Ostracism and Forgiveness

S. Nageeb Ali and David A. Miller*

February 8, 2016

Abstract

Many communities rely upon *ostracism* to enforce cooperation: if an individual shirks in one relationship, her innocent neighbors share information about her guilt in order to shun her, while continuing to cooperate among themselves. However, a strategic victim may herself prefer to shirk, rather than report her victimization truthfully. If guilty players are to be *permanently* ostracized, then such deviations are so tempting that cooperation in any relationship is bounded by what the partners could obtain through bilateral enforcement. Ostracism can improve upon bilateral enforcement if tempered by *forgiveness*, through which guilty players are eventually readmitted to cooperative society.

* Ali: Pennsylvania State University. Miller: University of Michigan. We thank Dilip Abreu, Susan Athey, Matt Elliott, Ben Golub, Avner Greif, Matt Jackson, Navin Kartik, Asim Khwaja, Bart Lipman, Meg Meyer, Markus Möbius, Paul Niehaus, Larry Samuelson, Andy Skrzypacz, Joel Sobel, Adam Szeidl, Joel Watson, and Alex Wolitzky. We especially thank our Co-Editor, Debraj Ray, and six anonymous referees for helpful and constructive comments, which substantially improved our paper. Aislinn Bohren, Erik Lillethun, Ce Liu, and David Yilin Yang provided excellent research assistance. This work is financially supported by NSF grant SES–1127643. In addition, Ali gratefully acknowledges financial support from NSF grant NetSe–0905645, as well as financial support and hospitality from Harvard, Microsoft Research, and UCSD; Miller gratefully acknowledges financial support and hospitality from Microsoft Research.

1 Introduction

Cooperation in society relies on players being punished if they cheat their partners. One form of enforcement is bilateral, where only Bob punishes Ann if she cheats him. Community enforcement enhances cooperation by strengthening the punishment: Ann is more willing to cooperate with Bob if her other partners would also punish her for cheating him. Ostracism is a form of community enforcement in which a guilty player is punished by all her partners, while innocent players continue to cooperate with each other. However, the community faces an informational challenge in ostracizing guilty players: the entire community cannot directly observe how each individual behaves in each relationship. If Ann's past behavior is observed only by her past partners, how do her future partners learn whether they should punish her?

Gossip is a realistic way for communities to spread this information. If Ann shirks on Bob, he should tell others what she has done, and after hearing these complaints, others will punish her while continuing to cooperate among themselves. Numerous case studies of communities and markets document that word-of-mouth communication plays this role in enforcing medieval trade (Greif 2006); moderating common property disputes (Ostrom 1990; Ellickson 1991); and facilitating informal lending, contracting, and trade in developing economies (McMillan and Woodruff 1999; Banerjee and Duflo 2000). Although this role of communication in sustaining cooperation is emphasized across the social sciences and legal scholarship,[1] a fundamental question remains unanswered: is it actually in the interests of Ann's victims to report her deviation?

We find that punishments that *permanently ostracize* deviators deter innocent players from truthfully revealing who has deviated. This most stringent form of ostracism is self-defeating, and performs no better than bilateral enforcement. By contrast, tempering ostracism with *forgiveness* enables innocent players to look forward to cooperating in the future with currently guilty players, giving them a stronger motive to testify truthfully.

2 An Example

We use a simple example to illustrate the logic of these results. Consider the society of three players depicted in Figure 1. Each link embodies an ongoing partnership between two players that meets at random times according to an independent Poisson process of intensity λ. When a partnership meets, each partner decides how much costly effort to exert. Partners perfectly observe everything that occurs within their partnership, but outsiders observe neither the timing

[1] Dixit (2004) surveys the literature on informal governance in economics, including the importance of communication. Bowles and Gintis (2011) discuss the roles of communication and ostracism more broadly in the evolution of social norms, and Posner (1996) discusses them in the context of law and economics.

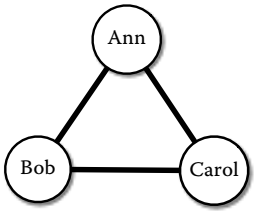

Figure 1

nor behavior within the partnership.

Ostracism requires that the innocent victims of a deviator be willing to report on her. Whenever a partnership meets, the partners have the opportunity to report on the behavior of others in the community, by communicating sequentially in random order before they choose their effort levels. Communication is "evidentiary" (Grossman 1981; Milgrom 1981), where players can reveal any subset of their past interactions: their messages contain nothing but the truth, but may not be the whole truth.

We model moral hazard using the variable stakes framework of Ghosh and Ray (1996): when a partnership meets, each player in that partnership simultaneously chooses an effort level $a \geq 0$, which comes at a personal cost of a^2 but generates a benefit of $a + a^2$ for her partner. Higher effort profiles are mutually beneficial, but increase the myopic motive to shirk, and therefore must be coupled with stronger incentives. This approach permits players to adjust the terms of their relationship based on who else is innocent or guilty, and faciltates a transparent comparison of equilibria for a fixed discount rate $r > 0$.

We first investigate two benchmarks to investigate how communication incentives impact cooperation: *bilateral enforcement*, in which players never use third-party punishments; and *permanent ostracism with mechanical communication*, in which players are mechanically forced to reveal the whole truth.

Bilateral enforcement: In bilateral enforcement, each partnership behaves independently. Consider the partnership between Ann and Bob, and a strategy profile in which each of them exerts effort a in their partnership if each has done so in the past; otherwise, each exerts zero effort. This behavior is an equilibrium if the one-time gain from shirking is less than the long-term gain from cooperation:

$$\underbrace{a + a^2}_{\substack{\text{Payoff from} \\ \text{shirking today}}} \leq \underbrace{a + \int_0^\infty e^{-rt} \lambda a \, dt.}_{\substack{\text{Payoff from working today} \\ \text{and in the future}}} \tag{1}$$

This incentive constraint is binding at effort level $\underline{a} = \lambda / r$.

Permanent ostracism with mechanical communication: Community enforcement enhances cooperation between Ann and Bob by leveraging their relationships with Carol. Suppose, hypothetically, that communication is mechanical: each player is constrained to reveal the full details of every prior interaction. Consider a strategy profile in which, on the path of play, each player exerts effort a whenever she meets a partner. If any player deviates on a partner, then each of them is mechanically constrained to report it to the third party. The two innocent players then permanently ostracize the deviator by exerting zero effort with that player; they continue to cooperate with each other at the bilateral enforcement effort $\underline{a}$, the highest effort supportable once the deviator is ostracized. Anticipating this, on the equilibrium path a player is motivated to work with a partner if

$$\underbrace{a + a^2}_{\substack{\text{Payoff from} \\ \text{shirking today}}} \leq \underbrace{a + 2\int_0^\infty e^{-rt}\lambda a \, dt.}_{\substack{\text{Payoff from working today} \\ \text{and in the future with both partners}}} \tag{2}$$

With mechanical communication, a player expects to forego cooperation with both partners after shirking on any one because she cannot conceal her deviation. This stronger punishment supports higher equilibrium path effort of $2\underline{a}$. Off-path incentive constraints are also satisfied. Thus, when players are forced to reveal all of their information, permanent ostracism supports more cooperation than bilateral enforcement.

Permanent ostracism with strategic communication: But what happens when individuals strategically choose which of their past interactions to reveal? At first glance, it might appear that even though a guilty player has every reason to conceal her own misdeeds, innocent players might have aligned interests in revealing and punishing the guilty. We show instead that innocent victims are tempted to conceal their victimization, and themselves to shirk on other innocent players. This strategic motive is so strong that it prevents permanent ostracism from improving upon bilateral enforcement.

