American Economic Review 2016, 106(8): 2329–2348
http://dx.doi.org/10.1257/aer.20130768

Ostracism and Forgiveness[†]

By S. Nageeb Ali and David A. Miller*

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. (*JEL* C73, D83, D85, O17, Z13)

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

* Ali: Pennsylvania State University, 411 Kern Building, University Park, PA 16802 (e-mail: nageeb@psu.edu); Miller: University of Michigan, 611 Tappan Avenue, Ann Arbor, MI 48109 (e-mail: econdm@umich.edu). 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 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. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[†] Go to http://dx.doi.org/10.1257/aer.20130768 to visit the article page for additional materials and author disclosure statement(s).

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.

I. 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 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 facilitates 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.

[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.

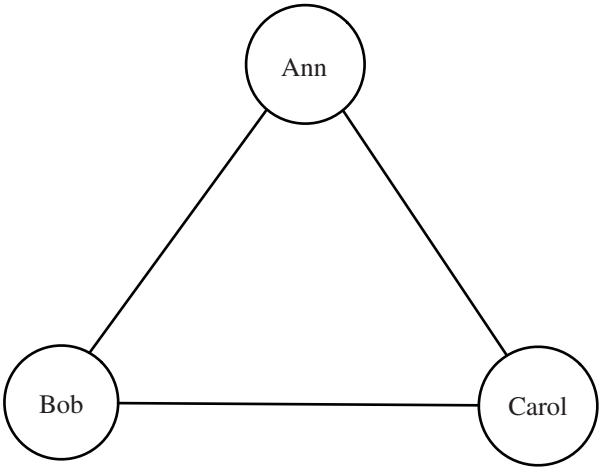

FIGURE 1

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:

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

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

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

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

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

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

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

where the inequality is Bob's truth-telling 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 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.

II. 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:

(i) *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.

(ii) *Stake Selection Stage*: Partners simultaneously propose *stakes*, and each proposal is a nonnegative real number. The minimum of their proposals is selected.

(iii) *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, nonnegative, 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

[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 player is the stakes and increasing these transfers increases the temptation

behavior at the limits of perfect patience ($r \to 0$), we 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 misaligned 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.

III. Main Results

Bilateral Enforcement.—We first formalize the lower bound on community enforcement that we introduced in Section I. 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 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 low stakes when their partners reveal little information, or work while guilty. The essence of permanent ostracism is to target permanent punishments toward the guilty and to communicate and

to shirk more than proportionally when players are risk-averse. In trade between a buyer and seller, the quality or 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.

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. 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 $\phi_{ij}\left(h_i^\tau, h_j^\tau\right)$ following truthful communication, and work at those stakes. We prove that for every pair of private histories $\left(h_i^\tau, h_j^\tau\right)$, the partners cooperate at stakes no greater than $\underline{\phi}$, the stakes from bilateral enforcement.[7]

Suppose otherwise. Consider a pair of private histories $\left(h_i^\tau, h_j^\tau\right)$ such that partners i and j are innocent and $\phi = \phi_{ij}\left(h_i^\tau, h_j^\tau\right) > \underline{\phi}$. 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 $\hat{\phi} = \phi_{ij}\left(\hat{h}_i^\tau, h_j^\tau\right)$, 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 bilateral 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. The online Appendix shows that the same bound applies approximately in discrete time: permanent ostracism can do better

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

than bilateral enforcement, but its advantage vanishes with the period length. In the online Appendix we also show that if partners communicate simultaneously, our results apply if certain unreasonable beliefs are ruled out and innocent players cooperate at stakes of 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 to 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 such 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 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

[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.

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 toward 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 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 the online Appendix, we optimally choose μ for four 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 the online Appendix for three 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 inherent lack of common knowledge about who is guilty.

IV. 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 the online Appendix, 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

[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.

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 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,

[10] For instance, if we expanded the stage game of Section II 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 $\underline{\phi}$.

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.

V. 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 toward 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]

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.

[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.

[13] See Hirshleifer and Rasmusen (1989); Bendor and Mookherjee (1990); Karlan et al. (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 the online Appendix, showing that if communication is mechanical, then permanent ostracism is optimal in the class of mutual effort equilibria.

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 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

[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.

