COOPERATION IN A DYNAMIC NETWORK

Abstract

We study the scope for cooperation in an evolving network. Under endogenous dynamics, cooperation leads to a growing neighborhood of long-term productive relationships, while defection exploits ephemeral relationships resulting in social marginalization. We characterize a class of simple stationary equilibria in which, even as players become patient, the coexistence of cooperation and defection naturally arises. This captures the observation that many social networks are characterized by a high, though not universal, level of cooperation. Low levels of cooperation can be sustained only through exclusivity among cooperating agents.

JEL codes: C73, D83, D85

Keywords: Cooperation, Networks, Social capital, Prisoner's dilemma, Repeated games

1 Introduction

Many social interactions and economic transactions are mediated through large evolving networks. The size of these systems and the transient nature of both individual membership and interpersonal relationships often confer a measure of anonymity to participants as they form new connections. In online contexts, there are technological aspects that reinforce anonymity, in that it is typically easy for a participant to create a new identity 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 less than universal, level of cooperation. We find that such an outcome can be explained as equilibrium behavior under a simple network formation process.

In our model, 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.[1] 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). A key feature of such platforms is anonymity: individuals typically have little information about the past behavior of their new partners. The discretion to sever a relationship thus plays a critical role in sustaining cooperation. It provides the only viable punishment mechanism in these anonymous settings, since defectors can easily abandon their identity and avoid future sanctions. As a result, the threat of link severance becomes the primary tool by which agents discipline opportunistic behavior.[2]

These dynamics are common to a variety of real-world settings. For instance, in decentralized online marketplaces such as eBay or Amazon Marketplace, buyers and sellers form an inherently bipartite network in which interactions strictly occur across two distinct groups (buyers and sellers). Sellers face a tradeoff between cooperation—delivering quality goods at fair prices—and defection—misrepresenting products, selling counterfeits, or offering inferior quality to boost profit margins. While some formal reputation mechanisms exist, these can often be circumvented through identity resets or the creation of new accounts. Participants thus operate under substantial anonymity, with cooperation incentivized by the potential for repeat business and positive reviews, and defection deterred by the threat of severed ties and lost future sales.

By contrast, the academic research community forms a non-bipartite network, as researchers freely choose collaborators from within a single group. Cooperation in this setting entails high effort towards the collective project, while defection involves shirking, free-riding, or disproportionately exploiting coauthors' efforts. Although the resulting research is public, the collaborative process remains largely opaque, with reputational consequences for defection often informal and diffuse. Nevertheless, the principal disciplining mechanism remains relational: defecting on one project may lead to exclusion from future collaborations while

[1] Jackson and Watts [2010] define and study social games that change 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. [2012] identify collaborative equilibria as a function of the social network that describes interactions.

[2] Because the defecting agent has the option of severing the relationship, one cannot threaten contingent play that reduces the defecting agent's continuation payoff from the relationship below its outside option.

cooperating may result in future opportunities. In both cases, cooperation is sustained primarily through the ability to selectively retain productive partnerships and sever exploitative relationships. While our model is formalized explicitly for peer-to-peer (non-bipartite) environments, we expect its main results to naturally extend to bipartite contexts as long as agents retain discretion over forming and severing partnerships.

Interestingly, there is now a growing pool of empirical studies that demonstrate the power of the incentives that derive from agents' discretion to initiate and sever relationships. Riedl and Ule [2002] provide empirical evidence from laboratory experiments that the ability to endogenously determine one's partners increases observed cooperation levels in repeated prisoner's dilemma interactions. Wang et al. [2012] show in an online experiment that cooperation is improved both by granting players the ability to avoid recent defectors and by punishing defectors through severance. Blackburn et al. [2012] show that cheaters in an online game are punished by an increased rate of losing partners, leading to more cooperation. These effects constitute crucial elements of the model we study.

There are three essential properties of the network process that drive our results. First, in order to induce strategic dependence across relationships, we assume that each agent takes the same action with each of his neighbors at any given time. In some applications, technological constraints limit an agent's ability to behave differently with different partners, or the chosen action may represent 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.[3] Further, if one introduces even a small amount of local information flow across the network, such that a defection with one partner is 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 on the profile of actions an individual chooses at a given moment in time.[4]

The second property 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. Since defection can be punished only by the severance of relationships,

[3] The common action approach is taken by most papers studying games on networks. See, e.g., Galeotti et al. [2010].

[4] One could view this construction as an implementation of "local community enforcement" along network connections, in the spirit of Kandori [1992] and Ellison [1994]. Ali and Miller [2013] is an important example in which individuals can take different actions with different partners, subject to imperfect monitoring across relationships. They show that community enforcement can still sustain cooperation through strategic communication.

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 local network of cooperative partners.[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 third property is that there is a limit to the number of relationships that an agent can maintain. In combination with the first property, this implies that a cooperator experiences decreasing marginal payoffs in the aggregate level of cooperation. This concavity is important because it creates the possibility that cooperating and defecting agents will co-exist in equilibrium. In fact, the coexistence result obtains for a wide range of parameters and we view it as having descriptive value.

The constraint on the number of relationships also 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 [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] Recent experimental work in this domain includes Fujiwara-Greve et al. [2023] and Lee [2020]. Both papers find

[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 to ours, Vega-Redondo [2006] studies social capital in a stochastically evolving network.

[6]MacLeod and Malcomson [1998] and Board and Meyer-ter Vehn [2015] contribute to the search literature in the context of relational contracts.

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

that increased costs of separation result in higher levels of cooperation.

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 constructions in which actions evolve in a non-trivial way throughout the course of a given relationship. In a network setting 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] Under 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 recurrent interactions occurring on the network. In particular, a relationship is severed when, and only when, a defection is observed. Such a norm is quite natural: defection is not tolerated, and cooperation is met with the opportunity for future interactions.[10]

We characterize (perfect Bayesian) equilibria that exhibit a steady-state in which agents take consistent actions on path. Perhaps surprisingly, it is relatively easy to achieve sequential rationality of consistent cooperation. Intuitively, agents cooperate in order to accumulate more cooperative partners in the future. Perpetual cooperation is sequentially rational in these equilibria because this incentive to attract more partners is robust to variation in an agent's current neighborhood. The condition that characterizes sequential rationality of cooperation requires that the temptation payoff to defect be small enough relative to the loss from being defected on. We demonstrate that the condition is relatively mild. Thus consistent behavior can be self-enforcing when paired with the norm of severing a link when, and only when, the partner defects.

We also 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 partic-

[8]In a very different framework, Jackson et al. [2012] 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 [2011] investigate relational contracts between a principal 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 of cooperators and defectors by employing dynamics based on imitation.

[10]The fact that defection results in the loss of one's links can be viewed as a form of ostracism. Ali and Miller [2016] show that ostracism, when temporary, can be used in a network setting to improve the amount of cooperation that can be sustained through bilateral enforcement.

ular circumstance: when he has very little network capital, and therefore little to sacrifice in terms of future links with cooperators. As the network becomes dense, such events become increasingly rare and, as a result, these strategies form an ϵ-equilibrium that generates essentially the same level of cooperation that we characterize under consistent strategies.

