Submission Number: PET11-11-00180

Cooperation in anonymous dynamic social networks

Brian W Rogers
MEDS, Kellogg, Northwestern

Abstract

We study the extent to which cooperative behavior can be sustained in anonymous, evolving social networks, such as online communities. Individuals strategically form relationships under a social matching protocol and engage in prisoner's dilemma interactions with their partners. An agent that defects escapes direct reciprocity by virtue of anonymity: when starting a new relationship, neither agent has available any information about the history of his partner. We demonstrate that cooperation is sustainable at equilibrium in such a model, and characterize a class of equilibria that support cooperation as a stationary outcome. The endogenous dynamics of the social network imply that cooperation allows an individual to interact with a growing number of other cooperators over time, potentially balancing the immediate gains from defection.

Submitted: March 02, 2011.

Cooperation in Anonymous Dynamic Social Networks*

Preliminary Working Paper

Nicole Immorlica† Brendan Lucier‡ Brian W. Rogers§

January 17, 2011

Abstract

We study the extent to which cooperative behavior can be sustained in anonymous, evolving social networks, such as online communities. Individuals strategically form relationships under a social matching protocol and engage in prisoner's dilemma interactions with their partners. An agent that defects escapes direct reciprocity by virtue of anonymity: when starting a new relationship, neither agent has available any information about the history of his partner. We demonstrate that cooperation is sustainable at equilibrium in such a model, and characterize a class of equilibria that support cooperation as a stationary outcome. The endogenous dynamics of the social network imply that cooperation allows an individual to interact with a growing number of other cooperators over time, potentially balancing the immediate gains from defection.

*We thank Wiola Dziuda, Christoph Kuzmics, Joel Sobel and Joel Watson for very helpful comments and conversations.

†Northwestern University, Evanston, IL USA. nicimm@gmail.com

‡University of Toronto, Toronto, ON Canada, blucier@cs.toronto.edu

§Northwestern University, Evanston, IL USA, b-rogers@kellogg.northwestern.edu

1 Introduction

There are many important social, political, and economic systems in which people face a choice between seeking immediate personal gain at the expense of others or cooperating to a lesser mutual benefit. In such contexts, it might be expected that an agent acting in his own interest ought to choose an uncooperative strategy. This intuition is amplified when people can act under pseudonyms, such as over the Internet, since an agent that develops a reputation for being uncooperative can, if he chooses, simply re-enter the system with a new identity. However, the continuing success of online interaction networks such as eBay (www.ebay.com) indicates that a group of anonymous agents need not devolve into a steady-state of predominantly uncooperative behavior. There is a large body of experimental evidence demonstrating that cooperation is a robust empirical phenomenon (see, for example, [2, 14, 18]). In fact, many systems with a high degree of anonymity and in which agents change partners over time are characterized by a high, but less than complete, level of cooperation. We find that such behavior can be explained as constituting a simple stationary equilibrium under a simple network formation model.

The main elements of our model are as follows. Agents enter the system over time and have finite lives. All strategic interactions are bilateral and described by a Prisoner's Dilemma (PD). A random matching process presents agents with opportunities to form new relationships. In every period, each agent chooses a behavior, cooperation or defection, and receives the sum of payoffs from the corresponding PDs with each of its partners. An agent plays the same action with each of his partners in a given round. This is the case when, e.g., the behavior under consideration is not relationship-specific, but a characteristic of an agent's general behavior. This behavior can also be motivated in the presence of local information, whereby a defection on a single partner would likely be recognized and reacted to by all of an agent's partners. In this case optimal behavior would amount to defecting on all partners or on no partners. At any time, each agent has the opportunity to sever any of its relationships.

The discretion to sever a relationship plays a key role by providing a mechanism with which to threaten punishment for uncooperative behavior. In fact, because of anonymity, this is the only effective mechanism for punishment: if at any point an agent becomes disconnected in the graph of relationships, other agents are not able to track its identity and thus in all future relationships this agent cannot be associated with its previous actions. In this way, agents are not able to credibly maintain a reputation for a particular behavior or to be punished due to a negative reputation.

For example, consider an online community. Agents seek partners with whom to profitably interact, such as for trading goods or engaging in joint activities. At any point in time agents can choose to conduct honest business (cooperate) or to cheat their partners for a gain (defect). If one of an agent's partners defects, the worst punishment that can be enacted is to sever the relationship. One might wish to, in addition, broadcast the agent's defection so as to enable further community punishment. But the agent who defected does not have an identity that can be tracked by his future partners, and so bears no negative consequence of his defection beyond the loss of those relationships.

The equilibria on which we focus involve simple and intuitive behavior. In particular, we exhibit equilibria in which agents are *consistent* in that they choose either to perpetually defect or to perpetually cooperate. In light of this behavior, optimal decisions regarding the formation and severance of relationships are also easy to describe. In particular, a relationship is severed when, and only when, a defection is observed. Such a social norm is very natural: defection is not tolerated, and cooperation is rewarded with the opportunity for future interactions. Under consistent behavior, this social norm is the only rational response governing the dynamics of relationships. Conversely, we are able to show further that, under appropriate conditions, consistent behavior is the only rational response to the social norms regarding relationships.

1.1 Summary of results

Our main goal is to analyze the dynamics of behavior and of the network that evolve according to this social norm. We show that there always exists an equilibrium with consistent behavior. Of course, as with any model of repeated PD interactions, universal defection remains one possible outcome. More interestingly, under certain natural assumptions on the parameters of the PD payoffs and the discount factor, these equilibria also support a second outcome with cooperative behavior. Depending on the parameters, the level of cooperation may or may not be universal. When it is not, the model predicts the coexistence of cooperative and uncooperative agents. In either case, the equilibrium outcome is stable, in that if the proportion of cooperative behavior is perturbed, optimal decisions of entering agents will eventually return the system to its original state. We provide an explicit characterization of stationary equilibria and comparative statics on the model's parameters.

The key mechanism that supports the possibility of cooperative behavior at equilibrium is that defection is punished by the loss of relationships. This natural punishment works as follows. In our model, each agent can immediately gain access to a certain (exogenously fixed) number of relationships by sponsoring them. Additional relationships can be formed only by waiting for relationships to be proposed and sponsored by others, which takes time. The incentive to cooperate, then, comes from the ability to attract a large network of profitable relationships over time. We think of these accumulated relationships as social capital.[1] An agent who defects builds no social capital because of these punishments. He is always able to sponsor some relationships and profit from them when he happens to find cooperators, but after every period he loses all his relationships and returns to being isolated. When the model supports the coexistence of cooperation and defection, it is because the gains from accumulated relationships that a cooperator expects are equal to the gains achievable by perpetually defecting on a smaller (and changing) set of partners.

The behavior that we have described thus far, while fairly general, requires some parametric assumptions to justify. After characterizing a class of equilibria that support a stationary

[1]There is a debate in the literature about how to define and measure "social capital". See [21] for a review. We want only to offer an intuitive term in the context of our model for the gradual accumulation of profitable relationships.

outcome with cooperative behavior, we turn to the issue of describing behavior more generally. We have two results that go some distance towards assessing optimal behavior when the requisite assumptions for the simple stationary equilibria do not hold.

First, the assertion that the behavior of consistently choosing to either cooperate or defect is the best response to the social norm governing relationship dynamics requires that the cost to a cooperator of being defected on is high enough relative to the gain to defecting on a cooperator. When this condition is not met, inconsistent behavior can be expected. We characterize the form that such behavior takes. In particular, an agent who has cooperated in the past may find it optimal to defect and re-enter the system under a new pseudonym. This strategy of building up a network of partners and then cheating them, however, appears only in a limited form. A cooperating agent has a profitable opportunity to defect only if the number of cooperators to whom he sponsors a link is high enough relative to the number of relationships sponsored by other cooperators. For many parameters, the bound on this ratio is tight enough that a cooperator defects only when he has, in fact, no relationships sponsored by others. This fact is consistent with our interpretation of relationships sponsored by others as social capital. Moreover, we find that it is never profitable for an agent to set out to build relationships with the intention of eventually exploiting her partners via defection; rather, the choice to defect is rationalized only for an agent who, through unexpected random circumstance, finds himself socially impoverished in that the majority of his relationships are sponsored by himself and not others.

Second, cooperating agents face uncertainty at the beginning of a relationship, since they have no information about the history of their new partner. If there are sufficiently few cooperators in the population, then the agent may be unwilling to accept new relationships. Even in a society with a non-trivial fraction of cooperators, when a new relationship is sponsored by another agent, the probability that it comes from a cooperator is lower than when the agent sponsors the relationship himself. This is because defectors, in every period, form a complete set of new relationships, whereas cooperators gradually accumulate stable relationships over time, and hence search for new partners less often. Thus the set of relationships being sponsored in any period is biased to those coming from defectors. It may therefore be rational to reject proposed links and participate only in relationships initiated by oneself. We characterize the condition under which it is optimal to accept new relationships from others, which is an important aspect for the equilibrium we describe. When that condition is not met, an interesting equilibrium emerges.[2] In particular, with the coexistence of cooperators and defectors, cooperating agents accept new relationships only with a certain probability. This has the effect of insulating cooperators to some extent from the outside world. It is also the necessary ingredient to properly incentivize cooperation. Without this barrier, defection becomes too tempting, and the equilibrium would collapse to a state of all defection.

[2]The full description and analysis of this equilibrium does not appear in this preliminary working draft, but will appear in a forthcoming version of the paper.

1.2 Literature studying related phenomena

Our model contributes to the long line of work seeking to understand the robust phenomenon of cooperation in repeated social interactions. When the partners in the game are fixed over time and the game is repeated indefinitely, an application of the Folk Theorem can explain cooperation (or any mutually beneficial payoff) [1, 10], via, e.g., trigger strategies, and can be extended to accommodate imperfect monitoring [9].[3] However when agents change partners over time, such a threat is no longer effective because a pair of agents may very well never meet again. Instead, community enforcement procedures can be used to sustain cooperation [12, 15]. In these equilibria, if the model allows for public reputations, then the community can always defect against an agent with a reputation of defection. However, this mechanism for sustaining cooperation is unsatisfactory in the domain of anonymous interactions: threat of retaliation is not a deterrent when an agent can re-enter the network as a new user at any time.

When agents are anonymous and histories are not publicly observable, a community can nevertheless still enforce cooperation by agreeing to defect on all partners as soon as any defection is observed [12, 7]. In this way, from a cooperative state, a deviating defector starts a contagion and will thus eventually be punished for the initial defection. This kind of community enforcement does not provide a natural explanation for the stable coexistence of cooperative and defective behavior. When anonymity is due to the use of pseudonyms that can be changed, an alternative approach to community enforcement is to penalize players using new pseudonyms (who are either genuinely new or using a recently acquired identity). This builds a level of trust in agents by forcing them to "pay their dues" [8]. This penalty de-incentivizes agents from taking new pseudonyms, at the cost of reduced efficiency in interactions with new participants.

In all of the work cited above, partnerships are formed exogenously (though possibly stochastically). When agents have choice over their partners or in the length of the relationship, new insights arise [4, 11, 13, 20]. In developing long-term relationships, agents have the opportunity to gradually build trust with their partners. This trust becomes an asset, and the threat of losing it produces incentives to cooperate. This mechanism operates without information flow (i.e., there are no public reputations), and without relying on contagious defection strategies.