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 targeted 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: 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 $\sigma_i = \left(\sigma_i^M, \sigma_i^S, \sigma_i^E\right)$, where in i's interaction with player j at time t given private history h_i^t, her message to player j is $\sigma_i^M(j, t, h_i^t, \emptyset) \in \mathcal{P}(h_i^t)$ if she communicates first and $\sigma_i^M\left(j, t, h_i^t, m_j^t\right) \in \mathcal{P}(h_i^t)$ if j communicates first (recall that a message is a set of interactions); her mixture over stakes proposals is $\sigma_i^S\left(j, t, h_i^t, m_i^t, m_j^t\right) \in \Delta[0, \infty)$; and her effort choice is $\sigma_i^E\left(j, t, h_i^t, m_i^t, m_j^t, \hat{\phi}_{ij}^t, \hat{\phi}_{ji}^t\right) \in \{W, S\}$. 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 $\sigma_j^E\left(k, t, \mathcal{E}_i^j(h_i^t, t^z) \cup m_j^{t^z}, m_j^{t^z}, m_k^{t^z}, \hat{\phi}_{jk}^t, \hat{\phi}_{kj}^t\right)$. 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:*

(*i*) *She sends the truthful message $m_i^t = h_i^t$.*

(*ii*) *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* $\sigma_j^S\big(i, t, m_i^t \cup m_j^t, m_j^t, m_i^t\big)$, *then i believes with probability* 1 *that j has not deviated, and i works.*

(*iii*) *If* $j \in \mathcal{G}_i\big(h_i^t \cup m_j^t\big)$, *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 constrain 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, $\mathcal{G}_i\big(h_i^t\big)$ 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 $\mathcal{E}_i\big(h_i^t\big)$ occurred, and $\{\omega^z\}_{z=0}^Z$ is a sequence of states such that $\omega^z \in \{0, 1\}^n$ for each z. Player $j \in \mathcal{I}_i\big(h_i^t\big)$ if $\omega_j^Z = 0$, and in $\mathcal{G}_i\big(h_i^t\big)$ 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 $\mathcal{E}_i\big(h_i^t\big)$ at time t^z, and j's effort choice is an observable deviation given what player i knows from his private history $\mathcal{E}_i\big(h_i^t\big)$ 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 $i \notin \mathcal{G}_i(h), j \notin \mathcal{G}_i\left(h \cup \hat{h}\right)$, 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

$$W(\phi, \mu, \ell) \equiv \phi + (\ell - 1) \int_0^\infty e^{-rt} e^{-\mu t} \lambda \phi \, dt + (n - \ell) \int_0^\infty e^{-rt} e^{-\mu t} \left(1 - e^{-\mu t}\right) \lambda \phi \, dt. \tag{A1}$$

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

$$S(\phi, \mu, \ell) \equiv T(\phi) + (\ell - 2)\mathcal{Z}(\phi) + (n - \ell)\left(\int_0^\infty e^{-rt} e^{-(2\mu+\lambda)t} \mu \, dt\right)\mathcal{Z}(\phi). \tag{A2}$$

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

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

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

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 $\mu \searrow 0$

$$\text{(A3)} \qquad \lim_{\mu \searrow 0} \frac{W(\underline{\phi}, \mu, 2) - S(\underline{\phi}, \mu, 2)}{\lambda\mu\underline{\phi}} = \frac{n-2}{r^2} - \frac{1}{r^2} - \frac{(n-2)(r+\lambda)}{r(r+\lambda)(r+2\lambda)}$$

$$= \frac{2\lambda(n-3) - r}{r^2(r+2\lambda)} > 0,$$

where the inequality follows from $r < 2\lambda(n-3)$. By L'Hôpital's Rule, the LHS is $\frac{1}{\lambda\underline{\phi}}\left(W_2(\underline{\phi}, 0, 2) - S_2(\underline{\phi}, 0, 2)\right)$. 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:

$$S(\underline{\phi}, 0, \ell) = T(\underline{\phi}) + \frac{(\ell-2)\lambda}{r+2\lambda} T(\underline{\phi}) = \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{r+\lambda}{r}\frac{(\ell-2)\lambda}{r+2\lambda}\underline{\phi}$$

$$< \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{(\ell-2)\lambda}{r}\underline{\phi} = W(\underline{\phi}, 0, \ell),$$

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 $\ell \in \{2, \ldots, n\}$ 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 from 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 afterward ensures that the new victim knows that he is not the first victim, so he will not spread news of i's guilt.

Because 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 $\ell \in \{2, \ldots, n\}$. 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

Ali, S. Nageeb, and David A. Miller. 2013. "Enforcing Cooperation in Networked Societies." http://www-personal.umich.edu/~econdm/resources/AliMiller-Networked.pdf (accessed June 16, 2016).

