Ostracism

S. Nageeb Ali and David A. Miller[*]

June 14, 2013

Abstract

Many communities rely upon *ostracism* to improve cooperation in bilateral relationships: if an individual deviates in one relationship, other innocent players come to learn about this deviation, and proceed to shun the guilty player while continuing to cooperate with each other. Typically, it is assumed that information spreads through word-of-mouth communication and that victims and third-parties have no incentive to lie about their experience with guilty players. We show, perhaps surprisingly, that innocent players may not have the incentive to communicate truthfully. Communication incentives are particularly severe in equilibria in which guilty players are permanently ostracized: such equilibria cannot support cooperation significantly beyond what each pair of players could obtain without community enforcement altogether. The challenge is that a victim of cheating prefers to cheat herself rather than to report her victimization to others. However, ostracism equilibria that feature forgiveness can foster truthful communication and thereby improve upon permanent ostracism. Our results suggest a new perspective on forgiveness and redemption in social norms.

[*] Ali: University of California, San Diego. 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, Debraj Ray, Larry Samuelson, Andy Skrzypacz, Joel Sobel, Adam Szeidl, Joel Watson, and Alex Wolitzky for helpful comments. Aislinn Bohren and Erik Lillethun 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 and Microsoft Research; Miller gratefully acknowledges financial support and hospitality from Microsoft Research, and thanks his former colleagues at UCSD for all their support.

Contents

1 Introduction

The value of a good reputation is a powerful motive for cooperation. Ann should be willing to cooperate more with each of her partners if she would lose her reputation with all of them should she shirk on any of them. *Ostracism* is a social norm 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 practicing ostracism: 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 that they should punish her?

A realistic way for communities to tackle this problem is to gossip. If Ann shirks on Bob, he will tell others about what she has done, and after hearing these complaints, others will punish her while continuing to work with each other. Numerous case studies of communities and markets document that word-of-mouth communication plays this role in enforcing medieval trade (Greif 1993, 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). Indeed, even modern electronic market platforms such as eBay rely on buyers communicating about their experience with sellers (Bajari and Hortaçsu 2004). The importance of communication and gossip in sustaining cooperation is emphasized across the social sciences and legal scholarship.[1]

Nonetheless, a fundamental question about this form of community enforcement remains unanswered: is it actually in the interests of Ann's victims to communicate truthfully about her deviation? Or does truthful reporting and continued cooperation among innocent players require further incentives? Existing approaches are not well suited to answer these questions because they rely upon centralized monitoring or assume truthful communication. While some approaches directly assume perfect monitoring across the community,[2] most approaches instead employ a "reputational label mechanism" in which 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.[3] Reputation label mechanisms correspond closely to centralized institutions that collect, store, and transmit information, but when such institutions are absent, information must spread by decentralized communication. Other approaches explicitly model the mechanics of word-of-

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

[2] For example, see Hirshleifer and Rasmusen (1989), Bendor and Mookherjee (1990, 2008), Karlan, Möbius, Rosenblat, and Szeidl (2009), and Jackson, Rodriguez-Barraquer, and Tan (2012).

[3] Reputation label mechanisms were formalized by Kandori (1992), Okuno-Fujiwara and Postlewaite (1995), and Tirole (1996). Apart from their prevalence in economics, sociology, and the law (Posner 1995, 1996), these mechanisms are also described in theoretical biology, where they are referred to as "image scoring" (e.g. Nowak and Sigmund 1998).

mouth communication, shedding light on how its speed and lag affect *cooperation incentives* (e.g. Raub and Weesie 1990; Klein 1992; Dixit 2003; Bloch, Genicot, and Ray 2008; Fainmesser and Goldberg 2011), but assume that communication is truthful. Our work is among the first to address the strategic incentives for truthful communication in community enforcement.

We study a networked society in which each link is an ongoing bilateral partnership with two-sided moral hazard. Each partnership meets at exponentially distributed arrival times to play a prisoners' dilemma at stakes that it chooses. Higher stakes generate greater payoffs for the partnership, but increase the temptation to shirk, and thus require stronger incentives. Each pair of partners perfectly observes everything that occurs within their own partnership, but third parties observe nothing—neither the timing of their meetings nor their behavior when they meet. Prior to selecting stakes, partners can communicate with each other about the behavior of others within the network. We study both communication based on "hard evidence," where players can conceal information but cannot fabricate or falsify it, and "soft," cheap talk communication. Throughout this paper, we focus on *ostracism* strategies, in which innocent players target punishments towards guilty players while working with those they believe to be innocent.

To understand the impact of strategic communication, we begin with two natural benchmarks. First, *bilateral enforcement* describes equilibria that do not use community enforcement or communication at all; for the prisoners' dilemma, this is the standard bilateral grim trigger punishment, played independently in each relationship. For the second benchmark, we consider an artificial setting in which players are constrained to reveal all their information truthfully, i.e., communication is mechanical. Permanent ostracism is easy to implement in this setting, since any player who has shirked must identify herself as guilty in all her future interactions, and all her partners can punish her. Since this equilibrium employs the harshest feasible threat against a deviator, it supports at least as much cooperation as any other Nash equilibrium, and is isomorphic to the most cooperative equilibrium in a model in which all interactions are publicly observed.

But what happens when individuals strategically choose what to reveal? At first glance, it might appear that at least some truthful communication should be incentive compatible: while a guilty player has every reason to conceal her own misdeeds, innocent players should have aligned interests in revealing and punishing the guilty. Our main result shows that this intuition is wrong. Instead, if guilty players are to be permanently ostracized, then their victims have a strong incentive to conceal evidence of their victimization, and themselves to then shirk on other innocent players. This strategic motive not only pushes permanent ostracism away from the mechanical communication benchmark but it also guarantees that the players are no better off than under bilateral enforcement. In other words, truthful communication is incentive compatible in permanent ostracism only if community enforcement is entirely redundant.

This stark negative result applies for every discount factor and every network, and even if communication takes the form of verifiable disclosure. Here is why: consider a permanent ostracism equilibrium and the relationship between two players, Bob and Carol. Suppose to the contrary that they cooperate at a higher level than is attainable under bilateral enforcement. Each player's incentives to work must then be partly driven by the threat of punishments from others. Now consider a private history at which Bob knows that everyone other than Carol has shirked and should be ostracized: should he tell her the truth? Because all other players are shirking, Bob's only remaining incentive to work arises from his continuation play with Carol, just as under bilateral enforcement. Therefore Bob strictly prefers to conceal his information from Carol and shirk at the equilibrium stakes, rather than tell the truth and then reduce their stakes to be able to continue cooperating. That is, permanent ostracism destroys Bob's incentives to communicate truthfully off the equilibrium path.

This result emerges with greatest clarity in prisoners' dilemmas, but an analogue applies for general bilateral stage games. We show that for permanent ostracism equilibria that are symmetric in each relationship, each player's equilibrium payoff in each relationship is bounded above by the highest payoff he can attain in any bilateral enforcement equilibrium in that relationship. Asymmetric permanent ostracism equilibria have more flexibility, but a bound on payoffs, arising from bilateral enforcement equilibria, still applies regardless of the network and the population size. Thus, the incentive to conceal information imposes a limit on how much can be attained through permanent ostracism even in general games.

There is a tension between our negative result for permanent ostracism in theory and the prevalence of ostracism in communities and markets. Yet, permanent ostracism omits an important feature of real-world community enforcement: players are often forgiven and only temporarily ostracized (Ostrom 1990; Greif 2006). We find that forgiveness of guilty players encourages innocent victims to communicate truthfully. We construct a temporary ostracism equilibria in which players are forgiven at random times. If players are sufficiently patient or society is sufficiently large, then innocent players communicate truthfully and cooperate with each other at levels beyond those attainable under bilateral enforcement. Our results identify a new motive for temporary punishments in community enforcement: it maintains "social collateral" that fosters communication and cooperation among innocent players even when others are guilty.

Our interest in communication and ostracism should be contrasted with community enforcement schemes without information transmission. The natural juxtaposition is to *contagion equilibria*, which were introduced for anonymous random matching environments by Kandori (1992) and Ellison (1994), and applied to social networks by Ali and Miller (2013).[4] In this equilibrium,

[4] Harrington (1995) uses contagion to show that relationships with low frequencies of interaction can be supported using relationships that interact more frequently. While not focusing on contagion, Takahashi (2010) shows that cooperation can be sustained in repeated prisoners' dilemmas if all that is observed are partners' past play. Deb (2012) offers a general folk theorem for anonymous random matching environments, building on notions of collective reputation

each player shirks on all others once any player shirks on her, and so an initial deviation triggers a contagion that spreads through society. Contagion offers a useful benchmark for attainable payoffs in the absence of institutions or communication networks, but it also represents a form of *collective reputation* in which cooperation is so fragile that a single infraction by Ann destroys Bob's trust in all of his partners. Ostracism, by contrast, reflects the principle that Bob trusts partners who has never shirked on anyone (to his knowledge) while punishing those who do so: reputations are entirely at the *individual* level. These are two extreme points, and one can envision community enforcement norms that blend individual and community responsibility. We model this spectrum as *permanent ostracism of depth d*: an innocent player communicates truthfully to other innocent players and ostracizes guilty players so long as he knows of no more than d guilty players, and otherwise, shirks on all his partners. Such equilibria improve upon permanent ostracism but are bound, in terms of average stakes, by contagion with $n - d$ players.

We revisit two commonly studied applications through the lens of communication incentives. Section 6.1 analyzes networked markets like eBay, where buyers and sellers can communicate about their experiences with the other side. Our logic on the futility of permanent ostracism applies here if there is moral hazard on both sides—e.g., if sellers can shirk on quality and buyers can shirk on payment. In contrast, permanent ostracism is efficient if only one side has an incentive to deviate from the trading arrangement. For online trading platforms as well as labor markets, this result supports the common practice of structuring payment and trade to take place sequentially rather than simultaneously. When a buyer has to pay first, he has no incentive to deviate, and can be counted upon to communicate truthfully about the seller. Our result also highlights how community enforcement can be complemented by legal institutions or for-profit enforcement intermediaries that can enforce trade for one side of the market.

Section 6.2 studies informal risk sharing on a network, building on analyses of self-enforcing arrangements in which deviations by any player are perfectly observed, and are punished by autarky while others continue to share risk (e.g. Kocherlakota 1996; Ligon, Thomas, and Worrall 2002). Punishing a player by excluding him from risk sharing and restricting him to autarky forever is an analogue of permanent ostracism.[5] We show that if a player's compliance with the risk sharing arrangement is observed only by the partner to whom she is meant to transfer wealth, permanent ostracism fails to improve upon bilateral enforcement. The implications can be quantitatively substantive: there exist ranges of discount factors for which permanent ostracism with mechanical communication can support full insurance, but permanent ostracism with strategic communication fails to support any risk sharing. Interestingly, these limits are similar to those

and community responsibility.

[5] Bloch, Genicot, and Ray (2008) study intermediate notions of exclusion in which players that are within a certain distance of the victim are those who cut ties with the deviating player.

that may arise when ostracism is forced to be robust to coalitional deviations (Genicot and Ray 2003; Ambrus, Möbius, and Szeidl 2013).

Our analysis throughout is simplified by several modeling innovations. We assume that players interact at random privately observed times, which contrasts with standard "random matching" games in which each player is known to have interacted in each period. This private information generates a non-trivial communication incentive because a player can conceal an interaction without her partner knowing that anything has been concealed. Incentives would differ if interaction times were public: the familiar force of unraveling would compel a player to reveal all details of her past interactions, since her partner could rationally consider her failure to disclose evidence to indicate that she has deviated. In that case, strategic communication is as effective as permanent ostracism with mechanical communication. However, we show that these equilibria are fragile: if there is even the slightest chance that some interactions happen at privately observed times, our negative result is restored.[6]

That players can endogenously choose their level of cooperation ("variable stakes") permits both a straightforward comparison of equilibria for a fixed discount rate and offers a more flexible technology for cooperation. The standard prisoners' dilemma (with fixed stakes) obscures incentives in ostracism by imposing severe technological restrictions. As we discuss in Remark 1, were the stakes of each relationship fixed, permanent ostracism could do no better than bilateral enforcement even if communication were mechanical—either players would be sufficiently patient to cooperate under bilateral enforcement or they would be unwilling to do so when only two innocent players remain. In contrast, variable stakes enable partners to adjust the terms of their relationship based on their mutual history, in particular to reduce their stakes once other players are ostracized. Since many commonly studied applications do permit players to endogenously adjust their relationships, the variable stakes framework shifts focus from constraints in the *technology* of cooperation to the challenge of providing *incentives* for truthful communication.[7]

Our focus on ostracism is motivated by the many papers that have used such equilibria with perfect monitoring. *Inter alia*, Bendor and Mookherjee (1990, 2008) study such social sanctions when each individual simultaneously interacts in a number of bilateral prisoners' dilemmas. Like multimarket collusion (Bernheim and Whinston 1990), third-party sanctions are effective when there is slack in bilateral incentives in some relationships that can be used to subsidize others, and

[6] The issue is reminiscent of how unraveling breaks once the sender may be uninformed (Shin 1994) . In our setting, as the period length vanishes, the probability of no interaction and thus of a sender having no information to transmit converges to 1.