We employ such an approach, in which agents have discretion over maintaining their relationships. As such, agents' behavior influences the network dynamics. Our mechanism for sustaining cooperation can be identified with the notion of social capital, taken to mean an agent's accumulated network of partners.[4] Here, the reason to cooperate comes from the fact that, through cooperation, one can gradually build up a social network consisting of other cooperators. Since the matching process we employ entails delay, the threat of severing a relationship can incentivize cooperative behavior.[5]

[3] There are less closely related papers that support cooperation through evolutionary approaches [16] or in stochastic models with boundedly rational agents [3].

[4] Vega-Redondo [22] studies social capital in a stochasticly evolving network.

[5] There is a large literature on matching models. See [6, 5] for important contributions.

This is an expression of a very general idea that can be traced at least to [19]. In their seminal work, firms incentivize workers to exert effort via the threat of termination. Termination is an effective punishment because, in equilibrium, there exists non-trivial unemployment. Thus, it is precisely the aggregate market outcome of unemployment that produces incentives for good behavior in specific worker-firm relationships. Many papers have built on the theme that losing a relationship has a cost that is determined endogenously along with the behavior within relationships. The cost can come in many forms, such as being cast into a matching market with frictions, having to start a new relationship that requires specific investment [17], or having to start small in a new relationship [23, 24].

While our work shares a common element with this literature, the threat of punishment in our model is effective precisely because agents benefit from accumulating *multiple partners*, i.e., our notion of social capital. This difference with the papers cited above is important. It allows for a novel interaction of strategic behavior with the dynamics of the social network. This, in turn, allows us to provide a characterization of the coexistence of cooperation and defection in equilibrium, and in that sense our analysis is the first of its kind.

The remainder of the paper is organized as follows. The model is described in Section 2. Section 3 characterizes simple stationary equilibrium outcomes. Section 4 develops the conditions under which our characterization of equilibrium behavior holds in a more general strategic setting, while Section 5 describes the way in which equilibrium outcomes change when these conditions are not met. We conclude and provide comments for further research in Section 6. An Appendix contains a formal development of the model.

2 A model of strategic interactions in a social network

For ease of exposition, we describe the basic elements of our model in this section. A formal development is presented in the Appendix.

All strategic interactions are governed by a prisoner's dilemma with the following payoff matrix.

	C	D
C	1,1	-b,1+a
D	1+a,-b	0,0

We take $a, b > 0$ and $a - b < 1$ so that, while mutual cooperation is the uniquely efficient outcome, defection is strictly dominant.

There is a continuum of agents, which we associate with points from the unit interval $N = [0, 1]$.[6] Agents interact repeatedly on an evolving directed network. Time is discrete.

[6]The primary reason to work with a continuum of agents is to guarantee that agents are "atomless", so that no individual can unilaterally affect expected aggregate behavior. This property is well-motivated in large networks, as one then expects a player to ignore the marginal effects that his behavior imposes on the system. We use a continuum of agents (as opposed to a countably infinite set) so that one can more readily define distributions over the set of players. Our results (i.e. equilibria) also hold in an approximate sense for a finite set of N players, with approximation errors vanishing as N grows large.

At each date each agent independently dies with a given probability $1-\delta$, in which case it is replaced by a new agent.[7] We speak of the age of an agent as the number of dates since its birth. Each agent i chooses an action $\alpha_i \in \{C, D\}$ at each date of its life. Agents observe the aggregate proportion, q, of C behavior in the population after each date.

Every agent sponsors a number $K \geq 1$ of connections to other agents. Thus an agent is generally involved both in relationships that it sponsors (outlinks) and also in relationships sponsored by others (inlinks), resulting in a directed graph of interactions. When a connection is proposed, the partner is chosen uniformly at random from the population, and the connection is then accepted or rejected by the chosen partner. Once accepted, each connection persists to the subsequent date unless one of the partners dies or chooses to sever the connection. When a connection is broken, the agent who sponsored it, provided he survives, re-matches with another agent, chosen uniformly at random, at the next date.

At each date an agent receives a payoff equal to the sum of the outcomes of the stage game played with each of his (in and out) partners, according to the chosen actions of the two agents and the payoff matrix given above. Agents seek to maximize the present value of expected lifetime payoffs.

To summarize, each time period proceeds according to the following order of events:

1. New agents are born.
2. Actions are chosen.
3. Outlinks are proposed to other agents.
4. Potential inlinks are accepted or rejected.
5. The stage game is played and payoffs are realized.
6. Agents sever any links that they choose to.
7. Death occurs.

3 Simple Behavior and stationary outcomes

Our goal is to understand equilibrium behavior. An agent's *strategy* maps observed histories to (probability distributions over) actions. The actions involve, on each turn: the choice of which proposed links to accept, whether to cooperate or defect, and which links, if any, to break. We focus on Markov behavior, in the sense that agents do not condition their actions on a common labeling of time. In other words, while each agent is aware of the number of rounds that have passed since he first entered the system, he does not use any universal description of time (e.g. the number of rounds since the system first began, etc.). A formal development of strategies is contained in the Appendix.

We begin the analysis by considering a setting in which three assumptions are imposed on strategies.

ASSUMPTION **1** *Consistent (C): An action from $\{C, D\}$ is chosen at birth (possibly mixing). At all future dates, the agent plays the action it chose at the previous date.*

[7] The conclusion that at each date, a proportion δ of the population survives, almost surely, relies on an exact law of large numbers for a continuum of random variables. See, e.g., Judd (1985).

ASSUMPTION **2** *Unforgiving (U): A connection is severed immediately whenever the partner chooses D.*

ASSUMPTION **3** *Trusting (T): All proposed inlinks are accepted and a link is not severed in a period when the partner chooses C.*

Consistency prohibits strategies that, for instance, allow an agent to cooperate until some stopping condition is met, and then defect. It also allows us to speak of "cooperators" and "defectors". The unforgiving and trusting assumptions completely determine the behavior of agents with respect to managing their set of relationships.

Throughout, we will restrict attention to symmetric equilibria. Under Assumptions 1-3, this means that the behavior of all agents can be completely described by a function $\phi : [0,1] \to [0,1]$ with the interpretation that $\phi(q)$ specifies the probability that an agent chooses C at its birth when it observes the state q. We will refer to strategies that satisfy Assumptions 1-3 as *simple strategies*.

These strategies are restrictive at the individual level, but they are flexible enough to permit interesting aggregate behavior. In particular, we will demonstrate that it is possible to sustain cooperation, sometimes (depending on parameters) in coexistence with defectors. We then show that under appropriate conditions these outcomes, with behavior satisfying Assumptions 1-3, can be supported by equilibria without a priori restrictions on the strategy space. Finally, we are able to say a fair amount about behavior and the evolution of the system in cases where these conditions do not hold.

3.1 Simple stationary equilibria

We are interested in determining when a particular level of cooperation q can be sustained as a stationary outcome of the system under Assumptions 1-3. Note, however, that a given value of q does not capture the entire state of the network, even in expectation. Other factors, such as the degree distributions and age distribution of the agents, impact expected payoffs because they influence the rates at which links with cooperators and defectors can be expected to arise. This motivates us to define a particular set of configurations of the system, the *steady-state at* q, L_q, to be the collection of states that result with positive probability when the fraction of cooperators is q, and agents have been applying a strategy for which $\phi(q) = q$ for an infinite number of rounds. A steady-state L_q captures all payoff-relevant information.

For a steady-state to be supported as an equilibrium outcome, it is necessary that the strategy $\phi(q) = q$ be optimal when the system is in state L_q. In this case, the fact that all agents apply ϕ implies that the system remains in state L_q. We will therefore say that $q \in [0,1]$ is a *simple stationary equilibrium* if, given that the system is in state L_q at all times, the application of a strategy that chooses cooperation with probability q at birth is optimal.

Let us briefly discuss the conditions under which there can exist a simple stationary equilibrium at q. Note first that to sustain a simple stationary equilibrium at $q = 0$, it must

be that all new agents choose defection as their consistent action. Thus, there is always a simple stationary equilibrium at $q = 0$ because the expected lifetime utility from defection is greater than the expected lifetime utility from cooperation, given that all agents in the system defect. A simple stationary equilibrium at $q = 1$ requires that all agents choose cooperation, and hence the expected lifetime utility of cooperation must be at least that of defection. Finally, a simple stationary equilibrium at $q \in (0, 1)$ requires mixing. Such a strategy is optimal if and only if the expected utilities of lifelong cooperation and lifelong defection are equal.

We conclude that, in order to characterize the stationary equilibria, it is enough to derive expressions for the expected utility of cooperation and of defection as a function of q, under the assumption that the state of the system is fixed at L_q. This is extremely useful for the analysis, since the state L_q pins down all aspects of the system that are relevant for expected utilities. There then exists a simple stationary equilibrium at each value of q for which the required relationship between these utilities holds.

3.2 Expected Utilities

We now derive the expected utilities associated with the (consistent) choices of cooperation and defection at an agent's birth. These utilities depend on the model's parameters, (a, b, δ). They depend as well on the proportion of cooperative agents in society, q. Since we are interested in simple stationary equilibria, we work under the assumption that the system is in state L_q and remains so over the agent's lifetime. If q is to be an simple stationary equilibrium it is rational for agents to compute their expected utilities assuming that the system remains in L_q.

The main task in computing expected utilities is to keep track of the expected number of inlinks and outlinks between agents of different behaviors, C and D, as a function of age. Define $n_{XY}^{Out}(s)$ as the expected number of outlinks from an agent of type X at age s to agents of type Y, $X, Y \in \{C, D\}$. The expected number of links from a cooperator of age s to other cooperators can be computed recursively according to

$$n_{CC}^{Out}(s) = \delta n_{CC}^{Out}(s-1) + q(K - \delta n_{CC}^{Out}(s-1)).$$

The first term retains the existing links with cooperators who remain alive, while the second term takes all links from the previous period that were broken (due to death or defection) and re-matches them, obtaining a fraction q of new cooperators. Setting $n_{CC}^{Out}(-1) = 0$ and solving produces

$$n_{CC}^{Out}(s) = qK\left(\frac{1 - (\delta(1-q))^{s+1}}{1 - \delta(1-q)}\right).$$

The remaining links sponsored by a cooperator go to defectors, so that $n_{CD}^{Out}(s) = K - n_{CC}^{Out}(s)$. For defectors, as mentioned, the case is much simpler, and depends only on the population frequency of cooperators. We have $n_{DC}^{Out}(s) = qK$ and $n_{DD}^{Out}(s) = (1-q)K$.

We derive next the expected number of inlinks from both types of nodes as a function of age. To do so, we compute the number of inlinks an agent expects to receive from agents of

either behavior at each date. Notice that the probability that a randomly selected node is age s is $p(s) = (1-\delta)\delta^s$. Then, the expected number of inlinks an agent will receive from cooperators and defectors at each date are, respectively,

$$\begin{aligned} r_C &= q\sum_{s=0}^{\infty} p(s)\left(K - \delta n_{CC}^{Out}(s-1)\right) = qK\frac{(1-\delta^2)}{1-\delta^2(1-q)}, \\ r_D &= (1-q)K. \end{aligned}$$

Notice that the calculation of r_C requires the assumption that the system is in a steady-state, since it presumes that for every age s, the proportion of age-s agents that cooperate is q. (The calculation for r_D, on the other hand, is valid for any state consistent with q since the number of outlinks sent by a defector is independent of age.)

