Deceit, Group Structure, and Cooperation

Working Paper*

Jennifer M. Larson

Harvard University

Abstract

Tight-knit groups of people can use decentralized punishment schemes to keep fellow group members cooperating with each other and with members of other groups. When groups are less tight-knit, members can still induce cooperation with decentralized punishment if they rely on the honesty of others (Larson, 2012*b*). Here, I generalize Fearon and Laitin (1996) to consider institutions that can sustain cooperation in imperfectly tight-knit groups when players may behave opportunistically and lie. I show that opportunities to lie abound when communication networks are incomplete. Groups can use community enforcement so long as lies are detectable, which depends on access to multiple sources of information and sufficiently long memory. By trusting in the face of uncertainty and reacting strongly to lies, groups can induce both honesty and cooperation with an institution that is robust to errors in behavior. Small clusters of contacts are more valuable than large sets of unconnected contacts in deterring liars. Inter-group cooperation is also possible, though interactions with an out-group admit greater opportunities to lie– not in order to defect against the out-group, but to frame in-group members and defect against them.

*Current version: April 20, 2012. First version: December 2011. This is still a working paper very much in progress; please do not cite or circulate without permission. I welcome any feedback: jmlarson@fas.harvard.edu

1 Introduction

Groups of people are often tasked with governing themselves. Decentralized punishment institutions maintain cooperation well when everyone knows and communicates with everyone else in a group. News spreads so rapidly that any misdeed can be punished swiftly and thoroughly. This logic is well-known and has been used to explain a wealth of examples of cooperation outside the 'shadow of the law,' from ethnic groups peacefully coexisting to traders keeping their word to ranchers minding where their livestock graze (Fearon and Laitin, 1996; Greif, 1993; Ellickson, 1991). When instead there is heterogeneity in the extent of direct communication between group members, so that some members are more peripheral than others, ensuring that everyone cooperates is more difficult. Some players are more likely to get away with offenses than others, and some players are more tempting targets than others, which ensures that universal cooperation is more difficult. In addition, some players could take advantage of others' ignorance, even if only temporary, and stand to gain from lying. This paper relates the exact network of communication in a group— a map of who communicates with whom— to how well that group can enforce cooperation *and* honesty.

Doing away with the standard assumption that everyone in a group stays perfectly informed about all others, I generalize the canonical model of inter-ethnic cooperation in Fearon and Laitin (1996) to consider the implications of an incomplete communication network (a network in which some people communicate directly with only some but not all other people). Incomplete communication networks are probably the more realistic depiction of groups with features like large populations, or time constraints, or high turnover, or limited communications technology, or sparse populations, to name a few.[1] In such cases, news may travel slowly enough that some are left out of the loop for at least some of the time. In Larson (2012*b*) and Larson (2012*a*), I show that when that is the case, exactly who communicates with whom affects how well decentralized punishment schemes can induce inter- and intra-group cooperation when people tell the truth.

Here I reconsider the role of networks in maintaining cooperation when people have incentives to lie as they spread information through the network. I show that opportunities to lie abound, even when networks contain nearly-but-not-quite all possible links. Groups can induce honesty and communication when they carry on as if people are truthful, eventually forgiving misdeeds,

[1] Even networks with low-cost links like Facebook friendship networks are incomplete Mislove (2009) and the network of user activity among friends is even sparser Wilson et al. (2009).

but react in outrage and terminate cooperation when they learn a lie has occurred. Some groups are better suited to enforce cooperation with such an institution. I show that groups with redundant paths between players are in a better position to verify lies early and so discourage them. Likewise, clustering among neighbors in the communication network prevents lying, making dense social cliques more valuable to honesty than a large number of neighbors.

2 Liars

In "environments of uncertainty" in which not everyone directly observes everyone else, communication becomes essential for enabling group enforcement of cooperation. Players can observe some actions, and can send messages to neighbors about what they have observed. If players send messages truthfully, cooperation depends on how well-integrated players are in the network Larson (2012*b*). To consider such a case, the content of the messages is by assumption not chosen strategically– players have no option to lie about their actions or motivations. This is an admittedly strong assumption, since, as I show below, players stand to gain from lies if they could get away with them.

The assumption of truthful communication may not be so misguided. Groups may develop a norm against lying or crying wolf in other contexts and apply that norm generally even if recalculation in the new context would show gains to lying. Other contexts might contain severe consequences to lying: since links in a social network are valuable, trustworthiness is particularly important among linked individuals and deceit risks severing those valuable links (Karlan et al., 2009). An evolutionary view supports the idea that truth-telling may be adopted because of its value in at least some contexts– for example, sociobiology identifies an evolutionary mechanism favoring truthful gossip given its role in fostering reputation mechanisms for cooperation in mobile and dispersed groups (Dunbar, 1998; Enquist and Leimar, 1993). Applying norms from one context to another may be especially likely in groups in which members interact in many different ways- business interactions and every-day interactions and social interactions, say.[2] Ethnic groups have this feature, and tend to be characterized by trust among members (see, for example, Horowitz, 1985) More simply, experimental evidence suggests that when given the opportunity to gossip about others, players in cooperation games do so truthfully (Sommerfeld et al., 2007). We might also imagine that this game is embedded in a larger game left unmodeled

[2] This is the idea of multiplex networks in Sociology. See, for example, Fischer (1982); McPherson, Smith-Lovin and Cook (2001), p. 437.

which contains high-enough risk of being discovered and punished for lying that players play this game truthfully.

Here I consider the possibility that none of this reasoning applies, and discuss the options for sustaining cooperation *and* truth-telling in light of incentives to lie. After all, we can easily imagine groups in which players would take opportunistic behavior if they could. Groups of legislators or traders, for example, might be less cohesive and may have developed less of a norm of honesty than groups formed around social or hereditary ties.[3]

3 Imperfect Private Monitoring with Communication

Groups of people often must make do with less-than-perfect information. In many real situations, people only interact with a subset of people, only observe a subset of people, only talk to a subset of people, or only hear about everyone else with noise. These possibilities have motivated a set of theoretical work that hunts for institutions that can induce cooperation despite missing, and sometimes erroneous information. When players do not have perfect information about everyone else, institutions incentivizing cooperation become complicated quickly, and the uncertainty renders the the usual tricks for finding cooperative equilibria useless.[4] Making players prefer cooperation to defection is difficult since the set of people who could target a perpetrator with punishment shrinks, and opportunities to profit from lying can be present.

The simplest approach in such an environment is to design an institution that does not require communication (see Kandori, 1992; Ellison, 1994). In the seminal work of Kandori (1992), it is shown that even if players are in an extremely-limited information environment in which they observe only the games in which they are involved, there exists an institution which can ensure cooperation without communication. Such an institution has players play the harshest, most alarmist strategy available, switching to defect forever once they observe a single defection. Of course such an institution is hardly robust since a single error or misinterpretation cascades into full defection and destroys cooperation forever. In many settings of interest, people *could* communicate, so the assumption of zero communication is overly restrictive for many purposes.

Acknowledging that people have the capacity to communicate, a second approach assumes players share the results of their games as widely as possible. Sometimes players are assumed to do so truthfully (as in Kandori, 1992; Larson, 2012*b*), or are assumed to be perfectly known

[3] College students report lying in 1 out of every 3 social interactions (DePaulo et al., 1996).

[4] For a careful exposition of the difficulties arising in such games, see Kandori (2002).

by some central, information-aggregating figure or mechanism (as in Anderlini and Lagunoff, 2006; Takagi, 2011). A related approach acknowledges that communication may be untruthful, but black-boxes the question of detection. In Harbord (2006), for example, players might lie but their lie will be detected (somehow) with a positive fixed probability.

Since players may face profits from lying (discussed below), for an institution to stand on its own to produce cooperation, it must also entice players to tell the truth (or account for the possibility that players may lie). When private monitoring is imperfect, a few means of inducing truth are available. One assumes that lies are either known to or detectable by some (see Calvert, 1995; Aoyagi, 2000), or that lies and the identity of the liar can be inferred Kandori and Matsushima (see 1998). In the latter approach, if the distribution of payoffs is just right, then players can use statistical inference to detect deviations, and if each player's deviation has a unique signature, players can know who the liars are. Such an approach is computationally taxing for the players. Another, computationally-easier solution uses the fact that players can determine when a lie has occurred because monitors' reports will disagree (see Ben-Porath and Kahneman, 1996, 2003). By ensuring that everyone is monitored by at least two others and punishing all who give conflicting reports, the monitors can coordinate on true reports.[5] If instead players can be insulated from the results of their lies so that the veracity of their communication is independent of their own payoffs, then players no longer have a strict incentive to lie (see Compte, 1998; Lippert and Spagnolo, 2011). The interest here is in the harder case of lies from which players could strictly profit.