To focus on the incentives that govern the *level* of cooperation, we first study a simplified model in which agents are restricted to use consistent strategies, allowing us to completely characterize steady-state equilibria. We find that the stage game payoffs 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 maximal *stable* equilibrium that supports cooperation, and it is such that either there is universal cooperation, or cooperators and defectors co-exist at a specific ratio. In this sense we provide an explanation for which societies permit universal 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 relative to the expected returns to cooperation.

The remainder of the paper is organized as follows. The model framework is presented in Section 2. Section 3 characterizes steady-state equilibrium outcomes under consistent strategies and compares the level of cooperation sustainable in our model with that in a bilateral indefinitely repeated prisoner's dilemma. Section 4 shows that consistency is a self-enforcing norm under the appropriate condition and describes outcomes that involve a minimal failure of consistency. Section 5 concludes. Most proofs are relegated to the online appendix available at `https://wustl.box.com/s/mfl9c4kdu4g5w0iskl4pwfyk7lphsgfj`.

2 A model of strategic interactions in a social network

Our model involves a continuum of agents, which we associate with points from the unit interval $[0, 1]$.[11] Time is discrete. In each round, each agent independently dies with a given probability $1 - \delta$, in which case it is replaced by a new agent. We speak of the age of an agent i, $s(i)$, as the number of rounds since its birth, so that $s(i) = 0$ in the round when i is born.

Agents interact repeatedly through an evolving directed network. Every agent is able to sponsor a number $K \geq 1$ of connections to other agents. Agents are involved both in relationships that they sponsor (outlinks) and also in relationships sponsored by others (inlinks), resulting in a directed graph of interactions.[12] 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 round 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, in the next round.

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

Table 1: Payoff matrix

We take $a, b > 0$ and $a - b < 1$ so that, while mutual cooperation is the uniquely efficient outcome, defection is strictly dominant. Each agent i chooses either action C (cooperate)

[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. Our model relies on discrete events in this continuum space, such as selecting a random point or a random set of a fixed measure. See, e.g., Judd [1985] for a formalization of these notions.

[12] 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 maintain.

or action D (defect) in each round $s(i)$ of its life and receives a payoff equal to the sum of the outcomes of the prisoner's dilemma games played with each of his (in and out) partners, according to the chosen actions of the two agents and the payoff matrix shown in Table 1. Agents seek to maximize the present value of expected lifetime payoffs.

Each round proceeds according to the following order of events:

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

2.1 Consistency

In order to gain traction on studying equilibrium outcomes, we begin the analysis by considering a setting in which agents optimize their strategies from among a restricted class of simple intuitive behaviors. Namely, agents are assumed to be consistent in their choice of cooperate (C) or defect (D) over the course of their lives.

Property **1** *An agent is* consistent *if an action from* $\{C, D\}$ *is chosen at birth (possibly mixing), and at all future rounds, the agent takes the same action it played in the previous round.*

Consistency allows us to speak of society as being comprised of "cooperators" and "defectors", and has a number of implications for optimal play. In particular, optimal strategies under this behavioral restriction can be completely described by a tuple in $[0, 1]^2$ with the interpretation that (p, q) specifies the probability q that an agent chooses C at birth and the probability p with which each inlink is accepted in the event the agent becomes a cooperator.[13] (Defectors, of course, have a dominant strategy to accept every inlink.)

Subject to consistency, the remaining optimal actions are straight-forward: if defection is observed, the defector's partner is *unforgiving* and severs the link; if cooperation is observed,

[13] Because of anonymity, all proposed inlinks are ex ante equivalent from the perspective of the agent receiving the proposal, so it is without loss to assume that a cooperator accepts each proposed inlink independently with a certain probability p.

the cooperator's partner is *trusting* and maintains the link (provided both agents survive). We refer to such strategies as (p, q)-*simple*.

2.2 Simple Stationary Equilibria

We focus our analysis on symmetric strategies and steady state equilibria.

DEFINITION **1** *(p, q) is a* steady state *if and only if (p, q) is a fixed point of the dynamics in which all agents play strategy (p, q).*

Observe that all (p, q)-simple strategies played symmetrically have exactly one associated steady state. Namely if all agents cooperate with probability q on birth and cooperators accept inlinks with probability p, then society will perpetually have proportion q of agents cooperating and p proportion of inlinks to cooperators will be accepted. Furthermore, in each steady state associated with a (p, q)-simple strategy, the age distribution of cooperators and defectors is the same. Let $u_i(p', q'|\hat{p}, \hat{q})$ denote the expected continuation utility of agent i of playing strategy (p', q') in the steady state associated with strategy $(\hat{p}, \hat{q})$ given that all other agents continue playing $(\hat{p}, \hat{q})$. Our solution concept is then defined as follows.

DEFINITION **2** *$(\hat{p}, \hat{q})$ is a* Simple Stationary Equilibrium *if and only if $(\hat{p}, \hat{q}) \in \arg\max_{p', q'} u_i(p', q'|\hat{p}, \hat{q})$ for all agents i.*

That is, $(\hat{p}, \hat{q})$ is a Simple Stationary Equilibrium (SSE) if and only if no agent has a profitable deviation in the space of (p, q)-simple strategies. As observed previously, any (p, q)-simple strategy, $(\hat{p}, \hat{q})$, has a unique associated steady state. Our equilibrium concept requires that, at its steady state, this strategy is (weakly) optimal for all agents. In order for $(\hat{p}, \hat{q})$ to be an SSE, we must check three additional conditions. First we must show that agents at birth want to become cooperators with probability $\hat{q}$. This is related to the expected lifetime utility of cooperating and defecting. Second, we must argue about when a cooperator should accept a proposed inlink. This will be determined in part by the probability that a proposed link comes from a cooperator. Notice that this probability is less than $\hat{q}$ in equilibrium since 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. Third, for any choice of p and q, we must confirm that it is indeed rational for cooperators to send outlinks. We perform these calculations in Section 3 wherein we fully characterize the set of SSE.

Consistency limits the complexity of strategic interactions in an important way. 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, such as elaborate punishment phases, it turns out to be a significantly milder condition than it may first appear. We formalize this assertion in Section 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 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 as a cooperator expects to build their network of social capital over time. 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, implying the SSE is in fact a steady-state equilibrium of the larger strategy space and generalized game.

3 Stationary outcomes under consistent behavior

In this section we characterize the set of SSE, i.e., the steady-state equilibrium outcomes under an assumption that all agents use consistent strategies. We view these results as interesting in their own right. Further, the subsequent section is devoted to showing that these outcomes, exhibiting consistent behavior, to large extent constitute optimal on-path play without any restrictions on strategies.

