \mathcal{G}_i(h_i^t)$. Suppose that player j reports h_j^t first. If player i truthfully communicates $\hat{h}_i^t$ to player j, they will continue with a bilateral enforcement equilibrium that satisfies (B.4). In contrast, communicating h_i^t and choosing a best response to α^*_{-i} guarantees a payoff of at least $\max_{a_i \in \mathcal{A}_{ij}} u_{ij}(a_i, \alpha_j^*) + \frac{\lambda_{ij}}{r} u_{ij}^{\min}$. Since we have imposed the constraint that α^* cannot depend on who was recognized to communicate first, the same incentive constraint applies even if player i is recognized to speak first.

We now prove the statement for a symmetric game $G_{\{ij\}}$ and an equilibrium in which the prescribed behavior α^* is symmetric on the equilibrium path. We claim that in the generalized permanent ostracism equilibrium σ, players are choosing on the equilibrium path only those action profiles α^* that satisfy

$$\frac{r+\lambda_{ij}}{r} u_{ij}(\alpha^*) \leq \overline{U}_{ij}. \tag{B.5}$$

We prove this claim by considering two cases that depend on the sign of

$$\frac{r+\lambda_{ij}}{r} u_{ij}(\alpha^*) - \max_{a \in A_{ij}} u_{ij}(a, \alpha^*_{-i}) - \frac{\lambda_{ij}}{r} u_{ij}^{\min}. \tag{B.6}$$

1. Suppose (B.6) is non-negative. Then in the repeated play of $G_{\{ij\}}$, there exists a bilateral equilibrium in which players i and j play α^* on the equilibrium path, and if either deviates, they revert to $(\underline{\alpha}_{ij}, \underline{\alpha}_{ji})$.[3] Since $\overline{U}_{ij}$ is the highest SPE payoff at the beginning of an interaction, the payoff from this SPE must be weakly lower resulting in the inequality in (B.5).

2. Suppose (B.6) is strictly negative. Then, (B.5) follows from (B.4) because:

$$\frac{r+\lambda_{ij}}{r} u_{ij}(\alpha^*) < \max_{a \in A_{ij}} u_{ij}(a, \alpha^*_{-i}) + \frac{\lambda_{ij}}{r} u_{ij}^{\min} \leq \overline{U}_{ij}.$$

Therefore, an upper bound for the expect payoff from interactions in $G_{\{ij\}}$ is $\frac{\lambda_{ij}}{r} \frac{r}{r+\lambda_{ij}} \overline{U}_{ij}$, resulting in the expression in Theorem 4. □

[3] Since the game and α^* is symmetric, neither player has an incentive to deviate.

C Temporary ostracism

C.1 Temporary ostracism with synchronized forgiveness

Proposition 3. *If the temporary ostracism construction used to prove Theorem 2 is modified to make all players' Poisson clocks perfectly correlated, then each player's expected equilibrium payoff never exceeds that of bilateral enforcement.*

Proof. If all Poisson clocks are perfectly correlated, then we must replace W (cf. A.1) with:

$$\hat{W}(\phi, \mu, \ell) \equiv \phi + (\ell - 1) \int_0^\infty e^{-rt} e^{-\mu t} \lambda \phi \, dt, \tag{C.1}$$

because the last term in (A.1) refers to payoffs that arise before a player's own forgiveness signal arrives but after her guilty partners are forgiven. Similarly, for the same reason we must replace S (cf. A.2) with:

$$\hat{S}(\phi, \mu, \ell) \equiv T(\phi) + (\ell - 2) \int_0^\infty e^{-rt} e^{-\mu t} e^{-\lambda t} e^{-\lambda t} \lambda T(\phi) dt. \tag{C.2}$$

A necessary condition for cooperation under temporary ostracism with synchronized forgiveness is that $\hat{W}(\phi, \mu, 2) \geq \hat{S}(\phi, \mu, 2)$, which is equivalent to:

$$\phi + \int_0^\infty e^{-rt} e^{-\mu t} \lambda \phi \, dt \geq T(\phi). \tag{C.3}$$

The stakes that bind this incentive constraint are the bilateral enforcement stakes when the discount rate is $r + \mu$. These stakes are maximized by setting $\mu = 0$; i.e., by making ostracism permanent, which by Theorem 1 is no better than bilateral enforcement. □