Define $n_{XY}^{In}(s)$ as the expected number of inlinks an agent of type X at age s has from agents of type Y, $X, Y \in \{C, D\}$. For CC links, we have the recursive relationship

$$n_{CC}^{In}(s) = \delta n_{CC}^{In}(s-1) + r_C.$$

Setting $n_{CC}^{In}(-1) = 0$ and solving produces

$$n_{CC}^{In}(s) = r_C\frac{1-\delta^{s+1}}{1-\delta}.$$

The remaining calculations are straightforward since they all involve defectors whose links are re-set every period. We have $n_{CD}^{In}(s) = n_{DD}^{In}(s) = r_D$ and $n_{DC}^{In}(s) = r_C$.

Finally, we can now define the expected lifetime utility of consistently cooperating or defecting. To that end we compute the expected payoff at a particular age s by summing the payoffs over the expected set of connections. We have

$$\begin{aligned} \pi_C(s) &= \left(n_{CC}^{Out}(s) + n_{CC}^{In}(s)\right) - b\left(n_{CD}^{Out}(s) + n_{CD}^{In}(s)\right), \\ \pi_D(s) &= (1+a)\cdot\left(n_{DC}^{Out}(s) + n_{DC}^{In}(s)\right). \end{aligned}$$

Expected normalized lifetime utilities are then simply $u_X = (1-\delta)\sum_{s=0}^{\infty}\delta^s\pi_X(s)$, $X \in \{C, D\}$. Simplifying the expressions and scaling by the factor $1/K$ delivers

$$\begin{aligned} u_C &= \frac{2q - b(1-q)(2-\delta^2(2-q))}{1-\delta^2(1-q)}, \\ u_D &= \frac{(1+a)q(2-\delta^2(2-q))}{1-\delta^2(1-q)}. \end{aligned}$$

We remark that δ plays two distinct roles in the model. First, it determines the turnover rate at which agents enter and leave the system. Because of this, δ has a direct effect on the evolution of the system, holding fixed the behavior of all agents. It is in this role only that δ appears in our analysis until we come to the computation of u_C and u_D. Second, δ affects the preferences of agents because it represents the effective temporal discount rate. Thus for any given system dynamics, δ influences optimal behavior.

3.3 Characterization of simple stationary equilibria

Each agent chooses at birth C or D so as to maximize his expected utility. In order to characterize optimal choices under Assumptions 1-3 we are interested in comparing u_C and u_D as a function of q under various parameterizations of the model. It is convenient to define $\Delta(q; a, b, \delta) = u_D - u_C$.

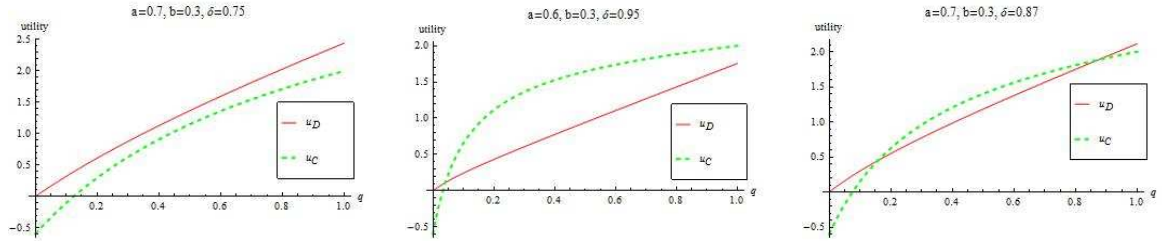
For given (a, b, δ), the values of q for which $\Delta(q; a, b, \delta) = 0$ are precisely the set of interior simple stationary equilibria.[8] This is so because, for such a value of q, mixing with probability q at birth is optimal provided the system remains in L_q, which it in fact will under the proposed strategy. For any other value of q, either u_C or u_D is strictly optimal under the assumption that the system remains in L_q, but then given that choices must be optimal, the system will not remain in L_q.

Notice that $\Delta(0; a, b, \delta) > 0$, so that $q = 0$ is always a simple stationary equilibrium. That is, if there are sufficiently few cooperators, then it cannot be optimal to commit to cooperation. On the other hand, if $q = 1$, then we will see that cooperation is sustained in certain settings, e.g., if the expected lifetime is sufficiently long and the gain from defecting against a cooperater is sufficiently small, then the long-term value from accumulated cooperator links is outweighed by the short-term gains from defection.

We are now able to characterize the mixtures of cooperation and defection that can be sustained in a simple stationary equilibrium. We are particularly interested in those simple stationary equilibria that are *stable*. A simple stationary equilibrium is stable if the system returns to it after sufficiently small perturbations of q, given that agents apply strategies in which they maximize utilities as calculated by u_C and u_D above. Note that this definition of stability can be viewed as bounding the rationality of the agents, as the utility calculations assume the system is at a steady-state at the current (perturbed) value of q, whereas the implied dynamics of the system in response to the perturbation are not, in fact, in steady-state. However, as long as the perturbation is sufficiently small, we assume that the error due to this approximation is small enough to justify our assumption that agents do not take it into account.

Generically, the set of stable simple stationary equilibria fall into three categories. First, it is possible that $\Delta(q; a, b, \delta) > 0$ for all q, in which case all-defection is the unique simple stationary equilibrium, and it is stable. For any a, b, this will be the case for sufficiently small δ. Second, it may be that there is a unique q^- for which $\Delta(q; a, b, \delta) = 0$, above which $\Delta(q; a, b, \delta) < 0$. In this case, all-defection ($q = 0$) and all-cooperation ($q = 1$) are the two stable simple stationary equilibria, while q^- is an unstable simple stationary equilibrium. Finally, it may be that $\Delta(q; a, b, \delta) < 0$ for an interior region of $q \in (q^-, q^+)$, and positive otherwise. In this case q^+ is a stable simple stationary equilibrium that involves the co-existence of cooperators and defectors (and q^- is again an unstable simple stationary equilibrium). See Figure 1 for an illustration of utility curves u_C and u_D corresponding to each of these scenarios. The following result fully characterizes these possibilities.

[8] This argument is readily extended to $q = 0, 1$.



(a) (b) (c)

Figure 1: Utility curves corresponding to different patterns of equilibrium occurance. (a) Only the all-defection state is an equilibrium. (b) The all-cooperate and all-defect states are at equilibrium, and there is an unstable equilibrium for some $q \in (0,1)$. (c) The all-defect state is an equilibrium, as well as two interior equilibria: the rightmost stable, the leftmost unstable.

Proposition 1 *For all (a,b,δ), $q=0$ is a stable simple stationary equilibrium. The remaining simple stationary equilibria are as follows:*
If $a>1$, then:

(i) if $b<2$ and δ is sufficiently large, there exist two interior simple stationary equilibria: one stable and one unstable, with the stable simple stationary equilibrium involving more cooperators.
(ii) otherwise ($b \geq 2$ or δ not large enough) there is only the $q=0$ simple stationary equilibrium.

If $a<1$, then:

(iii) if δ is sufficiently large then $q=1$ is a stable simple stationary equilibrium, and there exists an unstable interior simple stationary equilibrium.
(iv) if δ is sufficiently small then only the $q=0$ state is a simple stationary equilibrium.
(v) if $b<a(1+a)$, then there exists an intermediate range of δ for which there are two interior simple stationary equilibria: one stable, and one unstable, with the stable simple stationary equilibrium involving more cooperators.

Proof. First, $\Delta(0;a,b,\delta) = 2b > 0$, so that $q=0$ is always a stable simple stationary equilibrium. Next, internal simple stationary equilibria must satisfy the condition $\Delta(q;a,b,\delta)=0$. Solving for δ produces

$$\delta^*(q;a,b) = \sqrt{\frac{2\left(aq+b(1-q)\right)}{(2-q)\left((1+a)q+b(1-q)\right)}}.$$

To find interior simple stationary equilibria, we need to characterize for all a,b those δ such that $\delta^*(q;a,b)=\delta$ for some $q\in(0,1)$. In the following arguments, we derive the shape of the curve $\delta^*(q;a,b)$, proving that for any a,b, the shape will be similar to that drawn in Figure 2. Those values of δ that do not cross the curve describe systems with $q=0$ as

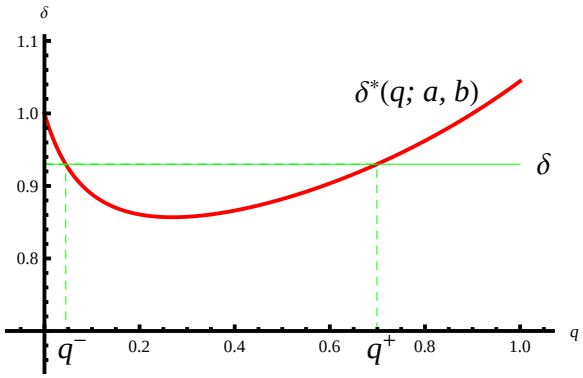


Figure 2: Graph of $\delta^*(q; a, b)$.

the only simple stationary equilibrium; those that cross it once result in a single unstable interior simple stationary equilibrium and a stable simple stationary equilibrium at $q = 1$; those that cross it twice result in two interior simple stationary equilibria, one stable and one unstable. Due to the shape of the curve, there are never more than two interior simple stationary equilibria.

To describe $\delta^*(q; a, b)$, we first note that given $a, b > 0$, $\delta^*(q; a, b)$ is bounded away from zero for all q. Furthermore, taking derivatives we see that for any $a, b > 0$ and $q \in [0, 1]$, $\Delta(q; a, b, \delta)$ is strictly decreasing in δ. Hence, for sufficiently small δ ($\delta < \sqrt{\frac{a}{1+a}}$ suffices), u_D dominates u_C for all q, and the only simple stationary equilibrium is $q = 0$, proving the second claim in part (ii) and part (iv). We next show that $\delta^*(q; a, b)$ is single-peaked (with the possibility of the peak at the boundary, in which case it is monotonic), and hence there can be at most two interior simple stationary equilibria $q \in (0, 1)$.

Claim **1** *The function $\delta^*(q; a, b)$ is single-peaked on the interval $[0, 1]$.*

Proof of Claim.
As is clear from the representation above, $\delta^*(q; a, b)$ has a unique point of discontinuity, and it is strictly greater than one. Thus, $\delta^*(q; a, b)$ is continuous on the unit interval. It is also continuously differentiable on the same interval. Thus to prove the claim, it is sufficient to show that $\delta^*(q; a, b)$ has at most one local optimum in $[0, 1]$.

Because $\delta^*(q)$ is bounded away from zero, its derivative has the same zeros as the derivative of $(\delta^*(q))^2$. Setting $\frac{\partial(\delta^*(q))^2}{\partial q} = 0$ produces a quadratic in q. Call the solutions q_1 and q_2. We must show that at most one solution falls inside the unit interval. If $a = b$ then it is easy to see that $q_1 = q_2 = 1 - b/2$. If $a = b + 1$ then $q_1 = -b - \sqrt{b(b+2)/2} < 0$. Otherwise, the solutions are

$$\frac{-b}{a-b} \pm \frac{\sqrt{(1+a-b)(2a-b)b}}{(1+a-b)(a-b)}.$$

Call this $Q_1 \pm Q_2$. If Q_2 is not real, we are done, so assume it is. If $a - b > 0$ then $Q_1 - Q_2 < 0$. On the other hand, if $a - b < 0$ then $Q_1 > 1$ so at least one of the solutions is greater than one. This proves the claim.

