Emergence of social networks via direct and indirect reciprocity

Steve Phelps

January 9, 2012

Abstract

Many models of social network formation implicitly assume that network properties are static in steady-state. In contrast, actual social networks are highly dynamic: allegiances and collaborations expire and may or may not be renewed at a later date. Moreover, empirical studies show that human social networks are dynamic at the individual level but static at the global level: individuals' degree rankings change considerably over time, whereas network-level metrics such as network diameter and clustering coefficient are relatively stable. There have been some attempts to explain these properties of empirical social networks using agent-based models in which agents play social dilemma games with their immediate neighbours, but can also manipulate their network connections to strategic advantage. However, such models cannot straightforwardly account for reciprocal behaviour based on reputation scores ("indirect reciprocity"), which is known to play an important role in many economic interactions. In order to account for indirect reciprocity, we model the network in a bottom-up fashion: the network *emerges* from the low-level interactions between agents. By so doing we are able to simultaneously account for the effect of both direct reciprocity (e.g. "tit-for-tat") as well as indirect reciprocity (helping *strangers* in order to increase one's reputation). This leads to a strategic equilibrium in the frequencies with which strategies are adopted in the population as a whole, but intermittent cycling over different strategies at the level of individual agents, which in turn gives rise to social networks which are dynamic at the individual level but stable at the network level.

1 Introduction

An understanding of the conditions under which cooperative outcomes are achieved by agents which maximise their own local objective functions is of great importance not only in economics and biology, but also in distributed artificial intelligence and multi-agent systems. Building large-scale systems comprised of many autonomous agents entails ensuring that the system as a whole is not undermined by incentives for uncooperative behaviours which are rewarding for individual agents, but which are harmful for the system as a whole.

Similar problems are faced in explaining the emergence large-scale cooperative structures in evolutionary biology; for example, genes cooperate to form regulatory networks; cells emerge from networks, multi-cellular organisms from cells and societies from organisms [26]. Hence there is a long tradition of modelling cooperation in social dilemmas using evolutionary models, such as the replicator dynamics, in which natural selection drives the choice of strategy.

A combination of both theoretical and empirical work has shown the importance of various forms of *reciprocal* behaviour in explaining cooperative outcomes. Reciprocity may be either direct, or indirect. Direct reciprocity entails rewarding or punishing other agents in order to elicit cooperation. When using direct reciprocity, agents condition their behaviour on personal experience of other agents — the archetypal example being Axelrod's tit-for-tat strategy [1]. On the other hand, indirect reciprocity occurs when agents cooperate with *strangers* in order to gain reputation (also referred to as an "image score"). This can lead to subsequent payoff from agents who cooperate with those with high reputation [14, 15].

Theoretical work has shown that large-scale cooperation can be sustained in large populations under natural selection through indirect reciprocity, provided that a critical threshold of the population adopts indirect reciprocity at the outset [14]. This is compatible with empirical studies which show that within groups of people, many have a tendency to be *strong* reciprocators [7]. In smaller populations, however, agents have a strong probability of repeatedly encountering each other and therefore we need to consider the possibility of competition between strategies for both direct and indirect reciprocity. In earlier work we showed that not only does the initial fraction of reciprocators determine the outcome, but that *both* direct and indirect reciprocity play a role in determining the evolutionary stable strategies that arise in smaller groups [20].

Earlier models of reciprocity considered a mean-field approach in which every agent has equal probability of interaction with any other agent [14, 15, 20]. In contrast, in reality many interactions occur in a more structured environment: for example, we might model agents as nodes on a graph who interact with their immediate neighbours capturing the fact that interactions occur within social networks. Thus there have been many models which analyse social dilemmas played on *graphs*.

Earlier research focused on exogenous graph-formation processes and analysed the conditions under which cooperation occurs, independently from the process which forms the graph itself. In contrast, empirical studies highlight that real world social networks are highly dynamic in nature [12]. This raises the question as to whether strategies themselves contribute to network structure in a scenario in which agents manipulate the network strategic for strategic advantage, and network structure in turn plays a role in determining which are the optimal strategies; network structure and strategy are both entangled in a co-evolutionary system [24]. In such models, agents play the usual normal-form social dilemma games with their immediate neighbours (for example the prisoner's dilemma), but they also have the

possibility to "rewire" their edges in order to strategically manipulate the network topology in their favour.