[7] Ghosh and Ray (1996) and Kranton (1996) are the first to study variable stakes frameworks in community enforcement, and they elucidate a different force: building cooperation over time screens out myopic players and deters patient players from shirking and re-matching. Their focus is not on ostracism or communication, but on generating cooperation incentives in the absence of information transmission.

when such slack is lacking, can do no better than bilateral enforcement.[8]

Three recent studies share our interest in understanding when information germane to community enforcement is credibly communicated, although none of them study ostracism. Lippert and Spagnolo (2011) study the role of networks and communication in an environment in which each individual plays separate, fixed-stakes prisoners' dilemmas with all of her partners simultaneously in each period, and the payoffs from these prisoners' dilemmas are heterogeneous. Bowen, Kreps, and Skrzypacz (2013) study favor exchanges when actions and messages are public but only its donor and recipient observe its payoffs. They focus on subtle timing issues: if the recipient of a favor can publicly excuse the donor from doing the favor in that period, should her message precede or succeed the action of her partner? Wolitzky (2013) studies the role of communication and money in settings with private monitoring, focusing in particular on when money endowments can be used towards community enforcement.

2 An Example

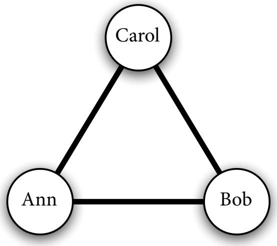

FIGURE 1. A SOCIETY OF THREE PLAYERS

We illustrate the failure of permanent ostracism in the society of three players depicted in Figure 1. Each "link" embodies a long-term partnership between a pair of players. In continuous time, each partnership is recognized according to an independent Poisson process with constant intensity $\lambda = 1$. Whenever a partnership (e.g., "AnnBob") is recognized, the partners (Ann and Bob) communicate sequentially in random order, and then choose effort levels simultaneously. They discount payoffs at rate $r = 1$. Each partnership tackles a two-sided moral hazard problem, as in Ghosh and Ray (1996): each player's effort choice, $a \geq 0$, comes at a cost of a^2 but confers a benefit of $a + a^2$ to her partner. The pair commonly observes their effort choices.

Bilateral Enforcement: The first benchmark to study is the level of cooperation that is incentive compatible if all enforcement is *bilateral*; i.e., each pair's behavior is measurable with respect to its mutual history and uninfluenced by the behavior in other partnerships. Consider a grim trigger

[8] Hirshleifer and Rasmusen (1989) study a public goods setting, and show that ostracism is effective even if costly.

profile in which each player exerts effort a with her partner if and only if (a, a) was chosen in every prior meeting between them; otherwise, each *shirks* (i.e., exerts an effort of 0). The equilibrium path incentive constraint is

$$a + a^2 \leq a + \int_0^\infty e^{-rt} \lambda a \, dt, \tag{1}$$

in which the left-hand side is the payoff from shirking to effort of 0, and the right-hand side is the sum of the immediate payoff from mutually working at effort of a and the discounted continuation payoff from staying on the equilibrium path. The highest effort enforceable in a bilateral enforcement equilibrium is $\underline{a} = 1$.

Permanent ostracism with mechanical communication: Community enforcement is generally believed to enhance cooperation by leveraging the ties that both partners have to the third party: if each player were to believe that she would be punished by both partners when she shirks on one, she has a stronger incentive to work. This logic manifests when players are constrained to communicate truthfully, regardless of their self interest. Consider a strategy profile in which on the path of play, each player is meant to exert effort a whenever she meets a partner. If Ann shirks on Bob, then both she and Bob are mechanically constrained to reporting it to Carol. Therefore, both Bob and Carol will exert zero effort in their relationships with her while continuing to work with each other at mutual effort level $\underline{a}$, which is incentive compatible given that they have ostracized Ann. The equilibrium path incentive constraint is

$$a + a^2 \leq a + 2 \int_0^\infty e^{-rt} \lambda a \, dt, \tag{2}$$

which binds at an effort level of $a = 2$. Permanent ostracism with mechanical communication raises the level of cooperation in each partnership.

Permanent ostracism with strategic communication: Our interest is in a setting in which monitoring information must be strategically transmitted by the players: only Ann and Bob observe what happens between them, and Carol learns about it *only if* it is voluntarily communicated to her. Specifically, whenever a pair meets, before they choose their effort levels, each of them can reveal any subset of her past interactions with anyone. Communication takes the form of a disclosure game (Grossman 1981; Milgrom 1981) in which players can either disclose or conceal, but neither fabricate nor distort, information about their past interactions. Although this communication with hard evidence permits fewer deviations than cheap talk, the incentive to manipulate information nevertheless is crippling.

To see why, consider a strategy profile similar to that before, in which innocent players are supposed to cooperate and communicate truthfully with each other. Consider Ann's incentives to work with Bob on the equilibrium path. If Ann shirks on Bob, she assumes that Bob will tell Carol after which point, she will obtain no payoffs from that partnership. Thus, if Ann shirks on Bob, she can expect to obtain payoffs from Carol if and only if she meets Carol before Bob and Carol meet, and she conceals from Carol that she had shirked on Bob in the past. This changes the equilibrium path incentive constraint to

$$a + a^2 + (a + a^2) \int_0^\infty e^{-rt} e^{-2\lambda t} \lambda \, dt \leq a + 2 \int_0^\infty e^{-rt} \lambda a \, dt. \tag{3}$$

The left-hand side is Ann's payoff from shirking immediately on Bob and her discounted expected payoff from possibly being able to shirk on Carol before Carol meets Bob. The right-hand side is her payoff from working immediately with Bob and in the future with both partners. The highest effort that satisfies this incentive condition is $\frac{5}{4}$. So the incentive constraints that arise on the equilibrium path allow for higher cooperation than is possible under bilateral enforcement.

While effort of $\frac{5}{4}$ is attainable in Nash equilibrium (wherein only equilibrium path incentive constraints apply), we find that it fails sequential rationality. The challenge is that Bob prefers to conceal the truth about Ann from Carol. If he discloses the truth, he and Carol will permanently ostracize Ann. Then, since Ann will no longer be available to assist them in enforcing effort of $\frac{5}{4}$ in their relationship, Bob and Carol can at best revert to bilateral enforcement, under which their effort level cannot exceed $\underline{a} = 1$. But Bob can deviate by conditioning what he reveals contingent on what Carol reveals, at least if Carol randomly speaks first. If Carol reveals that Ann has shirked on her, then Bob is indifferent between revealing and concealing the truth. On the other hand, if Carol does not reveal that Ann has shirked on her, Bob can conceal the truth. Then he expects Carol to work at the "on path" effort. So Bob will tell the truth only if his payoff from truthful disclosure and mutual effort at $\underline{a}$ outweighs that from concealing and shirking:

$$a + a^2 \leq \underline{a} + \int_0^\infty e^{-rt} \lambda \underline{a} \, dt = \underline{a} + \underline{a}^2. \tag{4}$$

Therefore, Bob will tell the truth to Carol if and only if the equilibrium path action is $a \leq 1$; i.e., truthful communication is sequentially rational only if the equilibrium effort level is no greater than that of bilateral enforcement.

Temporary Ostracism: The problem with permanent ostracism is that once Ann is ostracized, Bob and Carol are compelled to reduce the stakes of their partnership. This destruction of value impedes Bob's incentive to communicate truthfully to Carol. Temporary ostracism overcomes this challenge by re-admitting guilty players at a future random time. Public randomization devices offer a straightforward way to generate forgiveness: suppose that corresponding to each player i, there is a public Poisson signal with constant intensity μ that when realized, all others immediately forgive player i.[9] Using these public signals to improve cooperation beyond permanent ostracism is delicate: the forgiveness must be sufficiently fast that Bob and Carol can rely upon community enforcement even if all other players are currently guilty, but sufficiently slow that Bob himself lacks the incentive to shirk and be later forgiven.

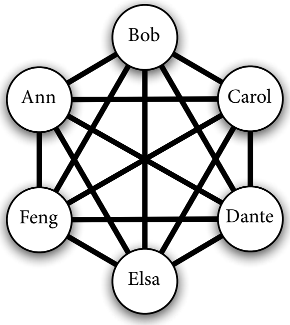

FIGURE 2. A SOCIETY OF SIX PLAYERS

Such forgiveness cannot improve upon permanent ostracism in three player examples, but can do so with more players. We show in the society of six players described in Figure 2 that temporary ostracism with even minimal communication performs better than permanent ostracism.

Suppose that if the first player that Ann shirks on is Bob, then Bob spreads that information to others, but other innocent victims and third-parties do not spread this information. Consider the incentive constraint analogous to (4): suppose that Bob privately knows that all players but Carol are guilty, and that Carol does not know this to be the case. With a forgiveness rate of μ, and substituting $\lambda = r = 1$, Bob's incentive compatibility condition can be shown to be

$$(a + a^2)\left(1 + \frac{4\mu}{(2+2\mu)(3+\mu)}\right) \leq a\left(1 + \frac{1}{1+\mu} + \frac{4\mu}{(1+\mu)(1+2\mu)}\right). \tag{5}$$

Equilibrium cooperation is maximized at $\mu \approx 0.15$, which generates $a \approx 1.1$, exceeding the cooperation from permanent ostracism. We view this as a new rationale for forgiveness in ostracism: it maintains social collateral that fosters truthful communication even when others are guilty.

[9] We use public signals for illustration; the same can be accomplished using verifiable randomization devices that are privately observed by each pair of players.

3 Model

Society is composed of a finite set of players $\mathcal{N} \equiv \{1, 2, \ldots, n+1\}$, with $n \geq 2$. Each pair of individuals engages in a bilateral partnership, and we denote the total number of partnerships by $G \equiv \frac{n(n+1)}{2}$. We refer to a partnership between players i and j as "link $\{ij\}$."

We study both discrete and continuous time games:

- In the discrete time game, players may interact at periods in $\mathcal{T}(\Delta) \equiv \{0, \Delta, 2\Delta, \ldots\}$, where $\Delta > 0$ specifies the period length, and $\lambda > 0$ is a parameter that specifies the frequency of interaction. In each period, society is *inactive* with probability $e^{-G\lambda\Delta}$, in which case no link is selected; or it is *active* with probability $1 - e^{-G\lambda\Delta}$, in which case a single link is selected. Conditional on society being active, each link is selected with equal probability. We write $p_\Delta \equiv \frac{1-e^{-G\lambda\Delta}}{G}$ for the probability that a particular link is selected.
- In the continuous time game, each link meets at random times in $\mathcal{T}(0) \equiv [0, \infty)$ distributed according to an independent Poisson process of constant intensity $\lambda > 0$.

The continuous time game is the limit of the discrete time game as $\Delta \to 0$. A feature common to both settings is that there is zero probability that multiple links are selected simultaneously.

When link $\{ij\}$ meets, partners i and j choose what to communicate to each other, what stakes to propose for this interaction, and whether to work or shirk. Payoffs accrued at real time t are discounted by e^{-rt}, and we write $\delta \equiv e^{-r\Delta}$ for the per-period discount factor. The extensive form of each interaction is:

1. *Communication Stage*: The partners send their messages sequentially. Independently of the past, each partner is equally likely to be selected to speak first.
2. *Stake Selection Stage*: The partners simultaneously propose their stakes, and the minimum of the two proposals is implemented. Player i's proposal is $\hat{\phi}^t_{ij}$ in $[0, \infty)$, and the selected stakes are $\phi^t_{ij} \equiv \min\{\hat{\phi}^t_{ij}, \hat{\phi}^t_{ji}\}$.
3. *Effort Stage*: Each partner simultaneously chooses to work (W) or shirk (S). If the stakes are ϕ, then their payoffs correspond to the prisoner's dilemma depicted in Figure 3.

Mnemonically, T is the *temptation* to shirk and V is the loss from being the *victim* of someone else's shirking. Both T and V are smooth, non-negative, and strictly increasing functions that satisfy $T(0) = V(0) = 0$, and $T(\phi) > \phi$ for $\phi > 0$. Central to our approach is

Assumption 1 (Increasing Temptation). *The temptation to shirk, T, is strictly convex and satisfies $T'(0) = 1$ and $\lim_{\phi\to\infty} T'(\phi) = \infty$.*

		Player j	
		W	S
Player	W	ϕ, ϕ	$-V(\phi), T(\phi)$
i	S	$T(\phi), -V(\phi)$	$0, 0$

FIGURE 3. The prisoners' dilemma with stakes ϕ

Information and Communication: Our setting features local monitoring: when players i and j interact at time t, they are the only ones to directly observe that they have met, their communications, the stakes that each proposes, and their effort choices. We endow players with a rich language that allows them to transmit their private histories to each other. An *interaction* between players i and j at time t, denoted by z^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 announces, and their effort choices. Player i's private history at time t, denoted h_i^t, is the set of all interactions that she has had strictly before time t. At history h_i^t, $M(h_i^t)$ denotes the set of available messages to player i. We denote by $\mathcal{P}(h)$ the power set of history h and $\mathcal{H}_i^t$ denotes the set of all feasible private histories for player i at time t. We focus on three different modes of communication:

Definition 1. *Communication is*

- ***mechanical*** *if for every history h_i^t, $M(h_i^t) = \{h_i^t\}$,*
- ***evidentiary*** *if for every history h_i^t, $M(h_i^t) = \mathcal{P}(h_i^t)$,*
- ***cheap talk*** *if for every history h_i^t, $M(h_i^t) = \mathcal{H}_i^t$.*

With mechanical communication, each player is constrained to communicate her entire past to her partner. Even though it involves some delay in information diffusion, mechanical communication is tantamount to public monitoring, since players always learn the information they need by the time they need to act on it. Evidentiary communication models the disclosure game introduced in Section 2: a player chooses which of her interactions to reveal. Information is "hard" insofar as players can neither distort their interactions nor fabricate them. The only evidence that a player lacks is being able to prove that she *did not* interact with anyone at a particular time. Cheap talk, by contrast, models a setting in which all information is "soft," and therefore, permits more deviations and nuanced messages. We focus on evidentiary communication since our negative results extend to cheap talk communication.