It is easily seen that $\delta^*(0;a,b) = 1$ and $\delta^*(1;a,b) = \sqrt{\frac{2a}{1+a}}$, which is less than one if and only if $a < 1$. Also,

$$\begin{aligned}\frac{\partial \delta^*(q;a,b)}{\partial q}|_{q=0} &= \frac{b-2}{4b},\\ \frac{\partial \delta^*(q;a,b)}{\partial q}|_{q=1} &= \frac{a(1+a)-b}{(1+a)^2\sqrt{\frac{2a}{1+a}}}.\end{aligned}$$

Recall that a stable interior simple stationary equilibrium occurs for some (a,b,δ) when $\delta^*(q;a,b) = \delta$ for two distinct values of q. From the above expressions, we see that stable interior simple stationary equilibria exist for some δ, provided that $b < 2$ and $b < a(1+a)$. We consider two cases based on the value of a.

1. If $a > 1$, the condition for existence of an interior stable simple stationary equilibrium reduces to requiring $b < 2$. Such a simple stationary equilibrium exists for all sufficiently large δ because $\delta^*(1;a,b) > 1$. This proves parts (i) and (ii).
2. If $a < 1$ then, since $a(1+a) < 2$, a stable interior simple stationary equilibrium exists for some δ whenever $b < a(1+a)$, proving part (v). However, now it is the case that $\delta^*(1;a,b) < 1$, which implies that a stable interior simple stationary equilibrium does not exist for $\delta > \delta^*(1;a,b)$. Rather, for $\delta > \delta^*(1;a,b)$, there is an unstable interior simple stationary equilibrium, and $q = 1$ is a stable simple stationary equilibrium, proving part (iii).

□

Each of the above conditions occurs for reasonable ranges of parameters; see Figure 1 for some typical examples.

We are particularly interested in stable equilibria that support (full or partial) cooperation. Roughly, there are two main factors driving the existence of these outcomes.

(a) In societies with nearly universal cooperation, the utility of cooperation cannot be far behind that of defection, i.e., $u_D(q)$ is not much greater (if at all) than $u_C(q)$ near $q = 1$).
(b) If $u_C(1) < u_D(1)$, then as defectors enter a mostly-cooperator system, the utility of defecting decreases faster than the utility of cooperating, i.e., the derivative of $u_D(q)$ is greater than that of $u_C(q)$ near $q = 1$.

A cooperator gains utility by building a network of relationships. Given sufficient time, the neighborhood of a cooperator limits to a particular size, at which point the death rate of neighbors matches the rate of finding other cooperators. A major factor in the payoff of a cooperator is the amount of time necessary to approach this limiting neighborhood size, relative to the expected lifetime. This quantity is influenced by the fraction q of cooperators in the system, but this influence suffers diminishing returns: when there are few cooperators present, a small increase has a large effect on the number of cooperators expected to meet

each other; when there are many cooperators, they will quickly approach their limiting neighborhood sizes, and thus the addition of more cooperators has little effect.

The utility of a cooperator is also affected by the losses incurred from interacting with defectors. This effect is roughly proportional to the number of defectors in the system (as is the total utility gained by a defector). Parameter b roughly determines the rate at which a cooperator's utility decreases with the proportion of defectors.

When $a > 1$, the expected utility obtained by a single defector in an otherwise all-cooperator environment will be greater than the expected utility of a cooperator who has a full neighborhood of other cooperators. That is, a $q = 1$ simple stationary equilibrium cannot exist for any δ. Starting from the $q = 1$ state, defectors begin to enter the system. As more defectors enter, the expected utility of each defector decreases roughly linearly. How is the utility of the cooperators affected?

First, if b is very large, the presence of more defectors degrades the utility of the cooperators heavily, due to losses that occur when interacting with defectors. If b is large enough (larger than two), this degradation will be so severe that defecting will always be the superior strategy, and the only simple stationary equilibrium of the system will be at $q = 0$.

Second, if the expected lifetime is sufficiently short, the presence of more defectors will make it substantially less likely that cooperators will form full neighborhoods of other cooperators within their lifetimes, again degrading their utility and destroying the $q = 1$ simple stationary equilibrium. If δ is small enough, the payoff due to forming a (partial) neighborhood will never overtake the utility of defecting, and again the only simple stationary equilibrium of the system will be at $q = 0$.

Third, if b is small and δ is sufficiently large, then an increase in the number of defectors will have a small effect on the expected welfare of a cooperator. Thus, as more defectors enter the system, the gap in welfare between defectors and cooperators will close, until at some interior point they become equal. This is precisely the stable interior simple stationary equilibrium described in the first half of the proposition.

When $a < 1$, the expected utility of a single defector in an otherwise all-cooperator utopia will be less than the expected utility of a cooperator who has a full neighborhood of other cooperators. That is, an all-cooperate simple stationary equilibrium exists provided δ is sufficiently large. In such a case, there must also be an unstable internal simple stationary equilibrium (since both the $q = 0$ and $q = 1$ states are stable, there must be some interior state where utilities are equal).

If δ is very small, then (as in the case $a > 1$) cooperators will not expect to develop large neighbourhoods of other cooperators during their lifetimes. In such a setting, it will always be better to defect than to cooperate, and only the $q = 0$ state will be stable.

Finally, consider a range of δ for which cooperators expect not to fully reach their limiting neighborhood size, but will come close. It may then be the case that a defector gains more utility than a cooperator in the $q = 1$ state. However, if the losses incurred due to exploitation are not too large, and if δ is large enough that cooperators expect to find many other cooperators over their lifetimes (though not as many as they could hope for), then an increase in the number of defectors will have more effect on the defectors' utilities than

on the cooperators' utilities. In this case, starting from $q = 1$ and adding defectors to the system, one reaches a state where the utilities of the defectors and the cooperators are equal.

Proposition 1 also enables us to state some comparative statistics relating changes in the simple stationary equilibrium fraction of cooperators to changes in the parameters. For given (a, b, δ), define q^- as the unstable simple stationary equilibrium and q^+ as the stable simple stationary equilibrium, when they exist.

Proposition 2 *Fix (a, b, δ) such that q^- and q^+ exist and are interior (i.e., $0 < q^- < q^+ < 1$). Then*

(i) as δ increases, q^- decreases and q^+ increases;
(ii) as a or b increase, q^- increases and q^+ decreases.

Proof. It is easy to verify that $\Delta(q; a, b, \delta)$ is strictly decreasing in δ, and strictly increasing in a and b, for all q. This, together with the fact that $\Delta(q; a, b, \delta)$ is strictly decreasing in q at q^- and strictly increasing in q at q^+, proves the result. □

That is, an increase in life expectancy, a decrease in gains from defecting against a cooperator, or a decrease in the penalty of cooperating with a defector, all result in more cooperation, in the sense that the stable simple stationary equilibrium increases and the basin of attraction for the simple stationary equilibrium with cooperation increases.

4 Robustness

The simple stationary equilibria identified in Proposition 1 are defined in a setting that requires agents to apply strategies that are consistent, unforgiving and trusting. This can be thought of as an equilibrium that arises under a very natural social norm, in the spirit of [11]. The social norm specifies how to behave in one's relationships as well as how to manage these relationships.

In this section we remove the assumptions from the previous section, and study optimal behavior in the absence of social norms that restrict strategies. We find that, under appropriate parametric conditions, the conventions described in Assumptions 1-3 are self-enforcing, in the sense that they constitute equilibrium outcomes.

Before stating the result, let us describe some parametric conditions that will be required. Notably, these conditions involve q as well as (a, b, δ). However, the analysis in this section concerns only steady-states of the system, where the fraction q of cooperation remains constant (though endogenously determined). The conditions below should be interpreted as requirements of a particular steady-state q under consideration. The two conditions, which we call the *consistency inequality* and the *trusting inequality*, are the following.

Definition **1** *The* consistency inequality *is*

$$\frac{1+b}{1+a} > 1 - (1-q)\delta^2. \tag{1}$$

DEFINITION **2** *The* trusting inequality *is*

$$b < \frac{q}{(1-q)(1-(1-q)\delta^2)}. \tag{2}$$

The result we provide is that any simple stationary equilibrium q that satisfies these two conditions also occurs as an equilibrium outcome without the assumptions of the previous section. Each agent applies a strategy that, at every observable history, optimizes his (time-discounted) continuation utility in expectation over future randomness and his beliefs about the state of the system given his observations. In turn, these beliefs are consistent with the strategy being employed (recall our focus on symmetric equilibria). A formal development of the solution concept is provided in the Appendix.

We demonstrate that there exists a equilibrium in which, on the equilibrium path, agents behave in a way that conforms with the norms of being consistent, trusting, and unforgiving. In other words, we find that these norms are in equilibrium even when they are not enforced. Recall that L_q denotes the steady-state of the system that occurs when agents apply strategies consistent with Assumption 1-3 and new agents choose cooperation with probability q.

THEOREM **1** *Suppose that $q \in [0,1]$ is a simple stationary equilibrium, and that the consistency and trusting inequalities are satisfied at q. Then there exists a equilibrium such that if the system is in state L_q, all agents apply actions that are consistent, trusting, and unforgiving. Moreover, q is a stationary level of cooperation under this strategy.*

Proof. See below. □

It is possible *a priori* that there exist equilibrium delivering steady-states in which agents do not behave according to Assumptions 1-3 on the equilibrium path. We demonstrate that if we *enforce* one of the norms, namely that all agents apply consistent strategies, and if the trusting inequality is satisfied at the steady-state q, then it must be that q is a simple stationary equilibrium.

THEOREM **2** *Suppose that agents play only consistent strategies. If there exists a equilibrium under which $q \in [0,1]$ is a stationary level of cooperation, and the trusting inequality is satisfied at q, then q must be a simple stationary equilibrium.*

Proof. See below. □

We interpret Theorems 1 and 2 as suggesting that our description of the likely level of cooperation in society, as given by Proposition 1, survives largely unchanged when considering the more general strategic setting of this section.

The remainder of this section is dedicated to the proof of Theorem 1 and Theorem 2. We discuss the optimality of each of the three assumptions separately, and conclude by combining these results.

4.1 Maintenance of relationships

We first assess the norm that individuals sever a relationship only upon observing a defection. If agents apply consistent strategies and commit to life-long cooperation or life-long defection, the beliefs of an individual regarding the future play of his partners are easy to describe. In fact, after a single interaction, the individual can perfectly forecast his partners' future play. Therefore, if other agents behave consistently, it is optimal to always maintain a relationship after observing cooperation, and it is optimal to sever a relationship after observing defection. Indeed, these decisions are *strictly* optimal: maintaining a link with a cooperator has positive expected utility, and maintaining a link with a defector has negative expected utility (for a cooperator). A link between two defectors must be severed because the sponsor of that link strictly prefers to re-match and have probability q of interacting with a cooperator. This behavior is sequentially rational and holds for off-path play in which an agent's partner behaves inconsistently, since consistency is defined as always taking the same behavior as was taken in the previous round. It is therefore the case that every best response has the property that a link is broken if and only if a defection is observed on that link.

4.2 Consistent Behavior

The analysis in Section 3 was conducted under the assumption that individuals have available to them only two (pure) strategies at their birth. Optimality, then, requires taking rational expectations over the implied outcomes of these two actions and choosing appropriately. There is no consideration of deviations from consistency; the choice is assumed to be made with commitment. We now want to show that if agents play consistent and unforgiving strategies, and the consistency inequality is satisfied, then consistent behavior is (part of) a best response.