3.1 Expected Utilities

We begin by deriving 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 by cooperators, p. Since we are interested in simple stationary equilibria, we work under the assumption that the system is in steady-state 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.[14] 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

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

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 new link rejection) 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 proposed inlinks an agent will receive from cooperators and defectors at each date are, respectively,

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

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

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

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

$$n^{In}_{CC}(s) = pr_C \frac{1-\delta^{s+1}}{1-\delta}.$$

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

Finally, we 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:

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

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.

3.2 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 must compare u_C and u_D as functions of (p, q) under given parameters of the model. It is convenient to define $V(p, q; a, b, \delta) = u_C - u_D$. We additionally define the cooperation region Γ:

$$\Gamma := \left\{(p, q) \in [0, 1]^2 : V(p, q; a, b, \delta) \geq 0\right\},$$

that is, the set of (p, q) pairs under which cooperation at birth is weakly optimal. The boundary of this region then identifies (p, q) pairs under which agents at birth are indifferent between perpetual cooperation and perpetual defection.

In steady-state, the expected value to a cooperator of a new relationship formed by accepting an inlink proposal 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

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

Define $p_t(q)$ to be the unique value (if any) for which a cooperator is indifferent to accepting or rejecting a proposed inlink when the population cooperation rate is q. That is, $p_t(q)$ satisfies:

$$v(q, p_t(q)) = 0.$$

We then define the inlink indifference curve $t(q)$:

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

This curve represents the acceptance probability such that cooperators are indifferent to accepting or rejecting additional inlinks truncated to $p \in [0, 1]$.

The following observations describe the various possibilities for SSE. First, notice that for any choice of parameters, there exists an SSE with $q = 0$, which we refer to as *trivial.* This is true because when all agents 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 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, so that cooperators are 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 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 depending 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 optimal behavior. 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 steady-state. Thus the fact that an inlink has non-negative value at such 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 steady-state. 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 be that outlinks have positive value.

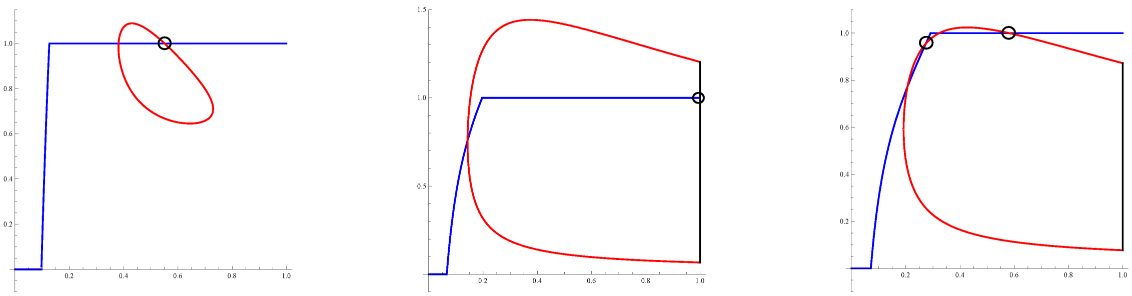
The main result of this section establishes conditions on the existence and quantity of nontrivial SSE. The autarky outcome ($q = p = 0$) constitutes a trivial SSE for all valid parameters of the game.

Proposition 1 *The following statements are true:*

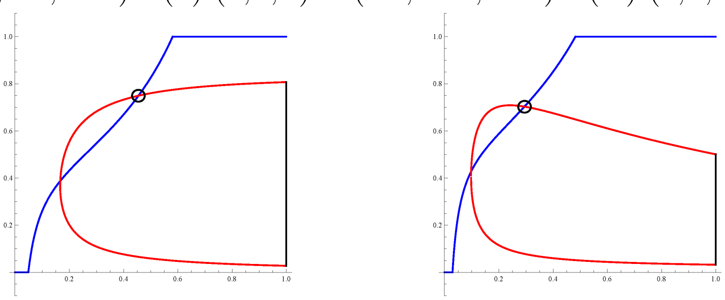
- *All SSE are ordered, in the sense that if (p, q) and (p', q') are two SSE with $q' > q$, then $p' \geq p$, with strict inequality when $p < 1$.*
- *In addition to the trivial outcome, there are generically either 0, 2, or 4 SSE.*

We outline briefly the main arguments relevant to Proposition 1. By definition, nontrivial SSE occur (i) at the intersections of $t(q)$ with the boundary of Γ and (ii) at $(p, q) = (1, 1)$ when $(1, 1) \in \Gamma$. The proof proceeds by bounding the number and type of possible intersections between Γ and $t(q)$ through explicit consideration of the utility functions. One can show 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 $p = 1$. 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. Examples for different parameterizations are shown in Figure 1.

Figure 1 depicts a representative set of examples for SSE configurations, computed numerically. In each case, the red curve represents the boundary of Γ, i.e., the steady-states (p, q) for which $u_C = u_D$, where in the interior $u_C > u_D$, while the blue curve represents $t(q)$, i.e., the steady-states for which the value of inlinks is $v = 0$, constrained to $p \in [0, 1]$, with steady-states below $t(q)$ having strictly positive value for inlinks ($v > 0$). Thus, SSE occur at the intersections of the two regions or when $(1, 1) \in \Gamma$.



(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 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 (interior); (iv) and (v) Two interior SSE. SSE occur at all intersections of Γ and $t(q)$. Stable SSE are circled.

3.3 Characterization of stable SSE

We are particularly interested in those SSE that are *stable*. Loosely, we call an SSE (p, q) stable if the system returns to it after small perturbations away from (p, q), when it is the case that agents apply optimal strategies at the observed perturbed values of p and q. Our primary motivation in considering stable SSE is to distinguish a certain type of inherently unstable equilibria which, due to their instability, are not meaningfully predictive of system state.

Important to our notion of stability is the observation that inlink acceptance p adjusts (infinitely) faster than cooperate/defect actions q. This is driven by consistency which allows only new entrants to choose q while permitting incumbent cooperators to update p throughout the course of their lives. We thus aim to evaluate incentives on cooperate/defect decisions after inlink acceptance has equilibrated.

This leads us to evaluate cooperate/defect incentives at the perturbed steady-state $(t(q'), q')$, as this is the unique steady state at q' for which inlink decisions have equilibrated. For any $p > t(q)$ inlinks have negative value and for any $p < t(q)$ inlinks have positive value. This ensures that any perturbation in q to q' yields a unique $p' = t(q')$ and $(t(q'), q')$ is well-defined. The proof of Proposition 2 then builds upon Proposition 1 by identifying intersections such that $t(q-\epsilon)$ is inside Γ and $t(q+\epsilon)$ is outside Γ. Examples are shown in Figure 1.

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

(i) if $q < 1$, *then for all* $q' \in (q, q+\epsilon)$, $u_C(t(q'), q') < u_D(t(q'), q')$, *and*

(ii) if $q > 0$, *then for all* $q' \in (q-\epsilon, q)$, $u_C(t(q'), q') > u_D(t(q'), q')$.