The approach taken here is in line with the usual solutions to private monitoring problems, though grapples with a slightly different lying problem. Here, players are observed by their neighbors and can communicate with them, who may pass along the messages to others who may pass them along, and so forth. Players have an incentive to misrepresent the history they know to defect without punishment (or at least with delayed punishment). This means their incentive to lie is greatest with second- or greater- hand information. Players' payoffs are realistically tied to their own messages (this tackles the heart of the lying problem, rather than rendering communication meaningless). I consider the possibility that players place trust in their group and presume that communication is truthful. Given truthful communication, sorting out defections is relatively simple. Any instance of untruthful communication is not tolerated,

[5] Annen (2011) suggests that this approach may be unsatisfying since coordinating on false reports would also be an equilibrium. If any collusion is possible, we might expect this instead of the truthful equilibrium.

though, and any indication that lying has occurred shatters players' confidence in their group and breaks down cooperation. Such a strategy is more robust than a trigger strategy used for everything (here a capitulation, or ”repentance” strategy is used for misdeeds).

This institution recognizes a qualitative difference between misdeeds and lies. Such a distinction comports with anecdotal evidence from everyday experience. A cheating significant other is usually punished more severely for lying about the scandal than committing it. Students playing a repeated prisoners dilemma can often recover from a rocky start of exchanging defections, but not from lying about what they will do. The approach here nests a game in which players assume messages are truthful in a broader game which very severely punishes lies. Lies are more difficult to detect since detection requires encountering both the true and the false information. Such an institution holds together cooperation and truthful communication, and only unravels if someone lies.

4 The Model

4.1 Model Setup

Consider two groups, A and B, each with a set of players $N = \{1, \ldots, n\}$. In each time period, all players play one round of prisoner's dilemma with a randomly assigned opponent. With probability p a player is paired with a member selected uniformly at random from the other group; with probability $1 - p$ a player is paired with a member selected uniformly at random from his own group. Each pair plays a single round of the prisoner's dilemma with common payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} 1,1 & -\beta,\alpha \\ \alpha,-\beta & 0,0 \end{pmatrix} \end{array}$$

where $\alpha, \beta > 1$ and $\frac{\alpha-\beta}{2} < 1$. Players are rematched in each round; rounds occur indefinitely. A player's total payoff is then a stream of discounted single round payoffs. Players have common discount factor $\delta < 1$.

Let $(N, g)_A$ be a fixed simple, undirected communication network for group A with nodes N

and $n \times n$ adjacency matrix g such that entry $g_{ij} = 1$ indicates the presence of a link between players i and j, $g_{ij} = 0$ indicates the absence of a link between players i and j. Let $(N, g)_B$ be the same for group B.[6] No links span $(N, g)_A$ and $(N, g)_B$.

Players can perfectly identify and recognize in-group members, but can only identify the group of out-group members. The network $(N, g)_A$ is common knowledge among players in A and $(N, g)_B$ is common knowledge among players in B; players know the shape of the other group's network but not who sits where (i.e. players only know the permutation class of the network of the other group).[7] The roster of random assignments and the actions played in each round are not observable to all players. Players observe only the matching and actions of their neighbors. This information spreads further through the network via messages passed from person to person governed by a rate of transmission, described below.

4.2 Communication

In each period t, after playing one round of prisoner's dilemma, a stage of communication occurs. Information about each game played between two players in each round is packaged in a message m containing the identity of the players (as specific as possible), the time period of the round, the actions of both players, and the motives of the players (relevant motives are determined by the strategy. For the strategy profile σ^{NWIGP} below, the relevant motive will be whether D was played out of punishment or defection.) Messages about games between same-group members i and j in t perfectly identify both players, $m_{i,j,t}$, and are sent to the neighbors of i and the neighbors of j in the communication network, e.g. to $N_g(i)$ and $N_g(j)$. Messages about games between players from different groups cannot perfectly identify the out-group opponent, and so if $i \in A$ and $j \in B$, a message $m_{i,B,t}$ is sent to neighbors of i and a message $m_{A,j,t}$ is sent to neighbors of j. A message $m_{i,j,t}$ expires after E rounds, in $t + E$. An expired message is no longer passed on. Let r govern the rate of communication spread so that player i does the following r times before $t + 1$ begins: send an unsent message about i's own game played in t and forward all unsent, unexpired messages to all of i's neighbors. This means that unexpired

[6] The assumption of identical networks for both groups is unnecessary but simplifies exposition. Both groups must be able to enforce cooperation which depends strictly on their own network. If both can do so, the equilibrium obtains.

[7] This assumption simply makes the coordination of the equilibrium more plausible. To carry out equilibrium strategies, technically players need know nothing about the other group. If they do know the shape, the know how well the other group could enforce their own, which might make it more likely that the two groups would have arrived at this equilibrium in the first place.

messages originating at i in time t are received by all players reachable in r degrees or fewer before $t+1$.

Actions are assumed to be observable to neighbors in the communication network. That is, when player i is assigned to play j in t, i's neighbors observe that he played j and who played which action. The first round of "communication," then, can be thought of as a message in which certain contents are true deterministically- the identity of the players, the time period of the round, and the actions of both players. First, note that this still leaves something in the message to lie about- the motives of both players. Seeing that i played D and j played C is insufficient to conclude that i is now guilty. It could be that i was rightfully punishing j. More on the ability to lie below.

If all players did observe all other players (as in a complete communication network), the information about motive would be redundant in the message. All players could observe a round in which the actions are D and C and could correctly infer guilt by using their past observations to determine if the person playing C had punishment coming from a past round, which would require rounds further in the past to establish, and so on. Here, some players do not observe the full history of everyone in the game and so players rely on others' assessment of guilt in lieu of missing history. Below I discuss the difficulty of verifying guilt and the most enticing lies.

Communication spreads deterministically in that each round, everyone passes the correct number of messages to the correct recipients, and all messages intended to be sent are received unchanged. In other words, there is no misunderstanding or accidental mistakes in sending or interpreting messages. There can, however, be intentional deceit. A player can include in his message false information. The case of honest-by-assumption communication is considered in Larson (2012*b*). Here, players may strategically choose the content of their messages.

4.3 The Value of Lying

When players choose the content of their gossip strategically, some players may have an incentive to lie. Assume that actions are observable to neighbors, so that information about who played which action is still conveyed truthfully to immediate neighbors.[8] Players would prefer to defect and not be punished, so players have an incentive to play D and claim they were punishing someone who deserved to be punished rather than defecting maliciously. Unless the

[8] This assumption seems to comport with the real issue with lying. Even among a decentralized group like the Nuer in the Sudan, when players seek adjudication for disputes, disagreement is rarely over who did which action; disagreement is over whether the actions were deserved (Evans-Pritchard, 1940).

player's neighbors know the complete history of every player in the game, some opportunity can arise in which they can not tell the difference between D the defection and D the punishment. This ambiguity is a problem because either defections are sometimes unpunished, or the group responds with a scheme which overpunishes and reduces incentives to cooperate.[9]

Specifically, a player i would like to send a message to each neighbor claiming that his current opponent j deserves punishment. Below I discuss exactly what lie would be most profitable.[10] Here I focus on the opportunity to lie. In the present setup, players observe neighbors' actions. If everyone were neighbor to everyone else (as in the complete network), players could perfectly distinguish D the defection from D the appropriate punishment because they observe all necessary information. It turns out that even small departures from the complete network admit the possibility of lying.

Call a player *discerning about* j if he can distinguish between j playing D the defection and j playing D the appropriate punishment using his observations of other players' actions.

Proposition 1 (Unverifiable Histories). *When a group interacts exclusively with in-group members, a player i is discerning about any $j \in N$ if and only if i has degree at least $n-2$. When a group also interacts with an out-group, a player i is discerning about any in-group player $j \in N$ if and only if player i has degree at least $n-1$.*

The proof can be found in appendix 1. For the intuition, consider the condition on players in groups interacting among themselves and with an out-group. The condition says that a player must observe all other in-group players. Suppose player i observes that j play D against k who plays C. To know whether j was punishing or defecting against k, i must know whether k deserved punishment or not. This requires knowing whether k always cooperated when required and punished when required, which requires knowing whether k's past opponents always cooperated when required and punished when required, which requires knowing whether k's past opponents' past opponents always cooperated when required and punished when required, and so on recursively. Since opponents are assigned at random, in order for i to be able to distinguish good Ds from bad Ds for any possible sequence of partner assignments, i must know everyone else's history of play.