First notice that for a defector in a steady-state, the calculation is identical at every round. This is so because, under the norm of unforgiving strategies, he loses all of his connections at every period. Thus, if he decides today that perpetual defection is better than perpetual cooperation, he will reach the same conclusion tomorrow.

For a cooperator the situation is complicated by the fact that his state (i.e. number of in-links and out-links) changes over time. At a simple stationary equilibrium $q > 0$, a cooperator is at least as happy with his choice, at birth, than he would be under the alternative plan of defection. But, in principle, there may arise interim situations in which a cooperator prefers to defect in a particular period, after which his optimization problem is identical again to the one at his birth.

We now introduce notation to describe the state of an individual of age s. For a given agent, let K_s^I denote the number of in-links from cooperators at the beginning of round s, and let K_s^O denote the number of out-links to cooperators at the beginning of round s.

The next result provides a sufficient condition to guarantee that cooperators never have a profitable deviation involving defection.

LEMMA **1** *Suppose $q \in (0,1]$ is a simple stationary equilibrium and the consistency inequality holds at q. Consider an agent that has K_s^I inlinks and K_s^O outlinks at the beginning of round*

s, with $K_s^I + K_s^O > 0$. Then the expected utility of cooperating on all rounds starting at s is strictly greater than the expected utility of defecting on round s and then cooperating on all subsequent rounds when other agents play simple strategies.

Proof. We focus attention on a fixed agent i. Let ϕ_C denote the simple strategy in which agent i cooperates each round, and let ϕ_D denote the simple strategy in which agent i defects each round. Let ϕ_F denote the strategy in which the agent defects for one round, then cooperates on every subsequent round (and is unforgiving and trusting on every round). For an arbitrary age s and a given strategy ϕ, write $u(\phi, k_I, k_O)$ for the expected utility, evaluated at the beginning of round s, of applying strategy ϕ when $K_s^I = k_I$ and $K_s^O = k_O$, and other players use simple strategies. To prove the lemma, we must show that $u(\phi_C, k_I, k_O) > u(\phi_F, k_I, k_O)$ whenever $k_I + k_O > 0$.

We will first show that $u(\phi_C, 0, 0) \geq u(\phi_F, 0, 0)$. To see this, note that ϕ_D and ϕ_F are identical on their first round of play, and at the end of that first round agent i will have no links (since other agents apply unforgiving strategies). After that first round, ϕ_F proceeds in the same way as ϕ_C. Moreover, since $q > 0$ is a simple stationary equilibrium, we know that $u(\phi_C, 0, 0) \geq u(\phi_D, 0, 0)$. Putting this together, we have

$$u(\phi_F, 0, 0) - u(\phi_D, 0, 0) = \delta(u(\phi_C, 0, 0) - u(\phi_D, 0, 0)) \leq u(\phi_C, 0, 0) - u(\phi_D, 0, 0)$$

from which we conclude $u(\phi_C, 0, 0) \geq u(\phi_F, 0, 0)$.

Write $\Delta u(\phi, k_I, k_O)$ for $u(\phi, k_I, k_O) - u(\phi, 0, 0)$, the utility gain due to adding k_I in-links and k_O out-links to agent i before applying strategy ϕ. We next show that $\Delta u(\phi_C, k_I, k_O) > \Delta u(\phi_F, k_I, k_O)$ for all $k_I + k_O > 0$, which will complete the proof. We note that these utility gains are additively separable in k_I and k_O, so that $\Delta u(\phi_C, k_I, k_O) = \Delta u(\phi_C, k_I, 0) + \Delta u(\phi_C, 0, k_O)$ and $\Delta u(\phi_F, k_I, k_O) = \Delta u(\phi_F, k_I, 0) + \Delta u(\phi_F, 0, k_O)$. We will therefore analyze these gains separately.

Consider first the utility gain due to in-links. We have $\Delta u(\phi_F, k_I, 0) = (1 + a)k_I$, since the agent gains $(1 + a)$ from each link and loses them after his first defection. When applying strategy ϕ_C, the gain is $\Delta u(\phi_C, k_I, 0) = \frac{k_I}{1-\delta^2}$. This is so because the cooperator gets extra utility for each period of the life of the relationship. We have that $\Delta u(\phi_C, k_I, 0) > \Delta u(\phi_F, k_I, 0)$ whenever $\frac{1}{1-\delta^2} > 1 + a$, which is necessary to sustain cooperation in a simple stationary equilibrium anyway.

We turn now to out-links, where a fraction k_O of the agent's out-links are already matched to cooperators, and the remaining out-links will be matched to the population at random. For strategy ϕ_F, $\Delta u(\phi_F, 0, k_O) = (1 + a)(1 - q)k_O$. To see this, note that the increase in the number of out-links to cooperators is $k_O + (1 - k_O)q - q = (1 - q)k_O$, and this gain is realized for exactly one period. For cooperators, $\Delta u(\phi_C, 0, k_O) = \frac{(1+b)(1-q)k_O}{1-(1-q)\delta^2}$. To see this, notice that per interaction, a cooperator gains $1 + b$ from interacting with a cooperator rather than a defector, and as discussed above the node gains $(1 - q)k_O$ extra out-links to cooperators. Finally, for a given outlink this gain is maintained as long as the node survives (probability δ), its cooperate partner survives (probability δ), and the outlink of the node in the scenario without the initial k_O cooperate outlinks is to a defector (probability $(1 - q)$).

These events happen independently and hence have a total probability of $\delta^2(1-q)$ yielding the above formula. Thus $\Delta u(\phi_C, 0, k_O) > \Delta u(\phi_F, 0, k_O)$ precisely when the consistency inequality holds, completing the proof. □

The condition in Lemma 1 guarantees that, as a node obtains more in-links or out-links with cooperators, the gain from those relationships is higher to a cooperator than to a defector. Thus, a node that found it optimal to cooperate at birth necessarily finds it optimal to cooperate at any future point in its lifetime.

4.3 Accepting links

We next address the norm that each agent, whether cooperator or defector, initially accepts every proposed inlink. For a defector, this behavior is strictly dominant whenever $q > 0$ (and weakly dominant when $q = 0$), since defectors necessarily obtain non-negative utility from any relationship. For cooperators, however, the rationality of this norm is not obvious, since they receive a negative payoff if they accept an inlink initiated by a defector. One might imagine a scenario in which there are many defectors in the population so that a proposed inlink is likely to have come from a defecting agent. In such a case, it may be that a cooperator suffers an expected utility loss from accepting an incoming link, and hence should refuse all inlinks. Of course, such decisions would have a severe impact on the network, as they prevent the formation of any profitable relationships, which are necessary to have any hope of sustaining cooperation in equilibrium. With this in mind, we wish to characterize the circumstances in which a cooperator's expected utility of a new inlink is positive at a given steady-state L_q.

Recall from Proposition 1 that we have a simple stationary equilibrium at $q = 0$ and possibly $q = 1$, depending on parameter values. However, the issue of accepting inlinks is not interesting in these cases, as either there are no cooperators to deviate from the norm (when $q = 0$), or the utility of accepting inlinks is trivially positive (when $q = 1$). We therefore focus on interior simple stationary equilibria.

We demonstrate that, whenever the trusting condition is satisfied at q, a cooperator has strictly positive expected utility from accepting an inlink. That is, all best responses at a simple stationary equilibrium involve agents accepting all inlinks.

LEMMA **2** *Suppose that $q \in (0, 1]$ is a simple stationary equilibrium, the trusting inequality is satisfied, and that agents apply consistent strategies. Then, for any proposed link, a cooperator has strictly positive expected utility for accepting the link.*

Proof. Given that agents apply consistent strategies, recall from Section 3 that the expected utility of accepting a proposed link is proportional to $r_C - b * r_D$. The result therefore holds if and only if $r_C - b * r_D > 0$, which is equivalent to $b < \frac{q}{(1-q)(1-(1-q)\delta^2)}$, as required. □

We remark that the act of re-matching an outlink upon the end of a previous relationship is built into the model. However, were this to become an endogenous choice, it would be strictly optimal to re-match an outlink immediately whenever the trusting condition is

satisfied. This is true because the probability of an outlink reaching a cooperator is q, whereas the probability that a new inlink comes from a cooperator is less than q, due to the fact that defectors send more outlinks per period than cooperators.

4.4 Equilibria with Unrestricted Behavior

We are now ready to complete the proofs of Theorem 1 and Theorem 2.

THEOREM **1** *Suppose that $q \in [0,1]$ is a simple stationary equilibrium, and that the consistency and trusting inequalities are satisfied. Then there is an equilibrium such that if the system is in state L_q, all agents apply actions that are consistent, trusting, and unforgiving on the equilibrium path. Moreover, q is a stationary level of cooperation under this strategy.*

Proof. We shall construct a symmetric equilibrium with the required properties. Recall that a formal definition of equilibrium appears in Appendix 7, but speaking briefly we require a strategy ϕ^* and a system of beliefs β^* about the state of the network, such that ϕ^* maximizes expected utility at all continuations given beliefs β^*, and β^* is consistent with observations under the assumption that other players apply strategy ϕ^*.

The strategy ϕ^* is as follows. First, if on any round an agent observes a fraction of cooperation other than q, the agent will choose to accept all proposed links, defect that round, and to break all links with observed defectors at the end of the round. Note that this behavior is optimal given that q is publicly observed and other agents also play according to ϕ^*, since these behaviors are optimal given the belief that all other agents will defect. Otherwise, if the observed fraction of cooperation is q, the agent will behave in accordance with Assumptions 1-3. On the agent's first round, upon observing the state q, he chooses to cooperate with probability q, otherwise he chooses to defect.

The associated belief system β^* is straightforward. At birth, the agent believes that the system begins in state L_q. The agent will continue to believe that the system is in steady-state L_q as long as the observed fraction of cooperation is q. Once a fraction of cooperation other than q is observed, the agent will believe that any other agent also alive on that round also observed this non-q fraction of cooperation, and will therefore defect on subsequent rounds. Note that we have not provided a full characterization of an agent's belief about the state of the network, but the properties discussed are sufficient to determine whether or not ϕ^* is an optimal strategy.

We note that ϕ^* satisfies the property that agents behave in accordance with Assumptions 1-3 in state L_q, and that q is a stationary level of cooperation under this strategy. It remains to show that applying strategy ϕ^* is optimal given that the observed fraction of cooperation is q and other agents play according to ϕ^*. Note first that, under the assumption that the system begins in state L_q and other agents play according to ϕ^*, it is consistent to believe that the system is in state L_q given that the observed fraction of cooperation is q. It is therefore sufficient to demonstrate that ϕ^* is optimal assuming the state of the system is L_q. Thus, for the remainder of the proof, we will assume that the system remains in state L_q at all times and describe our strategies only for this case.

We focus attention on a particular agent i. Write $u(\phi)$ for the expected lifetime utility of agent i when applying strategy ϕ. Let ϕ_{opt} denote a strategy that maximizes expected utility against the profile of all agents playing ϕ^* in state L_q, and suppose for contradiction that $u(\phi_{opt}) > u(\phi^*)$. Let ϕ_C denote the trusting, unforgiving and consistent strategy in which the agent chooses cooperation at birth, and let ϕ_D denote the similar strategy in which the agent chooses defection.

Note first that if $q = 0$, then no strategy obtains positive expected utility; thus strategy $\phi^* = \phi_D$ is optimal, since in this case $u(\phi_D) = 0$. We therefore assume $q > 0$ for the remainder of the proof.