The information contained in history h_i^t extends beyond the interactions that i has participated in *first-hand* up to that time; it also includes interactions that she learns about from others (including those interactions that they too have learned about through communication). We denote the set of all interactions recorded in history h_i^t by $\mathcal{E}(h_i^t)$.

Strategies and Equilibria: The above defines a dynamic game. A behavioral strategy is a function $\sigma_i = (\sigma_i^{\mathrm{M}}, \sigma_i^{\mathrm{S}}, \sigma_i^{\mathrm{A}})$ such that in player i's time t interaction with player j,

- $\sigma_i^{\mathrm{M}}\left(j, h_i^t, m\right)$ is his message to player j in which $m = \emptyset$ if player i communicates first, and $m = m_j^t$ if player j communicates first,
- $\sigma_i^{\mathrm{S}}\left(j, h_i^t, m_i^t, m_j^t\right) \in \mathcal{B}[0, \infty)$ is his stakes proposal,
- $\sigma_i^{\mathrm{A}}\left(j, h_i^t, m_i^t, m_j^t, \hat{\phi}_{ij}^t, \hat{\phi}_{ji}^t\right) \in \mathcal{B}\left\{W, S\right\}$ is his action choice.[10]

We study weak perfect Bayesian equilibria of the game, with the restriction that each player's stake proposal is uniformly bounded across histories.[11] Henceforth, we refer to these as *equilibria*. We describe an equilibrium as being *mutual effort* if all players work on the equilibrium path.

4 Permanent Ostracism

4.1 Benchmarks

Private Bilateral Enforcement: A lower bound for community enforcement is the level of cooperation that is incentive compatible without it. We say that enforcement is *bilateral* if behavior in the ij partnership depends only on the past interactions within that partnership. A notable bilateral enforcement scheme is "bilateral grim trigger," in which two partners always work at stakes ϕ as long as they have always done so in the past, but otherwise set stakes of 0. This strategy profile generates the following incentive constraint:

$$T(\phi) \leq \phi + \frac{\delta p_\Delta}{1-\delta}\phi. \tag{Bilateral IC}$$

By Assumption 1, there exists a unique strictly positive $\underline{\phi}(\Delta)$ that binds this constraint, and this is the highest level of cooperation that can be supported by any bilateral enforcement equilibrium.

Proposition 1. *There exists a bilateral enforcement equilibrium in which on the equilibrium path all players work at stakes $\underline{\phi}(\Delta)$ that solve*

$$\frac{T(\phi)}{\phi} = 1 + \frac{\delta p_\Delta}{1-\delta} \xrightarrow{\Delta \to 0} 1 + \frac{\lambda}{r}. \tag{6}$$

No bilateral equilibrium supports mutual effort at stakes exceeding $\underline{\phi}(\Delta)$ at any history.

[10] For a space X, we use $\mathcal{B}(X)$ to denote the set of probability measures on X.

[11] The restriction eliminates equilibria in which the stake proposals explode to infinity; it is equivalent to imposing an upper bound on feasible stakes at the stake selection stage.

Mechanical Communication: In contrast, suppose that players are mechanically constrained to reveal their entire past histories. Then monitoring is effectively public, because if a player shirks, both she and her victim are compelled to reveal it to all third parties.

Ostracism implements the greatest possible cooperation in this setting. Each player is initially considered to be *innocent*, but is deemed *guilty* at a history if she has deviated. A guilty player is punished by others setting zero stakes in their relationship with her, while innocent players work with each other at strictly positive stakes. To assess who is guilty, a player with history h examines her full record of interactions, $\mathcal{E}(h)$, and finds player i to be guilty if there is an interaction z^{τ} in $\mathcal{E}(h)$ in which player i shirked and her partner worked, or at which player i proposed the wrong stakes. The set of guilty players at history h is $\mathcal{G}(h)$, and its complement is $\mathcal{I}(h)$.

Definition 2. *In a* ***permanent ostracism*** *strategy profile, when players i and j meet at time t and following histories (h_i^t, h_j^t), they work with each other at strictly positive stakes if $\{i, j\} \subset \mathcal{I}(h_i^t \cup h_j^t)$; otherwise, each player announces stakes of* 0 *and works.*

Innocent players who learn that others have shirked must adjust their stakes so that they still have the incentive to work. A particular form of permanent ostracism is that in which for every link $\{ij\}$, the stakes at which innocent players i and j work when their joint history is $h_i^t \cup h_j^t$ depends only on $\mathcal{I}(h_i^t \cup h_j^t)$, and is invariant to time and other historical details; following Kandori (1992), we call such strategy profiles *straightforward*.

Incentives in a straightforward equilibrium are easily described. Consider a history at which the set of innocent players is $\mathcal{I}$, including players i and j. If $\phi_{ik}(\mathcal{I})$ are the stakes that players i and k set when the set of innocent players is $\mathcal{I}$, then player i's incentive constraint at history h_i^t when facing player j is:

$$T\big(\phi_{ij}(\mathcal{I})\big) \leq \phi_{ij}(\mathcal{I}) + \sum_{k \in \mathcal{N} \setminus \{i\}} \frac{\delta p_{\Delta}}{1-\delta} \phi_{ik}(\mathcal{I}). \qquad \text{(Mechanical IC)}$$

Using Assumption 1, we construct a straightforward permanent ostracism equilibrium with mechanical communication such that Mechanical IC binds at every history, and show that cooperation in this equilibrium is the highest of all mutual effort equilibria.

Proposition 2. *If communication is mechanical, there exists a straightforward permanent ostracism equilibrium in which on the equilibrium path, partners work at stakes $\overline{\phi}(\Delta)$ that solve*

$$\frac{T(\phi)}{\phi} = 1 + \frac{n \delta p_{\Delta}}{1-\delta} \xrightarrow{\Delta \to 0} 1 + n\frac{\lambda}{r}. \qquad (7)$$

No mutual effort equilibrium supports cooperation at stakes that exceed $\overline{\phi}(\Delta)$ in any history.

This result characterizes the greatest cooperation that is attainable when communication is mechanical. It coincides with the best mutual effort equilibrium in a setting with perfect monitoring, and it outperforms bilateral enforcement so long as there are three or more players. The optimality of this permanent ostracism equilibrium among all mutual effort equilibria emerges from combining the worst stick for shirking with the best carrot for working.[12]

Before proceeding to strategic incentives that arise in communication, we remark on the distinction between variable and fixed stakes that we alluded to earlier.

Remark 1 (Variable vs. Fixed Stakes)**.** In a "fixed stakes" environment in which the prisoner's dilemma involves some exogenously fixed ϕ, permanent ostracism could do no better than bilateral enforcement even with mechanical communication. If r is sufficiently low that Bilateral IC is satisfied, then bilateral enforcement is sufficient to support the same level of effort as permanent ostracism. Otherwise, permanent ostracism isn't an equilibrium since two innocent players would not cooperate when they are the only innocent players remaining. Thus, for every $r > 0$, permanent ostracism is exactly as effective as bilateral enforcement.[13] We view fixed stakes as a restrictive technological assumption, and relaxing it emphasizes that the challenge with permanent ostracism is offering incentives for truthful communication.

4.2 Limits of Strategic Communication

Mechanical communication requires players not only to report on others, but also to confess themselves if they are guilty. Naturally, once communication is strategic, Ann will not voluntarily confess to Carol that she shirked on Bob, since doing so would result in Carol immediately punishing her. One might hope that Bob would still be willing to report truthfully about Ann's deviation, but we find that this fails.

We begin by formally defining permanent ostracism with strategic communication. Given player i's private history h_i^t, and for a time $\tau \leq t$, let $m_j(h_i^t, \tau)$ denote the subset of interactions in $\mathcal{E}(h_i^t)$ that happened strictly before τ and in which player j was active. Player i knows that player j has concealed information if for some interaction z^τ, m_j^τ excludes an interaction from $m_j(h_i^t, \tau)$.

Definition 3. *In a permanent ostracism strategy profile, if player i meets player j at time t, following a history h_i^t, her behavior is*

1. *If $\{i, j\} \subset \mathcal{I}(h_i^t)$, then player i reveals history h_i^t. If $m_j(h_i^t, t) \subseteq m_j^t$ and $j \in \mathcal{I}\left(h_i^t \cup m_j^t\right)$, then i believes with probability* 1 *that j is innocent, proposes strictly positive stakes, and works.*

[12] While we have restricted our discussion to pure strategy equilibria, the same argument trivially applies to show that mutual effort equilibria in which players randomize in their stake proposals can also do no better. This benchmark also extends if the network is "incomplete": there, the maximal stakes in partnership $\{ij\}$ would correspond to substituting for n with the minimum of the degrees between players i and j.

[13] This issue persists even when relationships are heterogeneous, as in Lippert and Spagnolo (2011).

2. *If $j \in \mathcal{G}(h_i^t)$, player i sends message $m_i^t = \emptyset$, proposes stakes of* 0, *and works.*

In a permanent ostracism equilibrium, player i knows player j to be guilty if he has evidence that player j has ever deviated, whether by shirking, by proposing the wrong stakes, or by concealing information she should have revealed. If player i has no evidence of player j's guilt, he should believe that she is innocent with probability 1. An innocent player works with those she believes to be innocent, and permanently punishes those she knows are guilty. Were our interest merely in rational behavior on the path of play, we would find Nash equilibria in which players cooperate at stakes beyond those of bilateral enforcement, enforced by the threat of permanent ostracism. However, it is not sequentially rational for innocent players to communicate truthfully after being victimized, unless the stakes are at or below the bilateral enforcement benchmark. We first establish this contradiction for straightforward permanent ostracism equilibria.

Theorem 1. *No straightforward permanent ostracism equilibrium supports stakes in any partnership that exceed $\underline{\phi}(\Delta)$, regardless of players' patience r and period length Δ.*

Proof. We provide the argument for $\Delta > 0$, since the result for $\Delta = 0$ is a special case of Theorem 2. We construct a history whose communication incentive binds the equilibrium path stakes to be no greater than $\underline{\phi}(\Delta)$. Consider times t and $\tau = t + n\Delta$, a pair of players $\{i, j\}$ who meet at time t, and histories such that $\mathcal{I}\left(h_i^t \cup h_j^t\right) = \mathcal{N}$. Suppose that in the subsequent $(n-1)\Delta$ periods, player i meets every player other than j, all of whom shirk on him, and then players i and j meet at time τ. If player i reveals h_i^τ to player j, his best possible continuation play in a mutual effort equilibrium is cooperating at $\underline{\phi}(\Delta)$ forever with player j. If player i instead reports h_i^t and conceals $h_i^\tau \backslash h_i^t$, then he has the opportunity to shirk at the equilibrium path stakes ϕ_{ij}. Thus player i communicates truthfully if and only if

$$T(\phi_{ij}) \leq \underline{\phi}(\Delta) + \frac{\delta p_\Delta}{1-\delta}\underline{\phi}(\Delta).$$

Because $\underline{\phi}(\Delta)$ binds Bilateral IC (on p. 12), it follows that the right-hand side equals $T\left(\underline{\phi}(\Delta)\right)$. The conclusion then follows from T being a strictly increasing function. □

The result has a clear intuition: if two partners can support higher stakes in their partnership on the equilibrium path than through bilateral enforcement, then their other relationships must serve as social collateral for each of them to cooperate. Yet, when one of them privately observes third parties defecting while the other does not, the informed partner knows that social collateral is lost, giving her the incentive to conceal information and shirk. This force causes permanent ostracism in even large societies to unravel to stakes supportable by bilateral enforcement—for which community enforcement is entirely unnecessary.

Theorem 1 pertains to straightforward equilibria. But one may wish to consider equilibria in which players condition their behavior on more than simply the set of innocent players. For example, perhaps the stakes between innocent players should increase in the quantity of information shared. We show in an example that non-straightforward permanent ostracism equilibria can do better in discrete time than bilateral enforcement, but these gains vanish as $\Delta \to 0$.

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

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

For every $\Delta > 0$, this inequality is slack at $\phi = \underline{\phi}(\Delta)$, 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 $\underline{\phi}(\Delta)$. This permanent ostracism equilibrium can support cooperation at levels higher than bilateral enforcement.

Yet, as $\Delta \to 0$, these gains disappear since equilibrium path stakes exceed $\underline{\phi}(\Delta)$ 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.[14]

We establish that this collapse is general by showing that off-path communication incentives limit a pair of partners to work at no more than bilateral stakes $\underline{\phi}(\Delta)$ when neither reports any interaction in the previous $(n-1)$ time periods. Consider any permanent ostracism equilibrium in which on the equilibrium path, players may randomize in their stake proposals as a function of the history. For every pair of feasible messages at time t, m_i^t and m_j^t, $E[\phi_{ij}|m_i^t, m_j^t]$ denotes the expected equilibrium stakes that they select at time t when player i reveals m_i^t and j reveals m_j^t.