We view stability as an important refinement in our setting and the main result of this section establishes conditions on the stability of the SSE identified in Proposition 1.

Proposition 2 *The following statements are true:*

- *If there are 2 nontrivial SSE, the larger is stable and the smaller is unstable.*
- *If there are 4 nontrivial SSE,*
  - *the largest is stable and involves* $p = 1$.
  - *the third largest is stable and involves* $p < 1$.
  - *the second largest and smallest are unstable.*

Proof. See online appendix. □

Figure 1 depicts a representative set of examples for SSE configurations, computed numerically. In each case, the red curve represents the boundary of Γ while the blue curve represents $t(q)$. SSE occur at the intersections of the two regions. Stable interior SSE require further that $t(q)$ intersects the boundary of Γ from within the region. Case (iii) is of particular interest because these parameters generate two stable SSE. While such cases are an interesting result, they obtain only for a small range of parameters and have a particularly fragile, local stability. We focus our following analysis on the more cooperative stable SSE in such cases.

3.4 SSE for long-lived agents

The model allows for an explicit determination of the set of SSE for any parameters (a, b, δ). It is therefore possible to partition the parameter space into regions that map into the different configurations of SSE described by Propositions 1 and 2. However, as such an analysis is cumbersome, we instead present results for the limiting case in which players become perfectly patient (i.e., long lived), which captures much, but not all, of the richness of the model. We also leverage the simplicity of the limiting case to illustrate comparative statics of SSE with respect to parameter changes.

Proposition 3 *The following statements hold as $\delta \to 1$.*
There is a unique non-trivial ($q > 0$) stable SSE.
If $a < 1$ then $p = q = 1$ is the stable SSE.
If $a > 1$, then

- *all SSE have $q < 1$*
- *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 online appendix. □

Proposition 3 formalizes a number of previous observations. Namely, there is a rich set of qualitatively distinguished equilibrium outcomes, even in the limit as players become completely patient. See Figure 2 for a representation. Indeed, even when players are perfectly patient, full cooperation can be achieved only if the temptation payoff is small enough ($a < 1$).[15] 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, and has vanishingly few inlinks, for a total of $K(1 + a)$. 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 $2K$ 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 supported. In this case, i.e., when $b < 1 + a$, society consists of cooperators and defectors

[15]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 necessarily support full cooperation for patient players.

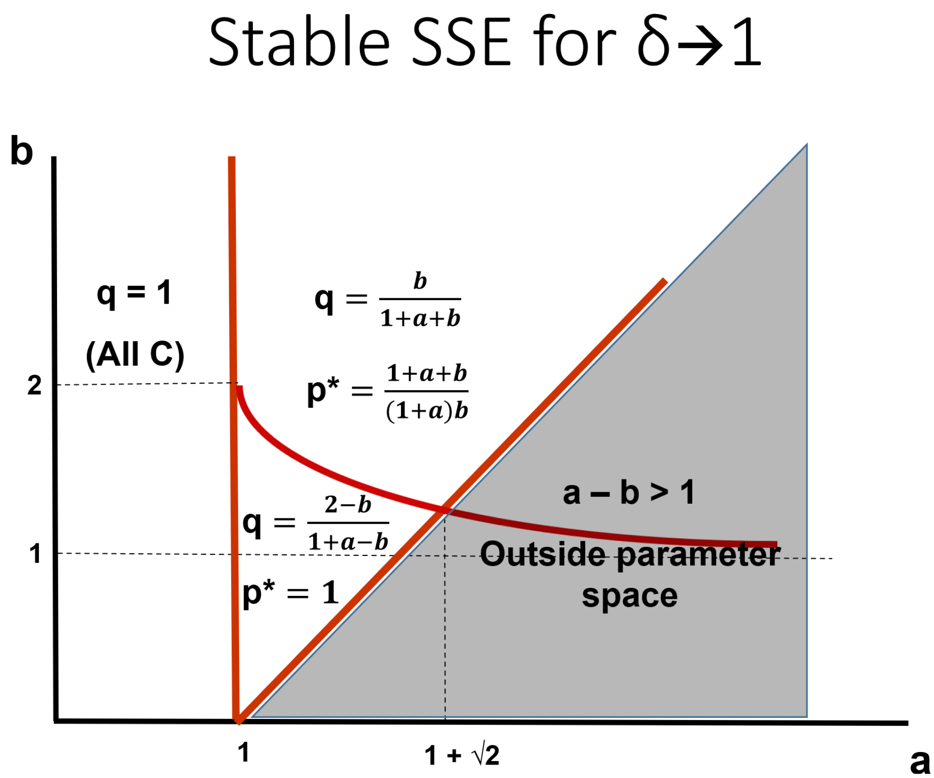


Figure 2: The partition in (a, b)-space corresponding to qualitatively distinct stable SSE for $\delta \to 1$.

co-existing at 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 lifetime, 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 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 $b = 1 + a$.

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, a, is high; patience is not necessarily enough to overcome this effect.

Finally, when $b > (1 + a)/a$, 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 from inlinks, and so total utility is captured by outlinks, which is equal to one per period 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 easily calculated as $1 - bp(1 - q)$. 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 claimed.

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 efficient relationships do not materialize. The idea behind exclusivity is the following. If cooperators accepted all proposed links, defection would be relatively more 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 SSE therefore involve the rejection of some proposed links. While this reduces the utility to all players, it also reduces 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 exclusivity is one half. That is, exclusivity is important to sustaining cooperation only when cooperators constitute a minority of society.[16]

In the limit case it is easy to explicitly solve for the stable SSE values of (q, p), so that we can make a few observations regarding the comparative statics of SSE outcomes as a function of parameters.

Proposition 4 *The following statements hold as $\delta \to 1$.*

- *In the region where there exists a nontrivial stable SSE with $p = 1$, q is decreasing in*

[16]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 $q \leq 1/2$.

a and b.

- *In the region where there exists a nontrivial stable SSE with $p < 1$, p and q are decreasing in a, p is decreasing in b and q is* increasing *in b.*

Proof. It is straightforward to compute the stable SSE values of (q, p) explicitly for $\delta \to 1$, obtaining the expressions in Figure 2. The conclusions are then immediate by inspection. □

For stable SSE without exclusivity ($p = 1$), the level of cooperation q behaves as expected, such that increases in either a or b reduce the sustainable level of cooperation. On the other hand, when cooperators are exclusive ($p < 1$), an increase in b actually increases the sustainable level of cooperation. But it also lowers p, so that cooperators form fewer links. It can be shown that the net impact on aggregate utility is negative.

All of the calculations upon which the limiting results rely are continuous in δ. That is, the boundaries depicted in Figure 2 shift continuously in δ. It is thus immediate that the bounds identified in Proposition 3 are approximately accurate for δ near one. We have also that the comparative statics in Proposition 4 are valid for δ near one. Finally, it can be shown that the stable SSE (p, q) are increasing in δ.

