COOPERATION IN ANONYMOUS DYNAMIC SOCIAL NETWORKS*

Nicole Immorlica† Brendan Lucier‡ Brian W. Rogers§

March 7, 2014

Abstract

We study how cooperation may be sustained in anonymous, evolving networks. Individuals form relationships under a matching protocol and engage in prisoner's dilemma interactions with their partners. We characterize a class of equilibria that supports cooperation as a stationary outcome. When cooperation is possible, its level is uniquely determined. The key mechanism is that the endogenous network dynamics allow an individual to gradually accumulate a neighborhood of profitable partnerships through cooperation, whereas defection results in marginalization. Even as players become patient, equilibrium allows for full cooperation, only autarky, or the coexistence of cooperation and defection, depending on payoffs. This captures, in particular, the observation that many large systems are characterized by a high, though less than universal, level of cooperation. Smaller levels of cooperation can be sustained only through exclusivity among cooperating agents, showing the reliance of strategic behavior on norms surrounding relationship formation.

JEL codes: C73, D83, D85

*We thank Simon Board, Wiola Dziuda, Drew Fudenberg, Matthew Jackson, Christoph Kuzmics, David Miller, Paolo Pin, Marzena Rostek, Bill Sandholm, Joel Sobel, Takuo Sugaya, Joel Watson, Simon Weidenholzer and Alexander Wolitzky for helpful comments and conversations.

†Microsoft Research, nicimm@gmail.com

‡Microsoft Research, brlucier@microsoft.com

§Corresponding author. Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130-4899, 314-935-8924 (office), 314-935-4156 (fax), brogers@wustl.edu

1 Introduction

In the modern world, many social interactions and economic transactions are mediated through large evolving networks of agents. The size of these systems along with the transient nature of both individual membership and the relationships amongst them often confer a measure of anonymity to participants. In online contexts, there are technological aspects that reinforce anonymity, in that it is typically possible to create a new identity at any point in time and thereby mimic a new entrant. In these settings, the scope for punishing uncooperative behavior and, hence, for sustaining efficient outcomes, is potentially limited. Nevertheless, many systems that are largely anonymous and in which agents change partners over time are characterized by a high, though often less than universal, level of cooperation. We find that such an outcome can be explained as equilibrium behavior under a simple network formation model.

One important aspect of social and economic networks is that an agent's incentives within the context of a particular relationship are influenced by the outcomes from other relationships. This could in fact be taken as a defining characteristic of any interesting network, as in the absence of some form of interdependence, one is essentially engaged in a set of separable relationships and the network, per se, has no influence on behavior.

We consider a strong form of strategic dependence in which at any given time each agent takes the same action with each of his neighbors. There are two reasons for adopting this approach. First, there are applications in which technological constraints limit an agent's ability to behave differently with different partners, or in which the chosen action represents a general characteristic of an agent that is not relationship-specific. This is the case if, for example, an agent must invest in a quality for each period and use that quality for each transaction.[1] Second, if there is even a small amount of local information flow in the network, such that a defection with one partner would be observed by the agent's other partners, then there will exist equilibria with the behavior that we describe here even if there are no restrictions at all on the profile of actions an individual chooses at a given moment in time.[2]

[1] This is the approach taken by most papers studying games on networks. See, e.g., Galeotti et al. (2010).

[2] One could view this as an implementation of "local community enforcement" along network connections, in the spirit of Kandori (1992); Ellison (1994), which we discuss in more detail below.

We now describe the main elements of our framework. Agents enter the system over time and have finite lives. All strategic interactions are bilateral and described by a prisoners' dilemma. 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 stage games with each of its current partners. After every period, each agent has the opportunity to sever any of its relationships.

Consider, for example, an online community in which agents seek partners with whom to profitably interact, such as for trading goods or engaging in joint production. At any point in time agents can choose to conduct honest business (cooperate) or to cheat their partners for a gain (defect). The discretion to sever a relationship has an important impact on behavior. In the model it 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 punishment mechanism since there is no scope for future partners to punish a defecting agent.[3] In addition to severing a link to a defecting agent, one might wish to broadcast the agent's defection so as to enable further punishment. But the agent who defected does not have an identity that can be tracked by future partners, and so bears no negative consequence of his defection beyond the potential loss of his current relationships.[4]

Interestingly, Riedl and Ule (2002) provide empirical evidence based on laboratory experiments that the ability to endogenously determine one's partners increases observed cooperation levels in repeated prisoner's dilmmea interactions, while Blackburn et al. (2011) show that cheaters in an online game are punished by an increased rate of losing partners. These effects constitue crucial elements of the model we study.

There are two essential properties of the matching process that drive our results. The first is that it takes (valuable) time to search for partners with whom to form *profitable* relationships. In an environment characterized by anonymity and the ability to sever relationships, this is generally a necessary feature to provide any possibility of cooperative behavior. The

[3]Jackson and Watts (2010) define and study social games that changes equilibrium outcomes by permitting players to choose with whom they interact. This produces very different insights than our work, since here players are matched through a random process. Dall'Asta et al. (2011) identify collaborative equilibria as a function of the social network that describes interactions.

[4]Because the defecting agent also has the option of severing the relationship, one can not threaten contigent play that reduces the defecting agent's continuation payoff from the relationship below its outside option.

intuition is that, because defection can be punished only by the severance of relationships, in order for this punishment to bite, it must be that future relationships are sacrificed as a result of defecting.

In light of this observation, our mechanism for sustaining cooperation can be interpreted as a particular implementation of the notion of social capital or, more specifically, network capital, here taken to mean an agent's accumulated network of cooperative partners with whom he is connected.[5] Indeed, the notion that cooperative behavior is determined in large part by social pressures has been studied at length in the sociology literature, including, for example, Krackhardt (1996). In our setting, the reason to cooperate comes from the fact that, through cooperation, one can gradually build up a large social network consisting of other cooperating agents.

The second property is that there is a limit to the number of relationships that an agent can maintain. In combination with the first property, this dictates that the maginal returns to cooperating be decreasing in the aggregate level of cooperation in the system. This concavity is important because it allows for the possibility that cooperating and defecting agents will co-exist in equilibrium. The coexistence result obtains for a wide range of parameters and we view it as having descriptive value.

Moreover, the constraint on the number of relationships has strategic importance in that it forces an agent to trade off the continuation value of each relationship against the outside option of initiating a new relationship. The value of the outside option is dictated by the aggregate system dynamics, and so provides a link between overall behavior in society and the incentives that govern behavior in a particular relationship.

Because of this strategic effect, our work is related to a branch of literature that studies repeated *bilateral* interactions when relationships can be endogenously terminated. The central idea in this literature is that the threat of severance disciplines behavior because being re-matched entails a cost. The cost can come in many forms, such as being cast into a matching market with frictions, as in the pioneering work of Shapiro and Stiglitz (1984), having to start a new relationship that requires a specific investment Ramey and Watson

[5]See Sobel (2002) for an overview of the large literature regarding the concept of social capital. With a related motivation, but very different analysis, Vega-Redondo Vega-Redondo (2006) studies social capital in a stochastically evolving network.

(2001), or having to start with small stakes in a new relationship Watson (1999, 2002).[6] New relationships may also entail a phase of gradual trust building, as in Ghosh and Ray (1996) and Datta (1993), or the payment of a bond, as in Kranton (1996).[7]

It is crucial to all of this research that an agent is involved in at most one relationship at any given time. All of the equilibria that are studied by this literature use strategies with the property that actions evolve in a non-trivial way throughout the course of a given relationship. In a networked society in which an agent continues to form new relationships while maintaining older relationships, these constructions have no natural analogue.[8] We focus instead on a particularly simple and intuitive kind of behavior that we call *consistent*. A consistent agent chooses to either perpetually defect or to perpetually cooperate.[9] We exhibit equilibria in which all agents take consistent actions. In light of this behavior, optimal decisions regarding the formation and severance of relationships become easy to describe, which allows us to precisely pin down the nature of the co-evolution of the network and the behavior in repeated interactions occurring on the network. 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 met with the opportunity for future interactions.

Assuming first that agents use consistent strategies, we provide a characterization of simple stationary equilibria (SSE). The first message is that the payoffs of the prisoner's dilemma have an important impact on the sustainable level of cooperation. Under adverse conditions, no cooperation can be sustained, even as players become perfectly patient. Otherwise, when agents are patient enough, there is a unique stable SSE that supports cooperation, and it is such that either there is universal cooperation, or cooperators and defectors co-exist at a specific ratio.[10] In this sense we provide an explanation for which societies permit universal

[6]MacLeod and Malcomson (1998) and Board and Meyer-ter vehn (2011) contribute to the search literature in the context of relational contracts.

[7]Sobel (1985) provides an earlier analysis of a related game.

[8]In a very different framework, Jackson et al. (2011) also leverage the idea that cooperation can be enforced through the use of multiple relationships, and so make a related connection to social capital. Levin (2002) and Board (2010) investigate relational contracts between a principle and many agents engaged in a repeated game.

[9]In a model where agents have only one partner at a time FUJIWARA-GREVE and OKUNO-FUJIWARA (2009) examines strategy profiles that involve agents taking different strategies, similar to our roles of cooperation and defection, and partially characterize equilibria with trust building via an evolutionary approach. In a very different model, Fosco and Mengel (2011) support a mixture being cooperators and defectors by employing dynamics based on imitation.

[10]The notions of SSE and stablity are defined below. The autarky equilibrium always exists and is stable.

cooperation, and which societies are necessarily subject to a fringe of cheating behavior. The model allows for simple comparative statics on the equilibrium level of defection, providing a first step towards assessing which kinds of policies can most effectively improve the welfare of the system.

The way in which cooperators manage their relationships takes one of two forms. First, it may be that they always accept new relationships when the opportunity arises. This case obtains when the pool of agents searching for a match is sufficiently cooperative. Otherwise, they accept new relationships only with some probability $p < 1$, thereby imposing a barrier on society to obtaining connections with cooperators, resulting in a form of *exclusivity*. This exclusivity implies that some agents will fail to find partners when given the opportunity to match. It can thus be interpreted in terms of an endogenous friction operating on the matching process. This friction is necessary to sustain lower levels of cooperation, which is sometimes all that can be achieved. While exclusivity is costly to all agents, it is effective because, under the right circumstances, it decreases the expected returns to defection by more than it decreases the returns to cooperation.

Having characterized SSE under consistent behavior, our seecond main result shows that under an appropriate condition on parameters, SSE outcomes describe play along an equilibrium path without imposing consistency on action choices. The condition, which requires that the temptation payoff to defect be small enough relative to the loss from being defected on, ensures that perpetual cooperation is sequentially rational at any possible history. This demonstrates that consistent behavior can be thought of as self-enforcing when paired with the social norm of severing a link when, and only when, the partner defects. We demonstrate that the required condition is relatively mild.

Finally, we examine behavior when this condition is not met. The only kind of profitable deviation from the consistent strategy profile involves a cooperator defecting under a particular circumstance: when he has very little social capital, and therefore little to sacrifice in terms of future links with cooperators. We show that as the network becomes dense, these events becomes increasingly rare and, as a result, these strategies form an epsilon equilibrium that generates essentially the same level of cooperation that we characterize under consistent strategies. In this sense, SSE remain a valid description of outcomes for the model.

When there is a positive stable SSE, there also exist either one or three unstable stationary equilibria.

The remainder of the paper is organized as follows. The model is described in Section 2. Section 3 characterizes stationary equilibrium outcomes under consistent strategies. Section 4 shows that consistency is a self-enforcing norm under the appropriate condition, while Section 5 describes outcomes that involve a minimal failure of consistency. We conclude and provide comments for further research in Section 6. An Appendix contains a formal development of the model and proofs.

2 A model of strategic interactions in a social network

For ease of exposition, we describe the main elements of the model in this section, relegating the full development of some technical aspects to Appendix A.