Consider a permanent ostracism equilibrium, in which innocent players are supposed to cooperate and communicate truthfully with each other. When Ann contemplates shirking on Bob, she anticipates that Bob will tell Carol at their next interaction, upon which time she will be ostracized by both of them. Then Ann's only opportunity to gain from her relationship with Carol would be to meet Carol before Bob does, and conceal that she has shirked on Bob. Therefore

Ann's incentive constraint for cooperating with Bob at effort a is

$$\underbrace{a+a^2}_{\substack{\text{Payoff from shirking}\\ \text{on Bob today}}} + \underbrace{\int_0^\infty e^{-rt}e^{-2\lambda t}\lambda(a+a^2)\,dt}_{\substack{\text{Payoff from possibly shirking on Carol}\\ \text{in the future}}} \leq \underbrace{a+2\int_0^\infty e^{-rt}\lambda a\,dt}_{\substack{\text{Payoff from working today}\\ \text{and in the future with both partners}}}. \tag{3}$$

Compared to the incentive constraint under mechanical communication (2), the new term on the left-hand side reflects Ann's potential opportunity to gain from shirking on Carol if she meets Carol first. Because Ann can gain from shirking on Carol at most once, and then only if Bob and Carol have not met, her payoff from shirking at time t must be weighted by $e^{-2\lambda t}$, which is the probability that by time t, Carol has met neither Ann nor Bob. The highest effort compatible with this incentive constraint is $\left(\frac{r+4\lambda}{r+3\lambda}\right)\underline{a}$—above the bilateral enforcement effort $\underline{a}$, but below the effort supportable under mechanical communication.

However, this strategy profile fails sequential rationality: Bob prefers to conceal from Carol that Ann shirked. Suppose Bob meets Carol, Carol speaks first, and Carol does not indicate that Ann has ever shirked on her. If Bob discloses the truth to Carol, Ann will be permanently ostracized, and thereafter Bob and Carol will revert to bilateral enforcement and cooperate at level $\underline{a}$. But Bob can conceal the interaction in which Ann shirked, in which case he expects Carol to work at the equilibrium path effort level a, and he can shirk while she does so. So Bob will disclose the truth about Ann only if

$$\underbrace{a+a^2}_{\substack{\text{Payoff from concealing Ann's guilt}\\ \text{and shirking on Carol today}}} \leq \underbrace{\underline{a}+\int_0^\infty e^{-rt}\lambda\underline{a}\,dt}_{\substack{\text{Payoff from disclosing Ann's guilt and}\\ \text{working today and in the future with Carol}}} = \underline{a}+\underline{a}^2, \tag{4}$$

where the inequality is Bob's truthtelling incentive constraint and the equality is by definition of $\underline{a}$. Hence Bob is willing to report on Ann's deviation only if the equilibrium path effort level is no greater than that of bilateral enforcement. In other words, truthful communication is incentive compatible under permanent ostracism only if it is redundant.

The underlying strategic force is that Bob no longer fears third-party punishment from Ann once she is ostracized. His loss of "social collateral" with Ann reduces his incentives in his remaining relationship with Carol to the level of bilateral enforcement. But his private information about Ann enables him to manipulate Carol into working at a level above that of bilateral enforcement, if she has not been shirked on by Ann. The challenge is not with Bob's credibility in punishing Ann, nor with his willingness for Carol to do so, but with his willingness to work with Carol when he privately knows that Ann is guilty.

Temporary ostracism with strategic communication: Forgiveness facilitates communication and cooperation: if Bob knows that guilty Ann will be forgiven in the future, he looks forward to subsequently working with her. Concealing information from and shirking on Carol only postpones that prospect, since Bob would then also have to wait for himself to be forgiven. Temporary ostracism tempers the threat players face on the equilibrium path but maintains the the threat of third-party punishment off the equilibrium path, so that communication among innocent players remains incentive compatible.

While this strategic logic is intuitive, the construction of an equilibrium is intricate, because players possess a tremendous degree of private information. We prove that if players are sufficiently patient or society is sufficiently large, there is a temporary ostracism equilibrium that strictly improves upon permanent ostracism.

3 The Model

Society comprises a finite number of players $1, 2, \ldots, n$, with $n \geq 3$, each of whom has a discount rate $r > 0$. Each pair of players i and j engages in a bilateral partnership, denoted "link ij." Each link meets according to an independent Poisson process of intensity $\lambda > 0$. Each time link ij meets, players i and j play the following extensive form stage game:

1. *Communication Stage*: First one partner is randomly selected to send a message; then the other partner sends a message in response. Each partner is equally likely to be selected to speak first. Their message spaces are defined below.
2. *Stake Selection Stage*: Partners simultaneously propose *stakes*, and each proposal is a non-negative real number. The minimum of their proposals is selected.
3. *Effort Stage*: Partners play the prisoner's dilemma with the selected stakes ϕ:

	Work	Shirk
Work	ϕ, ϕ	$-V(\phi), T(\phi)$
Shirk	$T(\phi), -V(\phi)$	$0, 0$

Both T and V are smooth, non-negative, and strictly increasing functions that satisfy $T(0) = V(0) = 0$, and $T(\phi) > \phi$ for all $\phi > 0$. Furthermore, T is strictly convex and satisfies $T'(0) = 1$ and $\lim_{\phi \to \infty} T'(\phi) = \infty$.

An *interaction* between players i and j at time t comprises the time t at which the pair meets, their names, the timing and contents of their communications to each other, the stakes that each proposed, and their effort choices. The interaction on link ij is perfectly observed by partners i and j, but not observed at all by any other player; others can learn of these interactions only

through communication. Player i's private history at time t, denoted h_i^t, is the *set* of all her interactions that occurred strictly before t. $\mathcal{H}_i^t$ is the set of all feasible private histories for player i at time t.

We focus on *evidentiary communication*: players may conceal interactions but cannot fabricate or distort them. Because a history h_i^t is a *set* of past interactions, any subset of h_i^t is a feasible message. The set of all feasible messages for player i at history h_i^t is $\mathcal{P}(h_i^t)$, the power set of h_i^t. Messages contain information about other interactions: the history h_i^t includes both those interactions that player i has experienced first-hand and those that others have disclosed to her. The set of all interactions that occurred or were mentioned in history h_i^t is $\mathcal{E}(h_i^t)$.

We study weak perfect Bayesian equilibria,[2] imposing the restriction that each player's stakes proposals are uniformly bounded across histories;[3] we refer to these as *equilibria*. A strategy profile has *mutual effort* if all players work on the path of play.

Discussion: Variable stakes, introduced by Ghosh and Ray (1996), offer a realistic depiction of many relationships, where players can choose how much effort to exert in a joint venture, or how much to trade in a contractual relationship, or how much wealth to transfer in a risk-sharing arrangement.[4] We study a variable stakes environment for two important reasons.

First, it permits partners to adapt their relationship to the set of players being ostracized, the flow of information within the partnership, and the dynamics of cooperation within the community. Were players instead constrained to play a fixed-stakes prisoners' dilemma, it would be mechanically true that permanent ostracism could do no better than bilateral enforcement: either the fixed stakes would be too high for two innocent partners to cooperate when all others were guilty, or cooperation could be attained through bilateral enforcement alone. Variable stakes shift the focus from constraints in the technology of cooperation to the challenge of providing incentives for truthful communication.

Second, variable stakes offer a convenient metric to compare equilibria at a fixed discount rate. Our assumptions on T ensure that players require proportionally stronger incentives to work at higher stakes, and that for a fixed λ and r, there is an upper bound to how much can be enforced by any equilibrium.[5] Instead of studying behavior at the limits of perfect patience ($r \to 0$), we

[2] Though this solution concept permits unreasonable beliefs off the equilibrium path, this property only strengthens our negative results, while our positive results do not exploit it.

[3] The restriction eliminates unreasonable equilibria in which the stakes always grow with further cooperation, eventually exploding to infinity. Bounding the space of stake proposals achieves equivalent results so long as the bound is sufficiently high (so that all equilibrium stakes we describe are interior).

[4] Also see Kranton (1996) and Watson (1999) for other studies employing variable stakes.

[5] Here are two applications in which this assumption holds: First, in self-enforced risk-sharing, the size of transfers from a wealthy player to a poor players is the stakes and increasing these transfers increases the temptation to shirk more than proportionally when players are risk-averse. In trade between a buyer and seller, the quality or

focus on how community enforcement negotiates the challenges of private information for a fixed discount rate.[6]

We model communication as evidentiary, rather than as cheap talk, for two reasons. First, because cheap talk expands the scope of deviations, our negative result necessarily extends. Second, cheap talk would focus attention on the issue of conflicting reports—"he-said, she-said"—which presumes that a victim and his victimizer have mis-aligned preferences on what to say to third parties. Eliminating that possibility permits us to focus on the more basic problem that a victim and his victimizer may have aligned preferences in concealing defections from third-parties.