3.5 Comparison with bilateral Prisoner's Dilemma

Consistent strategies invoke trusting and unforgiving behavior which provides the harshest possible punishment for defection in our networked game. Hence the indefinitely repeated bilateral Prisoner's Dilemma (RPD), supported by grim trigger punishments, is a natural benchmark for comparison. In the RPD with payoff matrix

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

grim trigger punishments sustain two canonical on-path patterns of cooperation, each governed by a distinct one-shot deviation constraint. Full cooperation every period is sustainable if and only if

$$\delta \ \geq \ \delta_{\text{full}} \ = \ \frac{a}{1+a},$$

and alternating "turn-taking"—cycling deterministically between (D, C) and (C, D)—is sustainable if and only if

$$\delta \;\geq\; \delta_{\text{turn}} \;=\; \frac{b}{1+a}.$$

Both inequalities follow directly from the one–shot deviation principle under grim–trigger punishments. Stahl [1991, Prop. 1] derives the full–cooperation threshold and Stahl [1991, Prop. 4] identifies the threshold at which the alternating on–path outcome becomes sustainable; in our notation these are exactly the δ_{full} and δ_{turn} identified above. Sorin [1986, Props. 14–15; Fig. 1] identifies a similar result. In the knife–edge case $a = b$, the two constraints coincide, yielding a single critical discount factor.[17]

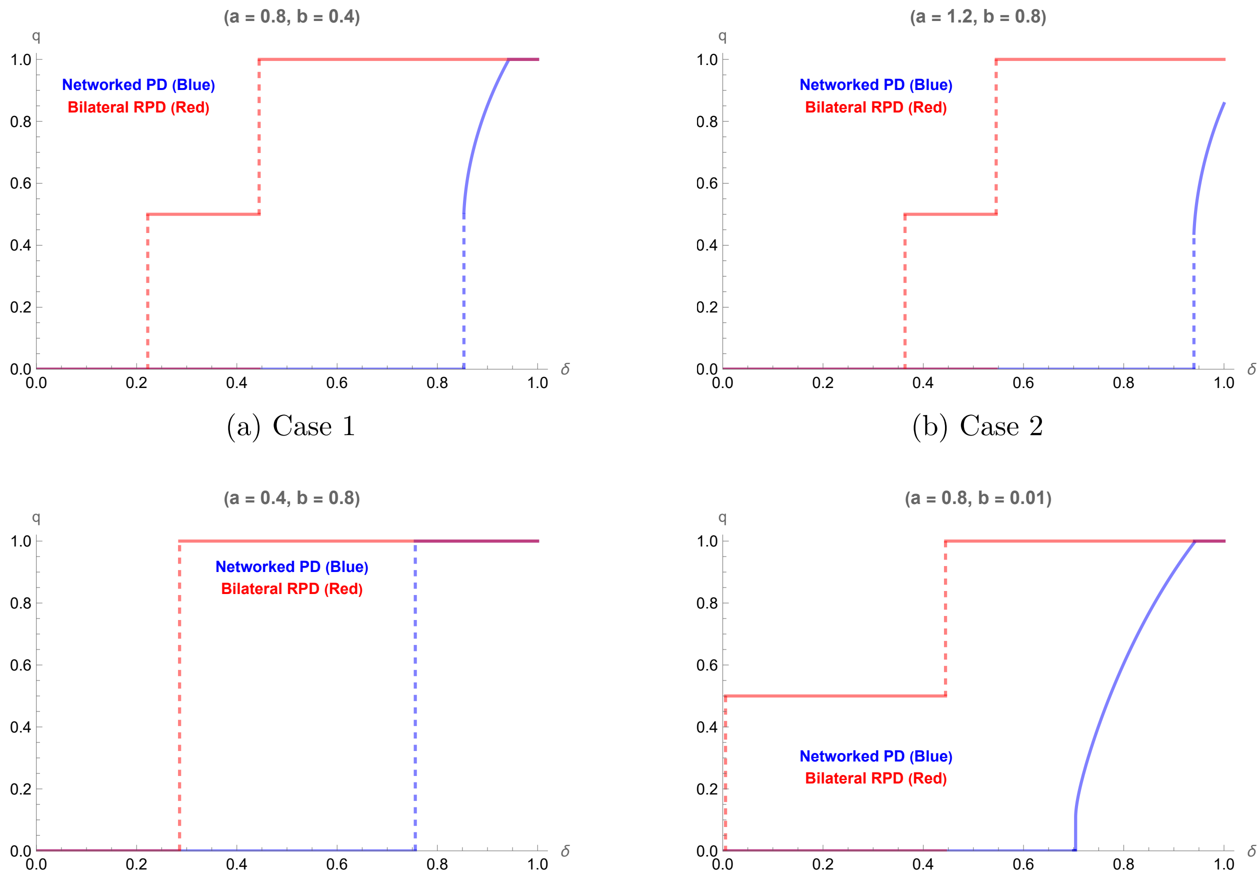
The resulting *cooperation frontier* of the RPD—that is, the maximal on-path cooperation rate achievable at equilibrium—then takes the form

$$q^{\max}_{\text{RPD}}(\delta) = \begin{cases} 0, & \delta < \min\{\delta_{\text{turn}}, \delta_{\text{full}}\}, \\ \frac{1}{2}, & \delta_{\text{turn}} \leq \delta < \delta_{\text{full}}, \\ 1, & \delta \geq \delta_{\text{full}}. \end{cases}$$

We use $q^{\max}_{\text{RPD}}(\delta)$ as a benchmark when comparing to our networked PD. For a given parameterization (a, b), let $q^*(\delta)$ denote the highest cooperation level supported by a stable SSE in our networked PD. Figure 3 plots $q^*(\delta)$ (blue) with $q^{\max}_{\text{RPD}}(\delta)$ (red) for a set of representative parameterizations.

In the bilateral RPD, maximal cooperation jumps discontinuously at the two critical values δ_{turn} and δ_{full}. In the networked PD there is likewise a critical threshold δ_{net} above which a stable SSE sustains positive cooperation. What happens at δ_{net} depends on (a, b). When b is large relative to a, the equilibrium can jump directly to full cooperation (Case 3). Otherwise, cooperation enters at an interior level $q_0 \in (0, 1/2)$—as in Cases 1, 2, and 4—and then increases continuously with δ, reaching 1 as $\delta \to 1$ if and only if $a < 1$ (Proposition 3). As $b \to 0$, this entry level q_0 can be made arbitrarily close to zero (Case 4). Intuitively, when the loss from being defected on is negligible, agents are willing to risk more defections while searching for cooperative partners. At the opposite extreme, as $a, b \to \infty$, both the networked and bilateral models require $\delta \to 1$ to sustain any cooperation.