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 $N = [0, 1]$.[11] Agents interact repeatedly on an evolving directed network. Time is discrete. At each date each agent independently dies with a given probability $1 - \delta$, in which case it is replaced by a new agent.[12] We speak of the age of an agent i, $t(i)$, as the number of dates since its birth, so that $t(i) = 0$ in the periods when i is born. Each agent i chooses an action $\alpha_i^{t(i)} \in \{C, D\}$ at each date $t(i)$ of its life. Agents commonly observe the aggregate proportion, q, of C behavior in the population after each date, and the aggregate proportion p of proposed inlinks that were accepted by cooperators.[13]

[11] We work with a continuum of agents so that players are "atomless", i.e., no individual can unilaterally affect aggregate behavior. This is a good approximation for large societies, as one then expects a player to ignore the marginal effect of her behavior on aggregate play.

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

[13] Common knowledge of these summary statistics is useful for our notion of equilibrium stability, in which agents respond to an exogenous shock to the state, as discussed in Section 3.3. In particular, this structure is not necessary to support equilibria.

Every agent is able to sponsor 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, is able to re-match with another agent, chosen uniformly at random, at the next date.

Notice that there is an implicit bound on the expected number of inlinks an agent will receive over the course of her life, due to the fact that every inlink corresponds to the outlink of some other agent. The fact that we explicitly bound the number of outlinks, while leaving the bound on inlinks implicit does not drive the results. Similar results would obtain if one instead bounds the total number of (in- and out-) links an agent is able to maintin.

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. (p, q) from the previous date is publicly observed
3. Actions are privately chosen.
4. Outlinks are proposed to other agents.
5. Potential inlinks are accepted or rejected.
6. The stage game is played and payoffs are realized.
7. Agents sever any links that they choose to.
8. Death occurs.

3 Consistent Behavior and stationary outcomes

An agent's strategy specifies at each date, as a function of everything the agent has observed, whether to cooperate or defect, how many links to propose, which proposed inlinks to accept, and which existing links to sever. A formal development is contained in Appendix A. Throughout, we maintain two assumptions on strategies. First, at the individual level, we assume a weak form of stationarity, in that agents do not condition their plan of action on a common labeling of time. In other words, while each agent is aware of his age and the history he has observed, he does not use any universal description of time. Second, at the collective level, we assume that strategy profiles are symmetric.[14] However, as we will see, agents will indeed take heterogeneous roles along a given path of play.

Even with these two assumptions in place, the analysis involves many agents, birth and death, an endogenously evolving network of relationships, and histories that are, to large extent, privately observed. In order to gain traction on studying equiibrium outcomes, we begin the analysis by considering a setting in which agents adopt a simple behavioral rule. Namely, agents are assumed to be consistent in their choice of action, cooperation or defection, over the course of their lives.

PROPERTY **1** *Consistency: An action from* $\{C, D\}$ *is chosen at birth (possibly mixing). At all future dates, the agent takes same the action it played at the previous date. That is, a strategy for* i *is consistent if for every* $t(i) \geq 0$, $\alpha_i^{t(i)+1} = \alpha_i^{t(i)}$.

Consistency allows us to speak of society as being comprised of "cooperators" and "defectors". This behavior is plausible as well as simple. Moreover, cooperation will not require the use of elaborate punishment phases in equilibrium construction.

Certainly, though, consistency limits the complexity of strategic interactions in an important way. In particular, it prohibits strategies that allow an agent to cooperate until a history with a certain property is reached, and then defect. Even though consistency rules out many of the standard constructions that permit cooperation, it turns out to be a significantly milder condition than it may first appear. We formalize this assertion in Section

[14]Notice that we thus are not concerned with measurability assumptions on strategy profiles with a continuum of agents.

4, but the intuition is as follows. If cooperation at a given round is part of an optimal strategy, then it is because the increased access to it provides to relationships tomorrow is enough to forgo the temptation payoff. But if that is true today, then tomorrow the same comparison is likely to hold true again. To the extent that connection is valid, optimality of consistent cooperation is not a much stronger requirement than the optimality of the initial act of cooperation.

We will first demonstrate that interesting aggregate outcomes obtain under consistent strategies. Second, and more importantly, the analysis of consistent strategies is a preliminary step to studying equilibrium outcomes more generally. We will argue below that consistent strategies are, to large extent, self-enforcing, and describe much about equilibrium behavior in our framework even when inconsistent behavior is permitted.

Under consistent strategies, it is straightforward to describe the optimal management of links. First, if a defection is ever observed, the best response of the defector's partner is to sever the link. We refer to this as *unforgiving* behavior. Second, when cooperation is observed, the best response is to maintain the link (provided both agents survive), which we refer to as *trusting* behavior. Third, it remains to be shown under which conditions a cooperator should accept a proposed inlink.[15] This will be determined by the probability that a proposed link comes from a cooperator. Notice that this probability is less than q in equilibrium since, as just argued, defectors lose their links at every period, and have a dominant strategy to propose all possible links, while cooperators generally maintain some links from previous periods, and so search less. Because of anonymity, all proposed inlinks are ex ante equivalent from the perspective of the agent receiving the proposal. It is therefore essentially without loss to assume that a cooperator accepts each proposed inlink independently with a certain probability p. Finally, for any choice of p and q, we must confirm that it is indeed rational for cooperators to send outlinks.

In light of these observations, behavior can be completely described by a function $\phi : [0,1]^2 \to [0,1]^2$ with the interpretation that $\phi(p,q)$ specifies the probability that an agent chooses C at its birth and the probability with which each inlink is accepted in the event the agent becomes a cooperator, upon observing the state (p,q). We refer to such strategies as *simple*.

[15] Defectors, of course, have a dominant strategy to accept every inlink

3.1 Simple stationary equilibria

We are interested in determining when a particular level of cooperation q can be sustained as a stationary outcome under consistent strategies. Note, however, that a given pair (p, q) does not, by itself, capture all of the payoff-relevant aspects of the system, even in expectation. Other factors, such as the amount of search by cooperators and defectors, which depends on the age distribution of those behaviors, impact expected payoffs. This motivates us to describe the state of the system under stationary behavior.

DEFINITION **1** *The* steady-state at (p, q), $L_{(p,q)}$, *is the limiting distribution over graphs that obtains when all agents have been applying the simple strategy described by* (p, q) *for* t *periods, as* $t \to \infty$.

From the steady-state we can extract all payoff-relevant quantities for expected utilities, such as the age distribution of cooperators and defectors, along with the expected amount of search agents of either type are conducting at each age.

For a steady-state to be supported as an equilibrium outcome, it is necessary and sufficient that the strategy $\phi(p, q) = (p, q)$ be optimal when the system is in state $L_{(p,q)}$. In this case, the fact that all agents apply ϕ implies that the system remains in steady-state $L_{(p,q)}$. This is useful for the analysis, since it is only under the steady-state assumption that we are able to derive expressions for expected utilities. We can now define equilibrium under consistent strategies.

DEFINITION **2** *A pair* (p, q) *is a* simple stationary equilibrium (SSE) *if, given that the system is in* $L_{(p,q)}$ *at all times, the application of a strategy that chooses cooperation with probability* q *at birth, and where cooperators accept each inlink independently with probability* p, *is optimal in the space of simple strategies.*

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, and the rate of

inlink acceptance, p. Since we are interested in simple stationary equilibria, we work under the assumption that the system is in state $L_{(p,q)}$ and remains so over the agent's lifetime.

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.[16] 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

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

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, p of whom accept the link. Setting $n_{CC}^{Out}(-1) = 0$ and solving produces

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

The number of links from a cooperator of age s to defectors can then be computed according to $n_{CD}^{Out}(s) = (1-q)(K - \delta n_{CC}^{Out}(s-1))$, since each link that is proposed at age s matches with a defector with probability $1-q$. For defectors the case is much simpler, as the property of unforgiving behavior implies that age dependency is trivial. We have $n_{DC}^{Out}(s) = pqK$ and $n_{DD}^{Out}(s) = (1-q)K$.

We turn now to the expected number of inlinks from both types of nodes as a function of age. To do so, we first compute the number of inlinks an agent expects to receive from agents of either behavior at each date. These are time-independent rates in steady-state, and from them the evolution of inlinks is easy to derive. The probability that a randomly selected node is age s is $f(s) = (1-\delta)\delta^s$. In steady-state, $f(s)$ also defines the age distribution of cooperators and the age distribution of defectors. Then, the expected number of inlinks an agent will receive from cooperators and defectors at each date are, respectively,

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

[16] See, e.g., Boylan (1992) and Alós-Ferrer (1999) for foundations of matching processes justifying laws of large numbers for continuous populations.

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.

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 and consistently defecting. To that end we compute the expected payoff at a particular age s by summing the payoffs over the expected sets 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

$$u_C = \frac{2pq - b(1-q)(1+p-\delta^2(1+p(1-pq)))}{1-\delta^2(1-pq)}, \tag{1}$$

$$u_D = (1+a)\left(pq + q\frac{1-\delta^2}{1-\delta^2(1-pq)}\right). \tag{2}$$

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 discount factor. 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 her expected lifetime utility. In order to characterize optimal choices under consistent strategies, we are interested in comparing u_C and u_D as a function of q and p under various parameterizations of the model. It is convenient to define $V(q, p; a, b, \delta) = u_C - u_D$.

For any given steady-state $L_{(p,q)}$, the expected value to a cooperator of a relationship from a new inlink is proportional to

$$v = \frac{r_C}{1 - \delta^2} - br_D.$$

Using the expressions above, we conclude that v is non-negative if and only if

$$b \leq \left(\frac{q}{1-q}\right)\left(\frac{1}{1-\delta^2(1-pq)}\right). \tag{3}$$

The following observations describe the various possibilities for SSE. First, notice that for any choice of parameters, there will exist an SSE with $q = 0$. This is true because when all agents in the system defect, defection is strictly optimal, i.e., $V(0, p; a, b, \delta) < 0$. For some parameters, this will in fact be the unique SSE, in which case it is not possible to sustain any level of cooperation.

Next, notice that at any SSE (p, q) where $q > 0$ (there exist cooperators) it must be that $p > 0$ (cooperators accept some inlinks). If the SSE is such that $p < 1$, then equation (3) must hold with equality, leaving cooperators indifferent to the acceptance of inlinks.

If the SSE is such that $q = 1$ (universal cooperation), then it must be that $p = 1$ (accepting inlinks is dominant). If the SSE involves an interior solution for q, then it must be that $V = 0$, as entering agents must be indifferent between cooperation and defection. In this case, the SSE might or might not involve an interior solution for p; this depends on the parameters.

We note that at any SSE with $q > 0$, it must be that the expected value of an outlink is strictly positive for cooperators, and so the specification that agents always search with available outlinks does not bind. To see this, consider two cases. First, when $p = 1$, the probability of a given outlink matching with a cooperator is q, which strictly exceeds the probability that a given inlink comes from a cooperator, since cooperators search less than

defectors in equilibrium. Thus the fact that an inlink has non-negative value at an SSE implies that an outlink has positive value. Second, when $p < 1$, it is generally possible that outlinks have lower value than inlinks in steady state, but not at an SSE. Consider an SSE with $0 < p < 1$, so that the value of an inlink is exactly zero. Since also $0 < q < 1$, we have that $u_C = u_D > 0$. As u_C consists of the separable values of inlinks and outlinks, the former of which is zero, it must therefore be that outlinks have positive value.

We are particularly interested in those SSE that are *stable*.

DEFINITION **3** *A simple stationary equilibrium* (p, q) *is stable if there exists* $\epsilon > 0$ *such that*

(i) if $q < 1$*, then for all* $q' \in (q, q + \epsilon)$*,* $u_C(p, q') < u_D(p, q')$*, and*

(ii) if $q > 0$*, then for all* $q' \in (q - \epsilon, q)$*,* $u_C(p, q') > u_D(p, q')$*.*

Our main result of this section establishes uniqueness of nontrivial stable SSE.[17] The autarky outcome ($q = p = 0$) constitues a stable SSE for all parameters.