4 Main Results

Bilateral Enforcement: We first formalize the lower bound on community enforcement that we introduced in Section 2. In a bilateral equilibrium, behavior in each partnership is independent. The most cooperative bilateral equilibrium is in grim trigger strategies, where partners work with each other at stakes ϕ on the equilibrium path, and otherwise mutually shirk. This profile generates the equilibrium path incentive constraint

$$T(\phi) \leq \phi + \int_0^\infty e^{-rt}\lambda\phi\, dt. \tag{5}$$

The above inequality holds with equality at a unique value of ϕ, which we denote as $\underline{\phi}$ and refer to as the *bilateral enforcement* stakes. We prove in the Online Appendix B.1.1 that no bilateral equilibrium supports mutual effort at higher stakes.

Permanent Ostracism: Permanent ostracism is a class of social norms in which "innocent" players cooperate and reveal their entire histories with each other, but permanently punish those who are "guilty" of deviating in the past. Player i deems player j to be *guilty* if he has directly observed player j shirk, or been informed of such a deviation by another player. He deems her *innocent* if he has no evidence that she has ever deviated. Permanent ostracism is formally defined in the Appendix.

Many social norms match this description—it places no restrictions on how innocent players adapt their stakes over time, on how guilty players should behave, or on whether players who merely conceal information are guilty. Conceivably, players could randomize their stakes proposals, reward their partners with higher stakes for sharing information, guard themselves by setting

quantity of trade offers a measure of stakes, and with increasing marginal cost, each side has a stronger temptation to shirk on larger trades.

[6] Our companion paper (Ali and Miller 2013) uses a similar framework, without communication, to compare the performance of different equilibria and networks for a fixed discount rate. Both papers leverage the flexibility of variable stakes and continuous-time, but the motivation and techniques differ.

low stakes when their partners reveal little information, or work while guilty. The essence of permanent ostracism is to target permanent punishments towards the guilty and to communicate and cooperate with those who are innocent. Our main result shows that permanent ostracism is no better than bilateral enforcement.

Theorem 1. *In every permanent ostracism equilibrium, each player's expected equilibrium payoff never exceeds that of bilateral enforcement.*

Proof. We restrict attention here to permanent ostracism equilibria in which players neither randomize their stakes proposals nor condition their stake proposals on who spoke first in the communication stage; we prove the general case, without these restrictions, in the Online Appendix B.1.2. A permanent ostracism equilibrium of this form is characterized by a stakes function ϕ_{ij} for each link ij, such that when innocent partners i and j meet at time τ with private histories h_i^τ and h_j^τ, they each propose stakes of $\phi_{ij}(h_i^\tau, h_j^\tau)$ following truthful communication, and work at those stakes. We prove that for every pair of private histories (h_i^τ, h_j^τ), the partners cooperate at stakes no greater than $\underline{\phi}$, the stakes from bilateral enforcement.[7]

Suppose otherwise. Consider a pair of private histories (h_i^τ, h_j^τ) such that partners i and j are innocent and $\phi = \phi_{ij}(h_i^\tau, h_j^\tau) > \underline{\phi}$. Consider another private history $\hat{h}_i^\tau$ that coincides with h_i^τ except that every other player has shirked on player i after the last interaction in h_i^τ. Suppose that player j communicates first and sends the message h_j^τ. In a permanent ostracism equilibrium, player i deems player j innocent, and so should report $\hat{h}_i^\tau$ truthfully. Then both partners should propose stakes $\hat{\phi} = \phi_{ij}(\hat{h}_i^\tau, h_j^\tau)$, which cannot exceed $\underline{\phi}$ since they must employ bilateral enforcement in their relationship while permanently ostracizing all the other players. A deviation for player i in which he reports h_i^τ rather than $\hat{h}_i^\tau$, proposes stakes ϕ, and shirks yields a payoff of

$$T(\phi) > T(\underline{\phi}) = \underline{\phi} + \int_0^\infty e^{-rt}\lambda\underline{\phi}\,dt \geq \hat{\phi} + \int_0^\infty e^{-rt}\lambda\hat{\phi}\,dt,$$

where the first inequality is implied by $\phi > \underline{\phi}$, the equality is by definition of $\underline{\phi}$, and the second inequality is implied by $\underline{\phi} \geq \hat{\phi}$. Since this deviation is strictly profitable, we have reached a contradiction. □

Our argument extends to incomplete networks, or more generally, if the frequency of interaction is link-specific: once the benchmark of bilateral enforcement is suitably redefined on a link-specific basis, the analogue of Theorem 1 binds permanent ostracism equilibria from supporting cooperation in a partnership by the level of cooperation that partnership could attain through bi-

[7] Our result here is slightly stronger than Theorem 1 because of the restriction to pure strategy equilibria.

lateral enforcement. The same conclusion applies if partners can interact more frequently when ostracizing other players.

We use continuous time and sequential communication to deliver an exact bound on all permanent ostracism equilibria. Online Appendix B.1.3 shows that the same bound applies approximately in discrete time: permanent ostracism can do better than bilateral enforcement, but its advantage vanishes with the period length. Online Appendix B.1.4 shows that if partners communicate simultaneously, our results apply if certain unreasonable beliefs are ruled out and innocent players cooperate at stakes at least $\underline{\phi}$.

Temporary Ostracism: Ostracism is more effective when guilty players are eventually "forgiven" and readmitted to productive society. Then an innocent player who deviates stands to lose not only his relationships with his innocent partners, but also his future relationships with his currently guilty partners. The additional social capital generated by forgiveness ensures that communication incentives are satisfied at levels of cooperation above bilateral enforcement, even when there are only two currently innocent players.

While this intuition is straightforward, construction of an equilibrium is complicated by challenges familiar in the study of private monitoring: players must coordinate the punishment and forgiveness of a guilty player, and a guilty player may profit from deviations at multiple histories. A guilty player may wish to slow the rate at which others learn of her guilt by employing a dynamic, history-dependent pattern of working in some interactions and shirking in others. We address these challenges by introducing two features to our model that do not change our negative result for permanent ostracism, but do simplify the construction of a temporary ostracism equilibrium. First, we coordinate the forgiveness of guilty players by embedding n public correlation devices. Each device is an independent "Poisson clock" that rings at rate μ, and each player is associated with one device. Second, we augment the stage game with an additional round of simultaneous communication immediately after the effort choices. We prove the following result.

Theorem 2. *If $r < 2\lambda(n-3)$, then there exists a temporary ostracism equilibrium that yields payoffs strictly higher than permanent ostracism.*

Our formal proof is in the Appendix; here we describe its essence. Players cooperate with those they deem innocent at a fixed stakes $\phi > \underline{\phi}$ both on and off the equilibrium path.[8] If a player learns that her partner is guilty, she sets zero stakes or shirks with him until his Poisson clock rings, at which time he is "forgiven" and deemed innocent. A key part of our argument is

[8] That stakes are history-invariant in this equilibrium implies that if stakes were exogenously fixed, then there would exist a range of discount rates at which temporary ostracism could enforce mutual effort, but bilateral enforcement and permanent ostracism could not.

showing that off the equilibrium path, if there are only two innocent players, each of them would strictly prefer to cooperate at bilateral equilibrium stakes $\underline{\phi}$ because she would like to be innocent when her currently guilty partners are forgiven. Because each of them strictly prefers to work at stakes $\underline{\phi}$, they also prefer to work at stakes slightly higher than $\underline{\phi}$. When there are more than two innocent players, each of them has even more to lose from shirking, implying that there exist stakes $\phi > \underline{\phi}$ at which innocent players are always willing to work.

But an equilibrium has to specify what guilty players also do off the path of play. To do so tractably, we construct an equilibrium in which a guilty player's first victim is the only one who spreads information about his deviation. Therefore, once a guilty player has already shirked on someone, he does not gain from working with others at stakes exceeding $\underline{\phi}$ because doing so would not slow down the rate at which information about his guilt diffuses. Therefore, once he shirks, he then has strict incentives to continue to shirk until he is forgiven. In order to testify to his later victims that they are not the first, he uses the round of communication immediately after the effort stage to reveal his past deviation.[9] Finally, because innocent players always cooperate at the same stakes both on and off the equilibrium path, the first victim suffers no harm from communicating truthfully.