In such models the network topology is represented explicitly, and is modelled in a top-down manner. In contrast, the pairwise public-goods games which are used to model indirect reciprocity suggest an intriguing alternative possibility in which the network topology is an *emergent* structure which arises from the donations made from one agent to another, which can be visualised as a directed graph (for an example, see Fig. 1). In this paper, we adopt such an approach in which the network arises directly from the strategies chosen by the agents. Agents are not restricted a priori from interacting with non-neighbouring agents. Rather, we use the emergent network structure to propagate reputation information in order to explore the "information hypothesis" [2] which conjectures that the value of reputation information is not contingent on its origin. In contrast, empirical evidence suggests that people place more trust information from direct sources [9].

The structure of this paper is as follows. In the following section we give an overview of the related literature. In Section 3 we formally describe our model of reciprocity and the reinforcement learning model which agents use to adjust their strategy. In Section 4 we describe how we analyse this model empirically, and we discuss our results in Section 5. Finally we conclude in Section 7.

2 Background Related work

The simplest task environment for studying trust and cooperation between payoff maximising agents is the so-called Prisoners' Dilemma (PD). Defection is the dominant strategy in the one-shot version of the game, however Axelrod [1] provided empirical evidence suggesting that cooperative strategies could survive in an evolutionary version of an *iterated* version of the game. Most notably a strategy called *tit-for-tat* which copies the last move made by its opponent performed extremely well in an evolutionary tournament.

There have been a number of psychological studies of PD with human subjects [31] in which *tit-for-tat* like strategies are commonly observed to be actually used. However, Roberts and Sherrat [22] noted that *tit-for-tat* like strategies are not always observed in ecological field studies.

Van Vught et al. [29] provide an overview of some of the central problems involved in building models of large-scale group formation that are both evolutionary and cognitively plausible. The key problem with larger groups is that it is more difficult to selectively retaliate against uncooperative behaviour, and thus cooperative equilibria are not stable [4, 21]; strategies like *tit-for-tat* or *raise-the-stakes* are not sufficient on their own to prevent free-riding in larger groups.

Van Vught et al. postulate that *reputation systems* are a necessary prerequisite of evolutionary-stable

cooperation in large groups. Reputation together with pressure to join profitable coalitions can result in "conspicuous altruism", also known as *indirect reciprocity*: that is, being generous to strangers in order to gain a good reputation, thus allowing entry into profitable coalitions. Nowak and Sigmund [14, 15] study the effect of reputation (which they call "image scoring") in a coalitional version of the prisoner's dilemma game using a combination of evolutionary simulation and mathematical analysis, and find that indirect reciprocity is likely to be widely adopted.

In these models randomly chosen pairs of agents are drawn from a larger population. One of these agents is designated as the donor and may choose to invest a certain amount in their partner. This results in a negative fitness payoff $-\gamma$ to the donor, and a positive fitness payoff $m \cdot \gamma$ to their partner.

Since we are interested in how cooperation can emerge in societies of selfish agents, we must analyse outcomes in which agents attempt to choose values of γ that maximise their own payoff. Provided that $m > 1$, over many bouts of interaction it is possible for agents to enter into reciprocal relationships that are mutually-beneficial, since the initial cost γ may be reciprocated with $m \cdot \gamma$ yielding a net benefit $mc - \gamma = m(\gamma - 1)$. Provided that the agents trust each other to reciprocate, they can increase their net benefit by investing larger values of γ. However, by increasing their donation they put themselves more at risk from exploitation, since just as in the alternating prisoner's dilemma [16], defection is the optimal strategy if the total number of bouts is known: the optimal behaviour is to accept the benefits without investing in return. In the case where the length of the game is *unknown*, and the number of agents is $n = 2$, it is well known that conditional reciprocation is one of several optimal solutions in the form of the so-called "tit for tat" strategy which copies the action that the opposing agent chose in the preceding bout. Roberts and Sherrat [22] demonstrate that a similar strategy, called *raise the stakes*, applies in the continuous game. This strategy plays cautiously against agents it does not trust, but invests generously in agents with a history of reciprocation. However, their result does not generalise to larger groups $n > 2$.