THEOREM **1** *There exists at most one stable SSE with* $q > 0$*.*

Proof. See Appendix. □

The proof is involved but expresses a graphical argument that we summarize here. The set of points $\Gamma = \{(q, p) \in [0, 1]^2 | V \geq 0\}$ is a connected (but not convex) region which we call Γ. The set of (q, p) at which $v = 0$ is defined by a strictly increasing function $p_t(q)$ that is negative for q near zero and greater than one for q near one. Define thus its restriction to the unit square by $t(q) = \min\{\max\{p_t(q), 0\}, 1\}$. See Figure 1 for illustrations. SSE occur (i) at the intersection of $t(q)$ with the boundary of Γ and (ii) at $(q, p) = (1, 1)$ when $(1, 1) \in \Gamma$. When $(1, 1)$ is an SSE it is generically stable, but otherwise stability involves the extra requirement that the intersection is on the right boundary of Γ and not the left boundary. The proof proceeds by limiting the number and type of possible intersections between Γ and t through explicit consideration of the utility functions.

[17] The stability notion is intentionally a weak requirement to emphasize that all other SSE fail even this basic requirement. Stronger notions of stability either agree with our notion or fail existence. Every SSE that is stable under our definition satisfies as well the analogous requirement for p.

We view stability as an important refinement in our setting. SSE that fail the stability requirement are unsatisfactory solutions, and in this sense the model makes a unique prediction (modulo the autarky outcome). Nevertheless the structure of the set of all SSE is informative. In the course of proving Theorem 1 we also prove the following result.

Proposition 1 *The following statements are true:*

- *All SSE are ordered, in the sense that if (p, q) and (p', q') are two SSE with $q' > q$, then $p' \geq p$, with strict inequality when $p < 1$.*
- *In addition to the autarky outcome, there are generically either 0, 2, or 4 SSE.*
- *If there are 4 nontrivial SSE, the largest one is stable and involves $p = 1$.*
- *If there are 2 nontrivial SSE, the smaller one is unstable.*

Proof. See Appendix. □

We outline here the main arguments relevant to Proposition 1. One can show explicitly that the boundary of Γ intersects any given horizontal line of the form $p = \kappa$ at most twice. This means that, in particular, there can be at most two SSE with $p = 1$. In this case, it is clear that the larger of the two SSE is stable, and the smaller is unstable. Through a separate argument using a change of variables we show that $p_t(q)$ can intersect the boundary of Γ at most twice, which in turn limits the number of interior equilibria to at most two. The most difficult part of the proof involves showing that the point in Γ with the smallest value of p lies below $p_t(q)$, which can then be used to show that all but the largest SSE are necessarily unstable.

Note finally that in the case of two interior SSE, the larger one may be stable or unstable, and it may involve $p < 1$ or $p = 1$; if it involves $p = 1$, then it must be stable. Importantly, in all cases when a stable SSE exists, it is the maximal and, therefore best, among all SSE.

The model allows for an explicit determination of the set of SSE for any parameters (a, b, δ). It is therefore possible in principle to partition the parameter space into regions that map into the different configurations of SSE described by Proposition 1. However, such an analysis is cumbersome. Instead, we present results for the limiting case in which players

becoming perfectly patient (i.e., long lived), which captures much, but not all, of the richness of the model. We then exhibit some examples with intermediate values of δ.

Proposition 2 *The following statements hold in the limit as δ approaches one.*
If $a < 1$ then $p = q = 1$ is a stable SSE.
If $a > 1$, then

- *all SSE have $q < 1$*
- *There exists a stable SSE if and only if $b < 1 + a$.*
- *If $(1 + a)/a < b < 1 + a$ the stable SSE has $p < 1$.*
- *If $b < (1 + a)/a$ the stable SSE has $p = 1$.*

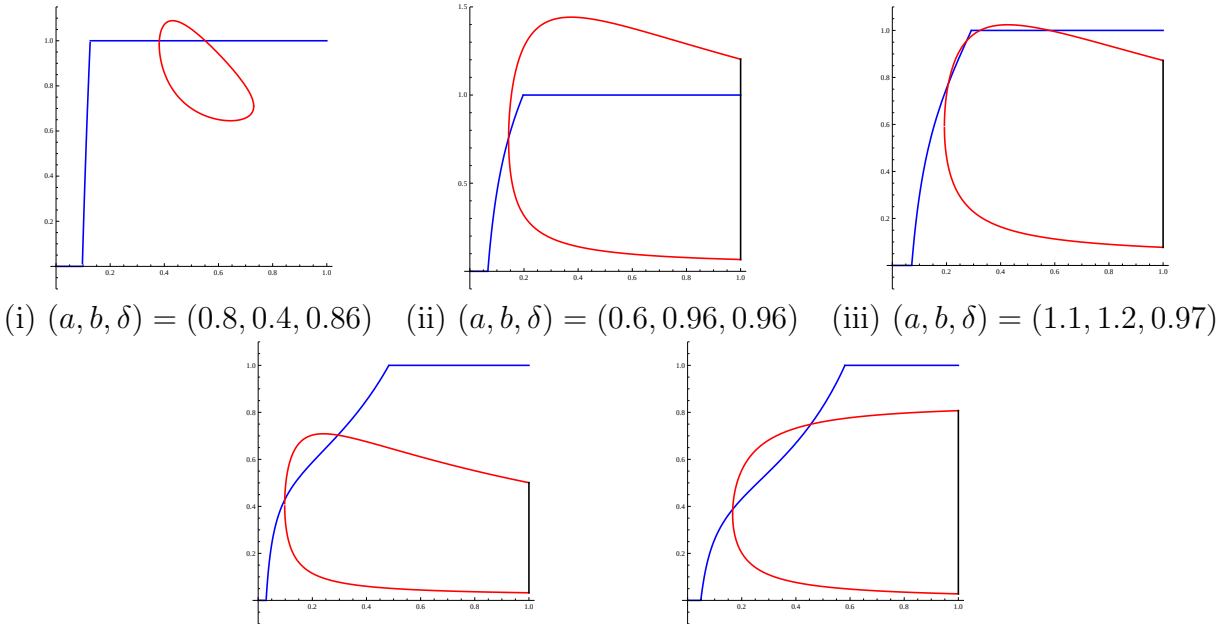
Proof. See Appendix. □

Proposition 2 formalizes a number of our previous observations. Namely, there is a rich set of qualitatively distinguished outcomes, even in the limit as players become completely patient. Indeed, full cooperation can be achieved for perfectly patient players only if the temptation payoff is $a < 1$.[18] The intuition for this bound comes from considering a world with full cooperation, $q = 1$, and a high discount factor. In this case, a defector earns $1 + a$ per outlink per period. A cooperator earns 1 from each outlink per period, and also expects to build an asymptotically complete neighborhood well before she dies, and in this case has as many inlinks as outlinks, for a total of 2 utils per outlink per period. Thus if $a > 1$, full cooperation cannot be sustained.

When full cooperation cannot be sustained, it is still possible that partial cooperation can be sustained. In this case, i.e., when $b < 1 + a$, society consists of cooperators and defectors co-existing in a specific ratio. To understand this bound, recall that q affects a cooperator's utility through two channels: (i) it determines how quickly, relative to the expected lifteme, a cooperator builds up her social network, and (ii) it determines how frequently a cooperator interacts with defectors. The first channel generally involves a non-linear interaction between

[18] This result stands in contrast to much of the work on repeated prisoner's dilemma with random matching, in which the goal is almost always to construct equilibria that always support full cooperation for patient players.

q and δ, but as $\delta \to 1$, a cooperator reaches its limit network very quickly, and so q has an effect only through the second channel. Specifically, an increase in q increases a cooperator's utility at rate pb. Defectors, on the other hand, benefit from an increase in q at rate $p(1+a)$. Stability of SSE requires precisely that a defector's utility is more sensitive to the level of cooperation than a cooperator, generating the threshold $b = 1 + a$.



(i) $(a, b, \delta) = (0.8, 0.4, 0.86)$ (ii) $(a, b, \delta) = (0.6, 0.96, 0.96)$ (iii) $(a, b, \delta) = (1.1, 1.2, 0.97)$

(iv) $(a, b, \delta) = (2.7, 1.9, 0.99)$ (v) $(a, b, \delta) = (1.39, 2.35, 0.99)$

Figure 1: Depiction of SSE possibilities in (q, p) space. (i) Two SSE with $p = 1$ and $q < 1$; (ii) One interior SSE and one with full cooperation; (iii) Four SSE, two of which involve exclusivity; (iv) and (v) Two interior SSE. In all but (v), the largest SSE is stable.

The possibility of co-existence of cooperation and defection allows for a novel description of some real-world communities such as, perhaps, eBay, in which most transactions are conducted in good faith but where one expects a fringe of cheaters. In our analysis, such an outcome obtains when the temptation payoff is high; patience is not necessarily enough to overcome this effect. If b is small enough that there is a stable SSE where all inlinks are accepted ($p = 1$), then the stable level of cooperation is easily determined to be $\frac{2-b}{1+a-b}$.[19]

Finally, when $b > (1+a)/a$ it is the case that the stable SSE requires exclusivity ($p < 1$). To understand this bound, recall that at an SSE with $p \in (0, 1)$, a cooperator gains zero value

[19]It is worth noting that because of the constraint $a - b < 1$, this possibility obtains only when $a < 1 + \sqrt{2}$.

from inlinks, and so total utility is captured by outlinks, which is equal to one for $\delta \to 1$. Defectors obtain $(1+a)pq$. When cooperators accept all inlinks ($p = 1$), this means that defection is preferred to cooperation for all $q > 1/(1+a)$, so that any SSE with $p = 1$ must involve a lower level of cooperation. On the other hand, the value of inlinks is increasing in q, such that sufficiently low levels of cooperation are inconsistent with SSE. For $\delta \to 1$, the expected value from inlinks is $1 - bp(1-q)$. For this to be non-negative when cooperators accept all inlinks requires $q \geq 1 - 1/b$. Thus, an SSE with $p = 1$ is possible only when $1 - 1/b < 1/(1+a)$, as desired.

We find exclusivity to be of particular interest, since the presence of cooperators is made possible only when those cooperators limit their exposure to society. Clearly this is costly, since some relationships are not materialized. The idea behind exclusivity is the following. If cooperators accepted all proposed links, defection would be relatively attractive, and so to balance the incentives between cooperating and defecting, the level of cooperation would have to be relatively low. But at low levels of cooperation, the expected value of an inlink to a cooperator is negative, and so cooperators would prefer to reject proposed links, leading to a breakdown of the network. The only equilibria therefore involve the rejection of some proposed links. While this reduces the utility to all players, it can reduce the relative incentive to defect, such that at strong enough levels of exclusivity, a positive level of cooperation becomes consistent with players' incentives.

We emphasize that exclusivity is important for sustaining low levels of cooperation. The intuition is that when cooperators are relatively rare, to incentivize cooperation in such a world one must penalize defection, and the rejection of inlinks by cooperators serves exactly this purpose. For the special case of $\delta \to 1$, the maximum level of cooperation at an SSE with exclusvity is one half. That is, exclusivity is important to sustaining cooperation only when cooperators constitute a minority of society.[20]

Figure 1 depicts a representative set of examples for SSE configurations. In each case, the red curve represents steady states (p, q) for which $u_C = u_D$, with $u_C > u_D$ inside the region, while the blue curve represents steady states for which the value of inlinks is zero,

[20]To see this, it is easy to solve for the stable interior SSE to obtain $q = b/(1+a+b)$. The largest value of q obtains when a takes its smallest value of $b-1$, as described above. Putting this together, we have that $q \leq 1/2$.

constrained to $p \in [0,1]$, with steady states below the blue curve having strictly positive value for inlinks. Thus, SSE occur at the intersections of the two regions, and stable SSE require further that the SSE involves the right hand portion of the red boundary. We emphasize that, by Proposition 2, all of these cases obtain for non-trivial sets of (a,b) pairs, even as $\delta \to 1$.