[9] This problem of ambiguity is well known in the study of overlapping generations models. See Takagi (2011) for a solution which makes use of a public signal that truthfully conveys information about the complete history.

[10] To preview, i's most profitable lie is one which he claims to one neighbor that he just learned a different neighbor was defected against by someone far away in the network. This must mean that the alleged culprit defected against one of i's neighbors that is rT^p degrees away rT^p periods ago. If the victim were a closer neighbor, i should have already known *and told his opponents* that j deserves punishment.

Note two features of this proposition. First, the presence of an out-group with whom players sometimes interact makes distinguishing good Ds from bad Ds more difficult. Since players observe their neighbors *and their neighbors' opponents*, so long as a player is connected to all but one other in-group members (a total of $n-2$ others), they will actually observe everyone every round. If i is connected to everyone but j and all rounds are between in-group members, j will always play someone i is connected to. In this way, i observes every in-group player's full history. If players sometimes play an out-group opponent, then part of j's history will be unverifiable to i. That is, j will sometimes play an out-group member and i won't see what happens in this round. Players must be connected to all other in-group members (all $n-1$ others) to make motives verifiable through observation. Below I further discuss the difference between groups exclusively interacting within their own groups and those interacting with out-groups.

Second, in order to guarantee that every motive will be verifiable to every player, all players must be discerning about all other in-group players. This implies a lower bound on the number of links that must present in a network for full verifiability.

Corollary 1 (**Verifiable Networks**). *When a group interacts exclusively with in-group members, all motives are verifiable to all players only if the network contains at least* $\frac{n^2-2n}{2}$ *links. When a group also interacts with an out-group, all motives are verifiable to all players only if the network contains all* $\frac{n(n-1)}{2}$ *links.*

That is, for solo groups, players can verify everything so long as only $\frac{n}{2}$ links are missing (out of a possible maximum of $\frac{n(n-1)}{2}$) and distributed in such a way that everyone is missing only one link. For groups interacting with out-groups, players can verify everything only if the network is complete. When the networks do not contain enough links, there is always a history such that the true motive of a D played will be unverifiable to at least one player.

Lies are possible when motives cannot be perfectly verified through observation. Corollary 1 implies that very small deviations from a complete network admit the possibility of lying.[11] To cope with missing information, players can communicate to fill each other in. If i doesn't observe j but k does, k could tell i what j did. If i doesn't know k but knows a mutual friend

[11] This makes the assumption of a complete network on the grounds that real world communication networks are "close enough" to complete especially suspect. Fearon and Laitin (1996) justify the assumption of perfect information within a group on the grounds that "ethnic groups are typically marked by relatively dense social networks and low-cost access to information about other group members' behavior [...] While it is not literally true even in small ethnic groups that all members observe all interactions [...], rumor, gossip and inquiry tend to be more developed and efficacious within than between ethnic groups" (p. 721). That news can eventually reach everyone does not imply equivalence to perfect observation.

l, i could learn from l who learns from k what j did. These second- or more- hand accounts are a natural way for people to fill in information gaps.

The difficulty with second- hand information is that players have an incentive to strategically choose what their messages contain. In particular, if a player could claim to have information that someone deserves punishment so that he can play D without himself incurring punishment, he would like to do so. To induce honesty, there must be consequences to lying. In particular, it must be possible to detect lies and punishment must be strong enough to make players prefer honesty.

4.4 Detecting Lies

Since opportunities for lying are present even in very densely connected networks and since there are gains to lying, players must be able to detect lies in order to punish and disincentivize them.[12]

Before specifying a sufficient punishment strategy, consider when detecting lies is even a possibility. Messages contain a variety of content that is all chosen by the sender, including who played whom in which round, who played which action, and who if any are defectors. If a player faces potential punishment for a detected lie, all else equal, if there are opportunities to better conceal a lie, a player would make use of them. This intuition gives rise to the following characterization of a detection scheme:

Remark 1. *Detection devices which rely on aspects of the message solely manipulable by a prospective liar are suboptimal because they present a greater opportunity to conceal a lie.*

To see a straightforward example of this, consider a poorly-designed detection device which deems any message about player i to be a lie. Obviously if j wanted to construct a lie, he could do so about a player other than i to thwart detection. While this detection scheme is far-fetched and obviously suboptimal, more complicated schemes that relied on information about which players can lie are also suboptimal for the same reason.[13]

[12] Technically, players could give up on trying to detect lies and treat all messages as ignorable cheap talk. Kandori (1992) gives an example of an institution that can ensure cooperation in such a way. Given the prevalence of gossip and interpersonal communication, here the goal is to characterize an institution that preserves the usefulness of communication.

[13] Consider a detection device which, once a lie has been detected, claims that the identity of the liar is the first person to profit from such misinformation. Any liar could include in his lie the claim that his neighbor profited from the misinformation first and again thwart detection. More on this below.

Two conditions must be met for a lie to be detectable. The first is that the liar's information must reach someone who also heard the true information. To ensure honesty in the community, this must hold for all possible liars and receivers of true information, leading to the following proposition:

Proposition 2 (3rd Party Reachability). *Lies are detectable on a network only if information from any pair of players can reach a third player without passing through the other of the pair. That is, for any $i, j \in N$, $\exists k \neq i, j$ such that there exists a path from i to k that does not include j, and there exists a path from j to k that does not include i.*

Proposition 2 ensures that however player i constructs a lie, the lie can eventually reach someone who will also know the true information. Even if this third party cannot tell which message is true and which is false, he will detect a conflict which indicates the presence of a lie.

Of course, the closest third party may be far away in the network, which could mean that the conflict will not be noticed for many rounds of play, the more so if the rate of information transmission, r, is small. If players only retain information for a finite number of rounds before wiping the slate clean (or plumb forgetting), the conflict may never be noticed or may be ignored. Hence, a second condition is necessary to ensure the detection of lies. Call the length of the shortest path from i and j to the nearest third party k the **Shortest Third Party Path ($STPP_{i,j}$)**, and let M be the length of memory, so that players remember information for M rounds of the game.

Proposition 3 (Good Memory). *Lies are detectable on a network only if the length of memory, M, is weakly longer than the number of rounds required to reach the nearest third party of any pair. That is, lies are detectable only if*

$$M \geq \left\lceil \frac{\max_{i,j}\{STPP_{i,j}\}}{r} \right\rceil.$$

Proposition 3 simply states that players must remember information for a sufficiently long time so that they notice conflicting information and hence can detect lies. All lies are detectable so long as even the most difficult to monitor player's lies can be detected. Notice that the slower information spreads through the network, the greater the demands on memory. When information takes a long time to reach others (as when communications technology is poor, or news is delivered in person over long distances, etc.), the lie may not reach anyone who knows conflicting information for a very long time. The required length of memory also increases in the distance to the farthest-away-closest third party[14] by the same intuition.

Take an example which violates the condition in Proposition 2, shown in figure 1. Pairs of players on the “spokes” of this star network are reachable by a third party, like players 1 and 2 who are directly reachable by 6. Any pair of players entailing player 6, though, are not. Take the pair 1 and 6: player 1 can only reach anyone else through player 6. This violates Proposition 2 and gives player 6 an advantage. He could lie to someone on a spoke about someone on a different spoke without that information clashing with true information. Hence, a central figure who knows all is not sufficient to prevent lying since *he* can have incentives to lie.[15]

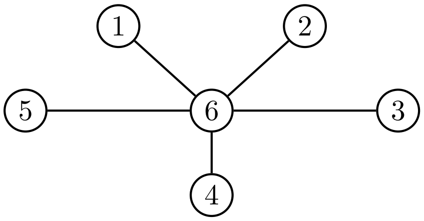

Figure 1: Example communication network with 6 players which violates Proposition 2

Likewise, consider an example which meets the condition in Proposition 2 but violates the condition in Proposition 3. Suppose the rate of communication is $r = 1$, so that news spreads a single degree each round. Suppose also that $M = 3$, so that players remember information for three rounds after the information is generated and then forget it[16], and that the communication network is as shown in Figure 2.