Lemma 1. *In every permanent ostracism equilibrium, $E\big[\phi_{ij} \bigm| m_i^t, m_j^t\big] \leq \underline{\phi}(\Delta)$ for any pair of reported histories (m_i^t, m_j^t) in which there is no interaction at or after $t - (n-1)\Delta$.*

Proof. Suppose otherwise: consider a pair of messages (m_i^t, m_j^t) such that $E\big[\phi_{ij} \bigm| m_i^t, m_j^t\big] > \underline{\phi}(\Delta)$, and there is no interaction at or before $t - (n-1)\Delta$. Let h_i^t be a history that is identical to m_i^t except that in the previous $(n-1)\Delta$ periods, player i has met every player other than j, all of

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

whom proceeded to shirk on player i. Suppose player j communicates m_j^t first. Once player i reveals history h_i^t, the maximal stakes that the two can work at are $\underline{\phi}(\Delta)$, resulting in an expected payoff of $\underline{\phi}(\Delta) + \frac{\delta p_\Delta}{1-\delta}\underline{\phi}(\Delta)$. Consider the expected payoff from a deviation in which player i reveals only m_i^t, chooses a proposal using the equilibrium strategy after histories (m_i^t, m_j^t), and chooses to shirk regardless of what stakes are selected:

$$E\big[T(\phi_{ij}) \mid m_i^t, m_j^t\big] > T\big(E\big[\phi_{ij} \mid m_i^t, m_j^t\big]\big) > T(\underline{\phi}(\Delta)) = \underline{\phi}(\Delta) + \frac{\delta p_\Delta}{1-\delta}\underline{\phi}(\Delta),$$

where the first two inequalities are implied by Assumption 1 and Jensen's Inequality, and the equality is by definition of $\underline{\phi}(\Delta)$. Since the payoff from deviation exceeds that from truthful communication, the strategy profile is not an equilibrium. □

This result constrains equilibrium payoffs because the probability that there is more than a single interaction in a time period of length $(n-1)\Delta$ vanishes as $\Delta \to 0$. Since these events occur with probability $O\left(\Delta^2\right)$ for small Δ and have payoffs that are uniformly bounded for all Δ, permanent ostracism collapses to private bilateral enforcement.

Theorem 2. *As the period length converges to* 0*, the maximal payoff achievable in any permanent ostracism equilibrium converges to that from private bilateral enforcement. For every $\varepsilon > 0$, there exists $\overline{\Delta} > 0$ such that for all discrete time games with period length $\Delta < \overline{\Delta}$, the expected continuation payoffs at any on-path history are within ε of those from private bilateral enforcement. No permanent ostracism equilibrium of the continuous time game supports stakes that exceed $\underline{\phi}(0)$.*

Our result demonstrates that even with hard evidence, cooperation much beyond bilateral enforcement cannot be supported using permanent ostracism equilibria. Because softer forms of communication expand the scope for deviations, our results apply in the absence of hard evidence, as well as if players can engage in both evidentiary and cheap talk communication.

This negative result has implications for our understanding of community enforcement. Prior work has advanced our understanding through public monitoring or reputational label mechanisms in which an individual's label—guilty or innocent—is automatically updated on the basis of her actions, or assumed that individuals communicate truthfully. In these works, once an individual is labeled guilty, she is punished by all those who meet her while innocent players continue to cooperate. Our result shows that individuals lack the incentive to communicate truthfully unless the level of cooperation is so low that community enforcement is not needed.

4.3 Robustness of Result

In establishing the limits of permanent ostracism, we made a number of simplifying assumptions, and below, we discuss how our result applies more broadly. An important generalization shows that permanent ostracism is limited even if the relationship on each link is not a prisoner's dilemma. We also show that our result continues to apply for general network structures, even if communication is simultaneous in each round, individuals have communication opportunities outside their interactions, and even if most interaction times are publicly observed.

4.3.1 General Network Structures

We have assumed a complete network with identical frequencies of interaction to simplify exposition. Suppose instead that each link $\{ij\}$ is recognized at a potentially different rate $\lambda_{ij} \geq 0$. Define $\underline{\phi}_{ij}(0)$ as the highest stakes that satisfy Bilateral IC (on p. 12), substituting the pair specific λ_{ij} for λ in that expression. An analogue of Theorem 2 binds permanent ostracism equilibria from supporting stakes in partnership $\{ij\}$ that exceed $\underline{\phi}_{ij}(0)$ in the continuous time setting. A similar approximation result would hold for discrete time games with short period length.[15]

Our result also applies to a setting in which when guilty players are ostracized, innocent players can increase their frequency of interaction. Without defining such a game in detail, suppose that the maximal frequency of interaction, $\overline{\lambda}$, is achieved between Ann and Bob when they do not work with anyone else. Defining the highest bilateral stakes that can support cooperation at this higher frequency of interaction as $\underline{\phi}$, our conclusion holds: permanent ostracism still cannot support cooperation that exceeds that from bilateral enforcement among isolated pairs.

4.3.2 General Bilateral Games

Here, we show that permanent ostracism equilibria are limited even when bilateral relationships are not prisoners' dilemmas. We operate in continuous time: partnership $\{ij\}$ is recognized at rate λ_{ij}, at which point players i and j sequentially communicate and then play the stage game $G_{\{ij\}}$. This stage game may differ across partnerships and be asymmetric. In $G_{\{ij\}}$, players i and j 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 $\underline{u}_{ij}$.

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 generalizing the notion of ostracism is straightforward.

Assumption 2. *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 2 guarantees that in each game, each player finds it incentive compatible to maximally punish the other in their bilateral relationship without requiring intertemporal incentives.

[15] In discrete time, Lemma 1 and Theorem 2 extend with a link-specific $p_{\Delta}^{ij} \equiv \frac{(1-e^{-\Lambda\Delta})\lambda_{ij}}{\Lambda}$, in which $\Lambda = \sum_{k,l\in\mathcal{N}} \lambda_{kl}$.

Apart from being satisfied in several moral hazard settings, Assumption 2 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.[16] For games in which Assumption 2 fails, our results pertain to equilibria in which guilty players are punished by Nash-reversion.

First, we describe private 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 payoff that player i attains in a sub-game perfect equilibrium at the beginning of an $\{ij\}$ interaction.

Now we describe permanent ostracism using the evidentiary information structure. A behavioral strategy for player i is a function $\sigma_i = (\sigma_i^M, \sigma_i^A)$, where σ_i^M specifies her reporting strategy and σ_i^A specifies her (mixed) action choice. A player is guilty when she takes an action that she should play with zero probability: in an equilibrium σ, denote the support of player i's equilibrium actions in $G_{\{ij\}}$ at history h_i^t, and after exchanging messages $\left(m_i^t, m_j^t\right)$ by

$$A(h_i^t, G_{\{ij\}}, m_i^t, m_j^t) \equiv \left\{a \in A_{ij} : \sigma_i^A(h_i^t, m_i^t, m_j^t)(a) > 0\right\}.$$

Player i considers player j to be *innocent* in history h_i^t—i.e., $j \in \mathcal{I}(h_i^t)$—if in every interaction z^τ, a_j^τ is in $A(m_j^\tau, G_{\{jk\}}, m_j^\tau, m_k^\tau)$. By contrast, player i considers player j to be *personally guilty*—i.e., $j \in \mathcal{G}_i(h_i^t)$—if there exists an interaction z^τ that involves players i and j in which a_j^τ is not in $A(m_j^\tau, G_{\{jk\}}, m_j^\tau, m_k^\tau)$. We define the analogue of permanent ostracism.

Definition 4. *In a* ***generalized permanent ostracism*** *strategy profile σ, if player i meets player j at time t, following a history h_i^t, her behavior is as follows:*

1. *If $\{i, j\} \subset \mathcal{I}(h_i^t)$, then player i reveals history h_i^t and follows $\sigma_i^A(m_i^t, m_i^t, m_j^t)$.*
2. *If $j \in \mathcal{G}_i(h_i^t)$, player i sends message $m_i^t = \emptyset$, and plays $\underline{\alpha}_{ij}$.*

A generalized permanent ostracism profile guarantees that a player continues to communicate and "cooperate" with those who are innocent, but requires that she suspend communication and 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.) That is, player i ostracizes player j while continuing to share information and "cooperate" with those who are innocent. Our result below extends Theorem 2.

Theorem 3. *Consider a generalized permanent ostracism equilibrium, σ. For any partnership $\{ij\}$, consider a mixed profile α^* in $G_{\{ij\}}$. If α^* is played on the equilibrium path, then:*

$$\max_{a \in A_{ij}} u_{ij}(a, \alpha_{-i}^*) + \frac{\lambda_{ij}}{r} \underline{u}_{ij} \leq \overline{u}_{ij}. \tag{8}$$

[16] Jackson, Rodriguez-Barraquer, and Tan (2012) and Fainmesser (2012) allow links to be severed during the game.

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 $\sum_{j\in\mathcal{N}\setminus\{i\}} \frac{\lambda_{ij}}{r+\lambda_{ij}}\overline{u}_{ij}$.

The incentive to conceal information constrains equilibrium path behavior: even if each player within the pair *ij* has frequent interactions with third parties, the set of actions played in $G_{\{ij\}}$ depends on the maximum payoffs $(\overline{u}_{ij}, \overline{u}_{ji})$ that can be sustained in bilateral enforcement equilibria of their game. Thus, permanent ostracism cannot escape all limits of bilateral enforcement.

This restriction is easier to interpret for symmetric games and equilibria that prescribe symmetric equilibrium path behavior: then each player's equilibrium payoff from partnership $\{ij\}$ can be no more than her best payoff from a bilateral equilibrium in that partnership. This is an improvement over the set of payoffs from bilateral equilibria insofar as each player in the relationship can attain such a payoff, but communication constraints still impose bounds based on bilateral enforcement. An implication of this result is that even if one were to use transfers (through money or continuation value) as rewards for communication in the game of Section 3, each player's payoff from each relationship are still bounded above by the highest payoff she attains in a bilateral equilibrium.

4.3.3 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. Although this uncertainty does not influence straightforward permanent ostracism equilibria for $\Delta > 0$ (so Theorem 1 remains unchanged), it can generate incentives to communicate in non-straightforward 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 $\underline{\phi}(\Delta)$; but if only one party reports on it, then they work at small stakes $\varepsilon > 0$ thereafter.
2. Consider a larger population, and suppose as in the history used in the proofs of Theorem 1 and Lemma 1, 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 $\underline{\phi}(\Delta)$; but if they believe that some but not all others are guilty, they set stakes $\varepsilon > 0$.

Characterizing the set of equilibria generated by this potentially rich set of first and second order off-path beliefs is beyond our scope here. Instead, we understand what ingredients would be needed to do better than private bilateral enforcement by imposing selection criteria that generate results similar to Theorem 2. We find that two natural selection criteria suffice: an adaptation of bilateral rationality from Ghosh and Ray (1996) and a "richness" condition on off-path beliefs.

Definition 5. *A permanent ostracism equilibrium is* ***bilaterally rational*** *if for histories* (h_i^t, h_j^t) *in which* $\{i, j\} \subset \mathcal{I}(h_i^t \cup h_j^t)$, *players i and j work at stakes* $\phi_{ij}(h_i^t, h_j^t) \geq \underline{\phi}(\Delta)$.

Bilateral rationality precludes a pair of innocent players from working at stakes strictly below $\underline{\phi}(\Delta)$ when each believes the other to be innocent. Its motivation may be phrased as "pairwise renegotiation among innocent players": since an innocent pair commonly believes that working at stakes of $\underline{\phi}(\Delta)$ is incentive compatible, they have a motive to renegotiate away from any equilibrium in which their stakes are strictly below $\underline{\phi}(\Delta)$.

The second condition restricts off-path beliefs. Let H_j^t be the set of private histories for player j in which she believes that all players are innocent. In contrast, let $H_i^t(j)$ be the set of private histories for player i in which the only innocent players are i and j, the past $n-1$ interactions are those in which some player k has successfully shirked on player i, and player i does not observe any interaction in which player j would have learned of any player being guilty.

Definition 6. *The off-path beliefs in a permanent ostracism equilibrium are* ***rich*** *if for every pair* $\{ij\}$, *for every time t, for every private history in* $H_i^t(j)$, *player i believes that with strictly positive probability player j's private history is in* H_j^t.

Richness fails if, when all players other than j have just shirked on i, player i believes that at least one of those players must have shirked on player j first. Perturbing beliefs of this form, and imposing bilateral rationality, makes permanent ostracism ineffective.

Proposition 3. *In every bilaterally rational permanent ostracism equilibrium with rich off-path beliefs,* $\phi_{ij}(m_i^t, m_j^t) \leq \underline{\phi}(\Delta)$ *for any pair of messages* (m_i^t, m_j^t) *in which there is no interaction at or after* $t - (n-1)\Delta$. *Therefore, the payoffs of these equilibria converge to that of private bilateral enforcement as* $\Delta \to 0$.