4 Consistency as self-enforcing behavior

Simple stationary equilibria are defined in a setting that requires agents to apply strategies that are consistent, with the immediate implication that agents are trusting and unforgiving. This can be thought of as an equilibrium that arises under a very natural social norm, specifying how to behave in one's relationships as well as how to manage these relationships, in the spirit of Ghosh and Ray (1996).

In this section we dispense with the presumption of consistency and study optimal behavior in the absence of social norms that restrict strategies. We find that, under an appropriate parametric condition, the behavior described above is self-enforcing, in the sense that it constitutes play on an equilibrium path.

Before stating the result, we present the condition that is utilized, which we call the consistency condition. It involves the strategy profile (p,q) as well as the parameters (a,b,δ). The consistency condition should thus be interpreted as a requirement of a particular steady-state (p,q) under consideration.

DEFINITION 4 *The* consistency condition *is*

$$\frac{(1+b)(1-pq)-bq(1-p)}{(1+a)(1-pq)} \geq 1-\delta^2(1-pq). \tag{4}$$

The result we provide is that every SSE (p,q) that satisfies the consistency condition arises as an equilibrium outcome. On the equilibrium path, agents behave in a way that conforms with the norms of being consistent, trusting, and unforgiving. Each agent applies a strategy that, at every history, maximizes his expected continuation utility given his beliefs about the state of the system. In turn, these beliefs are consistent with the strategy being employed (recall our focus on symmetric equilibria) and the agent's history. A formal development of the solution concept is provided in the Appendix A.

THEOREM **2** *Suppose that* $(p, q) \in [0, 1]^2$ *is an SSE at which the consistency condition is satisfied. Then there is an equilibrium such that if the system is in state* $L_{(p,q)}$*, all agents apply actions that are consistent, trusting, and unforgiving on the equilibrium path. Moreover, under this strategy* q *is a stationary level of cooperation and* p *is a stationary level of inlink acceptance by cooperating agents.*

Proof. Section 4.3 is dedicated to the proof. □

4.1 Maintenance of relationships

If agents apply consistent strategies, the beliefs of an individual regarding the future play of his partners are easy to describe. 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. A link between two defectors must be severed because the sponsor of that link strictly prefers to re-match and obtain probability pq of interacting with a cooperator at the next period. Importantly, this behavior is sequentially rational and holds for off-path play in which an agent's partner behaves inconsistently, since consistency is defined in terms of taking the same behavior as was taken in the previous period. Under the definition of equilibrium that is detailed in Appendix A, equilibrium beliefs require an agent to assign probability one to consistent behavior of his partners even after observing an off-path inconsistent choice, through use of a standard perfection requirement. It is therefore the case that every best response to a consistent strategy 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 governing their choices of cooperation and defection. Optimality, then, requires taking 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 condition is satisfied,

then consistent behavior is (part of) a best response. This requires showing that there is no history at which an agent can profitably deviate through the use of an inconsistent action.

First notice that for a defector in steady-state, the calculation is identical at every round. This is so because, under unforgiving strategies, he loses all of his connections at every period. Thus, if the continuation value of perpetual defection exceeds that of perpetual cooperation at some period, the same conclusion is true at subsequent periods.

For a cooperator the situation is complicated by the fact that the number of relationships with other cooperators changes over time. At an SSE with $q > 0$, 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, with positive probability there may arise histories at which a cooperator prefers to deviate by defecting (after which its optimization problem is identical again to the one at 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 (i.e., those links that are maintained from the previous period).

The next result provides the key implication of the consistency condition that we will use below to guarantee that cooperators never have a profitable deviation.

LEMMA **1** *Suppose $(p, q) \in (0, 1]^2$ is a simple stationary equilibrium and the consistency condition holds at (p, 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 (p, q)-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 accepts inlinks with probability p, and let ϕ_D denote the simple strategy in which agent i defects each round (and accepts all inlinks). 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, accepts all inlinks on the first round, and accepts inlinks with probability p on all subsequent rounds). 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 defined by (p, q). 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 (p, q) is a simple stationary equilibrium with $q > 0$, 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 now 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. Note that the 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, $\Delta u(\phi_F, 0, k_O) = (1 + a)(1 - pq)k_O$. To see this, note that the increase in the number of out-links to cooperators is $k_O + (1 - k_O)pq - pq = (1 - pq)k_O$, and this gain is realized for exactly one period.

For cooperators, $\Delta u(\phi_C, 0, k_O) = \frac{(1-pq+(1-q)b)k_O}{1-(1-pq)\delta^2}$. To see this, consider the expected loss experienced by a cooperator who does not have a link pre-formed to a cooperator. If he is unable to form a new link to a cooperator, then he suffers a loss of 1 (relative to a cooperator

who would derive a utility of 1 from a pre-existing link to another cooperator). Moreover, if he forms a link to a defector instead, he loses an additional b due to the interaction with the defector. The total loss is therefore $(1 - pq) + (1 - q)b$, per link. Finally, for a given outlink this gain is maintained as long as the node survives (probability δ), its cooperating partner survives (probability δ), and the outlink of the node in the scenario without the initial k_O cooperate outlinks is not to a cooperator (probability $(1 - pq)$). These events happen independently and hence have a total probability of $\delta^2(1 - pq)$ 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. □

By virtue of Lemma 1, under the consistency condition, as an agent accumulates relationships with cooperators, the marginal gain from those relationships is maximized by long-term cooperation, and not by defecting. Thus, if it is optimal to cooperate at birth it is necessarily optimal to cooperate at any future point in its lifetime.

4.3 SSE as equilibrium outcomes: Proof of Theorem 2

We shall construct a symmetric equilibrium with the required properties. Recall that a formal definition of equilibrium appears in Appendix A but, informally, what we require is a strategy ϕ^* and a system of beliefs β^* about the state of the network, such that ϕ^* maximizes expected utility at all histories 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, or an aggregate inlink acceptance other than p, the agent accepts all proposed links, defects that round, and breaks all links with observed defectors at the end of the round. Note that this behavior is optimal given that (p, 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 agent observes (p, q), the agent takes consistent, trusting, and unforgiving actions defined by (p, q). At birth, upon observing the state (p, q), he cooperates with probability q, and in that case accepts inlinks independently with probability p; otherwise he chooses to defect (and accept all inlinks).

The associated belief system β^* is straightforward. At birth, the agent believes that the

system begins in state $L_{(p,q)}$. The agent continues to believe that the system is in steady-state $L_{(p,q)}$ as long as he observes (p, q) at the end of each period. If a fraction of cooperation other than q is observed, or an aggregate inlinnk acceptance other than p, the agent believes that every other agent will defect on subsequent rounds, since (p, q) is public. Even though we have not provided a full description of an agent's belief about the state of the network, the properties discussed are sufficient to determine whether or not ϕ^* is an optimal strategy.

Under ϕ^* agents are consistent, unforgiving, and trusting provided the system remains in state $L_{(p,q)}$, so that (p, q) is indeed stationary under ϕ^*. It remains to show that applying strategy ϕ^* is optimal given the observation of (p, q) and the belief that other agents play according to ϕ^*. Note first that, under the belief that the system begins in $L_{(p,q)}$ and other agents use ϕ^*, it is rational to believe that the system remains in $L_{(p,q)}$ as long as agents observe (p, q). It is therefore sufficient to demonstrate that ϕ^* is optimal under the belief that the state of the system is described by $L_{(p,q)}$ at all times.

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 $L_{(p,q)}$, and suppose for a 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 accepts each incoming link independently with probability p, and let ϕ_D denote the similar strategy in which the agent chooses defection and accepts each incoming link.

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. Moreover, every optimal strategy accepts all inlinks on a round in which it prescribes defection so, 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. 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. Thus ϕ_C^r and ϕ_D^r differ only on their actions on round T_r, in which ϕ_C^r specifies cooperation (and accepting inlinks with probability p) and ϕ_D^r specifies defection (and accepting all inlinks). For notational convenience we 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. Therefore, Lemma 1 directly 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 number of maintained links with cooperators 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 through round T_{r-1}, after which both strategies prescribe cooperation on each turn, but ϕ_C^r does not necessarily accept every proposed inlink independently with probability p from round $T_{r-1}+1$ to round T_r. However, by virtue of the assumption that (p, q) is an SSE, it must be optimal to accept proposed links independently with probability p 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.

5 Inconsistent behavior

We have argued that when the consistency condition is satisfied, SSE outcomes are supported as on path behavior of an equilibrium without any presumption of consistency (Theorem 2). We now take up the task of analyzing behavior when the consistency condition fails.

5.1 Bank robbing

The first result demonstrates how a cooperator can profitably deviate at an SSE. To understand this behavior we first emphasize that the main mechanism at work in the model is the formation of social capital: valuable relationships with cooperating agents that take time to accumulate. In this regard, outlinks and inlinks are distinguished by the fact that inlinks,

if severed, are entirely unreplaceable. If there is a temptation to defect, the temptation is driven more by exploiting one's out-neighbors, since those links can be recast at the next period.

There are several possibilities for how incentives relate to the combination of inlinks and outlinks an agent has. Consider how an agent would behave in isolation with a given partner. First, it may be that an agent prefers to defect on an in-neighbor. If that is true he prefers to defect on all of his (in- and out-) neighbors and no SSE can support any level of cooperation. Second, it could be that an agent prefers to cooperate with an out-neighbor. In that case he would in fact prefer to cooperate with all of his (in- and out-) neighbors and consistent cooperation is sequentially rational. This is exactly the situation identified by the consistency condition.

The final possibility is that there is a tension between behavior with in- and out-neighbors, such that an agent would prefer to defect on an out-neighbor but cooperate with an in-neighbor. In this situation, optimal behavior is dictated by the relative number of inlinks and outlinks with cooperators. We formalize this intuition as follows.

Proposition 3 *Consider an SSE (p, q) with $q > 0$. A cooperator has a profitable inconsistent deviation if and only if*

$$\kappa_O \left[(1+a)(1-pq) - \frac{1-pq+(1-q)b}{1-\delta^2(1-pq)} \right] > \kappa_I \left[\frac{1}{1-\delta^2} - (1+a) \right]. \tag{5}$$

Proof. See Appendix C. □

Recall that Theorem 2 shows that an SSE at which the consistency condition is satisfied supports consistent cooperation in equilibrium. Proposition 3 demonstrates that the consistency condition is tight, in the sense that at an SSE where the consistency condition fails, with positive probability a cooperator reaches a history at which he has a profitable deviation.[21]

Proposition 3 suggests that the optimal play in an SSE at which the consistency condition is violated has a "bank robbing" flavor, in the sense that one begins life by cooperating and attempting to accumulate relationships, and continues to cooperate until such a time that

[21] To see this, note that with positive probability $\kappa_I = 0$ and $\kappa_O > 0$, in which case the failure of the consistency condition means that the left hand side of (3) is strictly positive.

there is little social capital to lose by defecting, and effectively starting the process again. In fact, this kind of strategy is used to construct an ϵ-equilibrium when the consistency condition fails in Proposition 5.

5.2 Failures of the consistency condition at SSE outcomes

Since every (a, b, δ) can be mapped into a unique stable SSE (p, q) (or to the trivial equilibrium), direct application of the consistency condition will verify whether or not any given set of parameters (a, b, δ) permits consistency of cooperation at its SSE. But the condition is complex enough that it is not transparent to immediately assess how stringent the condition is. The following result quantifies our assertion that the optimality of consistent cooperation is not a dramatically more demanding requirement then the optimality of any cooperation at all. In this sense, even though the consistency condition is tight, it is a relatively weak requirement.

Proposition 4 *The following statements are each (individually) sufficient for the consistency condition to hold.*

- (p, q) *is an interior SSE for* (a, b, δ)*, i.e.,* $(p, q) \in (0, 1)^2$
- $p = q = 1$
- $p = 1$ *and* $b \geq a$

Proof. See Appendix C. □

Notice in particular that if the stage game is supermodular ($b \geq a$) then at least in the limiting case as players become perfectly patient, all stable SSE satisfy the consistency condition. In particular, this verifies that under supermodularity, all of the stable SSE described by Proposition 2 survive as equilibria in which consistent, trusting, and unforgiving behavior is self-enforcing.