Nowak and Sigmund [15] demonstrated that reciprocity can emerge *indirectly* in large groups, provided that information about each agent's history of actions is summarised and made publicly available in the form of a reputation or "image-score" which summarises the propensity-to-cooperate of any given agent based on their history of actions. Provided that the initial population already contains a certain threshold of reciprocators, *discriminatory* strategies (that is, strategies that invest conditionally on a partner's image-score) are evolutionary-stable, and that this leads to indirect-reciprocity; agents help others not because they expect direct reciprocation from their partner, but because by increasing their image-score they will receive reciprocal donations indirectly from third parties.

Our work seeks to address two key questions arising from this kind of model. Firstly, Nowak and Sigmund's model assumes that the population is very large relative to the "viscosity" of interactions —

that is, the frequency with which agents interact with each other before reproducing. This makes their model more tractable to analytic techniques, and more importantly it implicitly rules out any possibility that strategies based on *direct* reciprocity might affect the outcome, since the probability of encountering the same agent before reproducing is negligible. In smaller groups, however, the possibility arises that strategies based on direct reciprocity and strategies based on indirect reciprocity might interact. In earlier work, we showed that within smaller populations *both* types of reciprocity contribute to the mix of evolutionary stable strategies in an evolutionary game-theoretic model [20].

Secondly, the transitive aspect of indirect reciprocity suggests an implicit *network* structure to the interactions within these models. For example, if A helps B who helps C who then helps A we can visualise these interactions as a directed graph. The natural question then arises as to whether this *emergent* network structure can be used to strategic advantage by the agents. For example, Granovetter [9] posits that the *source* of reputation information may be an important factor in an agent's decision, and thus it may be important to discount reputation information according to the social proximity of other agents in the emergent network.

The use of social network models in the analysis of social dilemmas is a well established area of research. Existing models can be classified according to whether the network formation process is exogenous or endogenous. In the former case, non-strategic models of network formation such as preferential attachment are used to initialise a network and subsequently agents play social dilemma games with their immediate neighbours. Santos et al. [23] showed that whether or not cooperation prevails depends on the topology of the network and that small-world networks formed using models such as preferential attachment lead to much greater cooperation. Ohtsuki et al. [17] generalised this result showing that natural selection favours cooperation if the benefit of the altruistic act divided by the cost exceeds the average number of neighbours on the network.

However, these models assume that the network itself is not subject to strategic manipulation — rather it is formed through some exogenous process and remains static during strategic interactions. In contrast, real-world social networks are highly dynamic [12]: allegiances and collaborations expire and may or may not be renewed at a later date. In order to account for this Santos, Pacheco and Lennaerts [24] analysed a model in which agents were also able to strategically rewire their connections resulting a co-evolution between the social network topology and the strategies used in the social dilemma: agents play the usual social dilemma with their neighbours and may make use of strategies such as tit-for-tat which are based on direct reciprocity, but can also strategically choose whether or not to rewire an edge replacing an existing partner X with one of X's neighbours selected at random.

These models are able to capture interactions based on direct reciprocity, but cannot incorporate the indirect nature of interactions based on reputation, since the social dilemma is restricted to intra-

network interactions. However, indirect reciprocity requires that agents are specifically able to seek out those with good reputation *regardless of their social proximity*; indeed, indirect reciprocity is the basis of modern electronic e-commerce systems which make use seller feedback in order to encourage people to trade with *strangers* [2].

In order to deal with these considerations we introduce a model in which boundedly-rational agents can choose between strategies based on *both* forms of reciprocity (in addition to unconditional defection or cooperation). Within this framework, we still allow for social network effects by modelling the network in a bottom-up fashion: the network *emerges* from the low-level interactions between agents, as formalised in the next section.

3 The Model

As in earlier models, agents invest in partners at a cost to themselves, but recipients receive a multiple $m > 1$ of the original donation. If agents reciprocate then all parties to the interaction are better off than they would have been acting alone. However, as with the prisoner's dilemma there is a temptation to defect by accepting donations from other agents without investing in turn.