Banerjee, Abhijit V., and Esther Duflo. 2000. "Reputation Effects and the Limits of Contracting: A Study of the Indian Software Industry." *Quarterly Journal of Economics* 115 (3): 989–1017.

Bendor, Jonathan, and Dilip Mookherjee. 1990. "Norms, Third-Party Sanctions, and Cooperation." *Journal of Law, Economics, and Organization* 6 (1): 33–63.

Bernheim, B. Douglas, and Debraj Ray. 1989. "Collective Dynamic Consistency in Repeated Games." *Games and Economic Behavior* 1 (4): 295–326.

Bloch, Francis, Garance Genicot, and Debraj Ray. 2008. "Informal Insurance in Social Networks." *Journal of Economic Theory* 143 (1): 36–58.

Bowen, T. Renee, David M. Kreps, and Andrzej Skrzypacz. 2013. "Rules with Discretion and Local Information." *Quarterly Journal of Economics* 128 (3): 1273–1320.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution.* Princeton, NJ: Princeton University Press.

Coleman, James S. 1988. "Social Capital in the Creation of Human Capital." *American Journal of Sociology* 94 (S1): S95–S120.

Deb, Joyee. 2012. "Cooperation and Community Responsibility: A Folk Theorem for Repeated Matching Games with Names." https://sites.google.com/site/joyeedeb/follow-me/3_CooperationandCommunityResponsibility.pdf (accessed June 16, 2016).

Deb, Joyee, and Julio González-Díaz. 2014. "Enforcing Social Norms: Trust-Building and Community Enforcement." https://sites.google.com/site/joyeedeb/follow-me/1_Community_Enforcement_beyond_PD.pdf (accessed June 16, 2016).

Dixit, Avinash K. 2003. "Trade Expansion and Contract Enforcement." *Journal of Political Economy* 111 (6): 1293–1317.

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

Ellickson, Robert C. 1991. *Order without Law: How Neighbors Settle Disputes.* Cambridge, MA: Harvard University Press.

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

Fudenberg, Drew, and Eric S. Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54 (3): 533–54.

Ghosh, Parikshit, and Debraj Ray. 1996. "Cooperation in Community Interaction without Information Flows." *Review of Economic Studies* 63 (3): 491–519.

Green, Edward J., and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52 (1): 87–100.

Greif, Avner. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade.* New York: Cambridge University Press.

Grossman, Sanford J. 1981. "The Informational Role of Warranties and Private Disclosure about Product Quality." *Journal of Law and Economics* 24 (3): 461–83.

Hirshleifer, David, and Eric Rasmusen. 1989. "Cooperation in a Repeated Prisoners' Dilemma with Ostracism." *Journal of Economic Behavior & Organization* 12 (1): 87–106.

Jackson, Matthew O., Tomas Rodriguez-Barraquer, and Xu Tan. 2012. "Social Capital and Social Quilts: Network Patterns of Favor Exchange." *American Economic Review* 102 (5): 1857–97.

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

Karlan, Dean, Markus Möbius, Tanya Rosenblat, and Adam Szeidl. 2009. "Trust and Social Collateral." *Quarterly Journal of Economics* 124 (3): 1307–61.

Kranton, Rachel E. 1996. "The Formation of Cooperative Relationships." *Journal of Law, Economics, and Organization* 12 (1): 214–33.

Lippert, Steffen, and Giancarlo Spagnolo. 2011. "Networks of Relations and Word-of-Mouth Communication." *Games and Economic Behavior* 72 (1): 202–17.

McMillan, John, and Christopher Woodruff. 1999. "Interfirm Relationships and Informal Credit in Vietnam." *Quarterly Journal of Economics* 114 (4): 1285–1320.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics* 12 (2): 380–91.

Nowak, Martin A., and Karl Sigmund. 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393 (6685): 573–77.

Okuno-Fujiwara, Masahiro, and Andrew Postlewaite. 1995. "Social Norms and Random Matching Games." *Games and Economic Behavior* 9 (1): 79–109.

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

Posner, Eric A. 1996. "The Regulation of Groups: The Influence of Legal and Nonlegal Sanctions on Collective Action." *University of Chicago Law Review* 63: 133–97.

Raub, Werner, and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96 (3): 626–54.

Tirole, Jean. 1996. "A Theory of Collective Reputations (With Applications to the Persistence of Corruption and to Firm Quality)." *Review of Economic Studies* 63 (1): 1–22.

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

Wolitzky, Alexander. 2015. "Communication with Tokens in Repeated Games on Networks." *Theoretical Economics* 10 (1): 67–101.