Our last result is relevant for SSE at which the consistency condition fails. By Proposition 4, this is possible only when $q < 1$, $p = 1$ and $b < a$. We construct an ϵ-equilibrium such

that as the network becomes dense, ϵ can be taken arbitrarily small, play becomes nearly consistent, and the associated steady state approaches that of the SSE.[22]

Proposition 5 *Take (a, b, δ) such that there exists a stable SSE (p, q) with $q < 1$, $p = 1$ and $b < a$. For every $\epsilon > 0$, there exists a $\bar{K}$ such that for all $K > \bar{K}$, there exists an ϵ-equilibrium with a stationary level of (p', q') with $p' = 1$ and $|q - q'| < \epsilon$.*

Proof. See Appendix C. □

This result establishes that, even in the case where the consistency condition fails, our description of SSE survives as a reasonable description of equilibrium behavior. The construction uses a bank robbing strategy as suggested by Proposition 3. The main idea is the following. Deviations from consistent cooperation are optimal only at unexpected histories such that the number of inlinks from cooperators is lower than anticipated. As the network becomes dense, such histories are reached with vanishing probability. Thus the optimal play becomes approximately consistent. This implies that the utility calculations that rely on simple strategies are approximately valid, and thus that the bank robbing strategy is approximately optimal when all players use it.

6 Conclusion

We have developed a model of interactions for a large anonymous community with turnover, in which agents are interconnected via an endogenously evolving network. The class of simple strategies that involve consistency of choices over time provides the foundation for our analysis. With consistent behavior, the social norm of ostracism is required in equilibrium, whereby links to an agent who defects are always severed immediately. We view this form of ostracism as capturing an empirically relevant phenomenon that is used to support cooperative behavior.

Under consistent strategies, we fully characterize stationary equilibria. Universal cooperation is sustainable for a non-trivial range of parameters, but not always. In particular, it requires not only that players are sufficiently long-lived (i.e., patient), but also that the

[22]The ϵ-best response is defined relative to the utility functions in equations (1) and (2), in which we normlize payoffs per-period and per-link.

temptation payoff for defecting not be too large. When these conditions are not both met, the presence of some level of non-cooperative behavior in a large anonymous system is unavoidable.

We believe this captures an important feature of a number of applications for which a fringe of exploitative behavior is observed. Our analysis offers new insights into thinking about how much cooperation can be sustained as a function of the underlying parameters of the system. When some level of defecetion persists, we can address through comparative statics what kinds of policies could be expected to improve the level of cooperation, and the total welfare of the system. Nearly all other related work falls either into the category of constructing equilibria for which full cooperation is sustainable, or else analyses models in which some form of inefficiency is inevitable. In a sense, our framework allows for a more balanced description of the acheivable level of cooperation in a society. This characterization is specifically due to the novel tradeoff in our model: that between immediate gains to defection and the gain to accumulating social capital in the form of additional links with cooperators over one's life.

When full cooperation does not obtain, the presence of defectors causes relationships among cooperators to be a scarce and valuable resource, which we identify as a form of social capital. In this case, the model provides for the possibility of a form of exclusivity to arise endogenously, in which cooperative players only occasionally agree to form new relationships with strangers. This exclusivity comes at a net welfare loss to society, but is necessary to incentivize cooperative behavior, as it slows down the rate at which cooperating partners can be found, thereby strengthening the penalty of ostracism due to defecting.

As it turns out, the characterization of equilibria under consistent strategies says a lot about equilibrium outcomes in which the possibility of inconsistent behavior is allowed. This demonstrates that the simple behavior we focus on can be self-enforcing, a finding that is not obvious a priori. In particular, under the appropriate condition, every simple stationary equilibrium has a corresponding equilibrium that supports consistent behavior on the equilibrium path with the same steady-state level of cooperation. The condition, which is satisfied for many parameter values, requires that the returns to links with cooperators are higher for cooperators than for defectors, thus implying that persistent cooperation is sequentially rational. Finally, we show that when this consistency condition is violated, a

"rob the bank" strategy, in which a cooperator deviates when he has a sufficiently low level of social capital, forms an epsilon equiibrium when players are patient and can maintain many links. We view these results as establishing a robust class of equilibria in a co-evolving network that generates a high degree of cooperation.

We have made some of the modeling choices to emphasize the point that it is possible to maintain cooperation even under unfavorable conditions. In thise sense, our results could be viewed as identifying a lower bound on what might be expected to obtain in related models. For example, one could imagine that new partners are found both at random, as we have modeled, and by searching the neighbors of current partners, as in Jackson and Rogers (2007). This would allow cooperators to preferentially find other cooperators more quickly, and would tilt incentives in favor of cooperation. It would also bring the degree distribution closer in line with the empirical observation that social networks tend to exhibit heavy tails. One could also imagine that agents have less than perfect access to anonymity. This generally has the consequence of making punishments for defection stronger, thereby increasing the scope for cooperation. Similarly, it might be reasonable to assume that cooperators have longer expected lives, which again makes cooperation easier to sustain.

We hope that the central elements in our framework will prove useful in further research. In particular, the combination of a random matching process,with endogenous link choices, in tandem with strategic actions taken with one's partners, is capable of describing important elements of other social phenomena. For example, Pin and Rogers (2013) use such a framework to study immigration dynamics in a nation.

We conclude with a remark about welfare. While we have not provided a complete characterization of equilibria in our framework, there is reason to be optimistic that the simple consistent equilibria that we identify generate a high level of average utility relative to other potential equilibria. The reason is that any strategy that incentivizes cooperation through inconsistent strategies has the difficulty that defecting with one partner requires defecting on all partners simultaneously. As such, the incentives of how to behave in the context different relationships are potentially in conflict, resulting in either diminshed incentives for cooperation or the inefficient loss of relationships. Consistent strategies have the unique property that an agent is never called on to change his behavior over the course of a relationship. We conjecture that the welfare associated with the stable SSE is maximal among all equilibria.

Author Affiliations

Immorlica: Microsoft Research

Lucier: Microsoft Research

Rogers: Washington University in St. Louis, Department of Economics

A For Online Publication Appendix A: formal development of the model

The model described in this paper is somewhat complex, incorporating a changing set of players, a large state space that is almost entirely unobserved by each individual player, and various sources of randomness. In the main text, 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, for the purpose of defining our equilibrium concept more precisely.

A.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 (p, 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 at any history 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 at any time 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 a 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 (p, q) determined by the choices at the previous round. Define (p^s, q^s) to be the proportion of cooperators and rate of inlink acceptance that i observes in the round when i is age s. At each s, i chooses an $\alpha_i \in \{C, D\}$. Also at each s, i has a set of partners, which includes both inlink and outlink partnerships. 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 to partner j was terminated in that round. The variable d_j^s equals 1 if j dies (0 otherwise), and the variables e_{ji}^s and e_{ij}^s 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_i^s 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_j^s. Due to anonymity, though, i does not know the value of j, but only the values of the elements in β_j^s. 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_i^s 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_i^s is empty. When agent i at age s receives a proposal for an inlink from agent j, it updates In_i^s as follows: if the link is accepted it appends β_j^s to In_i^s; if the proposed inlink is rejected outright, then we append a special symbol REJECT to In_i^s. After actions are realized, agent i updates each β_j^s in In_i^s appropriately. We define the *size* of list In_i^s, denoted by $|\text{In}_i^s|$ to be the number of *active* links contained in the list, i.e., the number of components of In_i^s for which $d_j^s = e_{ji}^s = e_{ij}^s = 0$.[23] Again, it is important that i not know the values of j corresponding to the various inlinks in In_i^s. Finally, denote by L_i^s 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

$$h_i^s = \{p^s, q^s, \alpha_i^s, L_i^s, \text{Out}_i^s, \text{In}_i^s\}.$$

The (private) history of i at age s is the vector $H_i^s = \{h_i^0, \ldots, h_i^s\}$. In a valid history it must be the case that the length of the list In_i^s grows monotonically with s and that if the k'th component of In_i^s is either REJECT or a β_j^s indicating a link termination (i.e., either d_j^s, e_{ji}^s, or e_{ij}^s 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_i^t equals the k'th component of In_i^s). Denote the space of feasible age-s histories for i by $\mathcal{H}_i^s$. The set of all histories for i is then $\mathcal{H}_i = \cup_s \mathcal{H}_i^s$.

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

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

(history dependent) action set of i at age s is $A_i^s(H_i^s) = [0,1] \times [0,1]^{L_i^s} \times [0,1]^{K+|\text{In}_i^s|}$, 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 $\mathcal{A}_i^s = \cup_{H_i^s \in \mathcal{H}_i^s} A_i^s(H_i^s)$ 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 $\phi_i(H_i^s) \in A_i^s(H_i^s)$ for all $H_i^s \in \mathcal{H}_i$. When i makes the choice of α_i^s, he has all the information in H_i^{s-1} as well as (p^s, q^s), but he has not observed the remainder of h_i^s. Similarly, when i makes his choice of accepting inlinks, he observes h_i^{s-1} and $(p^s, q^s, \alpha_i^s, L_i^s)$, but nothing else from round s. Last, when i makes the choice of severing active links, he has observed, additionally, the actions $\{\alpha_j^s\}$ in round 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.

A.2 Equilibrium concept

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 required that a strategy be measureable with respect to the information available for each action. 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 proceed with our discussion under such a convention, with the understanding that our notion of a history, strategy, etc. are equivalent to those developed in the previous section. Notice that the space of allowable histories and the space of action sets are the same for all agents; we can therefore define $\mathcal{H}^*$ and $\mathcal{A}^*$ so that $\mathcal{H}^* = \mathcal{H}_i$ and $\mathcal{A}^* = \mathcal{A}_i$ 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 states.

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 equilibrium, which will be a pair (ϕ, β) that satisfies certain properties. Informally, we require the following: at all valid histories ϕ maximizes expected utility given β when other agents apply ϕ; β is Bayes-consistent with an agent's observations and with the belief that all agents apply strategy ϕ; and, when faced with an unexpected (zero probability under β) history, β maps to a limit point of beliefs under a vanishing tremble probability on ϕ. 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 *self-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 self-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 a 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. We next incorporate an agent's history: given a distribution $\sigma \in \Delta(\Omega)$ over states, an agent i, and a history 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, shall require 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 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.

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 it is consistent with ϕ and if there exists $\{\beta^\epsilon\}$ such that:

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

We are now ready to formally 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 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 $\epsilon \to 0$.

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.

B Appendix B: Proofs of results for Section 3

The aim of this section is to prove Theorem 1, which states that there is a unique stable SSE with $q > 0$. The auxiliary results of Propositions 1 and 2 are proven in the course, and duly noted. Throughout we will write $d = \delta^2$ for convenience.

Roughly speaking, the strategy of our proof is to formalize the intuition conveyed by Figure 1. That figure illustrates the relationship between two indifference curves in (q, p) space: indifference between cooperating versus defecting, and indifference between accepting versus rejecting inlinks as a perpetual cooperator. In Section B.1 we develop the functional form of the inlink indifference curve, and study its properties. In Section B.2 we develop the functional form of the cooperation indifference curve, and describe some of its properties. In Section B.3 we describe the set of intersections between these curves. Finally, in Section B.4 we prove that there is at most a single intersection corresponding to a stable SSE.