As discussed in Section 4.1, we know that every optimal strategy breaks a link if and only if a defection is observed on that link. In particular, ϕ_{opt} must satisfy this property.

For all $r \geq 1$, define the random variable T_r as the age at which ϕ_{opt} prescribes that agent i defect for the r'th time. We then define strategy ϕ_D^r as the strategy in which agent i follows ϕ_{opt} up to and including round T_r, after which point he behaves according to ϕ_C. Note that ϕ_D^r is a well-defined strategy, since random variable T_r defines a stopping time. We also define strategy ϕ_C^r as the strategy in which agent i follows ϕ_{opt} up until round T_r, but on round T_r and all subsequent rounds he behaves according to ϕ_C. Note that ϕ_C^r and ϕ_D^r differ only on their actions on round T_r, in which ϕ_C^r specifies cooperation and ϕ_D^r specifies defection. Finally, for notational convenience we will define $\phi_D^0 = \phi_C^0 = \phi_C$.

We first claim that $u(\phi_C^r) \geq u(\phi_D^r)$ for all $r \geq 1$. Strategies ϕ_C^r and ϕ_D^r are identical until round T_r, at which point ϕ_C^r proceeds to cooperate on every subsequent round, whereas ϕ_D^r defects for a single round and then cooperates thereafter. Lemma 1 therefore implies that $u(\phi_C^r) \geq u(\phi_D^r)$, as agent i maximizes utility by cooperating on round T_r regardless of the configuration of incoming and outgoing links on round T_r.

We next claim that $u(\phi_D^{r-1}) \geq u(\phi_C^r)$ for all $r \geq 1$. Strategies ϕ_D^{r-1} and ϕ_C^r are identical until round T_{r-1}, after which both strategies prescribe cooperation on each turn, but ϕ_C^r does not necessarily accept every proposed inlink. However, Lemma 2 implies that it is optimal to accept proposed links when perpetually cooperating, and thus $u(\phi_D^{r-1}) \geq u(\phi_C^r)$.

Combining these two claims, we have that $u(\phi_D^{r-1}) \geq u(\phi_D^r)$ for all $r \geq 1$. But $\phi_D^0 = \phi_C$, and $\lim_{r\to\infty} u(\phi_D^r) = u(\phi_{opt})$ (noting that the limit must exist since utilities are time-discounted). We therefore conclude $u(\phi^*) \geq u(\phi_C) \geq u(\phi_{opt})$, which is the desired contradiction. □

Theorem 2 now follows easily.

THEOREM **2** *Suppose that agents play only consistent strategies. If there exists an equilibrium under which $q \in [0, 1]$ is a stationary level of cooperation, and the trusting inequality is satisfied at q, then q must be a simple stationary equilibrium.*

Proof. Assume that all agents behave consistently, and that their strategies form a equilibrium with a steady-state at q. Then, as discussed in Section 4.1, all optimal strategies must involve breaking links if and only if a defection is observed on that link. Additionally, Lemma 2 implies that all optimal strategies must accept all proposed links. We conclude that all optimal strategies are trusting and unforgiving, so these must be the behaviors that arise

at equilibrium. Moreover, it must then be that q is a simple stationary equilibrium, since it is a steady-state under an equilibrium of consistent, trusting, and unforgiving strategies. □

4.5 Remarks on Parameter Conditions

We now discuss the trusting inequality and the consistency inequality in more detail. It is easy to verify that the two conditions are independent in the space of paramters (a, b, δ). We begin by providing simple sufficient conditions that imply these inequalities. Then, we discuss the implications for behavior when the conditions are (separately) relaxed.

The consistency inequality is always satisfied when $b \geq a$.

Proposition 3 *If $b \geq a$, then the consistency inequality is satisfied for all $q \in [0, 1]$.*

Proof. Trivial. □

Next, any simple stationary equilibrium with $q > 2/3$ satisfies the trusting inequality.

Proposition 4 *If $q > 2/3$ is a simple stationary equilibrium, then the trusting inequality is satisfied at q.*

Proof. We note first that the trusting inequality is trivial when $q = 1$, so suppose $q < 1$. If $q > \frac{2}{3}$, then $\frac{q}{(1-q)(1-(1-q)\delta^2)} \geq \frac{q}{1-q} \geq 2$. Furthermore, Proposition 1 implies that $b \leq 2$ (either directly, if $a > 1$, or from the fact that $b < a(1 + a) < 2$ if $a < 1$). Thus $b \leq \frac{q}{(1-q)(1-(1-q)\delta^2)}$ as required. □

Our next result is that, if $a < 1$, then *every* simple stationary equilibrium $q > 0$ satisfies the trusting inequality. In other words, if the temptation to defect is not too large, then rational cooperators will choose to be trusting at equilibrium, regardless of the number of defectors in the population.

Proposition 5 *Suppose that $a < 1$ and that $q > 0$ is a simple stationary equilibrium. Then a cooperator obtains positive expected utility from accepting a proposed link.*

Proof. Fix a and b and choose δ such that a positive simple stationary equilibrium q exists. As in Proposition 4, it suffices to show that $b < \frac{q}{(1-q)(1-(1-q)\delta^2)}$.

We note first that the result is trivial when $q = 1$ (as cooperators can obtain only non-negative expected utility from any interaction), so suppose $q < 1$. Since $a < 1$, Proposition 1 implies that $b < a(1 + a)$, and hence $b < 2a$ and $b < 1 + a$. Recall from the proof of Proposition 1 that at a simple stationary equilibrium we have

$$\delta^2 = \frac{2(aq + b(1 - q))}{(2 - q)((1 + a)q + b(1 - q))}. \tag{3}$$

Define $Z(q; a, b)$ by

$$Z(q; a, b) := \frac{(2 - q)((1 + a)q + b(1 - q))}{(1 - q)(2 - q + aq + b(1 - q))}.$$

Substituting (3), it can be verified that $\frac{q}{(1-q)(1-(1-q)\delta^2)} = Z(q;a,b)$. It therefore suffices to show that $b < Z(q;a,b)$ at all internal SSE.

We first claim that $Z(q;a,b)$ is increasing as a function of q. This follows immediately from the following expression for the derivative of Z with respect to q,

$$\frac{\partial Z(q;a,b)}{\partial q} = \frac{q}{(1-q)^2} + \frac{4a-2b}{(2-q+aq+b(1-q))^2},$$

which is non-negative since $b < 2a$. Let q^*_{min} denote the smallest positive simple stationary equilibrium, over all possible choices of δ. Since $Z(q;a,b)$ is increasing in q, it is sufficient to show that $b < Z(a,b,q^*_{min})$.

We next derive an expression for q^*_{min}. Recall from the proof of Proposition 1 that $\delta^*(q;a,b)$, which relates δ to q at simple stationary equilibrium, is concave and single-peaked in the range $(0,1)$. Furthermore, whenever there exist $0 < \underline{q} < \overline{q} < 1$ such that $\delta = \delta^*(\underline{q};a,b) = \delta^*(\overline{q};a,b)$, $\overline{q}$ is a stable simple stationary equilibrium and $\underline{q}$ is an unstable simple stationary equilibrium. Thus q^*_{min} is precisely the value of q at which $\delta^*(q;a,b)$ achieves its minimum on $[0,1]$. Solving $\frac{\partial \delta^*(q;a,b)}{\partial q} = 0$ for q, we obtain the pair of solutions

$$q = \frac{-b \pm \sqrt{\frac{b(2a-b)}{1+a-b}}}{a-b}.$$

Write $r(a,b) := \sqrt{\frac{b(2a-b)}{1+a-b}}$. Using the facts that $a < 1$ and $b < a(1+a)$, it is a simple exercise to show that $\frac{-b+r(a,b)}{a-b} \in [0,1]$ and $\frac{-b-r(a,b)}{a-b} \notin [0,1]$. We conclude that

$$q^*_{min} = \frac{-b+r(a,b)}{a-b}.$$

Substition and simplification then yields

$$Z(q^*_{min};a,b) = \frac{b(2(1+a)-b)^2}{L(a,b)}$$

where

$$L(a,b) = 2(1+a)^2 r(a,b) - (1+a)b(2(1-a)+r(a,b)) + b^2(r(a,b)-(1+a)).$$

Thus, to show that $b < Z(q^*_{min};a,b)$, it suffices to show

$$\begin{aligned}(2(1+a)-b)^2 \;>\;& 4(1+a)^2\left(\frac{r(a,b)}{2}\right)\\ -\;& 4(1+a)b\left(\frac{2(1-a)+r(a,b)}{4}\right)\\ +\;& b^2(r(a,b)-(1+a)).\end{aligned}$$

We derive this inequality with the help of the following claim:

Claim: For all $a < 1$ and $b < a(1+a)$, it must be that $r(a,b) < 2$ and $r(a,b) < 1+a$.

To prove the claim, we note that for fixed a, $r(a,b)$ attains its maximum at $b = 1 + a \pm \sqrt{1-a^2}$. Since $b < 1+a$, the admissible solution is $b = 1 + a - \sqrt{1-a^2}$, which yields $r(a,b) = \sqrt{2 - 2\sqrt{1-a^2}} < 2\sqrt{a}$. But $2\sqrt{a} < 2$ since $a < 1$, and moreover $2\sqrt{a} < 1 + a$ by considering the fact that $(1 - \sqrt{a})^2 \geq 0$. Thus $r(a,b) < 2$ and $r(a,b) < 1 + a$ as required, completing the proof of the claim.

Our claim immediately implies that $\frac{r(a,b)}{2} < 1$, $r(a,b) - (1+a) < 1$, and $\frac{2(1-a)+r(a,b)}{4} \in (0,1)$. Taking

$$\lambda = \max\left\{\frac{r(a,b)}{2}, r(a,b) - (1+a), \frac{2(1-a)+r(a,b)}{4}\right\},$$

we conclude that

$$\begin{aligned} & 4(1+a)^2\left(\tfrac{r(a,b)}{2}\right) \\ & -4(1+a)b\left(\tfrac{2(1-a)+r(a,b)}{4}\right) \\ & +b^2(r(a,b)-(1+a)) \\ \leq\ & 4(1+a)^2\lambda - 4(1+a)b\lambda + b^2\lambda \\ =\ & \lambda(2(1+a)-b)^2 \\ <\ & (2(1+a)-b)^2, \end{aligned}$$

completing the proof. □

Finally, we note that Proposition 5 is tight, in that it fails to hold when we remove the assumption that $a < 1$. Indeed, for any given $a > 1$, there exists a simple stationary equilibrium at which rational cooperators would choose to reject in-links. This follows from the observation that, when $a > 1$, a simple stationary equilibrium $q > 0$ exists for any $b < 2$ and $\delta < 1$; however, as $b \to 2$ and $\delta \to 1$, the value of q at this simple stationary equilibrium becomes arbitrarily small. The quantity $\frac{q}{(1-q)(1-(1-q)\delta^2)}$ from Proposition 4 can then be made arbitrarily close to 1, and hence less than b.

In summary, the norm that cooperators accept all in-links is without loss for rational agents at a simple stationary equilibrium whenever there are sufficiently many cooperators in the network. If $a < 1$, it turns out that *any* simple stationary equilibrium must have enough cooperators to motivate that acceptance of in-links. For the case of $a > 1$, there exist simple stationary equilibria with arbitrarily few cooperators, and hence there are choices of parameters for which rational agents would choose not to accept incoming links.