Our use of n independent Poisson clocks contrasts with how Ellison (1994) uses a single Poisson clock to synchronize forgiveness across the community in contagion. Our construction leverages imperfectly correlated forgiveness towards incentives for innocent players: an innocent player recognizes that if she deviates, she may not be forgiven until after her currently guilty partners are forgiven. The Online Appendix B.2.1 shows that if instead all players were forgiven simultaneously, the analogous construction does not improve on bilateral enforcement.

One may envision more sophisticated variants of temporary ostracism to increase cooperation payoffs beyond our construction here. Modest gains emerge from optimally choosing the rate of forgiveness μ: in Online Appendix B.2.2, we optimally choose μ for 4 players, and graph how the optimal μ depends on patience. More dramatic gains may be possible by requiring guilty players to redeem themselves by "working for free" while an innocent partner shirks. We construct an efficient temporary ostracism equilibrium in Online Appendix B.2.3 for 3 players in which the cost of redemption perfectly offsets a guilty player's gain from being forgiven. This efficient equilibrium enforces cooperation at stakes weakly higher than that of contagion and more generally, any mutual effort equilibrium in which behavior on the path of play is stationary. Unfortunately, generalizing this construction to four or more players is challenging, because of the

[9] For our construction, the only point at which information needs to be verifiable is when a guilty player reveals the identity of his first victim, or is unable to produce such evidence. Without verifiability, he could fabricate a phantom first victim to stop his true first victim from spreading news of his guilt.

inherent lack of common knowledge about who is guilty.

5 Extensions

Rewarding whistleblowers through asymmetric play. Our definition of permanent ostracism involves symmetric mutual effort between two innocent partners after ostracizing all other guilty players. Relaxing this requirement enables a "whistleblower" to be rewarded with asymmetric payoffs for reporting a deviator. In principle, these rewards could be used to motivate truthful communication and raise the level of cooperation.

Nonetheless, our core argument delivers an upper bound on cooperation under permanent ostracism for a general class of stage games, including those in which such rewards may be used. In Online Appendix B.1.5, we consider "generalized permanent ostracism" strategies in which, when player i reports to an innocent partner j that other players have deviated, rather than reverting to symmetric bilateral cooperation, he may be rewarded with any continuation payoff compatible with any (potentially asymmetric) bilateral equilibrium on link ij.[10] Using the communication incentive constraint that arises when player i has seen every other player deviate since the last time he met player j, we derive a bound on action profiles implementable in permanent ostracism. This bound is straightforward to interpret when each stage game is symmetric and partners play symmetrically on the equilibrium path: we show then that player i cannot earn any more in the ij partnership in a generalized permanent ostracism equilibrium than the most she could earn in *any* bilateral enforcement equilibrium.

We note two qualifiers to this result. First, the result does not preclude a permanent ostracism equilibrium from simultaneously offering to each of players i and j the payoff she might attain in the best bilateral enforcement equilibrium. So, unlike the case of the Prisoner's Dilemma, permanent ostracism may conceivably offer the pair higher payoffs than it could achieve through bilateral enforcement. Second, because our existing toolkit falls short of being able to construct equilibria for a general class of stage games when the discount rate is fixed and monitoring is private, we do not know if there exists temporary ostracism equilibria that might surpass this bound.

Pure communication opportunities In some settings, players may have more opportunities to talk than to trade or cooperate. Adding opportunities for pure communication does not change our negative result: an innocent player who has been shirked on by everyone but his current

[10] For instance, if we expanded the stage game of Section 3 to allow players to transfer utility after communication, player i might receive a transfer from his partner after revealing that others are guilty, with both partners continuing to cooperate at bilateral stakes $\underline{\phi}$. Anticipating this reward, player i has a stronger motive to communicate truthfully than if he were to receive merely $\frac{\lambda}{r}\underline{\phi}$, enabling cooperation along the equilibrium path at stakes greater than $\underline{\phi}$.

partner lacks the incentive to reveal his full history.

A more subtle departure is one in which both partners can simultaneously broadcast public announcements to all the other players immediately following their effort choices. In principle, such a structure can enforce the same level of cooperation as public monitoring: if Bob is going to reveal that Ann just shirked on him, then Ann is indifferent and willing to reveal it as well, in which case Bob is also indifferent. Because both players would strictly prefer that the incriminating evidence remain concealed, if there were some lexicographic cost of revealing evidence, or if partners made their announcements sequentially, then both Ann and Bob would lie in order to shirk on Carol later. Even if Carol can simultaneously ask Ann and Bob whether either of them is guilty, each reveals information only if the other does so, but otherwise, Ann and Bob would have aligned interests in concealing that interaction.

Meeting times and voluntary evidence Our results leverage the fact that meeting times are private. Were all meeting times to be public, then with evidentiary communication, a standard unraveling argument implies that players can be compelled to reveal all details of their interactions. In such a setting, permanent ostracism is effective. A natural middle ground is a setting in which, prior to interacting, players can verifiably sign and time-stamp a document with their intent to interact, and send it to a public repository.[11] Their document makes their meeting time public, so by the same unraveling argument they are also compelled to reveal all verifiable details of their interaction. Therefore permanent ostracism is again effective at sustaining cooperation.

However, permanent ostracism falters if, in contrast to our assumption of evidentiary communication, meeting times constitute the *only* verifiable evidence. In that case, the partners cannot be induced to truthfully attest to the fact that one of them shirked on the other, because both the guilty partner and her victim would prefer to falsely attest that both of them worked.

6 Discussion

An extensive literature in the social sciences has studied mechanisms of community enforcement in which deviating players may be identified and punishments are targeted towards those deviants. Our main contribution is to highlight that when the identity of deviants needs to be voluntarily revealed, the most stringent forms of directed punishment may be self-defeating. By contrast, forgiveness fosters truthful communication and thereby facilitates cooperation.[12]

[11] We are grateful to an anonymous referee for this suggestion.

[12] Our insights complement existing motives for forgiveness, such as to avoid renegotiation (Bernheim and Ray 1989), tackle imperfections in monitoring (Green and Porter 1984), and support optimal punishments when mutual minmax is not a stage game Nash equilibrium (Fudenberg and Maskin 1986). Ellison (1994) uses temporary punishments to offer incentives for players to spread contagion, but we show in Ali and Miller (2013) that this is unnecessary in a variable-stakes environment.

We place our results within the context of the prior literature. Theoretical mechanisms proposed for such targeted punishments often feature public monitoring[13] or employ "reputational label mechanisms" wherein each individual carries a label of innocence or guilt that is automatically updated on the basis of her past history, and is observed by all those who interact with her.[14] This strand of the literature views public monitoring or reputational label mechanisms as proxies for the power of gossip, but neither models communication nor offers players the opportunity to conceal information. A different strand that models communication[15] elucidates how its speed and reach influence cooperation incentives, but abstracts from incentive issues that arise with strategic communication. When communication incentive constraints are set aside, the most stringent form of ostracism—featuring permanent punishments—emerges as the most cooperative equilibrium, and thus, is the natural focus of these prior studies.[16] But this theoretical focus on the most stringent punishments appears to conflict with qualitative evidence on the prevalence of forgiveness (Ostrom 1990; Ellickson 1991; Greif 2006). Our paper reconciles this conflict by demonstrating that when information must be voluntarily shared, these stringent punishments may be self-defeating and forgiveness may facilitate cooperation.

Three recent studies share our interest in understanding when information germane to community enforcement is credibly communicated. Lippert and Spagnolo (2011) consider a networked environment with fixed stakes and deterministic interaction timing. One of the social norms they consider is "multilateral repentance," which also involves temporary punishments and continued cooperation among innocent players. Bowen, Kreps, and Skrzypacz (2013) model favor exchanges with public actions and messages but privately observed payoffs, and they study whether messages should precede or succeed actions. Wolitzky (2015) studies when fungible tokens can outperform communication in community enforcement.