C.2 Optimal rate of forgiveness

Within the class of temporary ostracism equilibria we use to prove Theorem 2, the optimal equilibrium solves

$$\max_{\phi \geq 0, \mu \geq 0} \phi \quad \text{s.t. } W(\phi, \mu, \ell) \geq S(\phi, \mu, \ell) \; \forall \; \ell = 2, \ldots, n, \tag{C.4}$$

where we calculate that

$$W(\phi,\mu,\ell) = \phi + (\ell-1)\frac{\lambda}{r+\mu}\phi + (n-\ell)\frac{\lambda\mu}{(r+\mu)(r+2\mu)}\phi, \tag{C.5}$$

$$S(\phi,\mu,\ell) = T(\phi) + (\ell-2)\frac{\lambda}{r+2\lambda+\mu}T(\phi) + (n-\ell)\frac{\lambda\mu}{(r+2\lambda+\mu)(r+\lambda+2\mu)}T(\phi). \tag{C.6}$$

Let $n = 4$ for this example. Then the constraints for $\ell = 2, 3, 4$, respectively, rearrange to

$$\frac{T(\phi)}{\phi} \leq \frac{1+\frac{\lambda(r+4\mu)}{(r+\mu)(r+2\mu)}}{1+\frac{2\lambda\mu}{(r+2\lambda+\mu)(r+\lambda+2\mu)}} \tag{C.7}$$

$$\frac{T(\phi)}{\phi} \leq \frac{1+\frac{\lambda(2r+5\mu)}{(r+\mu)(r+2\mu)}}{1+\frac{\lambda(r+\lambda+3\mu)}{(r+2\lambda+\mu)(r+\lambda+2\mu)}} \tag{C.8}$$

$$\frac{T(\phi)}{\phi} \leq \frac{(r+2\lambda+\mu)(r+3\lambda+\mu)}{(r+\mu)(r+4\lambda+\mu)} \tag{C.9}$$

Observe that under our assumptions it suffices to choose μ to maximize $T(\phi)/\phi$ subject to these constraints. It can be shown that the global optimum subject to all three constraints is always the unique local optimum subject to only (C.7).

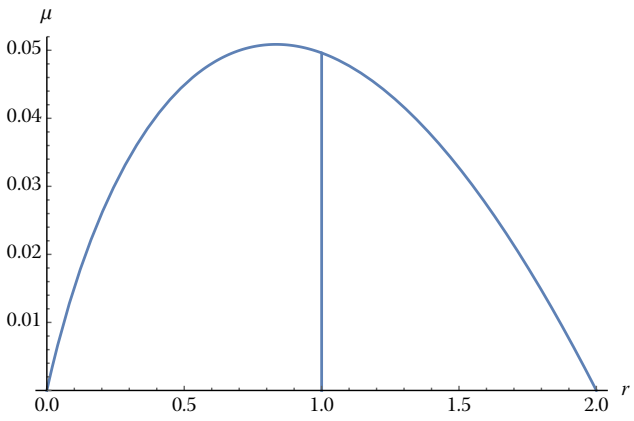

Figure 1. Optimal forgiveness rate μ vs. discount rate r, given $\lambda = 1$

This solution is the relevant root of a 6th degree polynomial, so unfortunately it does not have a closed form. However, it is relatively well behaved. To illustrate, let $\lambda = 1$; then the optimal rate of forgiveness, as a function of r, ranges from zero at the extremes of r to a maximum of about 0.05 at an interior discount rate, as shown in Figure 1. Not much forgiveness is needed to provide incentives when r is low, whereas not much forgiveness is incentive compatible when r is close to 2λ (cf. A.3).

C.3 Temporary Ostracism with Redemption

We prove in this section that for $n = 3$ players, there exists a temporary ostracism equilibrium that enforces the highest level of cooperation possible in any mutual equilibrium where the on-path stakes are stationary on the path of play. The construction is involved, and combines a number of features from our construction of contagion in Ali and Miller (2013) and temporary ostracism in Section III.

Define ϕ_C to bind the inequality

$$T(\phi) + \int_0^\infty e^{-rt} e^{-2\lambda t} \lambda T(\phi)\, dt \leq \phi + 2 \int_0^\infty e^{-rt} \lambda \phi\, dt. \tag{C.10}$$

Re-arranging, we see that

$$\frac{T(\phi_C)}{\phi_C} = \frac{(r+2\lambda)^2}{r(r+3\lambda)}, \tag{C.11}$$

which pins down the value of ϕ_C since $T(\phi)/\phi$ is strictly increasing in ϕ. We show in Ali and Miller (2013) that ϕ_C corresponds to the maximal stakes that can be supported by a contagion equilibrium, and indeed, any mutual effort equilibrium in which the distribution of stakes on the path of play is stationary.