4.3.4 Public Meeting Times

So far, we have assumed that Ann and Bob are the only ones to directly observe the timing and dynamics of their relationship. But in some settings, Carol may know *when* Ann and Bob meet to interact. With evidentiary communication, this setting differs from Section 3 in an important respect: when a player conceals information about an interaction, her partner knows that information has been concealed. If each player believes that her partner's failure to disclose indicates that he deviated, the familiar unraveling logic permits permanent ostracism to attain the mechanical communication benchmark. This logic applies more broadly to any setting in which it is common knowledge between partners when each of them has interacted with others, e.g., in a random matching environment in which each player interacts in each period. However, such permanent ostracism equilibria still have a surprising fragility to even slight possibilities for private meetings: permanent ostracism collapses again to bilateral enforcement.

Proposition 4. *If the timing of each interaction is publicly observed, there exists a permanent ostracism equilibrium in which partners work at stakes $\overline{\phi}(\Delta)$ on the equilibrium path. However, if the timing of each interaction is publicly observed with probability $1-\varepsilon$, and is privately observed by the interacting partnership with probability $\varepsilon > 0$, then analogues of Lemma 1 and Theorem 2 apply for every permanent ostracism equilibrium.*

4.3.5 Pure Communication Opportunities

In Section 3, players communicate only when they make effort choices. But some contexts may present opportunities for word of mouth communication at a faster rate than economic trade and exchange. Here, we describe the implications of additional communication possibilities.

Merely augmenting the game with additional opportunities for pure communication does not change our negative result because if the meeting is an "interaction meeting" rather than a pure communication meeting, an innocent player would lack the incentive to reveal that all others are guilty. This is transparent in the discrete time framework with both "interaction" and "pure communication" opportunities. Suppose that society is inactive in each period with probability $e^{-G(\lambda+\xi)\Delta}$, and otherwise it is active. Conditional on society being active, each link is recognized with uniform probability, and with probability $\frac{\xi}{\lambda+\xi}$ the partners meet purely to communicate; otherwise they meet to interact in the usual way (as described in Section 3). As $\Delta \to 0$, this discrete time game converges to a model in which interactions occur at rate λ and pure communication opportunities occur at rate ξ. Nevertheless, Lemma 1 and Theorem 2 apply: whenever no interaction is reported in the previous $n-1$ time periods, stakes cannot exceed $\underline{\phi}(\Delta)$.[17]

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 level of cooperation with mechanical communication: if the other partner also simultaneously reveals the interaction, then each partner is indifferent between revealing and concealing it. So players guilty of shirking are ostracized and those who are innocent have no incentive to conceal this information.

However, we find this approach unappealing. First, it seems like a stretch to depend on guilty

[17] An analogous bound applies if each player has opportunities to make public announcements at random times.

players to assist their victims in reporting their own guilt. In many cases, our interest is in studying whether victims and third parties can be counted to report on the misdeeds of others rather than for the guilty to self-incriminate. Second, it relies upon the ability for players to make public announcements, which is an institutional feature absent in many settings. Third, if there were some infinitesimal cost of revealing evidence, a guilty player would strictly prefer to conceal rather than reveal this information since she is ostracized in either case. Once a guilty player lacks the incentive to reveal her guilt, an innocent victim also prefers to conceal the interaction. Fourth, if public announcements are sequential rather than simultaneous, full revelation is not incentive compatible since if a guilty player conceals the interaction, her victim also does so.

5 Moving Beyond Permanent Ostracism

5.1 Forgiveness

Permanent ostracism fails to improve upon bilateral enforcement because severing relationships with guilty players destroys the social collateral that innocent players could use to cooperate beyond bilateral stakes. Temporary punishments may escape this challenge. Suppose that when a set of players is guilty, each player is forgiven at a random future time, independent of when other guilty players may be forgiven. Forgiveness has two opposing effects on the incentive to work. From the perspective of her relationship with an innocent Bob, Ann has less incentive to work since Bob eventually forgives her if she were to shirk instead. However, Ann would like to resume mutual effort with guilty players once they are forgiven, and if she shirks on Bob, then she has to wait longer to resume that cooperation. Which effect dominates depends on primitives, but we show that the latter dominates if players are sufficiently patient or society is sufficiently large (the exact condition is $r < 2\lambda(n-2)$). The channel by which temporary punishments enhances cooperation is new: forgiveness maintains (future) social collateral, and thereby fosters communication and cooperation among innocent players who know that many others are guilty.

We construct the equilibrium in continuous time using public correlation devices. Two features of our construction are that communication is minimal—players report only those interactions in which shirking has taken place—and equilibrium behavior is simple insofar as stakes at which players cooperate are identical across histories. In constructing this equilibrium, we are forced to tackle difficulties in coordination that stem from the failure of common knowledge. Because deviations are not publicly observed, a guilty player influences the rate at which others learn that she is guilty by working with some players while shirking with others. The most profitable deviation may then not entail shirking on all partners at the first opportunity, but instead a dynamic pattern of working in the present and shirking in the future. We take two steps to overcome this challenge. First, we study equilibria in which one's first victim is the one who spreads information about one's deviation; second, we augment the stage game with a round of communication immediately after effort choices so that a guilty player can inform her victim of whether he is the first victim.[18]

There are $n+1$ publicly observable payoff-irrelevant signal processes $(\theta_1, \ldots, \theta_{n+1})$, each of which is governed by an independent Poisson process of rate μ. The signal θ_i coordinates forgiveness for player i: if signal θ_i arrives at time t, player i is to be considered innocent by all unless and until she is found to have deviated after time t. That is, if player i is guilty, she is forgiven when her signal arrives. When players work, they always do so at stakes ϕ. When a player shirks, her first victim spreads information about the deviation to others.

We describe the cooperation payoffs that Ann foregoes by shirking, and having to wait for her forgiveness. Ann's expected payoff from cooperation with an innocent partner, say Carol, after Ann is forgiven is $\frac{\mu}{r+\mu}\frac{\lambda}{r}\phi$ if Carol is innocent; whereas for a guilty partner, say Dante, Ann has to wait for them both to be forgiven, which carries an additional discount of $\frac{2\mu}{r+2\mu}$. Thus, as she waits to be forgiven, if there are m innocent players then she foregoes

$$W(\phi, \mu, m) \equiv \phi + \frac{m\lambda}{r}\phi\left(1 - \frac{\mu}{r+\mu}\right) + (n-m)\frac{\mu}{r+\mu}\frac{\lambda}{r}\phi\left(1 - \frac{2\mu}{r+2\mu}\right) \tag{9}$$

The gain from shirking on an innocent partner, say Bob, is the ability to capture $T(\phi)$ immediately from Bob, as well as from another partner innocent Carol, if Ann meets Carol before Bob does and before Ann is forgiven. For Ann's guilty partner Dante, she can capture this payoff if Dante is forgiven before Ann is forgiven, and hasn't yet learned from Bob that Ann has deviated. Summing over all her partners yields her payoff from shirking:

$$S(\phi, \mu, m) \equiv T(\phi) + \frac{(m-1)\lambda}{r+2\lambda+\mu}T(\phi) + \frac{(n-m)\lambda\mu}{(r+\lambda+2\mu)(r+2\lambda+\mu)}T(\phi). \tag{10}$$

For Ann to find it in her interests to work, her gains from deviating must be exceeded by her foregone payoffs from future cooperation. We prove that this is the case for stakes that exceed those of permanent ostracism at low forgiveness rates. For expositional convenience, we let $\underline{\phi}$ denote the stakes from private bilateral enforcement.

Lemma 2. *If $r < 2\lambda(n-2)$, then there exists forgiveness rates and stakes greater than those of bilateral enforcement such that innocent players have no incentive to deviate: there exists $\overline{\mu} > 0$ such that for every $\mu \in (0, \overline{\mu})$, there exists $\phi_\mu > \underline{\phi}$ such that for every $\phi \in [\underline{\phi}, \phi_\mu)$, and $m \geq 1$, $S(\phi, \mu, m) \leq W(\phi, \mu, m)$.*

[18] This additional round has no effect on our results on permanent ostracism.

The above result is at the core of proving the existence of a temporary ostracism equilibrium. Fixing a forgiveness rate $\mu > 0$ and stakes $\phi > \underline{\phi}$ such that $S(\phi, \mu, m) \leq W(\phi, \mu, m)$ for every m, we construct an ostracism equilibrium such that players propose stakes of ϕ whenever they meet to cooperate, or stakes of 0 to punish. The result below follows.

Theorem 4. *If $r < 2\lambda(n-2)$, then there exists a temporary ostracism equilibrium that yields payoffs strictly higher than permanent ostracism.*

We find that so long as players are sufficiently patient, temporary ostracism is better because innocent players continue to have an incentive to communicate and cooperate despite the ostracism of others. While we have used public correlation devices to simplify analysis, it is straightforward to replicate this behavior by giving each pair of partners a verifiable private correlation device, which signals whether a guilty partner should be forgiven when they meet; once forgiven, he can prove to others that they should forgive him as well. We conclude our discussion here with a couple of remarks.

Remark 2 (Efficiency)**.** Our motive is to construct a simple temporary ostracism equilibrium that improves upon permanent ostracism. One can imagine more sophisticated forms in which forgiveness and the resumption of cooperation are gradual rather than immediate—analogous to Ghosh and Ray (1996)—and in which guilty players transfer continuation value to innocent players by working while letting innocent players shirk. Constructing these equilibria is challenging not least because such equilibria are in private strategies, and finding the efficient frontier for a fixed discount rate is an important question for future research.

Remark 3 (Individual vs. Collective Forgiveness)**.** In contrast to the collective forgiveness in Ellison (1994), our construction requires an individual's forgiveness to be imperfectly correlated with that of others. Were forgiveness at the collective level, Bob would lack the incentive to communicate truthfully to Carol when he knows that only the pair are still innocent because he would be forgiven for shirking at the same time as the others. That others may be forgiven before him offers an incentive to Bob to work.

Remark 4 (Fixed Stakes)**.** The history-invariance of stakes in this equilibrium suggests an immediate analogue for fixed stakes environments: for a prisoner's dilemma with exogenously fixed stakes, there is a range of discount rates such that mutual effort cannot be enforced by bilateral equilibria, but can be enforced by temporary ostracism.

5.2 Permanent Ostracism with Limited Depth

Our interpretation of ostracism is that it captures social relations in which trust is accorded to individuals as individuals: those who betray it are punished while those who cooperate continue to be trusted. This paints a sharp contrast to contagion equilibria in which once a single individual deviates, collective trust is eroded and players shirk on every other player. Contagion offers a theory of collective reputation and community responsibility. This section proposes hybrids of permanent ostracism and contagion in which permanent ostracism is practiced to a limited depth beyond which players shift to contagion behavior.

Suppose that when an innocent player knows of d or fewer deviators, she communicates truthfully to other innocent players, but if there are more than d deviating players, she shirks on all of her partners. The depth of 0 is contagion whereas that of $(n-1)$ is permanent ostracism. Intermediate depths capture both individual and collective reputations and the notion that once the number of deviators exceeds some threshold, collective reputations break down. Apart from realism, a virtue of these social norms is that backloading contagion to at least two steps off the equilibrium path addresses the concern that cooperation across the community is vulnerable to the deviations of a single individual.[19]

Nevertheless, communication incentives restrict equilibrium payoffs: a permanent ostracism equilibrium of depth d has average stakes that are bounded by what contagion can sustain in a smaller group of $n+1-d$ players. Although this bound improves over permanent ostracism of unlimited depth, cooperation is bounded away from the best contagion equilibrium.

The formal definition of permanent ostracism with limited depth is below.

Definition 7. *A strategy profile involves permanent ostracism of depth $d \leq n-1$ if, for every time t, player i, and history h_i^t, the following are satisfied:*

1. *If $|\mathcal{G}(h_i^t)| \leq d$ and $\{i,j\} \subset \mathcal{I}(h_i^t)$, then player i reveals history h_i^t. If $m_j(h_i^t, t) \subseteq m_j^t$ and $j \in \mathcal{I}\big(h_i^t \cup m_j^t\big)$, then player i believes with probability 1 that j is innocent, announces strictly positive stakes, and works.*
2. *If $|\mathcal{G}(h_i^t)| \leq d$, $i \in \mathcal{I}(h_i^t)$, and $j \in \mathcal{G}(h_i^t)$, player i sends any message, proposes strictly positive stakes, and then shirks.*
3. *If $|\mathcal{G}(h_i^t)| > d$, then player i sends any message, proposes strictly positive stakes, and shirks.*

In a permanent ostracism equilibrium with limited depth, players are willing to permanently ostracize up to d guilty players while maintaining cooperation among innocent players. When any player sees that more than d are guilty, she shifts to a contagion equilibrium in which she shirks with all remaining players. For simplicity, we restrict attention throughout to *straightforward* permanent ostracism equilibria of depth d, i.e., those in which the stakes in history h of players i and j are a function of $\mathcal{I}(h)$. So, for every set of innocent players $\tilde{\mathcal{I}}$, we can write the vector of stakes

[19] This approach to avoiding the destructiveness of contagious punishments differs from prior proposals to make contagion robust through the inclusion of a public correlation device (Ellison 1994) or designing the network to be incomplete (Jackson, Rodriguez-Barraquer, and Tan 2012).

at which pairs in $\tilde{\mathcal{I}}$ work as $\Phi(\tilde{\mathcal{I}}) \equiv \left(\phi_{ij}(\tilde{\mathcal{I}})\right)_{\{i,j\}\subseteq\mathcal{I}}$, and $\Phi_i(\tilde{\mathcal{I}}) \equiv \left(\phi_{ij}(\tilde{\mathcal{I}})\right)_{j\in\mathcal{I}\setminus\{i\}}$ as the vector of stakes that an innocent player i sets with his innocent partners. Let ϕ_i^{avg} be the average of $\Phi_i(\mathcal{N})$; i.e. player i's average equilibrium path stakes. Note that player i's equilibrium path payoffs are simply $\frac{n\lambda}{r}\phi_i^{\text{avg}}$.