The network in Figure 2 meets the condition in Proposition 2 since, for any two players, a third can be reached without needing to travel through the other of the two players.[17] The nearest third party to 1 and 5 is player 3 (or player 7), which can be reached by both 1 and 5 in 2 degrees ($STPP_{1,5} = 2$). Pairs with the longest paths to third parties are neighboring pairs, like players 1 and 2. The third parties that can be reached by both 1 and 2 in the shortest distance are 5 and 6. That is, if 1 tells a lie about 2 to his neighbor 8 but 2 tells the truth about himself to his neighbor 3, those messages will conflict and the first to notice the conflict will be

[14] That is, the third party closest to the pair who are farthest from their third party

[15] This observation is noteworthy in light of the special attention that “star networks” like the one in 1 receive in the strategic network formation literature (see Jackson, 2003). Such networks avoid redundant connections, but in settings where information needs to be verified to thwart lying, some redundancy is essential.

[16] This version of memory can also be thought of as the salience of information. Notice that players do not remember information for M rounds *after they first hear the news*, but rather for M rounds *after the news is generated*. Some news is about the too-distant past to be interesting or cared about. Even with this version of memory, it is still the case that the larger is M, the more information players must keep track of, so we would expect natural constraints on M.

[17] In fact, here, *all* others can be reached without needing to travel through one of the two players.

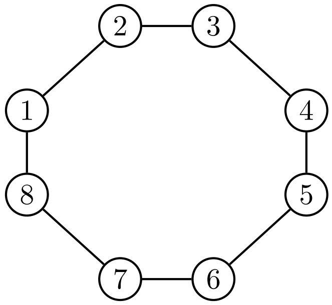

Figure 2: Example communication network with 8 players which meets Proposition 2 but violates Proposition 3 when $r = 1$, $M = 3$.

players 5 and 6 after 4 steps. The $\min_{i,j}\{STPP_{i,j}\} = 4$ here, and since $r = 1$, memory M must last at least 4 rounds. If news is only interesting 3 rounds after it is generated, by the time the lie reaches player 5 it is too old to pay attention to, and likewise for the truth reaching player 6. Neither would notice the conflict if news expired this quickly.

Speeding the spread of information (r), adding a link which jumps across the ring and shortens the maximum $STPP$, or increasing memory (M) would all allow the network in Figure 2 to meet both Propositions 2 and 3 and would make all lies detectable. These conditions do *not* guarantee that all *liars* can be identified, simply that the existence of a lie can be detected. It still may be that no player knows which of the conflicting accounts is true. As I discuss below, anonymous liars can still be incentivized to tell the truth with sufficient punishment.

5 Cooperation and Honesty

5.1 Strategies to Promote Cooperation, Taking Truth for Granted

When players ignore the urge to manipulate their messages and instead communicate truthfully, the following strategy profile results in full cooperation in equilibrium under certain conditions (shown in Larson (2012*b*)):

Definition 1 (Network Tit-For-Tat, NWTFT). *Always punish a player (play D) known to be in punishment phase. Always cooperate with a player (play C) not known to be in punishment phase.*

Punishment phase is a status defined by the strategy profile below- in general, being in punishment phase is undesirable and the threat of entering punishment phase is what drives players to cooperate. Here, a player can "know" because of experience in his own rounds, or because of received information. A player is "not known to be in punishment phase" by a player i if i has not experienced or heard gossip that the player is in punishment phase. Now consider a strategy profile in which players play NWTFT according to the following guidelines:

Definition 2 (Network In-Group Policing, σ^{NWIGP}). *All players play NWTFT using the following definitions of status: all players begin as cooperators (not in punishment phase). A player enters (or reenters) punishment phase for T^p periods when that player (1) defects against an out-group member, (2) defects against someone not known to be in punishment phase, or (3) cooperates with someone known to be in punishment phase. A player i is known by his opponent to be in punishment phase when his opponent was the victim of (2) or (3) committed by i in at least one of the past T^p rounds or his opponent has received a message that i committed (1), (2) or (3) in at least one of the past T^p rounds.*

Punishment here takes the form of capitulation (playing C against a punisher playing D) as it does in Fearon and Laitin (1996) and Calvert (1995). That defectors may be required to atone for their offense by capitulating or repenting for a certain number of rounds has the nice property that players like to punish defectors, and squares with real world observed punishment regimes (Harbord, 2006). The above is really a set of strategy profiles, each with a fixed length of punishment phase T^p. Larson (2012*b*) shows that some sparse networks and slowly communicating groups can only maintain full cooperation with a sufficiently high T^p.

Full cooperation is possible, assuming truthful communication, when the following conditions are met:

Proposition 4 (Full Cooperation ASSUMING Truth-Telling). *σ^{NWIGP} is sequentially rational for game G with networks $g_A = g_B$ iff*

$$\delta^{T^p} \geq \max\left\{\frac{\alpha - 1}{(1-p)z^{out}_{min,rT^p}(\beta+1)}, \frac{\beta}{(1-p)z^{in}_{min,rT^p}(\beta+1)}\right\}$$

and

$$p < \min\left\{\frac{z^{in}_{min,rT^p}(1+\beta) - \beta}{z^{in}_{min,rT^p}(1+\beta)}, \frac{z^{out}_{min,rT^p}(1+\beta) - \alpha + 1}{z^{out}_{min,rT^p}(1+\beta)}\right\}$$

where z^{out}_{min,rT^p} and z^{in}_{min,rT^p} are the probability that the least-punishable defection will be punished by an in-group member at the end of the punishment phase (in $t + T^p$) for defecting in t against an out-group member and in-group member, respectively.[18]

The binding player here is the one least observable by other players; in other words, the most "peripheral" player. In figure 3, player 4 is most peripheral. If news travels slowly enough and the punishment phase is short enough, he can expect to defect and face little punishment *even if news about him travels truthfully.*

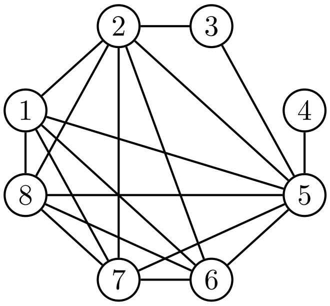

Figure 3: Example network where player 4 is most peripheral and so is the binding player for full cooperation, assuming truthful communication.

5.2 Strategies to Promote Cooperation AND Honesty

Players can induce cooperation and honesty by punishing lies as well as defections. The complication is that when a network is missing even a few links (as per Corollary 1), sorting out defections from punishment is difficult when players can strategically choose their messages. Even when the conditions in Propositions 2 and 3 hold, players may have difficulty determining *who* the liar is. Standard solutions like shoot-the-messenger approaches (see, for example, Ben-Porath and Kahneman (1996)) become quickly intractable when messages are passed along second- and greater-hand since there can be a string of messengers passing news before it reaches conflicting news.[19] The approach taken here blends the grim approach of Kandori (1992) with shoot-the-messenger approaches to capture an institution seemingly undertaken by real-world groups.

In particular, players carry on assuming that other players are trustworthy and react strongly if they receive evidence that their trust was taken advantage of. Players effectively play a game

[18] The proof of Proposition 4, a discussion of consistent beliefs to extend the strategies to a sequential equilibrium, and the full specification of the probabilities z^{in}, z_{out} as functions of network characteristics can be found in Larson (2012*b*).

[19] When messages are passed along indirectly, it becomes difficult (and sometimes impossible) for players to determine where the actual punishment for lying started, how long to carry out punishments, and how to coordinate these decentralized efforts so that they eventually end and players can return to cooperating.

in which they naively trust each other which is embedded in a broader game in which violations of this trust trigger the end of cooperation. Players are willing to trust until they learn a liar is afoot.

Consider the following strategy profile. It is a modification of σ^{NWTFT} above in which players also choose the content of their messages and are sensitive to conflicting information.

Definition 3 (Network In-Group Policing with Messages, σ^{NWIGPM}). *All players begin in "trustworthy" phase and regard all messages as truthful. So long as no information conflicts, play NWTFT using the following definitions of status: all players begin as cooperators (not in punishment phase). A player enters (or reenters) punishment phase for T^p periods when that player (1) defects against an out-group member, (2) defects against someone not known to be in punishment phase, or (3) cooperates with someone known to be in punishment phase. A player i is known by his opponent to be in punishment phase when his opponent was the victim of (2) or (3) committed by i in at least one of the past T^p rounds or his opponent has received a message that i committed (1), (2) or (3) in at least one of the past T^p rounds. At the end of a round, send all neighbors a true message containing the time, opponents, actions, and motives of the most recent round, and pass truthfully all unexpired, received messages r times. If a player receives information that conflicts with any of his previously known information or observes a "jaded neighbor," he switches to "jaded" phase and plays ALLD.*

In a nutshell, players presume truthful messages and punish defections in the usual way. Once someone learns that all have not been truthful, jadedness sweeps through the network, terminating cooperation.