A substantively different approach to community enforcement is that of *contagion*, proposed by Kandori (1992), in which a player shirks on all his partners after someone shirks on him. By punishing both innocent and guilty players for a single deviation, contagion operates in both anonymous and non-anonymous settings, but perhaps it is most suitable when players who cannot identify defectors lose trust in the community once anyone deviates. By contrast, a community enforcement scheme that uses targeted punishments and communication appears more

[13] See Hirshleifer and Rasmusen (1989), Bendor and Mookherjee (1990), Karlan, Möbius, Rosenblat, and Szeidl (2009), and Jackson, Rodriguez-Barraquer, and Tan (2012).

[14] Reputation label mechanisms were formalized by Kandori (1992), Okuno-Fujiwara and Postlewaite (1995), and Tirole (1996), and feature also in sociology (Coleman 1988; Raub and Weesie 1990), and evolutionary biology (Nowak and Sigmund 1998).

[15] For example, Raub and Weesie (1990), Dixit (2003), and Bloch, Genicot, and Ray (2008).

[16] We supplement this discussion in Online Appendix B.1.1, showing that if communication is mechanical, then permanent ostracism is optimal in the class of mutual effort equilibria.

relevant for markets and communities where defectors can be identified and excluded.

A difficulty in constructing contagion equilibria is that contagious players are tempted to work, rather than shirk, in order to slow the spread of contagion. Deb (2012) and Deb and González-Díaz (2014) show that this challenge applies across repeated games in anonymous environments: because players are anonymous, the only way to punish a defector is to punish all players, which makes innocent players hesitant to initiate punishment. Sophisticated schemes of community enforcement and communication may be needed to overturn this obstacle and support punishments that punish an entire community for the defection of a single individual.[17]

We study a complementary problem: when players can be identified and have a fixed level of patience, when can communication facilitate punishment of the defector alone and not her entire community? In studying targeted punishments, we see that a different strategic force is at play: because players act as each other's monitors and enforcers, an innocent Bob may not wish to reveal to an innocent Carol that Ann—who could otherwise act as a "stick" for Bob—has already defected. Instead of letting Carol know that he has less reason to behave, Bob would rather simply misbehave. Unlike the strategic challenge of anonymous interaction, Bob here faces no temptation not to punish Ann, nor does he fear retribution in the event that he recommends that Carol punishes Ann. He simply loses his motive to cooperate with other innocent players when Ann is permanently ostracized, but regains it if she may be forgiven.

Fundamentally, the incentive challenges here and in anonymous repeated interaction are illustrations of a broader principle: when players are on a punishment path, their continuation payoffs may be lower. In anonymous interaction, this principle manifests in tempting a victim to *work*, concealing others' deviations and preserving some future cooperation. In ostracism, by contrast, it manifests in tempting a victim to conceal others' deviations only because it expedites *shirking* at higher stakes on an unsuspecting partner, destroying future cooperation. Our results illustrate how ostracism involves not only a direct loss of continuation value from punishment, but also a loss of enforcement capability because the ostracized defector is no longer available as a monitor and enforcer. Cooperation is therefore facilitated by forgiveness that recoups that enforcement capability in the future.

Our attention has been devoted to cooperation in the absence of legal and other external enforcement mechanisms. Bounding the cooperation achieved through communication and tar-

[17] Deb (2012) uses cheap-talk communication to permit anonymous players to create "signatures." Since impersonation remains a possibility, she uses a *community responsibility* scheme that punishes a defector's entire community to prove a folk theorem. By contrast, our paper models a setting where players are non-anonymous and so neither communication nor community enforcement are needed for a folk theorem; bilateral enforcement alone engenders cooperation at arbitrarily high stakes as $\delta \to 1$. Communication, thus, serves different roles in her paper and ours: in hers, communication is a *substitute* for players' non-anonymity, whereas here we are using communication to *complement* players' non-anonymity to facilitate third-party punishment.

geted punishment offers an appreciation for the gap that may be filled by intermediaries and institutions. Even when such institutions are present, our motivating question remains of interest: when do victims truthfully report to an adjudicator that someone else has deviated?

Appendix A Definitions and proofs

Permanent Ostracism: A *permanent ostracism assessment* is a behavioral strategy profile and a system of beliefs. Player i's behavioral strategy in permanent ostracism is $\sigma_i = (\sigma_i^{\mathrm{M}}, \sigma_i^{\mathrm{S}}, \sigma_i^{\mathrm{E}})$, where in i's interaction with player j at time t given private history h_i^t, her message to player j is $\sigma_i^{\mathrm{M}}\big(j, t, h_i^t, \emptyset\big) \in \mathcal{P}(h_i^t)$ if she communicates first and $\sigma_i^{\mathrm{M}}\big(j, t, h_i^t, m_j^t\big) \in \mathcal{P}(h_i^t)$ if j communicates first (recall that a message is a set of interactions); her mixture over stakes proposals is $\sigma_i^{\mathrm{S}}\big(j, t, h_i^t, m_i^t, m_j^t\big) \in \Delta[0, \infty)$; and her effort choice is $\sigma_i^{\mathrm{E}}\big(j, t, h_i^t, m_i^t, m_j^t, \hat{\phi}_{ij}^t, \hat{\phi}_{ji}^t\big) \in \{W, S\}$. That is, player i's effort choice—either W or S—is conditioned on the identity of her current partner j, the time t, her private history h_i^t, both messages m_i^t and m_j^t, and both stakes proposals $\hat{\phi}_{ij}^t$ and $\hat{\phi}_{ji}^t$.

Let $\mathcal{E}_i(h)$ be the set of all interactions that i knows in history h, and let $\mathcal{E}_i^j(h, \tau)$ be the subset of $\mathcal{E}_i(h)$ that happened strictly before time τ and in which j participated. Let $\mathcal{G}_i(h)$ be the set of players that i deems guilty at h. Fixing player i and private history h_i^t, let $\{t^z\}_{z=1}^Z$ be an ordered list of the times at which the interactions in $\mathcal{E}_i(h_i^t)$ occurred. We now construct a state variable ω that tracks which players i' deems guilty; evolution of ω is governed by the interactions in $\mathcal{E}_i(h_i^t)$. Consider a sequence $\{\omega^z\}_{z=0}^Z$ of states such that $\omega^z \in \{0,1\}^n$ for each z. Player $j \in \mathcal{I}_i(h_i^t)$ if $\omega_j^Z = 0$, and in $\mathcal{G}_i(h_i^t)$ otherwise. The initial condition is $\omega_j^0 = 0$ for all j (all players start innocent), and if $\omega_j^{z-1} = 1$ then $\omega_j^z = 1$ (guilt is permanent). A transition from $\omega_j^{z-1} = 0$ to $\omega_j^z = 1$ occurs if player j and any neighbor k interact in $\mathcal{E}_i(h_i^t)$ at time t^z, and j's effort choice is an observable deviation given what player i knows from his private history $\mathcal{E}_i(h_i^t)$ and player j's message $m_j^{t^z}$; i.e., player j's effort choice is not $\sigma_j^{\mathrm{E}}\big(k, t, \mathcal{E}_i^j(h_i^t, t^z) \cup m_j^{t^z}, m_j^{t^z}, m_k^{t^z}, \hat{\phi}_{jk}^t, \hat{\phi}_{kj}^t\big)$. If $\omega_j^{z-1} = 0$, and j's communications, stake proposals, and effort choices conform to σ_j, then $\omega_j^z = 0$.

Definition 1. *An assessment is a* ***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. *She sends the truthful message $m_i^t = h_i^t$.*
2. *If j's message m_j^t satisfies $\mathcal{E}_i^j(h_i^t, t) \subseteq m_j^t$, and $j \in \mathcal{I}_i\big(h_i^t \cup m_j^t\big)$, then i believes with probability 1 that j has not deviated, and i proposes strictly positive stakes. If j also proposes stakes in the support of $\sigma_j^{\mathrm{S}}\big(i, t, m_i^t \cup m_j^t, m_j^t, m_i^t\big)$, then i believes with probability 1 that j has not deviated, and i works.*
3. *If $j \in \mathcal{G}_i(h_i^t \cup m_j^t)$, then i shirks.*

Our definition of permanent ostracism does not specify all details of the strategy profile or system of beliefs. For instance, it does not require that players who conceal information or propose off-path stakes be considered guilty, it does not constrains the behavior of guilty players, and it does not restrict players from dynamically adjusting their stakes.