We use *redemption payments* where a guilty player redeems herself by working while allowing her innocent partner to shirk. Let ϕ_R be defined implicitly through the equation

$$V(\phi_R) = \phi_C + 2 \int_0^\infty e^{-rt} \lambda \phi_C\, dt. \tag{C.12}$$

Notice that being forced to work while a partner shirks at stakes ϕ_R ensures that a guilty player's continuation value from resuming effort as an innocent player is 0.

In the interests of space, we offer a heuristic description of the strategy profile. Innocent players share their full history, propose stakes ϕ_C, and work with other innocent players. Broadly speaking, a single guilty player will try to shirk on as many innocent partners as possible and then "redeem herself" at stakes ϕ_R once she knows that all know she is guilty. Consequently, if guilty player i does not know if innocent player j knows that i is guilty, i would certainly conceal any interactions that indicate her guilt. Once a guilty player redeems herself, she is treated as innocent if she was the only guilty player. Once a player knows there are two guilty players, she shirks in every interaction. If there are two guilty players and one of them pays a redemption payment to the innocent player, then in the communication stage at the end of the interaction,

the innocent player reveals that they have transitioned to the phase with two guilty players. We verify all non-trivial incentives associated with such a scheme.

Given the value of ϕ_R, notice that once a guilty player i knows that both j and k know that i is guilty, her continuation value is 0. Therefore, (C.10) captures an innocent player's incentives to work on the equilibrium path.

We now turn to two relevant incentive constraints once player i has shirked on player j. Should player j communicate and cooperate with player k? The incentive constraint associated with this, if player j shirks before he has revealed to player k that i is guilty, is

$$T(\phi_C) \leq \phi_C + \frac{\lambda}{r}\phi_C + \underbrace{\int_0^\infty e^{-rt}e^{-2\lambda t}2\lambda\left(\frac{\lambda}{r}\phi_C\right)dt}_{\text{Resumption of Cooperation with } i} + \underbrace{\int_0^\infty e^{-rt}e^{-2\lambda t}\lambda T(\phi_R)\,dt}_{\text{Redemption Payment from } i}. \tag{C.13}$$

Re-arranging terms and using (C.10), we obtains the equivalent inequality

$$\frac{\lambda}{r+2\lambda}\phi_C \leq \frac{\lambda}{r+2\lambda}\left(T(\phi_C) + T(\phi_R)\right),$$

which is necessarily satisfied since $\phi_C < T(\phi_C)$.

We now tackle an additional incentive constraint, which is more challenging: once player j has shared information with player k that i has shirked and has the evidence needed for redemption, does he have any interest in further cooperating with player k? The relevant incentive constraint is

$$\begin{aligned} &T(\phi_C) + \int_0^\infty e^{-rt}e^{-2\lambda t}\lambda\left[T(\phi_R) + \int_\tau^\infty e^{-r(\tau-t)}e^{-2\lambda(\tau-t)}\lambda T(\phi_C)\,d\tau\right]dt \\ \leq &\phi_C + \frac{\lambda}{r}\phi_C + \int_0^\infty e^{-rt}e^{-2\lambda t}2\lambda\left(\frac{\lambda}{r}\phi_C\right)dt + \int_0^\infty e^{-rt}e^{-2\lambda t}\lambda T(\phi_R)\,dt. \end{aligned} \tag{C.14}$$

All terms involving $T(\phi_R)$ cancel out, and so simplifying (C.14), we obtain

$$T(\phi_C) + \left(\frac{\lambda}{r+2\lambda}\right)^2 T(\phi_C) \leq \phi_C + \frac{\lambda}{r}\phi_C\left(1 + \frac{2\lambda}{r+2\lambda}\right)$$

Re-writing $1 + \frac{2\lambda}{r+2\lambda}$ as $2 - \frac{r}{r+2\lambda}$, and using (C.10), we obtain

$$T(\phi_C) + \left(\frac{\lambda}{r+2\lambda}\right)^2 T(\phi_C) \leq T(\phi_C) + \left(\frac{\lambda}{r+2\lambda}\right)T(\phi_C) - \frac{r}{r+2\lambda}\frac{\lambda}{r}\phi_C.$$

We can re-write the above inequality as

$$\frac{T(\phi_C)}{\phi_C} \geq \frac{r+2\lambda}{r+\lambda},$$

which is satisfied by (C.11). Therefore, player j does indeed have an incentive to continue communicating and cooperating with player k.

References

S. Nageeb Ali and David A. Miller. Enforcing cooperation in networked societies. Working paper, 2013.

Parikshit Ghosh and Debraj Ray. Cooperation in community interaction without information flows. *Review of Economic Studies*, 63(3):491–519, 1996.