Note that such an alarmist strategy does not require knowing the identity of the liar. A player need only know that a lie has occurred, not the identity of the liar, to trigger the grim punishment. Being a grim punishment, a cooperative equilibrium will not be fully robust to mistakes. If someone mistakenly lies or sounds the liar alarm by switching to jaded phase, cooperation is terminated. However, such an equilibrium *will* be robust to mistakes in behavior. If someone accidentally plays D when they were supposed to play C, players observe misbehavior that deserves to be punished. So long as news spreads consistently about the defection, the liar alarm does not sound and cooperation can resume after the erring player incurs a finite punishment phase. Such an equilibria is more robust than that induced by fully-grim punishment as in Kandori (1992).

Observation suggests that people behave similarly to the prescription in σ^{NWIGPM} in environments in which truth is difficult to verify. Take romantic relationships. Not every late evening at the office can be verified, nor is a partner necessarily incentivized to be faithful at all times. Anecdotally, there does seem to be a difference between reactions to misbehavior and reactions to *lying* about misbehavior. The latter tends to elicit a stronger reaction, as in "worst of all, you weren't honest with me."

Likewise, any in-class run of a finitely repeated dilemma results in some students colluding despite instructions to the contrary. Pairs which never collude but get off to a rocky start can return to cooperating. Not so for the pairs who try to collude but take advantage of the other's trust. Once the communication is revealed to be a lie, game over- mutual defections finish the game. Comparing reactions to misbehavior to reactions to lies is ripe for an experimental setup. More on this below.

5.3 Most Profitable Lies

Since players pass along information second-hand, a player considering lying has options. He could lie about himself, or he could lie about other people. A liar benefits from playing D against a C without incurring punihsment for his misbehavior, but faces costs from tripping the alarm and converting his group to disillusioned non-cooperators.

A prospective liar's costs are greater when he trips the alarm more quickly, and when the cascade of jaded players unfolds more quickly. To delay tripping the alarm, the liar must maximize the time until his false information collides with the true information. Recall that player i is only supposed to punish people he *knows* deserve punishment, and also that player i passes along information he knows to his neighbors. If player i is assigned to play j, defects against him and tries to tell his neighbors that j had it coming, his neighbors should know i knew this if i supposedly learned about j's treachery long enough ago. The only way i's neighbors would think i might know about j's treachery without themselves knowing i knew it is if i just learned it and hasn't had a chance to pass the information along.

Remark 2. *A lie which only misrepresents one's own motive is not as profitable as a lie which misrepresents one's own motive and lies about someone else's history of play.*

In other words, the best lie i could make is to claim that his newly assigned opponent j defected against someone in the past such that i just learned about the defection at the end of last round and didn't yet share it with his neighbors. Any other lie would either create conflict immediately or earn i punishment for punishing someone he did not know to deserve punishment.

Lemma 1 (Most Profitable Lie). *Player i's most profitable lie is of the form: i tells neighbor j that i's current opponent defected against someone tk degrees away in round $t-k$ so long as j does not sit on the path between i and the supposed victim tk degrees away.*

If player i can tell his most profitable lie to a neighbor who does not sit on the path between i and the supposed victim tk degrees away, he can avoid punishment from the lied-to neighbor and the neighbor's neighbors and so on. If he can tell such a lie to all neighbors, which might entail telling different neighbors different lies, he can avoid any immediate punishment from players other than the victim (who experiences a defection against him and will start punishing the misbehavior).[20]

Proposition 5 (Most Profitable Bundle of Lies). *A player's most enticing set of messages about a particular opponent is the set which tells a most profitable lie to each neighbor.*

So the most likely liar is the player i who has an opponent that could be assigned such that i can tell a most profitable lie about the opponent's history and i's message will take as long as possible to collide with information from either the opponent or the opponent's alleged past victims.

Remark 3. *There exists an opponent which, if assigned to i, weakly maximizes i's gain to his most profitable bundle of lies and so is i's most tempting opponent to lie about. The person who stands to gain the most from his most profitable bundle in his most tempting pairing is the binding player: group honesty hinges on keeping this player telling the truth.*

Take a simple example, using the stylized network in Figure 4 and suppose that $r = 1$ so that news travels one degree after each round. Consider what happens when a player lies– he defects against someone who then wants to punish his bad behavior, and he lies to all those watching, which, if the conditions in Propositions 2 and 3 are met, will eventually reach someone who knows conflicting information and trigger the liar alarm. In this example Player 1 passes along information to 9 and 2, his two neighbors. The only way 1 would know that his current opponent deserves punishment without already having told 9 and 2 is if 1 *just* learned it, which

[20] The victim experiences a misdeed and not a lie- he will punish the defection with finite punishment, not set off the liar alarm.

would mean here that the offense targeted his other neighbor. 1 can tell 9 that he just learned 2 was victimized by the current opponent, and can tell 2 that he just learned 9 was victimized by the current opponent.

The most profitable opponent to lie about in this example is a neighbor and that neighbor's neighbor. Doing so maximizes the time to information conflict without increasing the expected punishment for his misdeed. For example, when assigned to play 2 in t, 1 can lie to 9 that 2 defected against 3 in round $t-1$, and tell the truth to 2, that 1 is defecting against 2. This lie will be discovered when the true information about 2 and 3 in $t-1$ reaches someone who receives the lie 1 started about 2 and 3 in $t-1$. Here, the conflict occurs 3 rounds after the lie. Any other lie takes a weakly shorter time to collide.

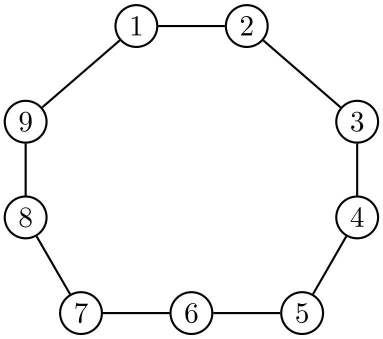

Figure 4: Any player's most profitable in-group lie is to lie about a neighbor to the other neighbor. 1 can tell 9 that 2 defected against 3 in $t-1$ when $r=1$ so that 1 can defect against 2.

Because true information creates a risk of colliding with false information and revealing a lie, situations with fewer true messages weakly reduce this risk. Messages spread deterministically between in-group players, but not between the in-group and the out-group. This leads to the following consequence of having inter-group relations:

Lemma 2 (Inter-group Relations Increase Opportunities to Lie). *The presence of an out-group with whom people sometimes interact but from whom no information reaches the in-group weakly increases the benefeis to lying.*

Corollary 2 obtains because when all interactions occur within a group, a lie about a past round must be about two people. These two have spread information about their history that will conflict with the false history the liar spreads. When some interactions occur between groups, a lie need only entail one in-group member and hence one person who spreads information which will conflict with the lie.

In Figure 4, 1 can tell 9 that 2 defected against an out-group member (instead of against 3 as in the earlier example). Now 2 spread the truth, 9 accidentally spreads the lie, and the information will conflict in 4 rounds (as opposed to 3).

Similar logic suggests that players whose neighbors are connected to each other (clustered) have a more difficult time lying than players whose neighbors are not connected to each other.

Lemma 3 (Clustered Neighbors Thwart Lying). *A player whose neighborhood is not a clique*[21] *faces a weakly more profitable lie than a player whose neighborhood is a clique.*

The best lie tells neighbors a player has private information that another player deserves punishment. If all neighbors are connected, though, they receive all information at the same time as the rest of their neighbors. Player i has no sources that the other neighbors do not also independently have. i can't claim to have privileged information that his opponent should be punished without his lie being detected immediately. In fact, this reasoning suggests that clustering matters more than degree.[22]

Corollary 2 (Clustering Matters More than Degree). *The presence of a small number of clustered neighbors does more to decrease the gains from lying than the presence of a large number of disjoint neighbors.*

Take Figure 5 in which player 1's neighbors are now also neighbors to each other:

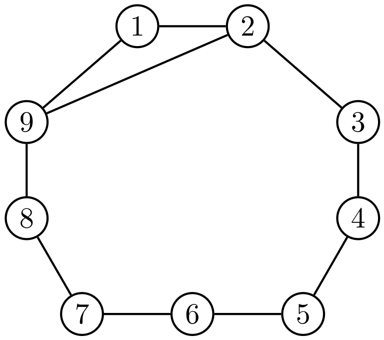

Figure 5: Player 1's neighborhood is a clique, so 1 will be immediately caught lying if he tries to tell any of his neighbors he has information that they have not received.