5 Additional Behaviors

We have discussed equilibria of the system subject to the consistency and trusting inequalities. In this section we consider the nature of behavior when the consistency inequality is

violated. Proposition 3 states that the consistency inequality is always satisfied when $b \geq a$. Thus, the only possibility for profitable deviation occurs when $a - 1 < b < a$.

Recall that we model a situation where an agent chooses one action in each period, and plays that action with each of his current partners. However, isolating the incentives from each relationship shows that a tension can arise in which an agent might prefer one action for some partners and one action for others. The proof of Lemma 1 shows that the consistency inequality implies that an agent with an existing out-link to a cooperator gains more utility from that relationship by cooperating than by defecting and forming a new relationship. If the consistency condition is violated, this no longer holds, and it would be profitable for an agent to defect in a relationship that she herself sponsors. However, again recalling the proof of Lemma 1, an agent always gains more utility from a relationship sponsored by another cooperator by cooperating than by defecting, provided the other agent is consistent. Thus, when the consistency condition is violated an agent would prefer to cooperate with in-link neighbors, but defect with out-link neighbors.

We conclude that the incentive to defect is strongest when the number of out-links to cooperators is high relative to the number of in-links from cooperators. Roughly speaking, it is better to defect when one's cooperating partners come from out-links, since those are the ones that are easier to replace.

When the consistency inequality is violated, a cooperator has a profitable deviation when the ratio of his out-links with cooperators to his in-links from cooperators is sufficiently high. Define this ratio to be $\bar{R} = k_O/k_I$. We note that under expected conditions, this ratio will not become large enough to rationalize a deviation. However, since agents maintain only a finite number of links, cooperators will reach a state that gives them a profitable deviation with positive probability. This happens to an agent, for instance, whenever all the cooperators maintaining links to him die simultaneously. In practice, these situations have significant probability only very early in the life of a cooperator, before it has had time to build a large network of in-links.

Proposition 6 *Assume that q is a simple stationary equilibrium. Assume that the trusting inequality is satisfied at q and the consistency inequality is violated at q. Then a cooperator has a profitable inconsistent deviation if and only if*

$$\bar{R}\left[(1+a) - \frac{1+b}{1-(1-q)\delta^2}\right] > \frac{1}{1-\delta^2} - (1+a).$$

Proof. If the right hand side is negative, then defection dominates cooperation and the only simple stationary equilibrium is $q = 0$, so assume otherwise. Then, the right hand side is the extra gain that a cooperator realizes from an in-link with a cooperator relative to the gain a defector realizes. The term in brackets is the extra gain a defector realizes from an out-link to a cooperator relative to the gain a cooperator realizes, which is positive by assumption. Then the result simply expresses that when the ratio of out-links to in-links is high enough, the net gain to defection is positive. □

We note the following corollary. Suppose q is a simple stationary equilibrium, the trusting condition holds at q, and agents apply strategies that are unforgiving, trusting, and

consistent. Then a best response by an agent i is the following "threshold" strategy: agent i will be unforgiving, trusting, and consistent, with the exception that if, on any round, $\bar{R}$ becomes larger than the threshold described in Proposition 6, the agent will defect on that round and restart the strategy on the following round.

What can we say about a setting in which *all* agents apply such a threshold strategy?[9] Note that as δ becomes large, the requisite threshold on $\bar{R}$ for profitable deviation increases without bound. Thus, in the limit, deviation is profitable only in cases where a node has *no* incoming links. Furthermore, due to the law of large numbers, the probability of any agent having no incoming links decreases exponentially with the number of links sponsered by each agent, K. Thus, for large values of δ, and K not too small, the difference in expected utility due to all agents applying threshold strategies (rather than consistent strategies) will be negligible. We therefore expect to find an ϵ-equilibrium of behavior in which all agents apply threshold strategies, using the threshold from Proposition 6.

One final observation about the threshold strategy is that, heuristically speaking, we cannot think of an agent as "setting out to exploit" relationships that are built up by virtue of cooperative behavior. Indeed, once an agent has built up social captial in the form of incoming links, he will strictly prefer to be cooperative in those relationships. However, if an agent's network becomes extremely poor due to the randomness inherent in the model, then the penalty for defection becomes weak enough that he may as well defect and restart the process of building capital in the subsequent round. In short, deviation from consistency arises from a lack of social ties, rather than premeditated exploitation of cooperators.

6 Conclusion

We have developed a model of interactions in an anonymous community with changing sets of partners. The class of simple strategies, which are unforgiving, trusting and consistent, provides the foundation for the first part of our analysis. Under simple strategies, we fully characterize stationary equilibria of the system. Full cooperation is sustainable for a non-trivial range of parameters, but not always. For some parameter choices, the presence of non-cooperative behavior in an anonymous system is unavoidable. We believe this captures an important feature of a number of applications.

Full cooperation requires not only that players are sufficiently patient, but also that the temptation payoff for defecting not be too large. When these conditions are not met, there necessarily exists some level of defection in society. The presence of defectors causes relationships among cooperators to be viewed as a scarce and valuable resource, which we identify as a form of social capital. This key mechanism, as well as its implications, are in sharp contrast to other models which focus on supporting only full cooperation, and have no natural way of describing a distinction between cooperators and defectors.

As it turns out, the characterization of steady-states under simple behavior, simple stationary equilibria, says a lot about outcomes in a more general setting where behavior is

[9] We give only an informal response to this question in this preliminary working paper. A formal description and analysis will appear in a forthcoming version of this manuscript.

not restricted to the class of simple strategies. This demonstrates that the simple behavior we focus on can be self-enforcing, in that it constitutes equilibrium behvaior more generally. In particular, under appropriate conditions, every simple stationary equilibrium has a corresponding equilibrium that supports simple behavior on the equilibrium pathand the same steady-state level of cooperation. Moreover, if an equilibrium with consistent behavior has a stationary level of cooperation, then that level of cooperation must be that of a simple stationary equilibrium.

We identify aspects of best responses when parameters are such that the conclusions regarding equilibrium fail. First, we demonstrate the optimality of a "rob the bank" strategy, in which a player cooperates initially, only to defect under a particular circumstance, burning their social capital. Notably, this behavior occurs *reactively* in response to a type of social empoverishment, and a cooperator never expects to find it optimal to carry out such a strategy ex ante. Finally, we are currently completing a result that exhibits an equilibrium with "exclusivity", i.e. where cooperators accept inlinks with some probability p.

7 Appendix A: formal development of the model

The model described in this paper is relatively complex, incorporating a changing set of players, a very large state space that is almost entirely unobserved by each individual player, and various sources of randomness. In the main text of this manuscript, we approached this model by handling the notions of strategies, equilibria, and beliefs in an informal manner. In this appendix we redescribe these concepts more formally, which will allow us to state the results more precisely.

7.1 Histories and Actions

The strategy of an agent is a mapping from its (private) history to (a probability distribution over) actions. The history encodes all the information the node has acquired during its life. In particular, the history of an agent contains its observation of q at each point during its life, all of its past actions, and the actions of each of its partners over time, together with how and when those relationships were initiated and ended. The action space is a choice of C or D, together with whether or not to sever any existing relationships and accept any new proposed links. We now develop these elements more formally.

The set of agents is the unit interval $N = [0, 1]$. Whenever an agent dies, it is replaced by an agent who takes the same name. We focus on an arbitrary agent i. Denote the age of i by s. In the period when i is born, $s = 0$; s increments by one in each subsequent round in which i remains alive. At each point in time, i observes the value of q determined by the choices at the previous round. Define q^s to be the proportion of cooperators that i observes in the round when i is age s. At each s, i chooses an $\alpha_i \in \{C, D\}$. For each partner j that i has at age s, the vector $\beta_j^s = \{\alpha_j^s, d_j^s, e_{ji}^s, e_{ij}^s\}$ defines the action that j takes, and whether and, if so, how the link was terminated in that round. The variable d_j^s equals 1 if j dies (0

otherwise), and the variables e^s_{ji} and e^s_{ij} record whether j or i respectively chooses to sever the link (a value of 1 corresponds to severing, 0 to not severing).

The collection of i's partners is recorded in two lists. The outlinks of agent i at age s are stored in a vector Out^s_i of length K. If i's k'th outlink at age s is to agent j, then the k'th element of this array is β^s_j. Due to anonymity, though, i does not know the value of j, but only the values of the elements in β^s_j. The inlinks of agent i require a bit more notation since there is not a fixed number of them. To account for this, we define a vector In^s_i representing the state of all current and past inlinks of agent i at age s. The k'th component of this list records information pertaining to the k'th inlink proposed to agent i over his life. Initially, In^s_i is empty. When agent i at age s receives a proposal for an inlink from agent j, it updates In^s_i as follows: if the link is accepted it appends β^s_j to In^s_i; if the proposed inlink is rejected outright, then we append a special symbol REJECT to In^s_i. After actions are realized, agent i updates each β^s_j in In^s_i appropriately. We define the *size* of list In^s_i, denoted by $|\text{In}^s_i|$ to be the number of *active* links contained in the list, i.e., the number of components of In^s_i for which $d^s_j = e^s_{ji} = e^s_{ij} = 0$.[10] Again, it is important that i not know the values of j corresponding to the various inlinks in In^s_i. Finally, denote by L^s_i the number of inlinks proposed to i in round s.

The information that i collects from the round in which he is age s is

$$h^s_i = \{q^s, \alpha^s_i, L^s_i, \text{Out}^s_i, \text{In}^s_i\}.$$

The (private) history of i at age s is the vector $H^s_i = \{h^0_i, \ldots, h^s_i\}$. In a valid history it must be the case that the length of the list In^s_i grows monotonically with s and that if the k'th component of In^s_i is either REJECT or a β^s_j indicating a link termination (i.e., either d^s_j, e^s_{ji}, or e^s_{ij} equals 1), then this component remains constant for the remainder of i's lifetime (i.e., for all $t > s$, the k'th component of In^t_i equals the k'th component of In^s_i). Denote the space of feasible age-s histories for i by $\mathcal{H}^s_i$. The set of all histories for i is then $\mathcal{H}_i = \cup_s \mathcal{H}^s_i$.

At each round, i takes three separate actions: (i) the choice of α_i, (ii) the acceptance or rejection of proposed inlinks, and (iii) the severance or continuation of each active link. The (history dependent) action set of i at age s is $A^s_i(H^s_i) = [0,1] \times [0,1]^{L^s_i} \times [0,1]^{K+|\text{In}^s_i|}$, with the interpretation that the first element specifies the probability that i chooses C at age s, the second element specifies the probability of accepting each proposed inlink, and the final element specifies the probability that i severs a link to each of his partners.

Let $\mathcal{A}^s_i = \cup_{H^s_i \in \mathcal{H}^s_i} A^s_i(H^s_i)$ denote the set of all age-s action sets, and let $\mathcal{A}_i$ denote the space of all action sets for i.

A strategy for i is a mapping $\phi_i : \mathcal{H}_i \to \mathcal{A}_i$, with the restriction that $\phi_i(H^s_i) \in A^s_i(H^s_i)$ for all $H^s_i \in \mathcal{H}_i$. When i makes the choice of α^s_i, he has all the information in H^{s-1}_i as well as q^s, but he has not observed the remainder of h^s_i. Similarly, when i makes his choice of accepting inlinks, he observes h^{s-1}_i and (q^s, α^s_i, L^s_i), but nothing else from round s. Last, when i makes the choice of severing active links, he has observed, additionally, the actions $\{\alpha^s_j\}$ in round

[10] Note that one can analogously define the size of Out^s_i; however as agents always replace outlink partners instantaneously, $|\text{Out}^s_i| = K$ for all i and s.

s of each of his active partners. We place the associated restrictions on strategies, so that actions depend only on the information observed at each of these times within a round.