We generalise earlier models by allowing agents to divide up their initial endowment γ into a *portfolio* of donations in the other agents.

At each time step t every agent $a_i \in \{a_1, a_2, \ldots a_n\}$ simultaneously chooses a portfolio vector:

$$\boldsymbol{P}^t_{i,*} = (w_1, w_2, \ldots w_n) \tag{1}$$

$$\begin{aligned} p^t_{i,j} &\in [0,1] \subset \mathbb{R} \; \forall_{i,j} && (2)\\ p_{i,i} &= 0 \; \forall_i && (3)\\ \sum_{j=0}^{n} p^t_{i,j} &\leq 1 \; \forall_i && (4) \end{aligned}$$

The weights $w_1, w_2 \ldots w_n$ represent how the initial endowment γ of agent a_i is to be split across the other agents $a_1, a_2 \ldots a_n$. Accordingly, the matrix of donations between agents at time t is given by

$$\boldsymbol{C}^t = \gamma \boldsymbol{P}^t \, , \tag{5}$$

and the payoff to agent a_i by

$$u_i^t = \sum_{j=1}^{n} m \cdot p_{j,i}^t - \sum_{k=1}^{n} p_{i,k}^t \ . \tag{6}$$

An agent i might choose to invest only a fraction, or none of their endowment in the rest of the community (equation 4). As in other models we make this information publicly available in the form of a reputation score $r_i^t \in [0,1] \subset \mathbb{R}$:

$$r_i^t = \sum_{j=1}^{n} \boldsymbol{C}_{i,j}^t \tag{7}$$

We allow agent a_i to condition their donation decision $\boldsymbol{P}_{i,*}^t$ on the reputation of other agents (indirect reciprocity) as well as the history of donations received (direct reciprocity). In each case, we use an exponential moving average to summarise this time series and give more weight to recent values.

The exponential moving average of the donations between agents is represented by the matrix $\boldsymbol{C}^t$:

$$\bar{c}_{i,j}^t = \max(\,\kappa,\ \alpha \cdot c_{i,j}^t + (1-\alpha) \cdot \bar{c}_{i,j}^{t-1}\,) \tag{8}$$

where κ is a threshold parameter normalised with respect to the population size, endowment and multiplier: $\kappa = \frac{\gamma \cdot m}{4n}$.

We can visualise the matrix $\bar{\boldsymbol{C}}^t$ as a weighted directed graph representing the social network that emerges from the donations between agents. Fig. 1 shows an example of the social network produced by our simulation model where the vertices represent agents and the labels on the directed edge between any two nodes i and j correspond to the value $\bar{c}_{i,j}^t$.

This emergent network plays an actual causal role in our model, since agents can use it to discount the reputation of other agents according to their distance in the social network, thus modelling the idea that information from direct sources may be implicitly more trustworthy more than information from strangers [9]. The vector $\bar{\boldsymbol{R}}^t$ contains exponential moving average of agents' reputations, without taking into account any effect of network distance:

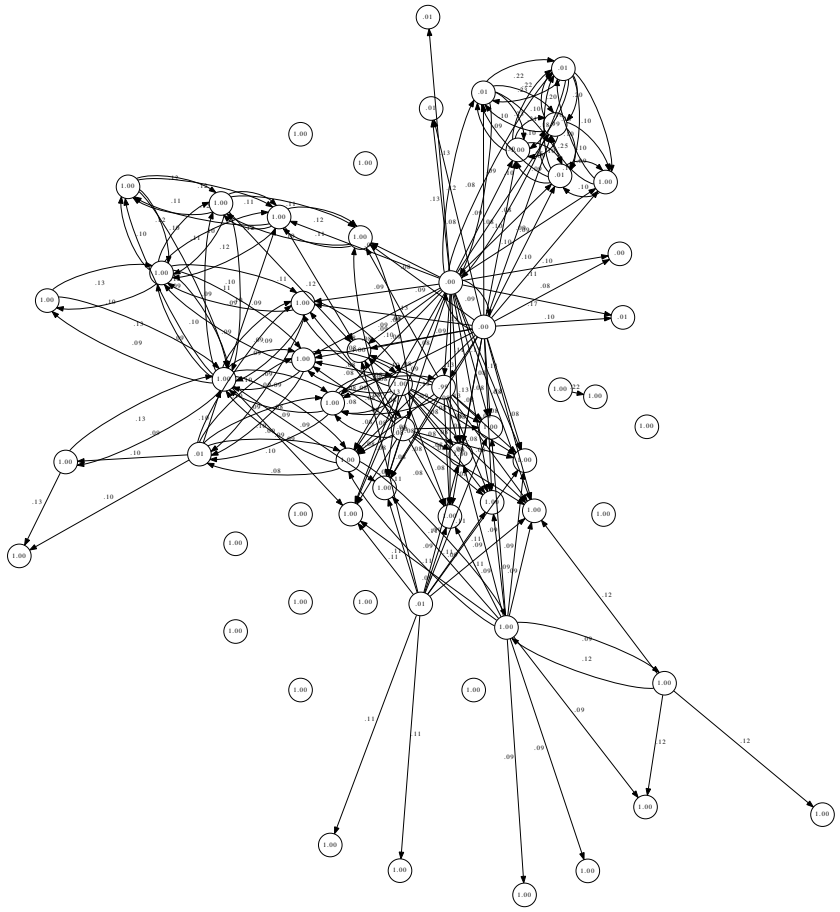
$$\bar{r}_i^t = \alpha \cdot r_i^t + (1-\alpha) \cdot \bar{r}_i^{t-1} \ . \tag{9}$$

On the other hand, the *networked* version of these reputation scores is given by the matrix $\boldsymbol{\Phi}^t$

$$\phi_{i,j}^t = \frac{\bar{r}_j^t}{d_{i,j}} \tag{10}$$

where $d_{i,j}$ is the shortest path from i to j on the graph defined by $\bar{\boldsymbol{C}}$. Agents can use either form of measure in making their donation decisions.

Figure 1: An example social network that arises through donations



The above figure shows a snapshot at one moment in time of the social network that arises from the donations between agents in a population of $n = 60$ agents. Each node on the directed graph represents an agent. The directed edges represent donations from one agent to another, with the edge label representing the current exponential moving average of the donation $\bar{c}^t_{i,j}$ (equation 8). The labels inside each node represent the current reputation score of the agent r^t_i.

3.1 Strategies

We analyse populations of agents choosing amongst the following set of strategies: the cooperate strategy (C), the defect strategy (D), the reputation weighted strategy (RW) and the tit-for-tat strategy (T4T). These are described and formalised below.

An agent a_i using the defect strategy (D) accepts donations without any reciprocation:

$$p_{i,j}^t = 0 \; \forall_{a_j \in A} \; . \tag{11}$$

An agent a_i using the cooperate strategy (C) unconditionally donates its endowment equally across the rest of the population:

$$p_{i,j}^t = \frac{1}{n-1} \; \forall_{a_j \in A : j \neq i} \; . \tag{12}$$

An agent a_i using the reputation weighted strategy (RW) distributes its endowment amongst the rest of the population in proportion to the other agents' reputation scores (as defined by Equation 9):

$$p_{i,j}^t = \frac{\bar{r}_{i,j}^{\,t-1}}{\sum \bar{\boldsymbol{R}}_{i,*}^{\,t-1}} \; \forall_{a_j \in A : j \neq i} \; .$$

An agent a_i using the reputation weighted networked strategy (RWN) distributes its endowment amongst the rest of the population in proportion to the other agents' *networked* reputation scores (as defined by Equation 10):

$$p_{i,j}^t = \frac{\phi_{i,j}^{t-1}}{\sum \boldsymbol{\Phi}_{i,*}^{t-1}} \; \forall_{a_j \in A : j \neq i} \; .$$

An agent a_i using the tit-for-tat strategy (T4T) distributes its endowment in proportion to the moving average of inward donations:

$$p_{i,j}^t = \frac{\bar{c}_{j,i}^{\,t-1}}{\sum \bar{\boldsymbol{C}}_{*,i}^{\,t-1}} \; .$$

3.2 Learning

Agents use a simple reinforcement learning algorithm based on Q-learning [30] in order to select between the above strategies. The central idea is that each agent attempts to estimate the expected payoff through an inductive sampling process in which the agent tries out different strategies and uses the payoff values thus obtained to estimate the expected payoff of each, and hence determine the strategy which will give the best long-term reward – the so-called greedy strategy. Similar models have been widely adopted in

modelling the behaviour that is empirically observed in strategic environments, both generally [6], and also specifically in the case of social dilemmas [8, 10].