As an interlude, we describe contagion equilibria when the set of innocent players is $\tilde{\mathcal{I}}$ such that $|\tilde{\mathcal{I}}| = n - d + 1$, i.e., the other d players have been ostracized. Once d players are ostracized, if an innocent player shirks on an innocent partner, then contagion spreads through the set of innocent players. For an innocent player to prefer working to shirking on all remaining innocent players, $\Phi(\tilde{\mathcal{I}})$ must satisfy the following incentive constraint for every pair $\{i,j\} \subseteq \tilde{\mathcal{I}}$:

$$T(\phi_{ij}(\tilde{\mathcal{I}})) + \sum_{k\in\tilde{\mathcal{I}}\setminus\{i,j\}} T(\phi_{ik}(\tilde{\mathcal{I}}))X_{n-d} \leq \phi_{ij}(\tilde{\mathcal{I}}) + \frac{\lambda}{r}\sum_{k\in\tilde{\mathcal{I}}\setminus\{i\}} \phi_{ik}(\tilde{\mathcal{I}}), \tag{11}$$

where X_{n-d} summarizes the rate at which contagion spreads. Ali and Miller (2013) show that

$$X_{n-d} = \frac{1}{n-d-1}\sum_{\ell'=2}^{n-d}\Big(\frac{1}{\ell'}\prod_{\ell=2}^{\ell'}\frac{\lambda\ell(n-d-\ell+1)}{r+\lambda\ell(n-d-\ell+1)}\Big).$$

Let ϕ^{n-d} be the symmetric solution that makes the above inequality bind, i.e., solves

$$T(\phi)\left(1+(n-d-1)X_{n-d}\right) = \phi\left(1+\frac{(n-d)\lambda}{r}\right). \tag{12}$$

The equality implies that if one were to set the stakes to ϕ^{n-d} in all partnerships between players in $\tilde{\mathcal{I}}$, each player would be indifferent between working and shirking if the set of innocent players were $\tilde{\mathcal{I}}$, and would have a strict incentive to shirk once any more players became guilty. While a permanent ostracism equilibrium of depth d may not use this particular form of contagion, we show that ϕ^{n-d} nevertheless bounds each player's average equilibrium path stakes.

Theorem 5. *In every permanent ostracism equilibrium of depth d, each player's average equilibrium path stakes are less than the average stakes of the binding contagion equilibrium with $n-d$ partners, and each player's equilibrium payoff is less than $\frac{n\lambda}{r}\phi^{n-d}$.*

Since ϕ^{n-d} is strictly decreasing in d, this result indicates that increasing the depth of permanent ostracism reduces the upper bound on average stakes. Generally, we do not know that this bound is tight because constructing permanent ostracism equilibrium with limited depth is challenging.[20] However, one can easily construct the permanent ostracism equilibrium of depth 1, in

[20] To slow down how quickly others learn of his guilt, a guilty player may choose to work with some players even after he has shirked. Accordingly, we cannot find the most profitable deviation in every permanent ostracism equilibria of limited depth.

which stakes equal ϕ^{n-1} throughout. These stakes are high enough that each player's most profitable deviation is to shirk everyone as soon as he meets them.[21] Yet, this most profitable deviation generates payoffs less than an innocent player's expected payoff on the equilibrium path and when at most one player is guilty. Thus, an innocent player has no incentive to deviate.

6 Applications

6.1 Communication and Cooperation in Markets

In many markets, buyers and sellers can renege on their promises. At the core of intertemporal incentives in medieval trade and modern reputation systems today (e.g., eBay) is the extent to which trading partners can communicate their trading experiences so that a deviating player finds it difficult to trade in the future. When are these communications truthful?

We answer this question in "networked markets" in which a participant trades with players on the other side, but can share information with both sides. We find that truthful communication in permanent ostracism is subtle. If both parties in each relationship have moral hazard, then permanent ostracism reduces again to bilateral enforcement; by contrast, if only one side has a myopic incentive to deviate, then permanent ostracism is effective.

We find this distinction interesting for several reasons. First, it supports the common practice in online trading platforms of structuring trade sequentially: when a buyer lacks the incentive to deviate because she moves first, she has no incentive to communicate non-truthfully. Second, it indicates that an enforcement intermediary who mitigates incentive issues on one side of the market can complement, rather than substitute for, community enforcement. Finally, our analysis enriches prior models of informal enforcement in markets that have focused on a *one-sided* prisoner's dilemma or a "product-choice" game by showing that permanent ostracism may be supported in these settings even with private monitoring and strategic communication.[22]

Society comprises a set of buyers $\mathcal{N}^B = \{1, \ldots, b\}$, and sellers $\mathcal{N}^S = \{1, \ldots, s\}$, and each buyer-seller pair meets at Poisson rate λ. When a buyer and a seller meet, they first communicate, and then simultaneously choose quantity (or quality) q and payment p respectively. When the buyer pays p for quantity q, the buyer's payoff is $q - p$ and the seller's payoff is $p - c(q)$. We assume that c is strictly increasing and strictly convex, that $c(0) = c'(0) = 0$, and $\lim_{q \to \infty} c'(q) = \infty$. We assume that the first-best efficient quantity is too high to be supported by mechanical communication so that players benefit from improvements in enforcement.

[21] Similarly, the stakes are sufficiently high that the contagion phase incentives are satisfied.

[22] Klein and Leffler (1981) and Greif (1993) study settings with public monitoring, whereas Klein (1992) and Ahn and Suominen (2001) assume mechanical communication. Fainmesser and Goldberg (2011) also studies mechanical communication but with the friction that the network is not common knowledge. Deb and González-Díaz (2011) study contagion in a random matching environment.

We first describe benchmarks of private bilateral and mechanical enforcement. In the former, each of the seller and buyer cooperate in a stationary profile if and only if

$$p \leq p - c(q) + \frac{\lambda}{r}(p - c(q)) \quad \text{and} \quad q \leq q - p + \frac{\lambda}{r}(q - p).$$

The highest level of trade consistent with these incentives solves $c(q)/q = \left(\frac{\lambda}{r+\lambda}\right)^2$, and no private bilateral enforcement equilibrium (including those with non-trivial dynamics on the equilibrium path) can support greater trade. In mechanical communication, a player is punished by all participants on the other side of the market if she deviates in any relationship. The incentive conditions in a symmetric and stationary profile are:

$$p \leq p - c(q) + \frac{\lambda b}{r}(p - c(q)) \quad \text{and} \quad q \leq q - p + \frac{\lambda s}{r}(q - p).$$

The highest incentive compatible trade under mechanical communication—denoted by $q_{b,s}$—satisfies

$$\frac{c(q)}{q} = \frac{bs\lambda^2}{(r + b\lambda)\,(r + s\lambda)}.$$

As before, no other equilibrium can support trade greater than $q_{b,s}$.

In studying communication incentives, we have to specify the protocol for players on the same side of the market. To starkly highlight the contrast between the two-sided and one-sided incentive cases, suppose each player can broadcast a message to all other players on her own side of the market whenever she likes, e.g., as in writing a public comment or review. Although communication is very permissive, permanent ostracism fails to improve upon bilateral enforcement with two-sided moral hazard, as we illustrate with even a single seller and a large number of buyers.

Example 2. Suppose towards a contradiction that with a single seller and b buyers, there is an equilibrium in which on the equilibrium path, sellers and buyers trade at volume $q \in \left(q_{1,1}, q_{1,b}\right]$ at a price of $\frac{\lambda}{r+\lambda}q$. While this satisfies the equilibrium path incentives for both buyer and seller, it violates communication incentives off the equilibrium path. Consider a history in which $b - 1$ buyers have reneged on their payments. Since buyers do not self-report their own defections to other buyers, the seller has no motive to do so: concealing the defection permits the seller to sell an object of zero quality at $\frac{\lambda}{r+\lambda}q$, generating a payoff that exceeds that from revealing the truth and perpetually trading objects of quality $q_{1,1}$ with the single innocent buyer.

This tension is general. We defer the formal definition of permanent ostracism to Appendix B, since it is analogous to Definition 3, with the additional wrinkle that each player reveals her interactions instantaneously and publicly to all participants on her side of the market, unless she is the one who has shirked (in which case she prefers to conceal that she has shirked).

Proposition 5. *No straightforward permanent ostracism equilibrium supports trading volume that exceeds $q_{1,1}$ in any interaction.*

The logic is identical to that of Theorem 1, and so we omit the formal proof: at histories in which all but one buyer has shirked on a seller, that seller lacks an incentive to communicate truthfully to the remaining innocent buyer. In the context of a networked market, a seller's incentive to trade beyond $q_{1,1}$ comes from how her behavior influences her relationship with other buyers. When those relationships are forfeit, she prefers to cheat the remaining innocent buyer, who wrongly believes that behavior is on the equilibrium path.

We contrast these results with what happens if the incentive challenge is one-sided in sequential trading. To fix ideas, suppose that the buyer has to make a payment first, and then upon receipt of the payment, the seller chooses whether to trade the object. Accordingly, there is a classic hold-up problem in which only the seller has a myopic incentive to deviate. Removing the buyer's incentives to deviate not only enhances bilateral enforcement, but also eliminates the challenge of communication incentives.

Proposition 6. *If trading is sequential, there exists a straightforward permanent ostracism equilibrium with strategic communication that supports a trading volume of $\hat{q}_b$ that solves*

$$\frac{c(q)}{q} = \frac{\lambda b}{r + \lambda b}.$$

This volume $\hat{q}_b$ strictly exceeds not only that sustainable with private bilateral enforcement ($\hat{q}_{1,1}$), but also the maximal volume that can be sustained by permanent ostracism with mechanical communication when trading is simultaneous.

Because all buyers lack an incentive to deviate, only a seller needs to be ostracized when she deviates. Ostracizing a deviating seller does not adversely affect the buyer's other relationships, and so a buyer has no incentive to conceal information from other buyers and sellers. Thus, the truthful communication of buyers can offer strong incentives to sellers to fulfill their trading obligations. Our logic applies to many other settings in which only one side has an incentive to deviate, e.g., creditors who lend the money up front and are thus willing to report truthfully to credit agencies, or employers who are contractually obligated to pay wages. In each of these cases, communication and permanent ostracism can be powerful tools for community enforcement.

6.2 Informal Risk-Sharing

The literature on limited enforcement in risk-sharing arrangements typically uses *autarky* as punishment: if an individual fails to make the transfer that she is expected to, she is forced to bear her idiosyncratic risk alone in the future. Implementing this punishment is straightforward if risk sharing is *centralized* (the community in its entirety observes the deviation) or if communication is *mechanical* (as described in Definition 1). Yet, much of risk sharing occurs only through bilateral interactions, in which the partners involved are the only ones to directly observe whether each of them followed the community's norm for risk-sharing behavior. We study how much can be transferred when players communicate strategically, and autarky-like punishments are used.

Suppose that in each period that lasts for $\Delta > 0$ units of real time, each player obtains a random endowment of either $\overline{y}$ or $\underline{y}$ units of a consumption good, in which $\overline{y} > \underline{y} > 0$. The utility of consumption is represented by $u(\cdot)$, a strictly increasing, strictly concave, and smooth utility function such that $u(0) = -\infty$. The distribution of the endowment is i.i.d. across individuals and over time, and the probability that a player obtains a high endowment in any period is $\eta \in (0,1)$. A rich player can transfer consumption to a poor player only when the pair meet.[23] When players meet, they first observe each other's endowments in that period, then they communicate, and finally, they choose how much to transfer between them. Players have no opportunities to save.

Full insurance involves transfers of $(\overline{y}-\underline{y})/2$ from a rich player to a poor player, equalizing their consumption in that period. Using $\alpha = \eta(1-\eta)$ to denote the probability that one player in an interacting pair is richer than the other, the maximal transfer ζ_n achieved by permanent ostracism with mechanical communication solves

$$u(\overline{y}) - u(\overline{y}-\zeta) = \frac{n\delta p_\Delta \alpha}{1-\delta}\Big(\big(u(\underline{y}+\zeta) - u(\underline{y})\big) - \big(u(\overline{y}) - u(\overline{y}-\zeta)\big)\Big).$$

A strictly positive solution to the above equation is not guaranteed to exist, and does so only if players are sufficiently patient, or society is sufficiently large. In contrast to mechanical communication, private bilateral enforcement sustains transfers of only ζ_1, and the incentive to conceal information constrains permanent ostracism with strategic communication to do no better.

Proposition 7. *No straightforward permanent ostracism equilibrium supports transfers that exceed ζ_1.*

The logic is similar to that before: if the transfer exceeds ζ_1, in any off-path history at which Bob knows that he and Carol are the only remaining innocent players, he prefers to not divulge the truth to her if she is richer. Concealing the truth induces Carol to make the higher equilibrium path transfer, whereas revealing it induces her to reduce the size of the transfer. Since Carol never learns the truth from others, Bob lacks the incentive to communicate truthfully to Carol.