Notice that, implicit in the construction of strategies is the Markovian property that, while actions generally depend on the age of an agent, they cannot be conditioned explicitly on time.

7.2 Equilibria

Recall that, in our definition of histories and actions, a single round involves a sequence of action choices to be resolved by an agent, where incremental observations are made between each choice. We then ensured that a strategy can use only the "currently available" information from the latest round of a history when defining action choices. While consistent with our informal game description, this point of view is notationally cumbersome. A change of variables would allow us to consider each step of a round as a separate information set, in which case a strategy is a mapping from histories to actions without restrictions. We will proceed with our discussion under this change, with the understanding that our notion of a history, strategy, etc. are fully equivalent to those developed in the previous section. Notice that $\mathcal{H}_i = \mathcal{H}^*$ and $\mathcal{A}_i = \mathcal{A}^*$ for all $i \in N$.

A *state of the world* ω is a directed graph with (labeled) vertex set $N = [0, 1]$, plus a history for each vertex. A state represents the links between players in a given round, along with each of their past observations. We write Ω for the set of all possible states of the world. In general, given any set S, we will write $\Delta(S)$ for the set of probability distributions over S.

A *belief* for agent i is a function $\beta_i : \mathcal{H}^* \to \Delta(\Omega)$ that maps each observed history to a distribution over possible world states. We interpret $\beta_i(H_i)$ as capturing agent i's beliefs about the state of the world given a sequence of observations.

We focus on strategy and belief profiles that are symmetric across agents, i.e., there is some strategy ϕ and belief β such that $\phi_i = \phi$ and $\beta_i = \beta$ for all $i \in N$.

Our goal is to define a notion of a symmetric equilibrium, which will be a pair (ϕ, β) that satisfies certain properties. Informally, we wish for the following: at all valid histories ϕ maximizes expected utility given β when other agents apply ϕ; β is consistent with an agent's observations and with the belief that all agents apply strategy ϕ; and, when faced with an unexpected history, β maps to a limit point of beliefs under a vanishing error probability. We now describe each of these desiderata in more detail.

We write $u_i(\bar{h}_i)$ for the expected continuation utility obtained by agent i, where $\bar{h}_i$ denotes a distribution over future histories that i will observe. Note that $\bar{h}_i$ captures any dependency on the strategy employed by agent i, as it is a distribution over future observations. Given strategies ϕ, ϕ_i' and state ω, we write $h_i^{\phi}(\phi_i', \omega) \in \Delta(\mathcal{H}^*)$ for the distribution over all future histories that will be observed by agent i when agent i applies strategy ϕ_i' and all other agents apply strategy ϕ, starting from state ω. We extend h_i^{ϕ} to accept a distribution over states in the natural way. We then say that ϕ is *optimal under belief* β if, for all $H_i \in \mathcal{H}^*$,

$$\phi \in \arg\max_{\phi_i'} \{u_i(h_i^{\phi}(\phi_i', \beta(H_i)))\}.$$

That is, for every history H_i, ϕ maximizes the expected utility of agent i given the distribution $\beta(H_i)$ over states, under the assumption that other agents apply strategy ϕ. We also say that ϕ is δ-approximately optimal if for all $H_i \in \mathcal{H}^*$, $u_i(h_i^\phi(\phi, \beta(H_i))) \geq u_i(h_i^\phi(\phi_i', \beta(H_i))) - \delta$ for all alternative strategies ϕ_i'.

Given ϕ, we now define the *progression function* $P^\phi : \Delta(\Omega) \to \Delta(\Omega)$. Given $\sigma \in \Delta(\Omega)$, $P^\phi(\sigma)$ is the distribution over states that results when all agents apply strategy ϕ for one round, starting from a state drawn from σ. Note that the resulting distribution is taken over randomness in strategy ϕ and the randomness inherent in the model, i.e. the death and matching processes. We next add the effects of an agent's observations to this distribution: given a distribution $\sigma \in \Delta(\Omega)$ over states, an agent i, and an observation h_i from a single round, we define $P^\phi(\sigma, h_i)$ to be the distribution over states that results after resolving a single round of play under ϕ, starting at a state drawn from σ, given that agent i observes h_i in that round. Note that this distribution is well-defined: one can consider the probability of observing h_i given each possible state and apply Bayes' rule.

We say that β is *consistent with strategy* ϕ if, for all i, s, H_i^{s-1} and h_i^s,

$$\beta(H_i^s) = P^\phi(\beta(H_i^{s-1}), h_i^s).$$

Observe that the requirement that β be consistent with strategy ϕ does not impose any restrictions on beliefs upon observation of a history that is inconsistent with ϕ. Thus, if this condition is taken to be sufficient for characterizing permissible equilibrium of beliefs, we have the undesirable feature that beliefs and, hence, behavior, is not appropriately restricted off the equilibrium path. This motivates us to require a form of perfection. Given an unexpected history H_i that has zero probability under ϕ, we would like (informally speaking) for agents to place belief in a minimal number of deviations from ϕ that yield a state consistent with H_i. To achieve this property formally, we will require not only that β be consistent with the application of strategy ϕ by all agents, but also that it maps to a limit point of beliefs under a vanishing trembling probability on actions.

We now formalize the intuition described above. Given any strategy ϕ and any $\epsilon \geq 0$, the ϵ-perturbation of ϕ is the strategy ϕ^ϵ that, independently for each action, follows ϕ with probability $1-\epsilon$, and with the remaining probability chooses an action uniformly at random. We say that β is robustly consistent with ϕ if

- β is consistent with ϕ,
- for all $\epsilon > 0$, there exists belief β^ϵ such that β^ϵ is consistent with ϕ^ϵ, and
- $\lim_{\epsilon \to 0} ||\beta^\epsilon - \beta||_{TV} = 0$ where $||\cdot||_{TV}$ denotes total variation distance.

Note that if ϕ is optimal given β, and β is robustly consistent with ϕ, then (taking β^ϵ as in the definition of robust consistency) ϕ^ϵ must be δ-approximately optimal for β^ϵ, where $\delta \to 0$ as $\epsilon \to 0$.

We are now ready to define our equilibrium concept. We say that (ϕ, β) is an equilibrium if ϕ is optimal given β, and β is robustly consistent with ϕ. Note that such an equilibrium always exists. For example, the ϕ that maps every history to "always defect" (formally, using

the notation from the previous section, for all $H_i^s \in \mathcal{H}_i^s, \phi(H_i^s) = 0 \times [1]^{L_i^s} \times [1]^{K+|\text{In}_i^s|})$, is a trivial equilibrium.

Theorem 1 demonstrates that there exists an equilibrium in which agents apply strategies of a particular form, as described informally in Section 3. We can now formally define this class of strategies.

- An agent is *unforgiving* if for every feasible history, ϕ_i sets $e_{ij}^s = 1$ whenever $\beta_j^s \in \text{Out}_i^s \cup \text{In}_i^s$, $\alpha_j^s = D$.
- An agent is *trusting* if for every feasible history ϕ_i sets $e_{ij}^s = 0$ whenever $\beta_j^s \in \text{Out}_i^s \cup \text{In}_i^s$, $\alpha_j^s = C$, and all proposed inlinks are accepted (i.e., REJECT $\notin \text{In}_i^s$).
- An agent is *consistent* if for every feasible history ϕ_i sets $\alpha_i^s = \alpha_i^{s-1}$ for all $s > 0$.

We can now discuss the equilibrium from Theorem 1 in more detail. This equilibrium (ϕ, β) has the following properties:

- The support of $\beta(\emptyset)$ (that is, an agent's beliefs before any observations are made) is contained in L_q, the subset of states consistent with agents applying strategy ϕ for an arbitrarily long sequence of rounds.
- For any history H in which the observed fraction of cooperation is q on every round, the support of $\beta(H)$ is contained in L_q (by consistency).
- For any history H in which the observed fraction of cooperation is not q on some round, belief $\beta(H)$ will be consistent with the appropriate fraction of the population erroneously changing their action from cooperation to defection on that round (by robust consistency).

Suppose that, in this equilibrium (ϕ, β), agent i observes a fraction of cooperation other than q on some round, say age s. Then, according to β, i believes that any other agents who were playing on round s also observed the change in q. Then, under strategy ϕ, all of these agents will defect on every subsequent round, leading to a state in which *every* agent defects every round (since the total fraction of cooperation must fall below q, which will be observed by all new agents). This rationalizes the decision for agent i to defect on every round after s.

References

[1] D. Abreu. On the theory of infinitely repeated games with discounting. *Econometrica*, 56(2):383–396, March 1988.

[2] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[3] O. Baetz. Cooperation in volatile networks. working paper, 2010.

[4] S. Datta. Building trust. Mimeo, Indian Statistical Institute, 1993.

[5] P. Diamond. Pairwise credit in search equilibrium. *Quarterly Journal of Economics*, 105:285–320, 1990.

[6] P.A. Diamond and E. Maskin. An equilibrium analysis of search and breach of contract, i: Steady states. *Bell Journal of Economics*, 10(1):282–316, Spring 1979.

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

[8] E. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.

[9] D. Fudenberg, D. Levin, and E. Maskin. The folk theorem with imperfect public information. *Econometrica*, 62(5):997–1039, September 1994.

[10] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, May 1986.

[11] P. Ghosh and D. Ray. Cooperation in community interaction without information flows. *The Review of Economic Studies*, 63(3):491–519, July 1996.

[12] M. Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59:63–80, 1992.

[13] R. Kranton. The formation of cooperative relationships. *Journal of Law, Economics, and Organization*, 12(1):214–233, April 1996.

[14] L.B. Lave. An empirical approach to the prisoners' dilemma game. *The Quarterly Journal of Economics*, 76(3):424–436, August 1962.

[15] M. Okuno-Fujiwara and A. Postlewaite. Social norms and random matching games. *Games and Economic Behavior*, 9:79–109, 1995.

[16] J.M. Pacheco, A. Traulse, H. Ohtsuki, and M.A. Nowak. Repeated games and direct reciprocity under active linking. *Journal of Theoretical Biology*, 250(4):723–731, February 2008.

[17] G. Ramey and J. Watson. Bilateral trade and opportunism in a matching market. *Contributions to Theoretical Economics*, 1(1), 2001. Article 3.

[18] A. Rapoport and A.M. Chammah. *Prisoner's Dilemma*. University of Michigan Press, 1965.

[19] C. Shapiro and J.E. Stiglitz. Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74(3):433–444, June 1984.

[20] J. Sobel. A theory of credibility. *Review of Economic Studies*, 52:557–573, 1985.

[21] J. Sobel. Can we trust social capital? *Journal of Economic Literature*, XL:139–154, March 2002.

[22] F. Vega-Redondo. Building up social capital in a changing world. *Journal of Economic Dynamics and Control*, 30(11):2305–2338, November 2006.

[23] J. Watson. Starting small and renegotiation. *Journal of Economic Theory*, 85:52–90, 1999.

[24] J. Watson. Starting small and commitment. *Games and Economic Behavior*, 38:176–199, 2002.