B.1 Acceptance of inlinks

We begin by exploring the indifference condition for accepting an inlink, for a cooperator. For a cooperator, indifference between accepting and rejecting a given inlink requires that $\frac{r_C}{1-d} - br_D = 0$. Solving this condition for p yields

$$p_t(q) = \frac{q - b(1-q)(1-d)}{bq(1-q)d}.$$

By inspection, this function is negative for small q (it approaches $-\infty$ as $q \to 0$), greater than one for large q (it approaches ∞ as $q \to 1$), and strictly increasing in q on the unit interval. Define, thus,

$$t(q) = \min\{\max\{0, p_t(q)\}, 1\}$$

as the constrained solution to the threshold for accepting inlinks. Note that $t(q)$ is increasing, and is strictly increasing when its value is strictly between 0 and 1. Furthermore, steady-states (p, q) for which $p < t(q)$ have the property that accepting inlinks is dominant for cooperators. Thus, if (p, q) is an SSE, then it must be that $p = t(q)$.

Since $t(q)$ is increasing, we conclude the first item of Proposition 1, which is that if (p, q) and (p', q') are SEE with $q' > q$, then $p' \geq p$, with strict inequality when $p < 1$.

B.2 Cooperation and Defection

We next explore the indifference condition for choosing to be a perpetual cooperator versus a perpetual defector, at birth. Indifference between cooperation and defection requires that $u_C = u_D$, as defined in (1).

Define $V = u_C - u_D$ and $\Gamma = \{(q,p) \in [0,1]^2 | V \geq 0\}$ as the set of steady states in which cooperation is optimal. The indifference curve of interest will be the boundary of Γ.

We claim that V is strictly increasing in d for all $(p,q) \in [0,1]^2$. To see this, differentiate V with respect to d to obtain

$$(q(1+a) + b(1-q) + 2(1-pq)) \frac{pq}{(1-d(1-pq))^2}, \tag{6}$$

which is strictly positive for $(p,q) \in [0,1]^2$. This implies that Γ is strictly increasing (in the sense of set inclusion) in d.

To describe the boundary of Γ, solve $V = 0$ for p to obtain:

$$p = \frac{A \pm \sqrt{B}}{C},$$

where

$$\begin{aligned} A &= q(1+d) - [aq + b(1-q)](1-d), \\ B &= [b(1-q)(1-d) + q(a-1-(a+1)d)]^2 - 4q(b+q(1+a-b))^2 d(1-d), \\ C &= 2q(b+q(1+a-b))d. \end{aligned}$$

Let us call these solutions $p_1(q)$ and $p_2(q)$ so that we have $p_1(q) \geq p_2(q)$ in the unit square when the solutions are real.

The solutions are real when $B \geq 0$. Notice that B is cubic in q with a leading coefficient of $-4(1+a-b)^2 d(1-d) < 0$. Thus, ignoring the constraint that $0 \leq q \leq 1$, $B \geq 0$ for sufficiently small q and possibly also for a finite interval of q. We have that $B(q=0) = b^2(1-d)^2 > 0$ and that $B'(q=0) = 2b(1-d)[a(1-d) - (1+b)(1+d)] < 0$ whenever $a - b < 1$, as we assume. But for q near zero, the solutions to p_1 and p_2 are not in the unit interval, which can be verified directly, and so these are not valid solutions to $u_C = u_D$; in this case $u_D > u_C$.

In fact, it is the values of q that lie between the second and third roots of B that define the boundary of Γ. To see this, note that as $d \to 1$, $p_1(q) \to \frac{2}{b+q(1+a-b)}$ and $p_2(q) \to 0$. Thus, in the limit as $d \to 1$, we have

$$\Gamma \to \{(q,p) \in [0,1]^2 | p \leq \frac{2}{b+q(1+a-b)}\}, \tag{7}$$

and it converges in a way that the leftmost point approaches $q = 0$ from the right, while the first root of B is strictly between 0 and the second root. Thus, for any (q,p) with small q, there is a d such that q is equal to the second root of d, and for slightly smaller d, q falls between the two roots, so that $u_D > u_C$ for that value of q and d, independent of p. Then, because Γ is increasing in d, it must be that Γ excludes all such values of q for all smaller d. So valid solutions occur between the second and third roots of B, after taking the intersection with the unit interval.

Ignoring the constraint to the unit square in (q,p)–space, we then see that $p_1(q)$ and $p_2(q)$ form the boundary of a connected region. Points inside that region correspond to (q,p) for which $u_C \geq u_D$, i.e., Γ.

We note that (7) directly implies the first claim in Proposition 2, which is that $a < 1$ implies that $(1,1)$ is a stable SSE. To see this, note that $a < 1$ implies that $\frac{2}{b+1\cdot(1+a-b)} < 1$, and hence $(1,1) \in \Gamma$.

B.3 Number of equilibria

We now describe the set of intersections between $t(q)$ and the boundary of Γ, without differentiating between stable and unstable intersections. Note that an intersection of the indifference curves corresponds to an SSE.

Consider the system of equations $V = 0$ and $p = \kappa$, for an arbitrary constant κ. It is easy to explicitly solve these equations and see that they have at most two solutions in q; they are of the form $q = \frac{A' \pm \sqrt{B'}}{C'}$. This means that the boundary of Γ intersects any given horizontal line at most twice. In particular, this implies that there are at most two equilibria involving $p = 1$. It also implies that $p_1(q)$ and $p_2(q)$ are single-peaked on the unit square.

We now show that, in addition to the $p = 1$ equilibria, there exist at most two interior equilibria. To accomplish this we use a change of variables from (q,p) to (x,y) where $x = q$ and $y = pq$. Interior equilibria must satisfy both $V = 0$ and $p = p_t(q)$. This system can be

written as

$$\frac{2y-(1-d)f(x)}{g(y)} - f(x)y/x = 0 \tag{8}$$

$$\frac{x}{g(y)} - b(1-x) = 0, \tag{9}$$

where $f(z) = b + (1+a-b)z$ and $g(z) = 1 - d(1-z)$.

Substituting (9) into (8) and simplifying produces

$$2y - (1-d)f(x) - (1+a)yg(y) - y = 0, \tag{10}$$

Interior equilibria are thus described by simultaneous solutions to (9) and (10). Equation (10) is a parabola in (y, x)-space with second derivative with respect to y equal to $(2(1+a)d)/((1+a-b)(-1+d))$. Equation (9) has second derivative equal to $-2(bd)^2/(1+b(1-d(1-y)))^3$. To complete the claim, we show that the second derivatives are never equal, so that the difference between the curves is strictly convex or concave, and thus has at most two roots. Equating the second derivatives and solving for a produces a solution $a = h(b, y, d)$ that is easily verified to be continuous and equal to -1 when, e.g., $d = 0$ and $d = 1$. Solving $h(b, y, d) = 0$ for y, it is easy to see that there is no solution for $y > 0$. Thus there are no values of $a > 0$ and $y > 0$ such that (9) and (10) have equal second derivatives.

We have now proved item 2 of Proposition 1, which states that generically there are either 0, 2, or 4 SSE (in addition to autarky).

As $d \to 1$, there are at least 2 equilibria. This can be verified from the limiting shape of Γ, given above, in particular the fact that the limiting upper boundary of Γ is $2/(b+q(1+a-b))$, along with the fact that the limiting shape of $p_t(q)$ satisfies $p_t(0) = 1/b < 2/b$.

B.4 Uniqueness of stable equilibrium

We will now analyze the form of the intersections between the two indifference curves, and determine the properties of an intersection corresponding to a stable SSE. This will allow us to conclude that there is at most one stable SSE.

It is obvious from the above that for any (a, b, d), $(q, p) = (0, 0)$ is an SSE.[24] We want to show that, if there exists another stable equilibrium with $q > 0$, there is a unique such one.

[24] If there is an SSE with $q = 0$, then it must be that $p = 0$, and if there is an SSE with $p = 0$, it must be that $q = 0$.

Stationary equilibria occur when (i) the boundary of Γ intersects $t(q)$ or (ii) $(q,p) = (1,1) \in \Gamma$.

Stability is captured more easily by solving $V = 0$ for q. This produces two solutions for q, call them $q_2(p) \leq q_1(p)$, that are an equivalent representation of the boundary of Γ. A stable SSE is an SSE such that $q_1(p)$ intersects $t(q)$ or (ii) $(q,p) = (1,1)$ lies in the interior of Γ. Note that an SSE that corresponds to an intersection of $q_2(p)$ with $t(q)$ is not stable, as argued in the discussion following the statement of Theorem 1 in Section 3.

Define q_0 as the solution to $p_t(q) = 1$. Then $t(q) = 1$ for $q > q_0$, and $t(q) < 1$ for $q < q_0$. What we will argue is that, at every point of intersection (p', q') — that is, such that $q_1(p') = q'$ and $t(q') = p'$ — it must either be that $q' > q_0$ or that $\frac{dq_1}{dp}(p') < 0$. This will imply uniqueness for the following reasons. First, there can be at most one intersection with $q' > q_0$, since there is at most one intersection of $q_1(p')$ with the line $p' = 1$ since $q_1(\cdot)$ is a function. Second, there can be at most one intersection at a point (p', q') with $\frac{dq_1}{dp}(p')$, since q_1 is single-peaked and p_t has non-negative derivative; thus, the portion of q_1 with negative derivative can intersect p_t at most once. Third and finally, there cannot be both an intersection (p', q') with $q' > q_0$ and an intersection (p'', q'') with $q'' < q_0$ and $\frac{dq_1}{dp}(p'') < 0$. To see why, suppose that both intersections did occur. We must have $p'' < 1$, since $q'' < q_0$. However, by the single-peakedness of q_1, we must have $q' = q_1(1) < q_1(p'') = q'' < q_0$, which contradicts the assumption that $q' > q_0$.

It remains to prove the claim, which is that at every point of intersection (p', q') it must either be that $q' > q_0$ or that $\frac{dq_1}{dp}(p') < 0$. Equivalently, we must show that there is no intersection (p', q') with $q' < q_0$ and $\frac{dq_1}{dp}(p') \geq 0$. Define (q_β, p_β) by $q_\beta = \arg\min_q p_2(q)$ and $p_\beta = p_2(q_\beta)$, i.e., (q_β, p_β) is the lowest point of the boundary of Γ in (q,p)-space. Then (recalling the single-peakedness of q_1), the segment of q_1 with $\frac{dq_1}{dp}(p') \geq 0$ lies entirely above value q_β; that is, $q_1(p') \geq q_\beta$ whenever $\frac{dq_1}{dp}(p') \geq 0$. Therefore, it is sufficient to prove that $q_\beta > q_0$.

LEMMA **2** $q_\beta > q_0$.

Proof. Notice that Γ is empty for sufficiently small δ and that, given (a,b) there is a smallest δ such that Γ is non-empty. We show two facts. First, as Γ increases with d, it

intersects the line $q = q_0$ from the right, so that $q_\beta > q_0$ at that d. Second, as Γ increases further, it does so in a way such that $q_\beta > q_0$ remains true for all $d < 1$.

Because p_2 is single-peaked, in order to prove these facts we show that $p_2(q)$ is decreasing in q at $q = q_0$ whenever d is large enough that $p_2(q_0)$ is defined. It is sufficient to show that the same properties hold for some $q^* > q_0$, again because of the fact that $p_2(q)$ is single-peaked. Using such an argument, we construct a set of exhaustive cases that prove $p_2(q)$ is decreasing at q_0 whenever it is defined.

From equation (3), note that $q_0 < b/(1+b)$.

Define $s(a, b, d) = \frac{\partial p_2}{\partial q}|_{q=\frac{b}{1+b}}$. It is easy but very messy to write s explicitly. We want to show that s is negative whenever it is defined.

One can check that $s(a, b, 1) = 0$ and $\frac{\partial s}{\partial \delta}|_{\delta=1} = (1+b)^2/(2b) > 0$. Thus $s < 0$ for large δ.

It remains to check that s remains negative throughout the range of δ for which it is defined. This can be done via an explicit analysis of the functional form of s: one can observe that s has a vertical asymptote at some value $\hat{d}$ that describes the lower bound of this range of δ, that the value of s tends to $-\infty$ as δ approaches $\hat{d}$ from above, and that s does not admit any roots for values of d strictly between $\hat{d}$ and 1. Details of this analysis are available upon request.