The failure of permanent ostracism can be quantitatively important for simple parameteriza-

[23] Our study is therefore analogous to *favor exchanges*, wherein one player has a random opportunity to do a favor for another (Möbius 2001; Jackson, Rodriguez-Barraquer, and Tan 2012), and our conclusions for ostracism equilibria apply virtually identically to that setting.

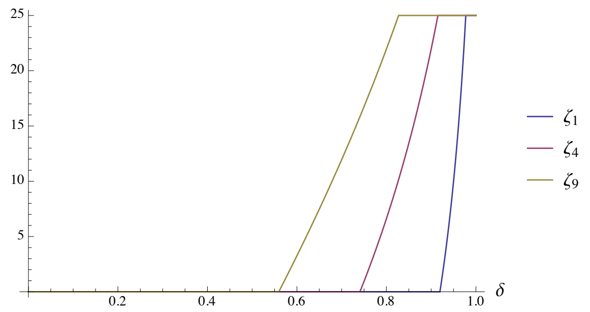

FIGURE 4. RISK SHARING TRANSFERS

tions, as we illustrate in Figure 4. Suppose that the primitives are $u(c) = -\frac{1}{2c^2}$, $\overline{y} = 100$, $\underline{y} = 50$, $\eta = \frac{1}{2}$, and $p_\Delta = \frac{1}{10}$. Regardless of community size, permanent ostracism with strategic communication implements only ζ_1 and requires a discount factor of $\delta \geq 0.919$ to implement any risk sharing at all, and requires $\delta \geq 0.977$ for full insurance. In contrast, were communication mechanical, a society of five players could attain full insurance at discount factors of $\delta \geq 0.915$, and ten players could do so at discount factors of $\delta \geq 0.827$. Although we have not studied their implications here, our earlier results indicate that temporary ostracism would be more effective than permanent ostracism. In general, our results suggest that in a setting that relies upon the strategic communication of private information, conclusions derived from assuming mechanical communication or public monitoring may significantly overstate the potential for risk-sharing.

7 Conclusion

The use of communication to ostracize deviating players is a compelling social norm. Our goal here has been to understand when truthful communication is incentive compatible. We find, perhaps surprisingly, that truthful communication cannot be incentive compatible when ostracism is permanent, unless the level of cooperation is sufficiently low that community enforcement is unnecessary. On the other hand, augmenting ostracism with either forgiveness or a sense of community responsibility improves cooperation while fostering truthful communication. Thus, the challenge of satisfying communication incentives can shed light on the motive for other practices salient in community enforcement. Moreover, our application to networked markets highlights why sequential trading may facilitate one side truthfully revealing their trading experiences.

We focus on ostracism's role in mitigating moral hazard and ignore its role in screening and ostracizing myopic or impatient players. Our results have direct implications for the continuation behavior among patient players once others are ostracized, but do not contradict the use of permanent ostracism to exclude those who appear impatient. We consider it important for future study to understand how players in a network may communicate and coordinate so as to exclude myopic players and recruit new players.

Our attention has been devoted to cooperation in the absence of legal and other institutions. Bounds on how much cooperation is achieved through communication on a network offers an appreciation for the gap that may be filled by intermediaries and institutions. Even when such institutions are present, our motivating question is of interest: when do victims truthfully report to an adjudicator that someone else has deviated?

Appendix A Main Proofs

A.1 Proofs for Section 4

Proof of Theorem 2 on p. 17. We prove the result immediately for the continuous time environment using an argument similar to Lemma 1:

Corollary 1. *For $\Delta = 0$, for every permanent ostracism equilibrium, and at every history, players work at stakes no greater than $\underline{\phi}(0)$.*

Proof. Consider a pair of messages (m_i^t, m_j^t) such that $E\big[\phi_{ij} \;\big|\; m_i^t, m_j^t\big] > \underline{\phi}(0)$. Consider a history h_i^t that coincides with m_i^t except that every other player has shirked on player i after the last interaction in m_i^t. Then player i has a profitable deviation from truthful communication once player j reports m_j^t. □

We proceed to argue that in the discrete time setting, payoffs collapse to bilateral enforcement $\underline{\phi}(\Delta)$ as $\Delta \to 0$, 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 periods into blocks of length $(n-1)\Delta$. In this profile, along the path of play, players cooperate

1. at stakes $\underline{\phi}(\Delta)$ when no interaction is observed in the previous or current block;
2. at stakes $\overline{\phi}(\Delta)$, otherwise.

Since any stakes that satisfy the incentives for permanent ostracism also satisfy the effort incentive for mechanical communication (Mechanical IC) with slack, it follows that in any permanent ostracism equilibrium with strategic communication, $E[\phi_{ij}|m_i^t, m_j^t] < \overline{\phi}(\Delta)$ for every ij and every pair of messages (m_i^t, m_j^t). Combined with Lemma 1, this implies that every permanent ostracism equilibrium with strategic communication has equilibrium path payoffs that are less than those of $\hat{\sigma}$.

For small Δ, we approximate the payoffs for $\hat{\sigma}$ by decomposing payoffs within each $(n-1)\Delta$ block, ignoring errors from discounting that are no more than $O(\Delta)$. Let π_H denote 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.

$$\begin{aligned}\pi_L &= (1 - Gp_\Delta)^{n-1} e^{-r\Delta(n-1)} \pi_L \\ &\quad + \sum_{k=1}^{n-1} \binom{n-1}{k} (Gp_\Delta)^k (1 - Gp_\Delta)^{n-1-k} \left(\frac{n}{G}\big(\underline{\phi}(\Delta) + (k-1)\overline{\phi}(\Delta)\big) + e^{-r\Delta(n-1)} \pi_H\right) + O(\Delta),\end{aligned}$$

and

$$\begin{aligned}\pi_H &= (1-Gp_\Delta)^{n-1}e^{-r\Delta(n-1)}\pi_L \\ &\quad + \sum_{k=1}^{n-1}\binom{n-1}{k}(Gp_\Delta)^k(1-Gp_\Delta)^{n-1-k}\left(\frac{n}{G}k\overline{\phi}(\Delta)+e^{-r\Delta(n-1)}\pi_H\right)+O(\Delta).\end{aligned}$$

Therefore,

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

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

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

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

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

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

Proof of Theorem 3 on p. 19. The argument is similar to Theorem 2. Suppose that α^* is the prescribed profile at equilibrium path history $h_i^t\cup h_j^t$ and after players have communicated messages $m_i^t=h_i^t$ and $m_j^t=h_j^t$. Consider a history $\tilde{h}_i^t$ 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 m_j^t first. If player i truthfully communicates $\tilde{h}_i^t$ to player j, her maximal payoff is $\overline{u}_{ij}$. In contrast, by communicating h_i^t and choosing a best response to α_{-i}^* guarantees a payoff of at least the left-hand side on (8), and thus, (8) specifies a necessary condition for truthful communication.

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. 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 consider two cases depending on the sign of

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

1. Suppose (14) is non-negative. Then in the repeated play of $G_{\{ij\}}$, there exists a sub-game perfect equilibrium in which players play α^* on the equilibrium path, and if either deviates, they revert to $\left(\underline{\alpha}_{ij}, \underline{\alpha}_{ji}\right)$.[24] 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 (8) because:

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

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 3. □

A.2 Proofs for Section 5

Proof of Lemma 2 on 24. Observe that for $m \geq 2$,

$$\begin{aligned} S(\underline{\phi}, 0, m) = T(\underline{\phi}) + \frac{(m-1)\lambda}{r+2\lambda}T(\underline{\phi}) &= \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{r+\lambda}{r}\frac{(m-1)\lambda}{r+2\lambda}\underline{\phi} \\ &< \underline{\phi} + \frac{\lambda}{r}\underline{\phi} + \frac{(m-1)\lambda}{r}\underline{\phi} = W(\underline{\phi}, 0, m). \end{aligned}$$

By continuity of S and W in μ, there exists $\overline{\mu}$ such that for every $\mu < \overline{\mu}$ and for every $m \in \{2, \dots, n\}$, $S(\underline{\phi}, \mu, m) < W(\underline{\phi}, \mu, m)$. By continuity of S and W in ϕ, there exists ϕ_μ such that $S(\phi, \mu, m) \leq W(\phi, \mu, m)$ for $\phi \in [\underline{\phi}, \phi_\mu)$.

A separate argument is needed for $m = 1$. Observe that for $\mu > 0$,

$$S(\underline{\phi}, \mu, 1) < W(\underline{\phi}, \mu, 1) \quad \iff \quad 1 < (n-1)\left(\frac{r}{r+2\mu} - \frac{(r+\lambda)(r+\mu)}{(r+2\lambda)(r+\lambda+2\mu)}\right).$$

If $r < 2\lambda(n-2)$, the above expression is satisfied at $\mu = 0$, and so by continuity, there exists $\overline{\mu}$ such that the expression remains satisfied for all $\mu \in (0, \overline{\mu})$. Therefore, the claim follows in this case. □

Proof of Theorem 4 on p. 25. Fix a $\mu > 0$ and $\phi > \underline{\phi}$ such that $S(\phi, \mu, m) \leq W(\phi, \mu, m)$ for every $m \in \{1, \dots, n\}$. We begin by describing the strategy profile. We denote by $\mathcal{E}_i(h)$ the set of interactions that player i knows in history h: this includes her interactions in which she has engaged as well as those she has learned from others. We let $\mathcal{G}_i(h)$ denote the set of players that player i considers guilty in history h, and we construct this recursively. Let $\tilde{T} = \{0, t_1, \dots, t_M\}$ be the set of all times of all interactions in $\mathcal{E}_i(h)$, recognition of publicly observable signals, and the starting time. We consider a sequence $(\omega^m)_{m=0,1,\dots,M}$, in which for each m, $\omega^m \in \{0,1\}^{n+1}$. Let ω^0 be the zero-vector, and we describe the transition rule for a generic $m < M$:

1. If θ_j is realized at time t_{m+1}, then $\omega^{m+1}_j = 0$;

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

2. If players j and k interact at time t_{m+1}, $\omega_k^m = 0$, player j proposes stakes not equal to ϕ, and player k proposes stakes ϕ, then $\omega_j^{m+1} = 1$;
3. If players j and k interact at time t_{m+1}, $\omega_k^m = 0$, both partners propose stakes ϕ, and player j shirks while player k works, then $\omega_j^{m+1} = 1$;
4. Otherwise $\omega_j^{m+1} = \omega_j^m$.

Player j is then in $\mathcal{G}_i(h)$ if $\omega_j^M = 1$.

In addition to defining who is guilty, we also describe an innocent player being the "first victim." Suppose that $\omega_i^m = \omega_j^m = 0$, players i and j interact at time t_{m+1} and $\omega_j^{m+1} = 1$. Then we say that player i is the *first victim* of player j. We let $\tilde{\mathcal{G}}_i(h)$ denote the set of players for whom player i is the first victim, and we let $\tilde{\mathcal{E}}_i(h)$ be the set of interactions in which player i became the first victim of another opponent.

Suppose that player i meets player j and has records $\mathcal{E}_i(h)$. Below, we describe the behavioral strategies followed by the players:

- *Communication pre-interaction*: Regardless of player j's message, and the timing of communication, player i sends the message $\tilde{\mathcal{E}}_i(h)$.
- *Stake selection*: Player i proposes ϕ.
- *Interaction*: Suppose player j has communicated message m. If $i \notin \mathcal{G}_i(h)$, $j \notin \mathcal{G}_i(h \cup m)$, and the selected stakes are ϕ, then player i works. Otherwise player i shirks.
- *Communication post-interaction*: If $i \in \mathcal{G}_i(h)$, and player i shirked at stakes ϕ in the previous stage, then player i sends the message $\mathcal{E}_i(h)$. Otherwise, send no message.

We now prove that this is an equilibrium. We have already verified the incentives for innocent players in communicating and cooperating with other innocent players. We first verify that a guilty player has an incentive to shirk immediately on all other innocent players at stakes ϕ. Since only the first victim communicates, when a guilty player i meets another innocent player j, working or shirking with player j affects no other relationship. Therefore, if π_{ij} represents player i's expected deviation payoff from $\{ij\}$ before player i is forgiven, then

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

Notice that if the second term above exceeds $T(\phi)$, then, because $\phi > \underline{\phi}$,

$$\pi_{ij} = \phi + \frac{\lambda}{r + \lambda + \mu}\phi < \phi + \frac{\lambda}{r}\phi < T(\phi),$$

which is a contradiction. Therefore, a guilty player's incentive is to shirk immediately on all other innocent players. Revealing the history afterwards ensures that each victim knows that he is not the first victim. □

Proof of Theorem 5 on p. 27. First, we describe a preliminary result. Let Ψ_{n-d} be the subset of $\mathbb{R}_+^{n-d}$ such that if $\Phi_i(\mathcal{I})$ is in Ψ_{n-d}, then (11) is satisfied for every $j \in \mathcal{I}\backslash\{i\}$. We prove in Lemma 4 of Ali and Miller (2013) that for every vector ψ in Ψ_{n-d}, the average of entries in Ψ is less than ϕ^{n-d}.