Temporary Ostracism: Each player is associated with a public Poisson clock that rings at rate μ. All Poisson clocks are independent of each other. Now that the stage game has an additional post-interaction

communication stage, an *interaction* between players i and j at time t comprises the time t at which the pair meets, their names, the timing and contents of their pre-interaction communications to each other, the stakes that each proposed, their effort choices, and the contents of their post-interaction communications. A history h is now a set of interactions, and Poisson clock rings of the form (i, t) specifying that the Poisson clock associated with player i rang at time t.

All players propose stakes ϕ on and off the equilibrium path, and work with innocent partners. As above, $\mathcal{G}_i(h_i^t)$ is the set of players that player i deems guilty at private history h_i^t. As above, $\{t^z\}_{z=1}^Z$ is an ordered list of the times at which the interactions and Poisson clock rings in $\mathcal{E}_i(h_i^t)$ occurred, and $\{\omega^z\}_{z=0}^Z$ is a sequence of states such that $\omega^z \in \{0,1\}^n$ for each z. Player $j \in \mathcal{I}_i(h_i^t)$ if $\omega_j^Z = 0$, and in $\mathcal{G}_i(h_i^t)$ otherwise. We modify the evolution of ω as follows to implement forgiveness. The initial condition is $\omega_j^0 = 0$ for all j. A transition from $\omega_j^{z-1} = 0$ to $\omega_j^z = 1$ occurs if and only if player j and any neighbor k interact in $\mathcal{E}_i(h_i^t)$ at time t^z, and j's effort choice is an observable deviation given what player i knows from his private history $\mathcal{E}_i(h_i^t)$ and player j's message $m_j^{t^z}$. A transition from $\omega_j^{z-1} = 1$ to $\omega_j^z = 0$ occurs (i.e., player j is forgiven) if and only if player j's Poisson clock rings at t^z. In all other cases, $\omega_j^z = \omega_j^{z-1}$.

We now define when an innocent player k is the "first victim" of a guilty player j. In history h, with associated times $\{t^z\}_{z=1}^Z$ as defined above, if $\omega_j^Z = 1$ (i.e., j is guilty) and there exists $z \in \{1, \ldots, Z\}$ and some player k such that $\omega_k^{z-1} = \omega_j^{z-1} = 0$, players k and j interact at time t^z, and $\omega_j^z = 1$, then we say that player k became the *first victim* of player j in the interaction at t^z. Let $\tilde{\mathcal{E}}_i(h) \subset \mathcal{E}_i(h)$ be the set of interactions in which player i became the first victim of another opponent. Similarly, let $\hat{\mathcal{E}}_i(h) \subset \mathcal{E}_i(h)$ be the set of interactions in which another opponent became the first victim of player i.

When player i meets player j with records $\mathcal{E}_i(h)$, his strategy specifies:

- *Communication pre-interaction*: Regardless of j's message and order of communication, send message $\tilde{\mathcal{E}}_i(h)$.
- *Stake selection*: Propose stakes ϕ regardless of the pre-interaction messages and order of communication.
- *Effort*: Let $\hat{h}$ be the message received from j. If $i \notin \mathcal{G}_i(h)$, $j \notin \mathcal{G}_i(h \cup \hat{h})$, and selected stakes are ϕ, then work; otherwise shirk.
- *Communication post-interaction*: If $i \in \mathcal{G}_i(h)$ and i shirked at stakes ϕ in the interaction stage, send message $\hat{\mathcal{E}}_i(h)$; otherwise, send no message.

Our construction involves minimal communication: at the pre-interaction communication stage, player i sends a non-empty message to player j only if player i was the first victim of some other player; and at the post-interaction communication stage, player i sends a message to player j only if player i shirked on player j and player j was not the first victim of player i.

To verify equilibrium incentives and construct an equilibrium, suppose that player i meets player j, and believes there are $\ell \geq 2$ innocent players (including i and j). For player i to work, her loss from shirking must exceed her gain. Her actions today do not affect her payoffs after her clock rings and so we include

only payoffs she expects before her clock rings. Her expected payoff from following equilibrium before her clock rings is

$$W(\phi,\mu,\ell) \equiv \phi + (\ell-1)\int_0^\infty e^{-rt}e^{-\mu t}\lambda\phi\,dt + (n-\ell)\int_0^\infty e^{-rt}e^{-\mu t}\left(1-e^{-\mu t}\right)\lambda\phi\,dt. \tag{6}$$

The first term is her immediate payoff from cooperating; the second and third are payoffs she accrues before her clock rings from working with other innocent players, and players who are currently guilty after they are forgiven. $e^{-\mu t}$ is the probability that her clock does not ring before t and $1-e^{-\mu t}$ is the probability that the clock for a currently guilty player rings before t.

We now consider the deviation where player i shirks on player j and every innocent partner she meets before her clock rings, and reveals the identity of her first victim to each. (We verify in Theorem 2, below, that this is her best deviation.) After shirking on player j, player i's expected payoff from possibly shirking on any currently innocent player k before i's clock rings is $\mathcal{Z}(\phi) \equiv \int_0^\infty e^{-rt}e^{-(\mu+2\lambda)t}\lambda T(\phi)dt$, where $e^{-(\mu+2\lambda)t}$ is the probability that neither has i's clock rung nor has k met either i or j before t. The total payoff that player i accrues from shirking until the next time her clock rings is

$$S(\phi,\mu,\ell) \equiv T(\phi) + (\ell-2)\mathcal{Z}(\phi) + (n-\ell)\left(\int_0^\infty e^{-rt}e^{-(2\mu+\lambda)t}\mu\,dt\right)\mathcal{Z}(\phi). \tag{7}$$

The first term is the immediate gain from shirking; the second is the payoff accrued from possibly shirking on all other innocent players, and the third is the payoff from possibly shirking on all currently guilty players. Fixing a guilty player k, $e^{-(2\mu+\lambda)t}$ is the probability that by t, neither i's nor k's clock has rung nor has k met j. Once k is forgiven, if she has not met j by then, i's expected payoff from shirking is $\mathcal{Z}(\phi)$.

Lemma 1. *If $r < 2\lambda(n-3)$, then there exist $\mu > 0$ and $\phi > \underline{\phi}$ such that $S(\phi,\mu,\ell) \leq W(\phi,\mu,\ell)$ for every $\ell \in \{2,\ldots,n\}$.*

Proof. We consider separately the case of $\ell = 2$ and $\ell > 2$. For $\ell = 2$, observe that for $\mu > 0$,

$$W(\underline{\phi},\mu,2) = \underline{\phi} + \frac{\lambda}{r+\mu}\underline{\phi} + \frac{(n-2)\lambda\mu\underline{\phi}}{(r+\mu)(r+2\mu)},$$

and

$$\begin{aligned} S(\underline{\phi},\mu,2) &= T(\underline{\phi}) + \frac{(n-2)\lambda\mu T(\underline{\phi})}{(r+2\mu+\lambda)(r+\mu+2\lambda)} \\ &= \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{(n-2)\lambda\mu}{(r+2\mu+\lambda)(r+\mu+2\lambda)}\frac{(r+\lambda)}{r}\underline{\phi}. \end{aligned}$$

For $\mu > 0$, observe that

$$\frac{W(\underline{\phi},\mu,2) - S(\underline{\phi},\mu,2)}{\lambda\mu\underline{\phi}} = \frac{(n-2)}{(r+\mu)(r+2\mu)} - \frac{1}{r(r+\mu)} - \frac{(n-2)(r+\lambda)}{r(r+2\mu+\lambda)(r+\mu+2\lambda)},$$

and therefore, taking limits as $\mu \searrow 0$

$$\lim_{\mu \searrow 0} \frac{W(\underline{\phi}, \mu, 2) - S(\underline{\phi}, \mu, 2)}{\lambda \mu \underline{\phi}} = \frac{n-2}{r^2} - \frac{1}{r^2} - \frac{(n-2)(r+\lambda)}{r(r+\lambda)(r+2\lambda)} = \frac{2\lambda(n-3) - r}{r^2(r+2\lambda)} > 0, \tag{8}$$

where the inequality follows from $r < 2\lambda(n-3)$. By L'Hôpital's Rule, the LHS is $\frac{1}{\lambda \underline{\phi}}\big(W_2(\underline{\phi}, 0, 2) - S_2(\underline{\phi}, 0, 2)\big)$. Therefore, we have established that $W_2(\underline{\phi}, 0, 2) > S_2(\underline{\phi}, 0, 2)$. By continuity, combining this inequality with $W(\underline{\phi}, 0, 2) = S(\underline{\phi}, 0, 2)$ implies that $W(\phi, \mu, 2) > S(\phi, \mu, 2)$ for $\mu > 0$ sufficiently small and $\phi = \underline{\phi} + \varepsilon$ for $\varepsilon > 0$ sufficiently small.