□

We conclude that there is at most one stable SSE, completing the proof of Theorem 1.

B.5 Remaining Items in Proposition 1 and Proposition 2

To complete the proof of Proposition 1, we remark that the fourth item of the proposition is clear from the shapes of $t(q)$ and Γ already discussed. Next, notice that if there are four SSE, then the largest two involve $p = 1$, since at most two can be interior. In this case it must be that exactly one of those two involves q_1 and is therefore stable. This proves the third item of Proposition 1.

Using the characterization of stable SSE in the proof above, Proposition 2 follows easily. It uses only the limiting shape of Γ already derived (in equation 7) and the fact that as $d \to 1$, $q_0 \to \max\{\frac{b-1}{b}, 0\}$.

C Appendix C: Proofs of results for Section 5

Proposition 3 *Consider an SSE (p, q) with $q > 0$. A cooperator has a profitable inconsistent deviation if and only if*

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

Proof. Notice that the left-hand term in brackets is zero when $(p, q) = (1, 1)$, while the right-hand term in brackets is non-negative at any SSE with $q > 0$ (this is necessary to make cooperating with an in-link partner rational). Thus it is only at an SSE with $q < 1$ at which Equation 3 can hold, in which case $u(\phi_C, 0, 0) = u(\phi_D, 0, 0)$. By immediate implication we have also that $u(\phi_C, 0, 0) = u(\phi_F, 0, 0)$ (recall that ϕ_F plays exactly like a cooperator except that it defects in the first period).

From the expressions in the proof of Lemma 1, we have that the left-hand term in brackets is the net gain in utility from defecting once and then cooperating forever, compared to cooperating forever, for each outlink to a cooperator that is inherited from the previous period. Similarly, the right-hand term in brackets is the net gain in utility from cooperating forever, compared to defecting once and then cooperating forever, for each inlink from a cooperator that is inherited from the previous period. Thus Equation 3 characterizes the set of histories (K_I, K_O) at which the temptation to defect is higher than at birth, i.e., in which

$$\begin{aligned} \Delta u(\phi_F, K_I, K_O) &> \Delta u(\phi_C, K_I, K_O) \\ u(\phi_F, K_I, K_O) - u(\phi_F, 0, 0) &> u(\phi_C, K_I, K_O) - u(\phi_C, 0, 0) \\ u(\phi_F, K_I, K_O) &> u(\phi_C, K_I, K_O), \end{aligned}$$

showing that a cooperator has a profitable deviation to defect at the current history. □

Proposition 4 *The following statements are each (individually) sufficient for the consistency condition to hold.*

- *(p, q) is an interior SSE for (a, b, δ), i.e., $(p, q) \in (0, 1)^2$*
- *$p = q = 1$*
- *$p = 1$ and $b \geq a$*

Proof. The last two items are easily verified by inspection of the consistency condition.

The first item is proved by contradiction. We define the following strategies. As before, let ϕ_C denote consitent cooperation where inlinks are accepted with probability p, and let ϕ_D denote defection, where all inlinks are accepted. Further, let ϕ_R denote the strategy that proceeds according to ϕ_C, except that it plays D (and accept all inlinks) at all histories satisfying Equation 3. Finally let $\tilde{\phi}_R$ denote the strategy that is identical to ϕ_R except that it never accepts inlinks at a stage in which it plays C.

To begin, assume that the consistency condition is violated. By Proposition 3, this implies that a cooperator reaches a history at which defection is strictly preferred with positive probability, so that $u(\phi_R) > u(\phi_C)$.

We now claim that under ϕ_R the expected value of a new inlink in a period when the agent plays C is negative. To show this, we observe that the value of a new inlink to an agent playing ϕ_C is exactly zero, since $0 < p < 1$, and we argue that the value of the new inlink is lower to an agent playing ϕ_R. To see this, note that if the new link comes from a defector than it's value is $-b$ in both cases. If it comes from a cooperator, then its value is uniquely maximized by perpetual cooperation provided $1/(1-\delta^2) > 1 + a$, which is necessarily true in an SSE with $q > 0$. Thus $u(\tilde{\phi}_R) > u(\phi_R)$.

We now show that $u(\tilde{\phi}_R) \leq u(\phi_D)$. Let $\tilde{\phi}_R^{(s)}$ be the strategy that defects for the first $s-1$ periods and then follows strategy $\tilde{\phi}_R$, let $u_s(\phi_D)$ be the utility obtained in the first $s-1$ periods by defecting every period, and let A be the event that an agent links to no cooperators in period s. Note that, conditional on A, $\tilde{\phi}_R^{(s)}$ and $\tilde{\phi}_R^{(s+1)}$ differ only in that, in period s, $\phi_R^{(s)}$ gets $-b$ from each outlink. We therefore have that $\mathrm{E}[u(\tilde{\phi}_R^{(s)}) \mid A] < \mathrm{E}[u(\tilde{\phi}_R^{(s+1)}) \mid A]$.

On the other hand, conditioning on $\neg A$ is, for the purpose of computing continuation payoffs, equivalent to assuming that the agent starts period s with at least one outlink to a cooperator. Then, applying Proposition 3, by the assumption that the consistency condition is violated, the agent prefers to defect in period s. Thus $\mathrm{E}[u(\tilde{\phi}_R^{(s)}) \mid \neg A] \leq \mathrm{E}[u(\tilde{\phi}_R^{(s+1)}) \mid \neg A]$. Putting this together, we get

$$\begin{aligned} u(\tilde{\phi}_R^{(s)}) &= u_s(\phi_D) + \delta^s \cdot \Big(\Pr[\neg A] \cdot \mathrm{E}[u(\tilde{\phi}_R^{(s)}) \mid \neg A] + \Pr[A] \cdot \mathrm{E}[u(\tilde{\phi}_R^{(s)}) \mid A]\Big) \\ &< u_s(\phi_D) + \delta^s \cdot \Big(\Pr[\neg A] \cdot \mathrm{E}[u(\tilde{\phi}_R^{(s+1)}) \mid \neg A] + \Pr[A] \cdot \mathrm{E}[u(\tilde{\phi}_R^{(s+1)}) \mid A]\Big) \\ &= u(\tilde{\phi}_R^{(s+1)}). \end{aligned} \tag{11}$$

Since this holds for every s, $\tilde{\phi}_R^{(0)} = \tilde{\phi}_R$, and $\phi_D = \lim_{s\to\infty} \tilde{\phi}_R^{(s)}$, we conclude that $u(\tilde{\phi}_R) < u(\phi_D)$. This yields the desired contradiction, since at an interior SSE, $u(\phi_C) = u(\phi_D)$. $\square$

Proposition 5 *Take (a, b, δ) such that there exists a stable SSE (p, q) with $q < 1$, $p = 1$ and $b < a$. For every $\epsilon > 0$, there exists a $\bar{K}$ such that for all $K > \bar{K}$, there exists an ϵ-equilibrium with a stationary level of (p', q') with $p' = 1$ and $|q - q'| < \epsilon$.*

Proof. Notation: Let ϕ_S denote the (p, q)-simple strategy. Let $\phi_R(T)$ be the bank robbing strategy that accepts inlinks with probability one and defects if and only if $K_O/K_I > T$. Let $\phi_R^q(T)$ denote the strategy that plays $\phi_R(T)$ with probability q and plays ϕ_D with probability $1 - q$. Let T^* be the threshold defined by Proposition (3).

Set any $\epsilon > 0$. We show that for large enough K, $\phi_R^q(T^*)$ constitutes an ϵ-equilibrium. More specifically, we will show that, for any strategy ϕ, $u(\phi_R^q(T^*), \phi_R^q(T^*)) \geq u(\phi, \phi_R^q(T^*)) - \epsilon$. Recall that the utility function u is as defined in equations (1) and (2), and in particular the utilities are normlized to be per-period and per-link.

The proof uses the following two claims, which are proved below.

Claim 1: For every $\epsilon > 0$, there exists a $\bar{K}_1$ such that for all $K > \bar{K}_1$, $u(\phi_R^q(T^*)), \phi_R^q(T^*)) \geq u(\phi_R^q(T^*)), \phi_S) - \epsilon$.

Claim 2: There exists a $\bar{K}_2$ such that for all $K > \bar{K}_2$, and for all ϕ, we have $u(\phi_R^q(T^*), \phi_S) \geq u(\phi, \phi_R^q(T^*))$.

Now, given ϵ, take $\bar{K} = \max\{\bar{K}_1, \bar{K}_2\}$. We then have the following for all $K > \bar{K}$:

$$\begin{aligned} u(\phi_R^q(T^*), \phi_R^q(T^*)) &\geq u(\phi_R^q(T^*), \phi_S) - \epsilon \\ &\geq u(\phi, \phi_R^q(T^*)) - \epsilon \text{ for all } \phi, \end{aligned}$$

which is the desired conclusion. This completes the proof of the first part of Proposition 5, subject to the proofs of Claim 1 and Claim 2. We will discuss the second part of Proposition 5, that the stationary level of cooperation q' at the ϵ-equilibrium is close to the SSE cooperation level q, after establishing Claim 1 and Claim 2.

Before proving the two claims, we provide a preliminary result:

LEMMA **3** *There is a constant $\gamma > 0$, depending only on a, b, δ, and q, such that if $K_O/K_I \leq T^*$ at age s, then $\frac{E_{s+1|s}[K_O]}{E_{s+1|s}[K_I]} < T^*(1 - \gamma)$.*

Proof. Let us fix a given (K_O, K_I) at age s, such that $K_O/K_I \leq T^*$. From the derivations of $n_{CC}^{Out}(s)$ and $n_{CC}^{In}(s)$ it is direct to write down the expected in- and out-links with cooperators at age $s+1$, from any given (K_O, K_I) at age s. Since $E_{s+1|s}[K_O]$ is increasing in K_O, it is enough to consider the case when K_O is maximal which, by assumption, is $K_O = K_I T^*$. Making this substituion into the desired conclusion produces an expression of the form

$$\frac{AK_I T^* + B}{CK_I + D} < T^*(1-\gamma),$$

for certain expressions A, B, C, D. We know that the conclusion is true when $K_I = 0$, as the ratio of the expected links one period after birth is strictly less than T^*, permitting the choice of a small enough $\gamma > 0$. This fact can be easily verified directly. This implies, substituting $K_I = 0$ in the above, that $B/D < T^*(1-\gamma)$. It is thus sufficient to show that $A < C(1-\gamma)$. Expanding the condition above, we have that $A = (1-q)(1-\delta^2(1-q))$ and $C = 1-\delta^2(1-q)$, so that $A = (1-q)C$. Since q is interior (see Proposition 4), we have $A < C$ as desired. This completes the proof of Lemma 3. □

We now proceed with the proofs of Claim 1 and Claim 2.

Proof of Claim 1:

Choose some sufficiently small $\epsilon_1' > 0$, and consider the strategy profile in which all agents are applying strategy $\phi_R^q(T^*)$. Consider an agent who plays the "cooperate" (i.e., bank-robbing) role at birth. We will show that there exists $\bar{K}_1$ so that for all $K > \bar{K}_1$, and given any history observed by this agent up to the agent's selection of an action on round s, the probability that the agent will defect at age $s+1$ is at most ϵ_1'.

To see why this fact would imply Claim 1, consider the difference in the history observed by an agent playing $\phi_R^q(T^*)$ against a population playing $\phi_R^q(T^*)$, as opposed to a population playing ϕ_S. These histories differ only in that the former agent may observe defections from neighbors who previously cooperated, whereas the latter agent does not observe such behavior. However, given the claim in the previous paragraph, each neighbor that previously cooperated has at most an ϵ_1' chance of deviating each round. Thus the expected difference in utility due to deviating neighbors is certainly at most $\epsilon_1'(1+a+b)$ per neighbor, since $1+a+b$ is an upper bound on the difference in utility due to observing a defection rather than a cooperation from a neighbor. Since utility is measured per-link, the difference in utility is $\epsilon_1'(1+a+b)$. This difference can be made smaller than ϵ_1 by setting ϵ_1' sufficiently small.