If in t player 1 tries to tell player 9 that 2 defected against the outgroup in $t-1$, this information will immediately clash with the true information 2 provided at the end of $t-1$.

[21] All players in a clique are connected to all other players. Here, a player's neighbors are all connected to each other.

[22] It can even be shown that adding a disjoint neighbor to a player with otherwise completely clustered neighbors *increases* his incentive to lie.

Any other player whose full neighborhood is not a clique can provide a lie that is not immediately detectable. Even a player with some neighbors forming a clique, like player 2 with neighbors 1 and 9 connected, can avoid detection for a while by telling the connected neighbors one lie and the other neighbors (3) something else.[23]

The usefulness of clustered neighbors is similar to the result in Ben-Porath and Kahneman (1996) which finds that the minimum number of monitors observing each player required to obtain truth-telling in equilibrium is 2. Here, having 2 monitors may be insufficient- players are all observed by 2 others in Figure 4, and yet there are gains to lying that could outweigh the costs (shown in the next section). The problem is not with keeping the monitors in line, but with verifying the information provided to the monitors. Clustering makes this possible.

5.4 Truth and Cooperation in Equilibrium

Under certain conditions, the prospective liar who would profit the most from his most profitable lie can be enticed to tell the truth (and with him, all other prospective liars). If all players also prefer to cooperate assuming truthful messages, and lies are detectable, then the strategy profile σ^{NWFTFM} results in true messages and fully cooperative behavior in equilibrium.

Proposition 6 (Fully Honest and Cooperative Equilibrium). σ^{NWTFTM} *is sequentially rational for game G given networks* $g_A = g_B$ *if all of the following hold:*

For any $i, j \in N, \exists k \neq i, j$ *such that there exists a path from* i *to* k *that does not include* j*,*
and there exists a path from j *to* k *that does not include* i (1)

$$M \geq T^p + \left\lceil \frac{\max_{i,j}\{STPP_{i,j}\}}{r} \right\rceil \tag{2}$$

$$\delta^{T^p} \geq \max\left\{ \frac{\alpha - 1}{(1-p) z^{out}_{min,rT^p}(\beta + 1)}, \frac{\beta}{(1-p) z^{in}_{min,rT^p}(\beta + 1)} \right\} \tag{3}$$

$$p < \min\left\{ \frac{z^{in}_{min,rT^p}(1+\beta) - \beta}{z^{in}_{min,rT^p}(1+\beta)}, \frac{z^{out}_{min,rT^p}(1+\beta) - \alpha + 1}{z^{out}_{min,rT^p}(1+\beta)} \right\} \tag{4}$$

[23] Such a lie would do best if it told neighbors 1 and 9 that the opponent defected against the other neighbor, 3, and told neighbor 3 that the opponent defected against one of 1 or 9.

$$\alpha - 1 \leq y_{cont,i}|_{t+T_i^E+1}^{\infty} - y_{pre,i}|_t^{T_i^X} - y_{post,i}|_{t+T_i^X+1}^{t+T_i^E} \ \forall i \in N \tag{5}$$

(1) and (2) are from Propositions 2 and 3, and (3) and (4) are from Proposition 4. T_i^X is the number of rounds before i's most profitable lie reaches someone who knows conflicting information and T_i^E is the number of rounds before i's most profitable lie results in full jadedness (the collapse of cooperation). Here, $y_{cont,i}|_{t+T_i^X-1}^{\infty}$ is the value i receives from continuing the game indefinitely starting from the round after a conflicting message would be discovered. $y_{pre,i}|_t^{T_i^X}$ is the value to i of taking full advantage of the peiod between the lie and its detection, and $y_{post,i}|_{t+T_i^X+1}^{t+T_i^E}$ is the value to i of taking full advantage of the period after the lie's detection but before complete breakdown of cooperation. Once again, the assumption of same networks in the two groups is unnecessary; the above conditions simply must hold within both groups, whatever their networks look like. A more precise characterization of these terms can be found in the Appendix.[24]

In words, Proposition 6 says that full cooperation and honesty are sequentially rational if (1) all pairs can reach an independent third party, (2) messages remain salient long enough, (3) & (4) conditions are right to induce cooperation if messages were truthful, and (5) the gains to continuing the game outweigh the gains from lying. The proof of Proposition 6 can be found in the Appendix, as well as a discussion of consistent beliefs which extend the behavior to a sequential equilibrium.

Condition (5) keeps potential liars telling the truth. Clearly whether (5) is satisfied or not depends on properties of the network and players communicating on it.

Corollary 3 (Keeping Liars Honest). *Condition (5) is easier to satisfy as detection of lies happens more quickly (closer third-party observers to shrink T^X), collapse occurs more quickly (the network has smaller diameter to shrink T^E)*[25]*, and players value future gains more highly (higher δ).*

Consider a simplified numerical example to confirm the existence of such an equilibrium for the group as pictured in Figure 6. It can be shown that, if members of the group occasionally interact with an out-group not pictured, the above conditions are met when $T^p = 2$, $r = 1$, $M = 6$, $\delta = .95$, $\alpha = 1.01$ and $\beta = 1.015$. Once again the most profitable lie is one in which

[24] Proposition 6 assumes that there are both intra- and inter-group interactions. A group playing σ^{NWTFTM} when there are no inter-group interactions would be fully cooperative in equilibrium under the same conditions, setting p to 0 in condition (3) and ignoring condition (4).

[25] Collapse occurs in $\lceil \frac{Diam}{r} \rceil$, so that if a lie occurs in t, cooperation will be expunged by $t + \lceil \frac{Diam}{r} \rceil$.

a player lies to one neighbor that the other neighbor defected against the out-group. That makes the time to information conflict, $T^X = 4$ and the time to the end of cooperation $T^E = 7$. Since $T^X > T^p$, a player could expect the group to be cooperative if he did not lie, earning $\sum_5^\infty \delta^{T^p}$. The amount a liar could expect to gain in the time before misinformation is detected is bounded above by earning α each round, and likewise for the amount gained in the time until cooperation ends.[26] The amount foregone by triggering the liar alarm outweighs even these optimistic gains. From Larson (2012*b*) we can calculate that the probability of punishment at the end of punishment phase when defecting against an outsider, z^{out}_{min,T^p} here is equal to $\frac{1}{2}$ and the same when defecting against an in-group member, z^{in}_{min,T^p} is $\frac{5}{8}$. If players encounter the out-group infrequently enough, say with $p = .1$, players prefer to cooperate with everyone when messages are truthful. Since memory is long enough and any pair can be reachable by a third independent player, $\sigma NWTFTM$ results in a fully cooperative and honest equilibrium.

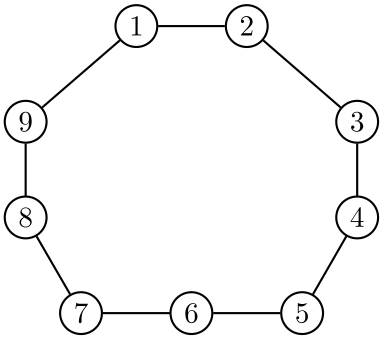

Figure 6: A fully cooperative and honest equilibrium is possible even for the above network.

Consider what could happen if condition (1) were violated. Figure 7 shows one such network, reproduced from above.

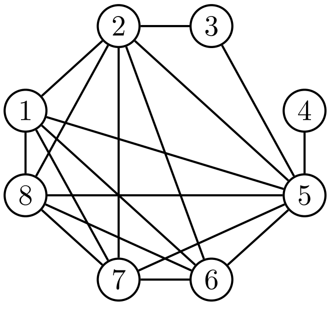

Figure 7: Example network where information from player 4 only reaches others via player 5.

[26] Of course, the player would be overly optimistic if he expected this full payoff- there is a chance he would encounter the player he wronged and fail to gain α from him, for example.

Player 5 controls access to information about player 4. Without a third party reachable independent of player 5, 5 can claim that 4 defected against the out-group, say, and no news to the contrary ever reaches anyone else. 4 could punish 5 in return, but 4 could always spin this to his neighbors as a defection against him and avoid further punishment. Controlling access points to information opens opportunities to lie.[27]

6 Discussion

The above shows that groups for which verifying information through observation is difficult can still maintain cooperation and honesty with decentralized institutions. Players who stand to gain from lying are discouraged from doing so when the costs of lying will be realized and quickly enough.

One interesting result is really the absence of a result: a player's degree is not what matters. That is, it is not the number of connections a player has that makes the difference per se but how those connections are arranged. Increasing the number of neighbors for a player has a very small effect on the incentive to lie, and the net effect can even be to increase the value of lying.[28] Typical monitoring stories focus on the strength of oversight, but here it is not the volume of oversight that matters. Instead, here what matters is the extent to which the overseers receive information from the same sources that the person being monitored does and at the same time.