The estimated payoff may depend not only on the strategy selected, but also the *state* of the system. In our model we approximate the state of the environment $\theta_{i,t}$ from the perspective of agent a_i at time t by rounding the agent's reputation score $\bar{r}_i$ into four possible values $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. This allows for the possibility that different strategies may be effective depending on whether or not the agent currently has a good or poor reputation.

The payoff estimates are held in a table of Q values which gives the current estimate for each strategy in each possible state, and these are updated according to the following equation:

$$\begin{aligned} Q_{i,t}(s_{i,t'}, \theta_{i,t'}) = \quad & \alpha \cdot [U_{i,t'} + \beta \cdot Q_{i,t}(s*_{i,t}, \theta_{i,t})] \\ & + (1 - \alpha) \cdot Q_{i,t'}(s_{i,t'}, \theta_{i,t'}) \end{aligned}$$

where $s_{i,t'}$ is the strategy that agent a_i played in period $t - 1$, α is the learning-rate parameter, β is the discount parameter and $s*_{i,t}$ is the greedy strategy of agent a_i. The above equation is simply a discounted exponential moving average of historical payoff samples. The recency parameter gives more weight to more recent samples and this takes into account that the environment may be highly dynamic, and thus we should give more weight to more recent information. In our case, the environment consists of other agents who are dynamically changing their behaviour which in turn will determine the expected payoffs.

If the true expected payoff to each strategy in each possible state were known a priori then a rational agent would always choose the greedy strategy $s*_{i,t}$ in order to maximise its payoff. However, because agents estimate payoffs through sampling, there is in an inherent trade-off in exploiting the current greedy action as opposed to exploring alternatives which may be proved to be more profitable once further samples are collected.

We perform experiments with two commonly-used exploration methods: epsilon-greedy selection versus softmax [28]. When using the epsilon-greedy method, at time t agent a_i plays the greedy strategy with probability $1 - \epsilon$. If the greedy strategy is not chosen the agent chooses at random between all available strategies with equal probability.

On the other hand, when using softmax action selection the probability of agent i choosing strategy a at time t' is given by

$$P(s_{i,t'} = a) = \frac{\exp(Q_{i,t}(a, \theta_{i,t})/\tau)}{\sum_b e^{Q_{i,t}(b)/\tau}} \tag{13}$$

The steady-state outcomes arising from the dynamics defined by this learning model are not necessarily equivalent to the game-theoretic equilibria implicit in the payoff structure. However, reinforcement learning models are attractive from a modelling perspective since they can both be grounded in theories of learning from cognitive psychology, and they have also been able to explain many deviations from game-theoretic that are empirically observed with real subjects [6]. Additionally, the learning-theoretic equilibria *can* be related to game-theoretic equilibria in certain cases [11]. This is an important point which we shall return to later.

4 Methodology

We analyse this model using empirical methods by simulating the agent-based model described in the previous section.

At the beginning of each simulation a minority fraction of the agents in the population sr are initialised without learning and are configured to adopt the reputation-weighted strategy (RW) unconditionally irrespective of the payoff received. In line with [3] we refer to this fraction of the population as *strong reciprocators*; these agents do not interact with those of poor reputation even if this leads to a reduction in their own payoff. The remainder of the population are configured to use the learning algorithm described in Section 3.2 and are free to switch between any strategy according to payoff.