[17] Full cooperation: deviating from C to D yields a one–period gain a and triggers (D, D) forever, so $a \leq \delta/(1-\delta)$, i.e., $\delta \geq \frac{a}{1+a}$. Turn–taking: on a C–turn, deviating to D avoids $-b$ (gain b) but forfeits next period's scheduled D–turn payoff $1 + a$ under grim trigger, so $b \leq \delta(1 + a)$, i.e., $\delta \geq \frac{b}{1+a}$.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 3: Maximal cooperation (q) in the Networked PD and the bilateral RPD as a function of the discount factor (δ).

This comparison highlights two distinctions. First, the RPD's cooperation frontier is governed by exactly two cutoffs—one for on-path alternation and one for full cooperation—whereas the networked environment typically admits interior cooperation levels that vary smoothly with patience. Second, because grim-trigger reversion in the bilateral game is a harsher punishment than the link severance discipline available in the network, the RPD supports cooperation at lower discount factors. This result is intuitive as the RPD allows for punishments that can force the value of a defector below the outside option of ending the relationship. Such punishments are not enforceable with voluntary separation.

4 Consistency as self-enforcing behavior

Simple stationary equilibria are defined under the restriction that agents must apply strategies that are consistent, which immediately implies the optimality of trusting and unforgiving behavior. 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. In order to do so, we allow for strategies that depend on an agent's history and beliefs. Because such strategies can not be characterized by a (p, q) pair, we necessarily adopt a more robust equilibrium concept. We begin this section by formalizing these ideas. We then show, under an appropriate parametric condition, that consistent behavior is self-enforcing in the sense that it constitutes play on an equilibrium path. In other words, behavior consistent with an SSE is also supported at equilibrium in the unrestricted game.

We then address the converse and characterize the set of equilibria that exhibit consistency on path. We show that in any equilibrium with consistent on-path strategies, the fraction of cooperators in society and the fraction of relationships accepted by cooperators must match those of an SSE. Finally, we analyze outcomes when our parametric condition fails and SSE outcomes can not be supported at equilibrium. In such cases, we construct an epsilon equilibrium with outcomes which approach stable SSE outcomes as the network becomes sufficiently dense.

4.1 Histories and actions

The global history H in a given round consists of all information generated in the course of the game that is payoff relevant to agents alive in that round. That is, for each living agent and each round of that agent's life, the global history records the cooperate/defect choices of the agent and each of his partners and the linking decisions involving the agent including: his proposed outlinks and whether they were accepted/rejected, inlinks proposed to him and whether he accepted/rejected them, and links involving him that were severed and whether this was because the corresponding partner died. Let $\mathcal{H}$ be the set of all possible global histories.

The private history H_i^s of an agent i at age s is a mapping from the global history H to a masked version containing only the information acquired by i up to age s. In particular,

i does not observe the identity of his relationship partners, just the cooperate/defect and linking actions they have taken.

Let $\mathcal{H}_i^s$ be the set of all potential private histories of i at each age s and let $\mathcal{H}_i = \cup_s \mathcal{H}_i^s$ be the set of all potential private histories.

The strategy of an agent i is a mapping from its private history to (a probability distribution over) actions. In each round, i takes three separate actions: (i) the choice of cooperate or defect, (ii) the acceptance or rejection of proposed inlinks, and (iii) the severance or continuation of each active link. Let $A_i^s(H_i^s)$ be the action set of agent i at age s with history H_i^s, $\mathcal{A}_i^s = \cup_{H_i^s \in \mathcal{H}_i^s} A_i^s(H_i^s)$ denote the set of all history-dependent age-s action sets,[18] and $\mathcal{A}_i = \cup_s \mathcal{A}_i^s$ 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 chooses whether to cooperate or defect, and whether to accept or reject inlinks, he has the information H_i^{s-1}, but he has not observed the cooperate/defect actions in the current round. When i makes the choice of severing active links, he has observed, additionally, the cooperate/defect actions in the current round of each of his active partners, inlinks that were proposed to him, and the accept/reject decisions on his outlinks if he proposed any. 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.

4.2 Equilibrium concept

We adopt a refinement of perfect Bayesian equilibrium (PBE) as our solution concept in this unrestricted game. Roughly speaking, a PBE is a pair of strategy and belief for each agent that is self-consistent. That is, at all valid histories, (1) an agent's strategy should maximize his expected utility given his belief when other agents follow their equilibrium strategies, (2) the belief should be Bayes-consistent with his observations and with the belief that all agents follow their equilibrium strategies, and (3) when faced with an unexpected (zero probability under the belief) history, the agent's belief should map to a limit point of beliefs under a vanishing tremble probability on the equilibrium strategies. Our refinement additionally requires that the beliefs of agents at birth be consistent with the limiting distribution of global histories induced by the dynamics where all agents apply their equilibrium strategies.

[18]The action space is history dependent as links change over time.

We now describe this notion and our refinement in more detail, tailored to our setting. Given any set S, we will write $\Delta(S)$ for the set of probability distributions over S. Note that $\mathcal{H}_i = \mathcal{H}_j$ for all agents i and j, i.e., the potential private histories are uniform across the population. Denote this set by $\mathcal{H}^*$. A *belief* for agent i is a function $\beta_i : \mathcal{H}^* \to \Delta(\mathcal{H})$ that maps each observed history to a distribution over possible global histories. We focus our analysis 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 agents i.

Given strategies ϕ, ϕ'_i and global history H, we write $h_i^\phi(\phi'_i, H) \in \Delta(\mathcal{H}^*)$ 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 global history H. We extend h_i^ϕ to accept a distribution over global histories in the natural way. Let $u_i(h_i^\phi(\phi'_i, \beta(H_i^s)))$ denote the expected continuation utility of agent i, according to his beliefs. We then say that ϕ is *optimal given belief* β if, for all $H_i^s \in \mathcal{H}^*$,

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

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

Given a strategy ϕ, belief $\beta(H_i^{s-1})$, and the private observation h_i^s of agent i corresponding to the current round s only, let $P^\phi(\beta(H_i^{s-1}), h_i^s)$ be the distribution over global histories that results after a single round of play of strategy ϕ starting from a history $H \sim \beta(H_i^{s-1})$ and conditioned on agent i observing h_i^s in the current round. Note that this distribution is well-defined: one can consider the probability of h_i^s given H and apply Bayes' rule. We say that β is *consistent with strategy* ϕ if, for all agents i, ages $t(i) = s$, private histories H_i^{s-1}, and current round observations 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 incompatible 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, are not appropriately restricted off the equilibrium path. This motivates us to require a form of perfection. Given an unexpected history H_i^s that has zero probability under ϕ, we shall require agents to place belief in a minimal number of deviations from ϕ that yield a state consistent with H_i^s. 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 β^ϵ 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 perfect Bayesian equilibria (PBE) in our setting.