When a player's neighbors are all themselves connected, lying is detected immediately. This is because lies take the form "in the past, person j defected against person k." If neighbors all receive the same accounts of the past that i receives and at the same time, they will know immediately if i tries to change his telling of the past for his own gain. A small number of clustered neighbors goes farther toward preventing lying than a large number of un-clustered neighbors.

The above results suggest that there are some social structures which are more conducive to honesty and cooperation than others in environments of uncertainty. Given evolutionary pressure to cooperation, we might expect an evolutionary preference for certain structures or

[27] The same is true for "brokers" or "bridges" in a network, the players who connect two otherwise disjoint components. These players control which information reaches which component of the network and have a greater opportunity to deceive.

[28] Increasing the degree of a player without increasing the total number of links must mean that paths to a third party weakly decrease in length which weakly increases the cost to lying. Increasing the degree also increases the probability that a player will be assigned to play one of his most tempting victims to lie about.

features of groups over others.

For example, as per condition (1) in Proposition 6, groups without ubiquitous third-party reachability (i.e. groups in which one person can control access to others' information) pose problems for honesty and therefore cooperation in equilibrium. As seen in Figure 7 person tenuously connected to a group relies on the good nature of their sole contact. Likewise, someone serving as a bridge between two sections of a network that would be disjointed without the bridge's ties control information and make lying to one section about people in the other section a tempting possibility. Given the difficulty of enforcing honesty in such settings, we might expect few examples of true bridges.[29]

Similarly, we should expect greater pressure to build clusters of friend groups rather than amass a large number of possibly unacquainted friends. Clustered friends can verify information sources and encourage truth-telling in ways that a laundry list of isolated friends cannot. It may be no surprise that human networks (as opposed to networks of objects or ideas) tend to feature high clustering (Watts and Strogatz, 1998).

If groups rely on in-group policing mechanisms, we might also expect pressure to interact almost exclusively with in-group members. The presence of an out-group provides opportunities to lie by siphoning off potential sources of information. While both parties to an interaction share information about it when meeting in-group members, an out-group member's account of the interaction never permeates the in-group. This does not create opportunities to defect against the out-group – since everyone is always instructed to cooperate in those interactions and actions are observable, no lie could excuse an observed D against the out-group– but does create opportunities to claim that *others* defected against the out-group.

Whether groups truly make use of the institution described here is an open question. A first step toward empirically verifying the institution could be taken in the laboratory. A key component to the institution is a difference in reaction to misdeeds and to lies. Players are willing to eventually excuse misdeeds but not lies. Anecdotal evidence suggests that people do tend to react more strongly to lies, but more rigorously verifying this reaction is possible and poses a promising direction for future research.

[29] Granovetter (1973) makes a similar claim, observing that when bridges do exist, they should not be comprised of a strong tie. The reasoning here is different, and suggests another reason why groups may experience pressure to build in redundancy and to close triads.

7 Conclusion

When communities are tasked with enforcing behavior, information is essential. People must know who misbehaves and who doesn't. When doling out punishment looks like misbehavior, and community members are supposed to punish misdeeds committed against anyone, the informational demands are extremely strong. If only a few links are missing from the network, rightful punishment cannot always be distinguished from wrongful actions.

This suggests that lying is an important hurdle to cooperation in environments lacking perfect information. This also suggests that assuming complete networks on the grounds that real networks are "close enough" may in fact be a big leap that overlooks difficulties arising in networks very close to complete ones.

I have established when lies can be detected, characterized the most profitable lies that must be disincentivized if honesty is to obtain, and established an institution that can sustain both honesty and cooperation. The institution embeds trusting behavior in a game which treats lies gravely. Players can eventually forgive misbehavior but not lies. This institution is more robust than institutions which respond gravely to both actions and lies.

Here, players may lie because they stand to gain personally. Another lying scenario is possible. Someone may lie because they are interested in changing the group behavior. For example, someone may lie because they have a vendetta against someone or want to incite conflict with another group. While this case isn't taken up here, an interesting area of future research entails the role of networks at thwarting or magnifying incendiary rumors.

Appendix 1: Proofs

Proof of Proposition 1. If i observes j playing D against player k in t, i must determine whether k deserved punishment, which depends on k's actions in the last T^p rounds. Any single round can establish k's guilt, so if i observes all but 1 of k's last T^p rounds, $\exists$ history such that k is guilty because of the unobserved round and appears innocent otherwise. Therefore the maximum number of people who must be observed to know k's relevant actions is T^p. k's guilt against any of the T^p opponents depends on *their* guilt and hence their past T^p opponents, any one of whom could determine their guilt. Their status depends on *their* past T^p opponents, and so on back. In other words, the maximum number of people who must be known is $(T^p)^t$. Since the game is infinitely repeated, $t \to \infty$, there is a point in the game beyond which the maximum number of people who must be known exceeds $n-1$ and $\exists$ history of sufficiently-mixed round matches which maximizes the required number of people known. Therefore, $n-1$ people must be known to safeguard against a history in which guilt cannot be determined. When matches are all in-group, $n-2$ links suffices since the one person not linked to will always be matched to someone connected to i. When some matches are with the out-group, all $n-1$ links are required since any person not connected could be matched with the out-group and escape observation. □

Proof of Corollary 1. The proof follows immediately from Proposition 1 and properties of simple, undirected networks. For all i to be discerning about all possible opponents in exclusively in-group interactions, all must be connected to $n-2$ others. The largest complement network in which all are connected to at most 1 other has $\frac{n}{2}$ links, so the smallest network which excludes the complement network contains $\frac{n(n-1)}{2} - \frac{n}{2}$ links. When some interactions are with an out-group, all $\frac{n(n-2)}{2}$ links must be present. □

Proof of Proposition 2. Players can detect a lie by knowing the truth through observation and hearing a lie, or by not knowing the truth but perceiving conflicting information. By Proposition 1, all can verify truth if they have degree $\geq n-2$, which implies all pairs have paths to an independent third party, establishing necessity for the first detection method. If truth cannot be verified through observation, a lie can be detected if some player receives both the lie and the conflicting true information. Suppose $\exists$ pair i, j such that all paths to all other players from j go through i. Then i can prevent lies about j from reaching conflicting information. If, however, $\exists$ a player l who can be reached by j without crossing i and from i without j, neither player can prevent conflict. □

Proof of Proposition 3. The nearest 3rd party is the first location of information conflict, i.e. the first opportunity to discover the presence of a lie. If players forget information before it even reaches this 3rd party, conflict will never be noticed and the lie will be undetected. □

Proof of Lemma 1. Taking advantage of a lie in the future rather than immediately decreases the window of exploitation and makes the gain probabilistic (since a player may or may not be randomly assigned to play the person lied about), so a liar cannot do better than to lie about a current opponent.

Given the same punishment for deviating, no deviation is as profitable as playing D when instructed to play C, so a liar cannot do better than to claim his opponent deserved punishment.

Lying about a round involving both players more than rk degrees away in $t - k$ means lying about players whose information should not have reached i yet, so the lie is immediately detectable to i's neighbors. Lying about a round involving a player fewer than rk degrees away in $t - k$ means lying about a round i already told his neighbors about, so again the lie is immediately detectable. Hence, the best lie is about a round that took place in $t - k$ which involved one player rk degrees away from i.

If neighbor j is on a path $\leq rk$ degrees from the alleged victim, j already knows the truth when i tells him the lie, so the conflict is detected immediately.

□

Proof of Proposition 5. The proof is immediate from Lemma 1. Each player does best by picking a most profitable lie to tell each neighbor. □

Proof of Lemma 2. Take i's most profitable lie when interactions are in-group only, which entails a person rk degrees away from i and another ingroup member in an alleged round in $t - k$. The lie can be modified to entail the person rk degrees away and a member of the out-group when an out-group exists. Such a lie is no more detectable but entails only one source of the true information (as opposed to two sources when both subjects of the lie are in-group members). Decreasing the number of sources of information that can conflict with the lie never increases the costs of a lie, all else equal. □

Proof of Lemma 3.] Suppose i's neighbors, $N(i)$, are all connected to each other. This implies that any path from a player rk away from i can reach all $n(i)$ at least as quickly as i. Therefore, any lie which claims i knows something $N(i)$ do not is discoverable immediately. If $\exists$ a neighbor $l \in N(i)$ who is not connected to the other neighbors $N(i) \setminus l$ and $N(i) \setminus l$ form a clique, then $\exists$ lie i could tell which is not discoverable immediately. By construction, information from the clique and the other neighbor will not collide until, at the earliest, one additional round. □

Proof of Proposition 6. (1) Given that players spread messages to neighbors and detect lies based on conflicting information, the sufficiency of (1) is clear.