Now we consider $\ell > 2$. Evaluating $S(\phi, \mu, \ell)$ and $W(\phi, \mu, \ell)$ at $\mu = 0$ and $\phi = \underline{\phi}$ yields:

$$\begin{aligned} S(\underline{\phi}, 0, \ell) = T(\underline{\phi}) + \frac{(\ell-2)\lambda}{r+2\lambda} T(\underline{\phi}) &= \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{r+\lambda}{r}\frac{(\ell-2)\lambda}{r+2\lambda}\underline{\phi} \\ &< \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{(\ell-2)\lambda}{r}\underline{\phi} = W(\underline{\phi}, 0, \ell), \end{aligned}$$

where the first equality is by definition of S, the second is by substituting $T(\underline{\phi}) = \underline{\phi} + \frac{\lambda}{r}\underline{\phi}$, the inequality is from $\frac{r+\lambda}{r+2\lambda} < 1$, and the final equality is by definition of W. Since S and W are continuous in their first two arguments, and ℓ takes finitely many values, the system of inequalities $S(\phi, \mu, \ell) < W(\phi, \mu, \ell)$ for every $\ell \in \{2, \dots, n\}$ holds on an open neighborhood of $(\mu, \phi) = (0, \underline{\phi})$. □

Proof of Theorem 2. Consider $\phi > \underline{\phi}$ and $\mu > 0$. We first verify that a guilty player i has an incentive to shirk immediately on all other innocent players at stakes ϕ. Since only the first victim communicates, when guilty i meets another innocent player j, working or shirking with j affects no other relationship. Therefore, if π_{ij} represents i's expected payoff from activity on ij before i is forgiven, then

$$\pi_{ij} = \max\left\{ T(\phi), \phi + \frac{\lambda}{r + 2\lambda + \mu}\pi_{ij} \right\}.$$

The first term in the maximand is from shirking immediately, and the second is form working immediately and then possibly earning π_{ij} the next time link ij is recognized (if i has not been forgiven and i's first victim has not met j in the meantime). If $\pi_{ij} > T(\phi)$, then it follows that $\pi_{ij} = \phi + \frac{\lambda}{r+\lambda+\mu}\phi$, which is strictly less than $\phi + \frac{\lambda}{r}\phi$. But $\phi > \underline{\phi}$ implies that $\phi + \frac{\lambda}{r}\phi < T(\phi)$, yielding a contradiction. Therefore, guilty i has the motive to shirk on all other innocent players. Revealing the history afterwards ensures that the new victim knows that he is not the first victim, so he will not spread news of i's guilt.

Since this behavior is optimal for a guilty player, it follows that the most profitable deviation for an innocent player meeting another innocent partner is to shirk immediately and then on all others she meets before her clock rings, revealing the identity of her first victim to each. Now consider by Lemma 1 $\mu > 0$ and $\phi > \underline{\phi}$ such that $S(\phi, \mu, \ell) \leq W(\phi, \mu, \ell)$ for every $\ell \in \{2, \dots, n\}$. At such (ϕ, μ), innocent players are willing to cooperate at stakes ϕ regardless of the number of guilty players.

Finally, an innocent player is willing to shirk on guilty players and willing to communicate truthfully

when he is the first victim, because there is no penalty for doing so. □

References

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

Abhijit V. Banerjee and Esther Duflo. Reputation effects and the limits of contracting: A study of the Indian software industry. *Quarterly Journal of Economics*, 115(3):989–1017, 2000.

Jonathan Bendor and Dilip Mookherjee. Norms, third-party sanctions, and cooperation. *Journal of Law, Economics, and Organization*, 6(1):33, 1990.

B. Douglas Bernheim and Debraj Ray. Collective dynamic consistency in repeated games. *Games and Economic Behavior*, 1(4):295–326, 1989.

Francis Bloch, Garance Genicot, and Debraj Ray. Informal insurance in social networks. *Journal of Economic Theory*, 143(1):36–58, November 2008.

T. Renee Bowen, David M. Kreps, and Andrzej Skrzypacz. Rules with discretion and local information. *Quarterly Journal of Economics*, 2013.

Samuel Bowles and Herbert Gintis. *A cooperative species: Human reciprocity and its evolution*. Princeton Univ Press, Princeton, N.J., 2011.

James S. Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94(S1): S95–S120, 1988.

Joyee Deb. Cooperation and community responsibility: A folk theorem for repeated matching games with names. Working paper, September 2012.

Joyee Deb and Julio González-Díaz. Enforcing social norms: Trust-building and community enforcement. Working paper, 2014.

Avinash K. Dixit. Trade expansion and contract enforcement. *Journal of Political Economy*, 111(6):1293–1317, 2003.

Avinash K. Dixit. *Lawlessness and Economics: Alternative Modes of Governance*. Princeton University Press, 2004.

Robert C. Ellickson. *Order without law: How neighbors settle disputes*. Harvard University Press, Cambridge, MA, 1991.

Glenn Ellison. Cooperation in the prisoner's dilemma with anonymous random matching. *Review of Economic Studies*, 61(3):567–588, July 1994.

Drew Fudenberg and Eric S. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, May 1986.

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

Edward J. Green and Robert H. Porter. Noncooperative collusion under imperfect price information. *Econometrica*, 52(1):87–100, January 1984.

Avner Greif. *Institutions and the path to the modern economy: Lessons from medieval trade*. Cambridge Univ Press, New York, N.Y., 2006.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, December 1981.

David Hirshleifer and Eric Rasmusen. Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior & Organization*, 12(1):87–106, 1989.

Matthew O. Jackson, Tomas Rodriguez-Barraquer, and Xu Tan. Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5):1857–1897, 2012.

Michihiro Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59(1):63–80, 1992.

Dean Karlan, Markus Möbius, Tanya Rosenblat, and Adam Szeidl. Trust and social collateral. *Quarterly Journal of Economics*, 124(3):1307–1361, August 2009.

Rachel E. Kranton. The formation of cooperative relationships. *Journal of Law, Economics, and Organization*, 12(1):214–233, 1996.

Steffen Lippert and Giancarlo Spagnolo. Networks of relations and word-of-mouth communication. *Games and Economic Behavior*, 72:202–217, 2011.

John McMillan and Christopher Woodruff. Interfirm relationships and informal credit in Vietnam. *Quarterly Journal of Economics*, 114(4):1285–1320, 1999.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, Autumn 1981.

Martin A. Nowak and Karl Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393 (6685):573–577, 1998.

Masahiro Okuno-Fujiwara and Andrew Postlewaite. Social norms and random matching games. *Games and Economic Behavior*, 9:79–109, 1995.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge, UK, 1990.

Eric A. Posner. The regulation of groups: The influence of legal and nonlegal sanctions on collective action. *University of Chicago Law Review*, 63:133–197, 1996.

Werner Raub and Jeroen Weesie. Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, pages 626–654, 1990.

Jean Tirole. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies*, 63(1):1–22, 1996.

Joel Watson. Starting small and renegotiation. *Journal of Economic Theory*, 85(1):52–90, 1999.

Alexander Wolitzky. Communication with tokens in repeated games on networks. *Theoretical Economics*, 10(1):67–101, January 2015.