Definition 4 *A pair (ϕ, β) is a* perfect Bayesian equilibrium *if ϕ is optimal given β, and β is robustly consistent with ϕ.*

Such a PBE always exists. For example, the strategy ϕ that maps every history to "always defect", together with the belief β that in the past everyone has always defected and rejected/severed all links, is a trivial PBE. Also, 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 $\epsilon \to 0$. Hence there is a natural notion of approximate PBE in our setting, which we will use in subsection 4.5.

Equilibria as defined above do not explicitly restrict the beliefs of agents at birth. We will focus on a refinement of PBE in which the beliefs at birth are required to be consistent with the global history in the limit. More formally, given a strategy ϕ, let a *steady-state distribution with respect to ϕ*, $\Delta^\phi(\mathcal{H})$, be a distribution of global histories induced by a fixed point of the dynamics in which all agents apply ϕ indefinitely. Note that some strategies may have either zero or multiple fixed points. However, all (p, q)-simple strategies have exactly one fixed point as shown in Section 2. This is sufficient for our purposes of analyzing the survival of our SSE as equilibria of the unrestricted game. We can now define our refinement.

Definition 5 *A* steady-state perfect Bayesian equilibrium *is a perfect Bayesian equilibrium (ϕ, β) such that beliefs at birth are a steady-state distribution with respect to ϕ.*

Again, such an equilibrium always exists. For example, the equilibrium described in the preceding paragraph is a steady-state PBE. Furthermore, we have shown in Section 3 that the SSE derived from the consistency restricted game are optimal amongst (p, q)-simple strategies. Therefore, they constitute steady-state PBE of the consistency-restricted game when paired with appropriate beliefs. We know turn to deriving the conditions under which (p, q)-simple behavior remains optimal amongst all history-dependent strategies.

4.3 Existence of consistent equilibria

We begin by establishing conditions under which consistent play can be implemented on the path of an equilibrium. 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 **6** *The* consistency condition *is*

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

Since every (a, b, δ) can be mapped into a unique maximal 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. Moreover, we claim that the consistency condition is relatively mild, in the sense that it is not a much more demanding requirement than the sustainability of any level of cooperation. In this sense, even though the consistency condition is tight, it is a relatively weak requirement. The following result quantifies this assertion, by providing sufficient conditions under which (4) will hold.

Proposition 5 *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 online appendix. □

Notice in particular that if the stage game is supermodular ($b \geq a$) then all stable SSE satisfy the consistency condition. Since the consistency condition is sufficient for consistent play to be implementable on the path of an equilibrium (as we will show below), this verifies that under supermodularity, all of the SSE described by Proposition 2 survive as equilibria in which consistent, trusting, and unforgiving behavior is self-enforcing.

We now turn to showing that every SSE (p, q) that satisfies the consistency condition can also arise as an equilibrium outcome of the unrestricted game. 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.

Theorem **1** *Suppose that $(p, q) \in [0, 1]^2$ is an SSE at which the consistency condition is satisfied. Then there is an equilibrium in which 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. See online appendix. □

We prove the result in three parts. First, we show that trusting and unforgiving behavior is always a best response to consistent behavior. Second, we prove a technical lemma: in an SSE, for a cooperator with at least one link, it is strictly more profitable to continue cooperating than to defect and re-enter the market as a cooperator. It is here that we use the consistency condition, on which this lemma crucially depends. Finally, we complete the proof of Theorem 1 by showing that the existence of a profitable deviation away from consistent behavior must imply the existence of a profitable deviation in which an otherwise consistent cooperator defects exactly once.

4.4 Characterizing Consistent Equilibria

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 1). We next consider the converse: are these the only equilibria with consistent behavior on path? As it turns out, the answer is yes, regardless of whether the consistency condition holds.

THEOREM **2** *Suppose that there is an equilibrium such that, on path, all agents apply actions that are consistent, trusting, and unforgiving. Moreover, suppose that at this equilibrium q is a stationary level of cooperation and p is a stationary level of inlink acceptance by cooperating agents. Then (p, q) is an SSE.*

Proof.

We can assume that $q > 0$, as otherwise the equilibrium is necessarily the all-defect equilibrium which is also an SSE. Suppose (p, q) is not an SSE. Then there is a profitable deviation within the context of strategies constrained to being consistent. We will use this to construct a deviation within the game that does not presuppose consistent play. Suppose the profitable deviation is to $(p', q') \neq (p, q)$. Consider then the strategy that selects cooperation at birth with probability q', is trusting and unforgiving, plays consistently thereafter, and subject to selecting cooperation accepts each inlink independently with probability p'. Since $q > 0$, this play remains on path for all other agents, and therefore (by assumption) all other agents will play in a manner consistent with the SSE (p, q). We conclude that the utility enjoyed by the deviating strategy is precisely the utility in the game that imposes consistent play, and is therefore a profitable deviation within the unconstrained game. □

4.5 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 1). We now take up the task of analyzing behavior when the consistency condition fails.

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 6 *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 online appendix. □

Recall that Theorem 1 shows that an SSE at which the consistency condition is satisfied supports consistent cooperation in equilibrium. Proposition 6 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.[19]

Our last result is relevant for SSE at which the consistency condition fails. By Proposition 5, this is possible only when $q < 1$, $p = 1$ and $b < a$. Proposition 6 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 there is little social capital to lose by defecting, and effectively starting the process again. We formalize this intuition in Proposition 7 below, by showing that such a bank-robbing strategy constitutes an ϵ-equilibrium when the consistency condition fails. This ϵ-equilibrium is 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.[20]

[19] 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 (5) is strictly positive.

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

Proposition 7 *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 online appendix. $\square$

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 6. 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 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.

5 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 agents refuse to maintain connections with those who defect. 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 agents are sufficiently long-lived (i.e., patient), but also that the temptation payoff for defecting not be too large. When these conditions are not both met, the presence of some level of non-cooperative behavior 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 defection persists, we can address through comparative statics what kinds of policies increase the level of cooperation, and the total welfare of the system, whereas most related work falls either into the category of constructing equilibria for which full cooperation is sustainable, or else analyzes models in which some form of inefficiency is inevitable. In a sense, our framework allows for a more balanced description of the achievable level of cooperation in a society. This characterization is specifically due to the novel tradeoff in our model: that between the immediate gains to defection and the long-run benefit of accumulating social capital in the form of sustained relationships with cooperators.

When full cooperation does not obtain, the presence of defectors causes relationships between cooperators to be a scarce and valuable resource, which we identify as a form of social capital. 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 inconsistent behavior is allowed. In other words, 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. The condition, which is satisfied for many parameter values, requires that the returns to links with cooperators are higher for cooperators than for defectors, so that persistent cooperation is sequentially rational. Finally, we show that when this consistency condition fails, a "rob the bank" strategy, in which a cooperator deviates when he has a sufficiently low level of social capital, forms an epsilon equilibrium when players are patient and can maintain many links.

We have made some of the modeling choices to emphasize the point that it is possible to maintain cooperation even under unfavorable conditions. 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 are not fully anonymous. This generally allows harsher punishments for defection, thereby increasing the scope for cooperation.