(2) From Proposition 5, we know the most profitable lie entails a lie which delays conflict. Players can take advantage of short memory and construct a lie about someone rT^p away, who supposedly defected in $t - T^p$. Such a lie ensures that people forget it after i says it while still allowing him to punish within the punishment window. Players remember long enough to detect any profitable lie when (2) holds.

(3) & (4) When players are in trusting phase, these conditions are sufficient to discourage any deviation in behavior. The proof can be found in Larson (2012*b*).

(5) Players prefer to keep cooperating and avoid triggering jadedness when the gains to the most profitable lies are smaller than the costs to remaining trustworthy. If i complies and

refrains from lying, he expects to earn

$$1+\sum_{l=1}^{\infty}\delta^l\left[(1-p)\left[prob_l(c^*)+prob_l(d^*)\alpha\right]+p\right]$$

where $prob_l(c^*)$ and $prob_l(d^*)$ are the believed probabilities of being randomly matched with a cooperator and a defector, respectively, in $t+l$.

If instead i lies and takes full advantage of the opportunities presented by lying and the impending end of the game, he expects to earn

$$\alpha+(y_{exploit}-y_{pun})|_{t+1}^{t+T^X}+(y_{exploit}-y_{pun})|_{t+T^X+1}^{T^E}-\sum_{l=T^E+1}^{\infty}\delta^l\left[(1-p)\left[prob_l(c^*)+prob_l(d^*)\alpha\right]+p\right]$$

where, once i breaches trust, he faces possible gains from exploiting the impending end of the game and costs from punishment he expects to incur from his victim who thinks he committed a misdeed and spreads the word that i deserves finite punishment. The exact specification of these values is not essential for present purposes- for a numerical example confirming existence, see section 5.

$(y_{exploit}-y_{pun})|_{t+1}^{t+T^X}$ are the net gains from exploitation before anyone detects the lie and $(y_{exploit}-y_{pun})|_{t+T^X+1}^{T^E}$ are the same after the lie is detected but before jadedness fully overtakes the group. These terms can be rewritten $y_{pre}|_{t+1}^{t+T^X}$ and $y_{post}|_{t+T^X+1}^{T^E}$ for convenience. Likewise, the value to continuing the trustworthy cooperation game, can be rewritten $y_{cont}|_{t+T^E+1}^{\infty}$. A player prefers to play honestly and keep players in trusting phase when (5) holds.

Hence, conditions (1) through (5) establish sufficiency for sequential rationality.

The behavior in equilibrium is of greater interest than the beliefs in equilibrium. The above shows that σ^{NWIGPM} forms half of a sequential equilibrium since the strategy is sequentially rational in any information set with any beliefs over nodes in the information set. Consider a set of beliefs, μ, which places probability one on all events expected in equilibrium, probability zero on out-of-equilibrium events, and which are updated according to Bayes' rule. Now consider a perturbation of these beliefs in which players are assumed to tremble with small independent probability $\epsilon > 0$. Given the perturbation, all information sets are reached with positive probability and Bayes' rule pins down beliefs everywhere. Let μ^* be the limiting beliefs derived from Bayes' rule as $\epsilon \to 0$. Since sequential rationality holds for any specification of beliefs given the above conditions, it holds for μ^*. Hence, the assessment (σ^{NWIGPM}, μ^*) is a sequential equilibrium.

□

Proof of Corollary 3. $y_{pre}|_{t+1}^{t+T^X} \geq 0$ and $y_{post}|_{t+T^X+1}^{T^E} \geq 0$ since player i could always block defections against himself while the game is ending. Therefore, (5) is weakly easier to satisfy as T^X decreases because smaller T^X means fewer opportunities to gain from exploitation. The same applies to T^E. As the future is valued more highly, the continuation value increases faster than the two short term gains from the impending end of cooperation, making (5) easier to satisfy as δ increases. □

References

Anderlini, L. and R. Lagunoff. 2006. "Communication in Dynastic Repeated Games:Whitewashes and Coverups." *Rationality and Equilibrium* pp. 21–55.

Annen, K. 2011. "Lies and Slander: Truth-Telling in Repeated Matching Games with Private Monitoring." *Social Choice and Welfare* 37(2):269–285.

Aoyagi, M. 2000. "Communication Equilibria in Repeated Games with Imperfect Private Monitoring.".

Ben-Porath, E. and M. Kahneman. 1996. "Communication in Repeated Games with Private Monitoring." *Journal of Economic Theory* 70(2):281–297.

Ben-Porath, E. and M. Kahneman. 2003. "Communication in Repeated Games with Costly Monitoring." *Games and Economic Behavior* 44(2):227–250.

Calvert, R.L. 1995. Rational Actors, Equilibrium, and Social Institutions. In *Explaining Social Institutions*, ed. J. Knight and I. Sened. University of Michigan pp. 57–94.

Compte, O. 1998. "Communication in Repeated Games with Imperfect Private Monitoring." *Econometrica* pp. 597–626.

DePaulo, B.M., D.A. Kashy, S.E. Kirkendol, M.M. Wyer and J.A. Epstein. 1996. "Lying in Everyday Life." *Journal of Personality and Social Psychology* 70(5):979.

Dunbar, R. 1998. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.

Ellickson, R.C. 1991. *Order Without Law: How Neighbors Settle Disputes*. Harvard Univ Pr.

Ellison, G. 1994. "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *The Review of Economic Studies* 61(3):567–588.

Enquist, M. and O. Leimar. 1993. "The Evolution of Cooperation in Mobile Organisms." *Animal Behaviour* 45(4):747–757.

Evans-Pritchard, E.E. 1940. *The Nuer: A Description of the Modes of Livelihood and Political institutions of a Nilotic People*. Clarendon Press Oxford.

Fearon, J.D. and D.D. Laitin. 1996. "Explaining Interethnic Cooperation." *The American Political Science Review* 90(4):715–735.

Fischer, C.S. 1982. *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago Press.

Granovetter, M.S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* pp. 1360–1380.

Greif, A. 1993. "Contract Enforceability and Economic Institutions in Early Erade: The Maghribi Traders' Coalition." *The American Economic Review* 83(3):525–548.

Harbord, D. 2006. "Enforcing Cooperation among Medieval Merchants: The Maghribi Traders Revisited." *working paper* .

Horowitz, D.L. 1985. *Ethnic Groups in Conflict*. Univ of California Pr.

Jackson, M.O. 2003. "The Stability and Efficiency of Economic and Social Networks." *Advances in Economic Design* 6:1–62.

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

Kandori, M. 2002. "Introduction to Repeated Games with Private Monitoring." *Journal of Economic Theory* 102(1):1–15.

Kandori, M. and H. Matsushima. 1998. "Private Observation, Communication and Collusion." *Econometrica* pp. 627–652.

Karlan, D., M. Mobius, T. Rosenblat and A. Szeidl. 2009. "Trust and Social Collateral*." *Quarterly Journal of Economics* 124(3):1307–1361.

Larson, Jennifer M. 2012*a*. "Cheating Because They Can." *Working Paper* .

Larson, Jennifer M. 2012*b*. "A Failure to Communicate: The Role of Networks in Inter- and Intra-Group Cooperation." *Working Paper* .

Lippert, S. and G. Spagnolo. 2011. "Networks of Relations and Word-of-Mouth Communication." *Games and Economic Behavior* 72(1):202–217.

McPherson, M., L. Smith-Lovin and J.M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* pp. 415–444.

Mislove, A.E. 2009. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems.* ProQuest.

Sommerfeld, R.D., H.J. Krambeck, D. Semmann and M. Milinski. 2007. "Gossip as an Alternative for Direct Observation in Games of Indirect Reciprocity." *Proceedings of the National Academy of Sciences* 104(44):17435.

Takagi, Y. 2011. "Local Gossip and Inter-Generational Family Transfers: Comparative Political Economy of Insurance Provision." *Working Paper* .

Watts, D.J. and S.H. Strogatz. 1998. "Collective Dynamics of Small-World Networks." *nature* 393(6684):440–442.

Wilson, C., B. Boe, A. Sala, K.P.N. Puttaswamy and B.Y. Zhao. 2009. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems.* Acm pp. 205–218.