B Supplemental Appendix (for online publication)

This Supplementary Appendix contains several results for both permanent and temporary ostracism. Section B.1 contains all results for permanent ostracism not in the main paper, namely:

- Section B.1.1 proves that permanent ostracism is efficient when communication is mechanical. Corollary 1 describes the best bilateral enforcement equilibrium.
- Section B.1.2 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.1.3 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.1.4 describes the possibilities that arise when communication is simultaneous and for which selection of equilibrium our negative result applies.
- Section B.1.5 describes our extension to general games and the extent to which "rewards for whistleblowers" can improve the outlook for permanent ostracism.

Section B.2 contains all of our results for temporary ostracism not in the main paper, namely:

- Section B.2.1 proves that if forgiveness of all players were synchronized, then temporary ostracism could do no better than bilateral enforcement.
- Section B.2.2 considers the optimal rate of forgiveness for 4 players.
- Section B.2.3 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.1 Permanent ostracism

B.1.1 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

$$T(\phi) = \phi + (\ell - 1)\int_0^\infty e^{-rt}\lambda\phi\, dt; \tag{9}$$

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

$$\frac{T(\phi)}{\phi} \leq 1 + (n-1)\frac{\lambda}{r} = \frac{T(\overline{\phi}_n)}{\overline{\phi}_n}. \tag{10}$$

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.1.2 Permanent ostracism

Proof of Theorem 1 on p. 8, 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

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

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.1.3 Permanent ostracism in discrete time

In the discrete time game, players may interact at times $0, \xi, 2\xi, \ldots$, 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. Eq. 9), 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

$$T(\phi) \leq \phi + (n-1)\frac{\delta p_\xi}{1-\delta}\phi. \tag{11}$$

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

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

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.[18]

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

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

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

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:

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

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 (11) 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,

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

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.1.4 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.1.5 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.[19] Our definition does not restrict how 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

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

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:*

$$\max_{a \in A_{ij}} u_{ij}(a, \alpha^*_{-i}) + \frac{\lambda_{ij}}{r} u_{ij}^{\min} \leq \overline{U}_{ij}. \tag{12}$$

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 Eq. 12 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 Eq. 12. 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

$$\frac{r+\lambda_{ij}}{r}u_{ij}(\alpha^*) \leq \overline{U}_{ij}. \tag{13}$$

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

$$\frac{r+\lambda_{ij}}{r}u_{ij}(\alpha^*) - \max_{a\in A_{ij}} u_{ij}(a,\alpha^*_{-i}) - \frac{\lambda_{ij}}{r}u_{ij}^{\min}. \tag{14}$$

1. Suppose (14) 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 $\left(\underline{\alpha}_{ij}, \underline{\alpha}_{ji}\right)$.[20] 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 (13).

2. Suppose (14) is strictly negative. Then, (13) follows from (12) 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. □

B.2 Temporary ostracism

B.2.1 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. Eq. 6) with:

$$\hat{W}(\phi,\mu,\ell) \equiv \phi + (\ell-1)\int_0^\infty e^{-rt}e^{-\mu t}\lambda\phi\, dt, \tag{15}$$

because the last term in Eq. 6 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. Eq. 7) with:

$$\hat{S}(\phi,\mu,\ell) \equiv T(\phi) + (\ell-2)\int_0^\infty e^{-rt}e^{-\mu t}e^{-\lambda t}e^{-\lambda t}\lambda T(\phi)dt. \tag{16}$$

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

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:

$$\phi + \int_0^\infty e^{-rt} e^{-\mu t} \lambda \phi \, dt \geq T(\phi). \tag{17}$$

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. □

B.2.2 Optimal rate of forgiveness

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

$$\max_{\phi \geq 0, \mu \geq 0} \phi \quad \text{s.t. } W(\phi, \mu, \ell) \geq S(\phi, \mu, \ell) \; \forall \, \ell = 2, \ldots, n, \tag{18}$$

where we calculate that

$$W(\phi, \mu, \ell) = \phi + (\ell - 1)\frac{\lambda}{r + \mu}\phi + (n - \ell)\frac{\lambda\mu}{(r + \mu)(r + 2\mu)}\phi, \tag{19}$$

$$S(\phi, \mu, \ell) = T(\phi) + (\ell - 2)\frac{\lambda}{r + 2\lambda + \mu}T(\phi) + (n - \ell)\frac{\lambda\mu}{(r + 2\lambda + \mu)(r + \lambda + 2\mu)}T(\phi). \tag{20}$$

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

$$\frac{T(\phi)}{\phi} \leq \frac{1 + \frac{\lambda(r+4\mu)}{(r+\mu)(r+2\mu)}}{1 + \frac{2\lambda\mu}{(r+2\lambda+\mu)(r+\lambda+2\mu)}} \tag{21}$$

$$\frac{T(\phi)}{\phi} \leq \frac{1 + \frac{\lambda(2r+5\mu)}{(r+\mu)(r+2\mu)}}{1 + \frac{\lambda(r+\lambda+3\mu)}{(r+2\lambda+\mu)(r+\lambda+2\mu)}} \tag{22}$$

$$\frac{T(\phi)}{\phi} \leq \frac{(r + 2\lambda + \mu)(r + 3\lambda + \mu)}{(r + \mu)(r + 4\lambda + \mu)} \tag{23}$$

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 Eq. 21.

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 2. 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. Eq. 8).

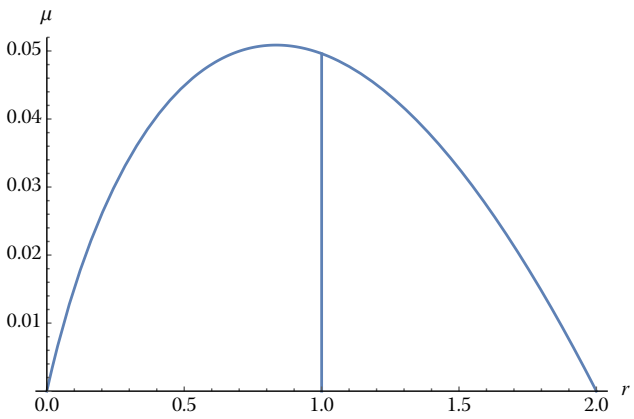

Figure 2. Optimal forgiveness rate μ vs. discount rate r, given $\lambda = 1$

B.2.3 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 4.

Define ϕ_C to bind the inequality

$$T(\phi) + \int_0^\infty e^{-rt} e^{-2\lambda t} \lambda T(\phi)\, dt \leq \phi + 2 \int_0^\infty e^{-rt} \lambda \phi\, dt. \tag{24}$$

Re-arranging, we see that

$$\frac{T(\phi_C)}{\phi_C} = \frac{(r+2\lambda)^2}{r(r+3\lambda)}, \tag{25}$$

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

$$V(\phi_R) = \phi_C + 2 \int_0^\infty e^{-rt} \lambda \phi_C\, dt. \tag{26}$$

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, (24) 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

$$T(\phi_C) \leq \phi_C + \frac{\lambda}{r}\phi_C + \underbrace{\int_0^\infty e^{-rt}e^{-2\lambda t}2\lambda\left(\frac{\lambda}{r}\phi_C\right)dt}_{\text{Resumption of Cooperation with } i} + \underbrace{\int_0^\infty e^{-rt}e^{-2\lambda t}\lambda T(\phi_R)\,dt}_{\text{Redemption Payment from } i}. \tag{27}$$

Re-arranging terms and using (24), 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

$$\begin{aligned} &T(\phi_C) + \int_0^\infty e^{-rt}e^{-2\lambda t}\lambda\left[T(\phi_R) + \int_\tau^\infty e^{-r(\tau-t)}e^{-2\lambda(\tau-t)}\lambda T(\phi_C)\,d\tau\right]dt \\ \leq &\phi_C + \frac{\lambda}{r}\phi_C + \int_0^\infty e^{-rt}e^{-2\lambda t}2\lambda\left(\frac{\lambda}{r}\phi_C\right)dt + \int_0^\infty e^{-rt}e^{-2\lambda t}\lambda T(\phi_R)\,dt. \end{aligned} \tag{28}$$

All terms involving $T(\phi_R)$ cancel out, and so simplifying (28), 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 (24), 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 (25). Therefore, player j does indeed have an incentive to continue communicating and cooperating with player k.