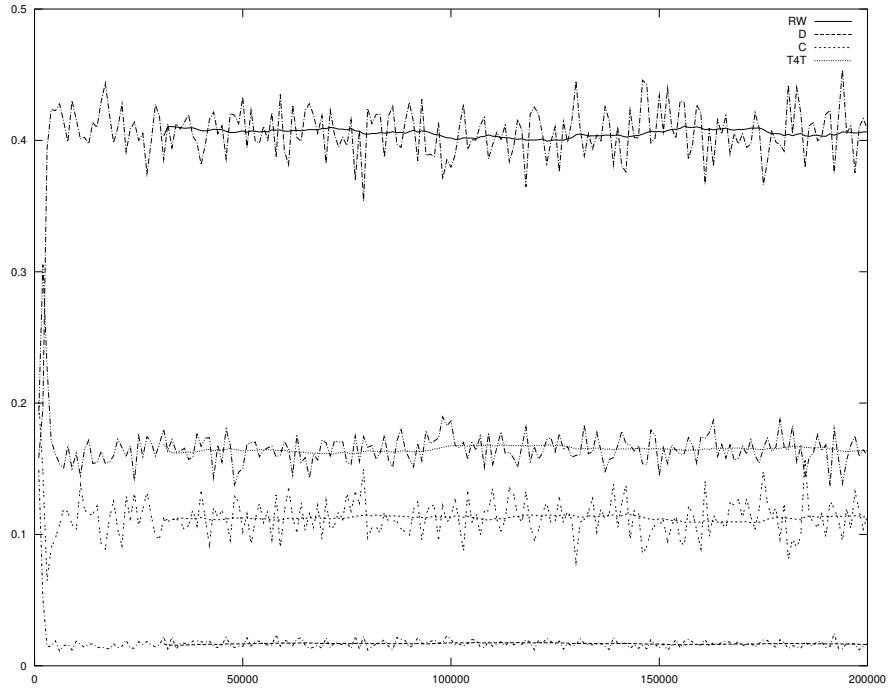
Parameter	*Distribution*	*Description*
ϵ	$\sim U(10^{-4}, 10^{-2})$	Experimentation
α	$\sim U(10^{-4}, 1 - 10^4)$	Recency
β	$\sim U(0.9, 1 - 10^4)$	Discount rate
Q_0	$\sim N(0, 100)$	Initial value estimate
n	$\in \{20, 60, 100\}$	Number of agents in the population
sr	$\in \{0,\ 0.05,\ \ldots,\ 0.4\}$	Proportion of strong reciprocators
m	$\in \{1.5,\ 2,\ 2.5,\ 3\}$	Multiplier

Table 1: Parameter settings
For each combination of the discrete parameter values below the line, we run 10^3 simulations with real valued variables drawn i.i.d. from the distributions specified above the line.

We run a total of 3.6×10^5 independent simulations with parameters configured according to Table 1; for each combination of the discrete parameter values, we run 10^3 simulations with real valued variables drawn *i.i.d.* from the distributions specified in the table.

We run each individual simulation for $t = 2 \times 10^5$ periods, taking the average reputation across the last 5×10^4 periods as our estimate of the level of cooperation in steady-state. We justify these time

Figure 2: Mean frequency of each strategy as a time series



Mean frequency of each strategy executed by the fraction of learning agents (excluding strong reciprocators) as a time series from $t = 0$ to $t = 2 \times 10^5$, sampled at intervals of 10^3. The strategies are, from top to bottom: the reputation weighted strategy (RW), the defect strategy (D), the cooperate strategy (C) and the tit-for-tat strategy (T4T). The population rapidly converges to a steady-state in which the frequencies fluctuate around a static mean value as illustrated by the 30-period moving average shown for each series.

intervals on the grounds that the population quickly converges to a steady-state: Fig. 2 shows how the frequency with which each strategy is executed varies over time for a typical run of the simulation.

We treat each independent run as a single observation in our data set. For each observation we record: all the values of random variates, the frequency with which each type of strategy (section 3.1) is adopted over the entire population of agents, and the mean reputation of the population in the steady-state time period $\bar{r}$. In order to separate the contribution of strong reciprocators, the resulting level of cooperation is also measured as the frequency with which cooperative strategies are chosen by the learning fraction of the population, which we denote Γ.

The model was implemented using the Java Agent-Based Modelling (JABM) framework [19], and the Mersenne Twister algorithm was used to draw all random values in the simulation [13]. All of the code required to run the simulations described in this paper is freely available under an open-source license [18].