We hope that the central elements in our framework will prove useful in further research.

In particular, the combination of a random matching process and 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 [2015] 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 conjecture that the simple consistent equilibria that we identify generate maximal welfare among all equilibria.

By Theorem 2 any equilibrium with more cooperation than in an SSE must involve inconsistent play. The possibility of inconsistent play lowers the value of mutually cooperative partnerships. Inconsistent play may also require departures from trusting and unforgiving behavior which similarly lower the expected duration (and value) of mutually cooperative relationships and increase the expected value of defection respectively. The conjecture is difficult to prove with our techniques as, in general, equilibria may not entail a steady-state.

References

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

S. Nageeb Ali and David A. Miller. Ostracism and forgiveness. *American Economic Review*, 106(8):2329–2348, 2016. doi:10.1257/aer.20130768.

Carlos Alós-Ferrer. Dynamical systems with a continuum of randomly matched agents. *Journal of Economic Theory*, 86(2):245–267, 1999. doi:10.1006/jeth.1999.2522.

Jeremy Blackburn, Ramanuja Simha, Nicolas Kourtellis, Xiang Zuo, Matei Ripeanu, and Adriana Iamnitchi. Cheaters in the steam community gaming social network. *arXiv preprint arXiv:1112.4915*, 2012.

Simon Board. Relational contracts and the value of loyalty. *American Economic Review*, 101(7):3349–3367, 2011. doi:10.1257/aer.101.7.3349.

Simon Board and Moritz Meyer-ter Vehn. Relational contracts in competitive labor markets. *Review of Economic Studies*, 82(2):490–534, 2015. doi:10.1093/restud/rdu036.

Richard T. Boylan. Laws of large numbers for dynamical systems with randomly matched individuals. *Journal of Economic Theory*, 57(2):473–504, 1992. doi:10.1016/0022-0531(92)90046-K.

Luca Dall'Asta, Matteo Marsili, and Paolo Pin. Collaboration in social networks. *Proceedings of the National Academy of Sciences*, 109(12):4395–4400, 2012. doi:10.1073/pnas.1105757109.

S. Datta. Building trust. Mimeo, Indian Statistical Institute, 1993.

G. Ellison. Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies*, 61(3):567–588, July 1994. doi:10.2307/2297904.

Constanza Fosco and Friederike Mengel. Cooperation through imitation and exclusion in networks. *Journal of Economic Dynamics and Control*, 35(5):641–658, 2011. doi:10.1016/j.jedc.2010.12.002.

Takako Fujiwara-Greve and Masahiro Okuno-Fujiwara. Voluntarily separable repeated prisoner's dilemma. *Review of Economic Studies*, 76(3):993–1021, 2009. doi:10.1111/j.1467-937X.2009.00539.x.

Takako Fujiwara-Greve, Naoko Nishimura, Masahiro Okuno-Fujiwara, and Nobue Suzuki. Voluntary partnerships and cooperation: An experimental study. *SSRN Electronic Journal*, 2023. doi:10.2139/ssrn.4521316.

Andrea Galeotti, Sanjeev Goyal, Matthew O. Jackson, Fernando Vega-Redondo, and Leeat Yariv. Network games. *The Review of Economic Studies*, 77(1):218–244, 2010. doi:10.1111/j.1467-937X.2009.00570.x.

P. Ghosh and D. Ray. Cooperation in community interaction without information flows. *The Review of Economic Studies*, 63(3):491–519, July 1996. doi:10.2307/2297892.

Matthew O. Jackson and Brian W. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007. doi:10.1257/aer.97.3.890.

Matthew O. Jackson and Alison Watts. Social games: Matching and the play of finitely repeated games. *Games and Economic Behavior*, 70(1):170–191, 2010. doi:10.1016/j.geb.2008.02.004.

Matthew O. Jackson, Tomás Rodríguez-Barraquer, and Xu Tan. Social capital and social quilts: Network patterns of favor exchange. *American Economic Review*, 102(5):1857–1897, 2012. doi:10.1257/aer.102.5.1857.

Kenneth L. Judd. The law of large numbers with a continuum of IID random variables. *Journal of Economic Theory*, 35(1):19–25, 1985. doi:10.1016/0022-0531(85)90059-6.

M. Kandori. Social norms and community enforcement. *Review of Economic Studies*, 59: 63–80, 1992. doi:10.2307/2297925.

D. Krackhardt. Social networks and the liability of newness for managers. In C. Cooper and M. Rousseau, editors, *Trends in Organizational Behavior*, volume 3, chapter 9, pages 159–173. John Wiley & Sons Ltd, New York, NY, 1996.

R. Kranton. The formation of cooperative relationships. *Journal of Law, Economics, and Organization*, 12(1):214–233, April 1996. doi:10.1093/oxfordjournals.jleo.a023358.

Natalie Lee. An experiment: Voluntary separation in indefinitely repeated prisoner's dilemma game. *SSRN Electronic Journal*, 2020. doi:10.2139/ssrn.3207669.

Jonathan Levin. Multilateral contracting and the employment relationship. *The Quarterly Journal of Economics*, 117(3):1075–1104, 2002. doi:10.1162/003355302760193968.

W. Bentley MacLeod and James M. Malcomson. Motivation and markets. *American Economic Review*, 88(3):388–411, 1998.

Paolo Pin and Brian W. Rogers. Cooperation, punishment and immigration. *Journal of Economic Theory*, 160:72–101, 2015. doi:10.1016/j.jet.2015.07.002.

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

Arno Riedl and Aljaz Ule. Exclusion and cooperation in social network experiments. Mimeo, University of Amsterdam, 2002.

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

J. Sobel. A theory of credibility. *Review of Economic Studies*, 52:557–573, 1985. doi:10.2307/2297732.

J. Sobel. Can we trust social capital? *Journal of Economic Literature*, 40(1):139–154, March 2002. doi:10.1257/jel.40.1.139.

Sylvain Sorin. On repeated games with complete information. *Mathematics of Operations Research*, 11(1):147–160, 1986. doi:10.1287/moor.11.1.147.

Dale O. Stahl. The graph of prisoners' dilemma supergame payoffs as a function of the discount factor. *Games and Economic Behavior*, 3(3):368–384, 1991. doi:10.1016/0899-8256(91)90035-D.

F. Vega-Redondo. Building up social capital in a changing world. *Journal of Economic Dynamics and Control*, 30(11):2305–2338, November 2006. doi:10.1016/j.jedc.2005.07.004.

Jing Wang, Siddharth Suri, and Duncan J. Watts. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36): 14363–14368, 2012. doi:10.1073/pnas.1120867109.

J. Watson. Starting small and renegotiation. *Journal of Economic Theory*, 85:52–90, 1999. doi:10.1006/jeth.1998.9999.

J. Watson. Starting small and commitment. *Games and Economic Behavior*, 38:176–199, 2002. doi:10.1006/game.2001.0857.