Step 1 Consider a set of innocent players $\tilde{\mathcal{I}}$ of cardinality $n-d+1$, and a player $i \in \tilde{\mathcal{I}}$. Let $\tilde{\Phi}_i$ be that the vector of player i's equilibrium path stakes restricted to partners in $\tilde{\mathcal{I}}\backslash\{i\}$, in other words, $\tilde{\Phi}_i \equiv \left(\phi_{ij}\right)_{j\in\tilde{\mathcal{I}}\backslash\{i\}}$. We argue using communication constraints that $(1,\ldots,1)\cdot\tilde{\Phi}_i \leq (n-d)\phi^{n-d}$.

Consider the communication constraints when player i meets player j at history h_i^t in which player j is in $\mathcal{I}(h_i^t) = \tilde{\mathcal{I}}$, and player j sends a message m_j^t such that $\mathcal{G}(m_j^t) = \emptyset$. If player i communicates truthfully to player j and all other innocent players, then each pair of innocent players would work according to $\Phi(\tilde{\mathcal{I}})$, while shirking on guilty players. On the other hand, if player i reports a message m_i^t in which $\mathcal{G}(m_i^t) = \emptyset$ to every innocent partner, he can shirk on the equilibrium path stakes on any innocent partner who is still working. Therefore, player i has an incentive to communicate truthfully if and only if

$$T(\phi_{ij}) + \sum_{k\in\tilde{\mathcal{I}}\backslash\{i,j\}} T(\phi_{ik})X_{n-d} \leq \phi_{ij}(\tilde{\mathcal{I}}) + \frac{\lambda}{r}\sum_{k\in\tilde{\mathcal{I}}\backslash\{i\}} \phi_{ik}(\tilde{\mathcal{I}}).$$

An analogous communication constraint must be satisfied for every $j \in \tilde{\mathcal{I}}\backslash\{i\}$. Adding up all of these $n-d$ constraints implies that

$$\begin{aligned}\big(1+(n-d-1)X_{n-d}\big)\sum_{j\in\tilde{\mathcal{I}}\backslash\{i\}} T(\phi_{ij}) &\leq \left(1+\frac{(n-d)\lambda}{r}\right)\sum_{j\in\tilde{\mathcal{I}}\backslash\{i\}} \phi_{ij}(\tilde{\mathcal{I}}) \\ &\leq \left(1+\frac{(n-d)\lambda}{r}\right)(n-d)\phi^{n-d} = (1+(n-d-1)X_{n-d})(n-d)T(\phi^{n-d}),\end{aligned}$$

where the second inequality follows from $\Phi_i(\tilde{\mathcal{I}}) \in \Psi_{n-d}$, and the equality follows from the definition of ϕ^{n-d}. Using Jensen's Inequality,

$$T\left(\frac{\sum_{j\in\tilde{\mathcal{I}}\backslash\{i\}}\phi_{ij}}{n-d}\right) \leq \frac{\sum_{j\in\tilde{\mathcal{I}}\backslash\{i\}} T(\phi_{ij})}{n-d} \leq T(\phi^{n-d}).$$

The monotonicity of T implies that $(1,\ldots,1)\cdot\tilde{\Phi}_i \leq (n-d)\phi^{n-d}$.

Step 2 Now we prove that $\phi_i^{\text{avg}} \leq \phi^{n-d}$. By the previous step, we know that if we take the $(n-d)$ highest entries of $\left(\phi_{ij}\right)_{j\in\mathcal{N}\backslash\{i\}}$, the average is less than ϕ^{n-d}. Including the remaining d entries cannot increase the average. □

References

Illtae Ahn and Matti Suominen. Word-of-mouth communication and community enforcement. *International Economic Review*, 42(2):399–415, 2001.

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

Attila Ambrus, Markus Möbius, and Adam Szeidl. Consumption risk-sharing in social networks. *American Economic Review*, 2013.

Patrick Bajari and Ali Hortaçsu. Economic insights from internet auctions. *Journal of Economic Literature*, 42(2):457–486, 2004.

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.

Jonathan Bendor and Dilip Mookherjee. Communitarian versus universalistic norms. *Quarterly Journal of Political Science*, 3(1):1–29, 2008.

B. Douglas Bernheim and Michael Whinston. Multimarket contact and collusive behavior. *Rand Journal of Economics*, 21(1):1–26, 1990.

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.

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. Community enforcement beyond the prisoner's dilemma. Working paper, 2011.

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.

Itay P. Fainmesser. Community structure and market outcomes: A repeated games in networks approach. *American Economic Journal: Microeconomics*, 4(1):32–69, 2012.

Itay P. Fainmesser and David A. Goldberg. Bilateral and community enforcement in a networked market with simple strategies. Working paper, January 2011.

Garance Genicot and Debraj Ray. Group formation in risk-sharing arrangements. *Review of Economic Studies*, 70(1):87–113, January 2003.

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

Avner Greif. Contract enforceability and economic institutions in early trade: The Maghribi Traders' coalition. *American Economic Review*, 83(3):525–548, 1993.

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.

Joseph E. Harrington, Jr. Cooperation in a one-shot prisoners' dilemma. *Games and Economic Behavior*, 8: 364–377, 1995.

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.

Benjamin Klein and Keith B. Leffler. The role of market forces in assuring contractual performance. *Journal of Political Economy*, pages 615–641, 1981.

Daniel B. Klein. Promise keeping in the great society: A model of credit information sharing. *Economics & Politics*, 4(2):117–136, 1992.

Narayana R. Kocherlakota. Implications of efficient risk sharing without commitment. *Review of Economic Studies*, 63(4):595–609, 1996.

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

Ethan Ligon, Jonathan P. Thomas, and Tim Worrall. Informal insurance arrangements with limited commitment: Theory and evidence from village economies. *Review of Economic Studies*, 69(1):209–244, 2002.

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.

Markus Möbius. Trading favors. Working paper, May 2001.

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. Law, economics, and inefficient norms. *University of Pennsylvania Law Review*, 144:1697, 1995.

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.

Hyun Song Shin. The burden of proof in a game of persuasion. *Journal of Economic Theory*, 64(1):253–264, 1994.

Satoru Takahashi. Community enforcement when players observe partners' past play. *Journal of Economic Theory*, 145(1):42–62, 2010.

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.

Alexander Wolitzky. Communication with tokens in repeated games on networks. Working paper, 2013.

B Supplemental Appendix (for online publication)

B.1 Proofs for Section 4

Proof of Proposition 2 on p. 13. Let $\overline{\phi}_d(\Delta)$ correspond to the solution in (7) in which n is replaced by d. In the permanent ostracism equilibrium, at every history h where $|\mathcal{I}(h)| = d+1$ each innocent player announces stakes $\overline{\phi}_d(\Delta)$ and works at those stakes when encountering another innocent player, and announce stakes of 0 with any guilty player. Players at every history are indifferent between working and shirking, and can never do strictly better by announcing any other stakes; 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: We begin by arguing 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 profile defined above.

Step 2: By Step 1, it suffices to establish that no permanent ostracism equilibrium supports cooperation at higher stakes than $\overline{\phi}(\Delta)$. In principle, stakes may be asymmetric (across partnerships) and history dependent on the equilibrium path. Take any such equilibrium, and let $\phi_{ij}(h^t)$ denote the stakes that players i and j would have after the full history h^t. Notice that the payoff from working at history h^t is increasing in $\phi_{ik}(h^\tau)$ for every equilibrium path history h^τ that follows h^t. Let $\phi = \sup_{ij,h}\phi_{ij}(h)$: because for an equilibrium, there is a uniform bound on the stakes across all pairs and all histories, the existence of ϕ is guaranteed. It follows that for every equilibrium path history h^t and every player i, the continuation payoff from working is at most $\frac{n\delta p_\Delta}{1-\delta}\phi$. Since there is some history along which $\phi_{ij}(h)$ is arbitrarily close to ϕ, it follows that

$$\frac{T(\phi)}{\phi} \leq 1 + \frac{n\delta p_\Delta}{1-\delta} = \frac{T(\overline{\phi}(\Delta))}{\overline{\phi}(\Delta)}. \tag{15}$$

Assumption 1 implies that $\phi \leq \overline{\phi}(\Delta)$, and so every mutual effort equilibrium supports stakes less than $\overline{\phi}(\Delta)$ in every history. □

Proof of Proposition 3 on p. 21. Consider a bilaterally rational permanent ostracism equilibrium with rich off-path beliefs. Let (m_i^t, m_j^t) be a pair of messages such that there is no interaction reported in the previous $(n-1)$ time periods, and suppose towards a contradiction that the equilibrium path stakes are$\phi_{ij}(m_i^t, m_j^t) > \underline{\phi}(\Delta)$. Without loss of generality, consider a history h_i^t that coincides with m_i^t before time $t-(n-1)\Delta$, and that in the previous $n-1$ periods, some other neighbor has shirked on player i. Notice that $h_i^t \in H_i^t(j)$ and suppose, without loss of generality, that player i attributes strictly positive probability to H_j^t. If player i reveals history h_i^t, then the pair work at stakes $\underline{\phi}(\Delta)$ whereas if player i conceals the previous $n-1$ interactions, then regardless of player j's history, they will chooses stakes that are at least $\underline{\phi}(\Delta)$, and with strictly positive probability, choose $\phi_{ij}(m_i^t, m_j^t) > \underline{\phi}(\Delta)$. Concealing negative interactions and shirking when the stakes exceed $\underline{\phi}(\Delta)$ is a profitable deviation. □

Proof of Proposition 4 on p. 22. First, consider the game in which each interaction time is publicly monitored. Consider a strategy profile in which a player is considered guilty either if he shirks while a partner works or if he conceals an interaction in this or any preceding period. At every history h such that $|\mathcal{I}(h)| = d+1$, each innocent player announces stakes of $\overline{\phi}_d(\Delta)$, and works at all $\phi \leq \overline{\phi}_d(\Delta)$ when encountering another innocent player, and announces stakes of 0 with any guilty player. Innocent players are thus indifferent between working and shirking, and no player—innocent or guilty—has a strictly profitable deviation from concealing an interaction or announcing any other stakes.

Second, suppose that independently of the past, each interaction is privately observed with probability $\varepsilon > 0$. An argument identical to Lemma 1 implies that stakes have to be no more than $\underline{\phi}(\Delta)$ if no interactions are reported in the last $n-1$ time periods, and therefore, Theorem 2 follows as a consequence. □

B.2 Proofs for Section 6

First, we define permanent ostracism in our setting with networked markets. We say that a player "works" if she makes the "promised" payment as a buyer, or delivers the "promised" quality as a seller, and we say that she shirks otherwise. As before, a player j is in $\mathcal{G}(h)$ if there exists an interaction in $\mathcal{E}(h)$ such that player j shirks in $\mathcal{E}(h)$ or conceals information.

Definition 8. *A strategy profile involves* ***permanent ostracism*** *if for every time t and history h_i^t,*

1. *If $\{i,j\} \subset \mathcal{I}(h_i^t)$, then player i reveals history h_i^t. If $m_j(h_i^t, t) \subseteq m_j^t$, and $j \in \mathcal{I}\left(h_i^t \cup m_j^t\right)$, then player i works.*
2. *If $j \in \mathcal{G}(h_i^t)$, player i sends message $m_i^t = \emptyset$, and makes a payment of* 0 *(if i is a buyer) or sets quality to* 0 *(if i is a seller).*
3. *After the interaction, player i sends a public message to her side of the market revealing it unless she shirks.*

Proof of Proposition 6 on p. 30. We first construct the permanent ostracism equilibrium that supports trade of quality $\hat{q}_b$. Suppose that regardless of the number of innocent sellers, each innocent seller produces an object of quality $\hat{q}_b$, and each buyer pays $\hat{q}_b$. Every incentive constraint is satisfied: each innocent seller is indifferent between deviating and producing at quality $\hat{q}_b$. Whenever a seller shirks, every buyer (weakly) prefers to communicate this publicly to the other buyers and to every seller that she meets. It follows from straightforward algebra that $\hat{q}_b$ exceeds $\hat{q}_1$ for $b > 1$ and that $\hat{q}_b$ strictly exceeds $q_{b,s}$. It is clear that $\hat{q}_b$ is the maximal trade in any symmetric and stationary equilibrium; it remains to be shown that this is greater trade than any asymmetric or non-stationary equilibrium. The argument is identical to the logic of Proposition 2: a seller's incentive to trade is strictly increasing in her surplus from future trade in that relationship and in other relationships, and so the equilibrium with maximal trade is symmetric and stationary. □

Proof of Proposition 7 on p. 31. Suppose otherwise. Suppose that player i meets player j and the latter is richer while the former is poorer. Consider a history for player i, h_i^t in which all players other than j have failed to make their transfers, and player j communicates a history h_j^t in which $\mathcal{I}(h_j^t) = \mathcal{N}$. If player i communicates truthfully, then he obtains transfers ζ_1, and so his expected payoff today is $u(\underline{y} + \zeta_1)$ plus the continuation value from bilateral enforcement. In contrast, consider a deviation in which he conceals the information today and in all subsequent meetings. He obtains $u(\underline{y} + \zeta_n)$ today. In the future, other players do not reveal that they failed to make the transfer to player i and so player j never learns that player i deviated. Thus, in the future, whenever players i and j meet, they transfer at least ζ_1, and thus sustain a higher continuation value than when player i reveals the truth to player j. □