Note that this difference in utility does not take into account any change in the likelihood that the given agent would deviate himself; however, since the probability of deviation is at most ϵ_1' on each round, the impact of this difference is at most ϵ_1' times the normalized per-round and per-link utility of the agent, which is again $\epsilon_1'(1+a+b)$.

It therefore suffices to show that, for any history observed by the agent up to the action-selection in round s, the probability that the agent will defect at age $s+1$ is at most ϵ_1'. To see this, we first observe that there exists a positive value $\gamma > 0$ such that $\frac{E_{s+1|s}[K_O]}{E_{s+1|s}[K_I]} < T^*(1-\gamma)$. This follows because either $K_O/K_I \leq T^*$ at age s, in which case the bound follows from Lemma 3, or else the agent will choose to defect on round s, in which case we effectively have $K_O = K_I = 0$ and Lemma 3 again applies.

Choose $\epsilon_1' = \gamma/4$. Consider a certain fixed agent, and suppose that every other agent in the population deviates with probability at most ϵ_1' in each round. We must show that our fixed agent also deviates with probability at most ϵ_1'. Since he observes each in-neighbor deviating with probability at most ϵ_1' between rounds s and $s+1$, the ratio of expected inlinks to expected outlinks *after* deviations are taken into account is at most $\frac{1}{1-\epsilon_1'} \cdot \frac{E_{s+1|s}[K_O]}{E_{s+1|s}[K_I]} < T^* \frac{1-\gamma}{1-\epsilon_1'}$.

Finally, for sufficiently large K, the law of large numbers implies that, on round $s+1$, K_O and K_I, as well as the number of observed deviations, will be concentrated around their expectations. Thus we can take K large enough so that, with probability at least $1-\epsilon_1'$, the ratio K_O/K_I at age $s+1$ will be at most $T^* \frac{1-\gamma}{1-\epsilon_1'} \cdot (1+\epsilon_1')$. Because of our choice of ϵ_1', this quantity is less than T^*. Thus, with probability at least $1-\epsilon_1'$, the ratio K_O/K_I will be less than T^* at age $s+1$, and hence the agent will not deviate at age $s+1$. This completes the proof of Claim 1.

Proof of Claim 2:

We show first that for high enough values of K, $\phi_R(T^*)$ is a best response to ϕ_S, i.e., for all ϕ, $u(\phi_R(T^*), \phi_S) \geq u(\phi, \phi_S)$.

Observe that at a history for which $k_I > 0$ and $k_O/k_I \leq T^*$, and also at birth, there is a best response that cooperates. This follows from the assumption that $0 < q < 1$ is an SSE by application of Proposition 3. Next, Lemma 3 shows that if an agent is at a history where $k_O/k_I \leq T^*$ then it expects to remain below the threshold at the next period. Iterating the argument, it in fact expects to cooperate at all future periods. By a law of large numbers,

as K is taken large, the probability of such an agent defecting at any point in his life can be made as small as desired.[25] Since, by assumption, a consistent cooperator has a strict preference for accepting inlinks, i.e., $v > 0$, it follows that at such a history, and for large enough K, it must be optimal to set $p = 1$ at that period.

We now argue that application of the T^* threshold characterizes the optimal cooperation behavior.

We claim first that whenever $k_O/k_I \geq T^*$, defecting immediately is better than cooperating and then defecting in the next period. It is straightforward to show that defecting immediately is better if and only if $Z = D_k - C_k - \delta(D_{k^+} - D_0) > 0$, where D_k (C_k) represents the one-period payoff from defecting (cooperating) with existing links given by $k = (k_O, k_I)$, and k^+ represents the expected links after one period of cooperation, given k today. Take a ratio of links that equals the threshold, i.e., $k_O = k_I T^*$. We use the fact that at an SSE with $p = 1$ and $0 < q < 1$, $\delta^2 = \frac{2(b(-1+q)-aq)}{(2-q)(b(-1+q)-(1+a)q)}$. Making the substitions into Z for k_O and δ^2, and multiplying by $(1-\delta^2)(1-\delta^2(1-q))$ produces an expression Z' that is an increasing (linear) function of k_I. Thus, the worst case for showing that $Z > 0$ is when $k_I = 0$. Setting $k_I = 0$ in Z' and multiplying by $(1-\delta)$ produces an expression that is positive by simple inspection. This proves the claim for all ratios equal to T^*. To prove the claim for ratios greater than T^*, it is sufficient to show that Z is decreasing in k_I. In fact, this derivative is negative if and only if $1 + a < 1/(1-\delta^2)$, which is necessarily true in any SSE with $q > 0$.

Thus, if $k_O/k_I \geq T^*$, defecting today is better than defecting tomorrow. Iterating this argument, at all future periods one can compute the expected ratio of links. At some finite period in the future it will be less than T^*. From that point on, defecting is clearly dominated, via the calculations in Lemma 1 in the main text. Until that time, iterating the argument above shows that defecting immediately is the optimal time to defect. Thus, the relevant calculus in computing the optimal action at a given period described by k, is the comparison between defecting immediately, and cooperating always. But this is exactly the calculation that leads to the threshold of T^* describing when it is optimal to defect. This proves that $\phi_R(T^*)$ is the best response to ϕ_S.

Now, suppose by way of contradiction that there exists a strategy ϕ^* and an $\epsilon' > 0$ such

[25] For example, take S such that $\delta^S < \epsilon'/2$ and set K large enough that $Pr(K_O^{s+1}/K_I^{s+1} > T^* | K_O^s/K_I^s = T^*) < \epsilon'/(2S)$. Then the probability of cooperating at every period of the agent's life is at least $1 - \epsilon'$.

that, for a diverging sequence of $\{K_j\}$, $u(\phi^*, \phi_R^q(T^*)) > u(\phi_R^q(T^*), \phi_S) + \epsilon'$. From the proof of Claim 1, we have that population dynamics under $\phi_R^q(T^*)$ converge to those under ϕ_S as $K \to \infty$. We thus have a contradiction to the fact that $\phi_R^q(T^*)$ is a best response to ϕ_S, completing the proof.

Now, suppose by way of contradiction that there exists a strategy ϕ^* such that, for a diverging sequence of $\{K_j\}$, $u(\phi^*, \phi_R^q(T^*)) > u(\phi_R^q(T^*), \phi_S)$. Construct a strategy ϕ' that, for each neighbor, simulates a history of play according to the distribution of a world in which that neighbor plays $\phi_R^q(T^*)$. That is, ϕ' plays like ϕ^* except that it breaks links to neighbors with histories of pure cooperation, each independently with a probability equal to the probability that neighbor would defect were it playing $\phi_R^q(T^*)$, conditional on the history of the agent playing ϕ'. Now consider an agent playing ϕ' against a poopulation playing ϕ_S. Clearly, ϕ' does not change the behavior of ϕ_S agents. Furthermore, the proportion of cooperators in ϕ_S is equal to the proportion of bank robbers in $\phi_R^q(T^*)$. Thus this agent observes the same distribution over histories as one playing ϕ' against $\phi_R^q(T^*)$ with the exception that in the former case a neighbor cooperates on the last round of play before the agent breaks a link whereas in the later case it defects. This difference can only increase the utility of the agent, and so $u(\phi', \phi_S) \geq u(\phi^*, \phi_R^q(T^*)) > u(\phi_R^q(T^*), \phi_S)$, contradicting the fact that $\phi_R^q(T^*)$ is a best response to ϕ_S. This completes the proof of Claim 2.

Finally, we must establish the second part of Proposition 5: that the stationary level of cooperation q' at the ϵ-equilibrium can be made arbitrarily close to the SSE cooperation level q, given sufficiently large K. This follows from the following fact, which was established in the proof of Claim 1: for any ϵ_1' there exists a sufficiently large K such that, if all agents play strategy $\phi_R^q(T^*)$, then given any history observed by an agent who plays "cooperate" at birth, up to the agent's selection of an action on round s, the probability that the agent will defect at age $s+1$ is at most ϵ_1'. Since a fraction q of agents cooperate at birth, this fact implies that the level of cooperation on each round is at least $q(1-\epsilon_1')$. That is, if q' is the level of cooperation at the bank-robbing ϵ-equilibrium, then $|q'-q| < \epsilon_1' q \leq \epsilon_1'$. Since ϵ_1' can be taken to be as small as desired, the result follows.

□

References

Alós-Ferrer, C. (1999): "Dynamical systems with a continuum of randomly matched agents," *Journal of Economic Theory*, 86, 245–267.

Blackburn, J., R. Simha, N. Kourtellis, and X. Zuo (2011): "Cheaters in the Steam Community Gaming Social Network," *Arxiv preprint arXiv:*.

Board, S. (2010): "Relational contracts and the value of loyalty," *American Economic Review*.

Board, S. and M. Meyer-ter vehn (2011): "Relational Contracts in Competitive Labor Markets," 1–39.

Boylan, R. T. (1992): "Laws of large numbers for dynamical systems with randomly matched individuals," *Journal of Economic Theory*, 57, 473–504.

Dall'Asta, L., M. Marsili, and P. Pin (2011): "Collaboration in Social Networks," *Arxiv preprint arXiv:1104.2026*, 1–18.

Datta, S. (1993): "Building Trust," Mimeo, Indian Statistical Institute.

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

Fosco, C. and F. Mengel (2011): "Cooperation through imitation and exclusion in networks," *Journal of Economic Dynamics and Control*, 35, 641–658.

FUJIWARA-GREVE, T. and M. OKUNO-FUJIWARA (2009): "Voluntarily Separable Repeated Prisoner's Dilemma," *Review of Economic Studies*, 76, 993–1021.

Galeotti, A., S. Goyal, M. O. Jackson, F. Vega-Redondo, and L. Yariv (2010): "Network games," *The Review of Economic Studies*, 77, 218–244.

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

Jackson, M., T. Rodr\'\iguez Barraquer, and X. Tan (2011): "Social capital and social quilts: Network patterns of favor exchange," *American Economic Review*.

Jackson, M. and B. Rogers (2007): "Meeting Strangers and Friends of Friends: How Random are Social Networks?" *The American Economic Review*, 97.

Jackson, M. O. and A. Watts (2010): "Social games: Matching and the play of finitely repeated games," *Games and Economic Behavior*, 70, 170–191.

Judd, K. L. (1985): "The law of large numbers with a continuum of {IID} random variables," *Journal of Economic Theory*, 35, 19 – 25.

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

Krackhardt, D. (1996): "Social Networks and the Liability of Newness for Managers," in *Trends in Organizational Behavior, Volume 3*, ed. by C. Cooper and M. Rousseau, New York, NY: John Wiley & Sons Ltd, chap. 9, 159–173.

Kranton, R. (1996): "The Formation of Cooperative Relationships," *Journal of Law, Economics, and Organization*, 12, 214–233.

Levin, J. (2002): "Multilateral contracting and the employment relationship," *The Quarterly Journal of Economics*, 1075–1104.

MacLeod, W. and J. Malcomson (1998): "Motivation and markets," *American Economic Review*, 388–411.

Pin, P. and B. Rogers (2013): "Cooperation in a Society with Differential Treatment of Immigrants," *Working Paper*.

Ramey, G. and J. Watson (2001): "Bilateral Trade and Opportunism in a Matching Market," *Contributions to Theoretical Economics*, 1, article 3.

Riedl, A. and A. Ule (2002): "Exclusion and cooperation in social network experiments," *Unpublished Paper, CREED, University*.

Shapiro, C. and J. Stiglitz (1984): "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, 74, 433–444.

Sobel, J. (1985): "A theory of credibility," *Review of Economic Studies*, 52, 557–573.

——— (2002): "Can we trust social capital?" *Journal of Economic Literature*, XL, 139–154.

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

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

——— (2002): "Starting Small and Commitment," *Games and Economic Behavior*, 38, 176–199.