We study our model under two different treatment conditions. Firstly, we analyse outcomes when learning is stateless and agents cannot condition their strategy on reputation scores, which we denote the

"stateless treatment". This treatment corresponds to the earlier models of Nowak and Sigmund, which do not allow agents to switch to an alternative strategy depending on whether they currently have good standing. In order to deal this criticisms we analyse our model under a second treatment in which each agent's reputation is discretized and used as a state value as described in Section 3.2, which we denote the "stateful treatment".

5 Results

We first analyse the stateless treatment in which agents choice of strategy depends only on payoff estimates and not reputation score. In line with other studies, we find that cooperation can be sustained in this model provided that the proportion of strong reciprocators sr is sufficiently high. Along with the multiplier m, these two parameters have the strongest effect on steady-state cooperation (Γ) with linear correlation coefficients of ≈ 0.54. In contrast, all other parameters have correlation coefficients $\leq 10^{-2}$ apart from the discount rate β. A multiple regression gives:

$$\Gamma = 0.29 \times m + 1.23 \times sr + 0.02 \times \beta - 0.44 \tag{14}$$

as the best linear model with $R^2 \approx 0.57$, suggesting that the sensitivity of our results to the β parameter is very small as compared with m and sr.

Closer inspection shows that the relationships are in fact non-linear. Fig. 3 shows the interaction between m and sr in determining the final level of cooperation which is shown as the mean value of Γ.

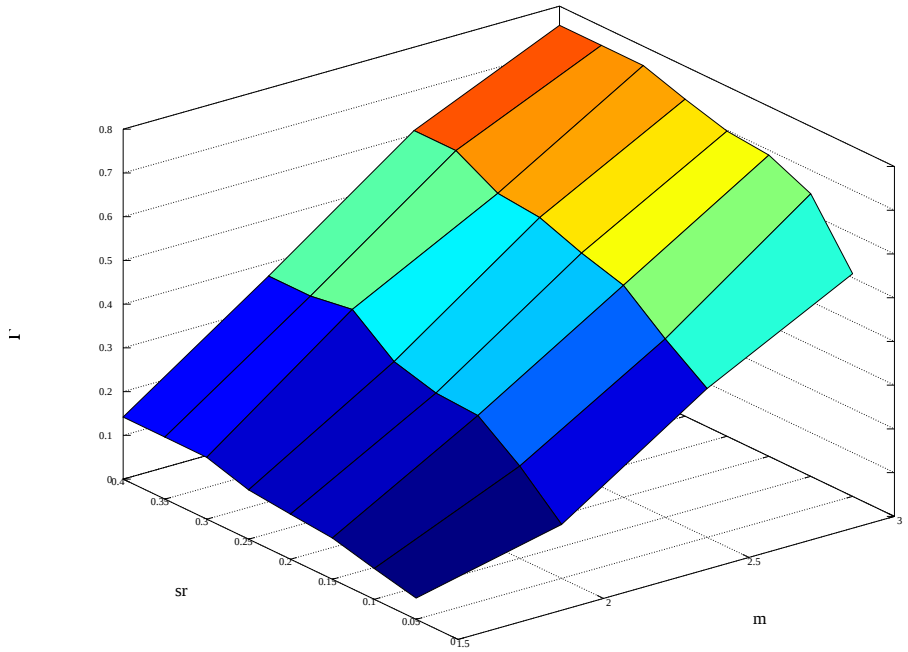
Fig. 4 shows a cross section of this surface showing $\bar{\Gamma}$ against sr when the multiplier is held constant at $m = 3$. The error bars show the confidence intervals for $p = 0.05$, and the additional plots below show the contribution made by each cooperative strategy.

As the proportion of strong reciprocators — sr — is increased the remaining fraction of the population responds by becoming more cooperative. Initially there is a relatively large response but this tails off at $sr \approx 0.2$. The results show that provided $sr \geq 5 \times 10^{-3}$ cooperative strategies will be chosen on average more than half of the time by the rest of the population.

Under this treatment it is apparent that the network appears to play relatively little role in determining agents' strategies; the frequency with which network-discounted reputation — RWN — is chosen over the standard reputation weighted strategy — RW — is virtually identical over the entire range of sr. This is also the case for all other values of the multiplier m. We also see that indirect reciprocity is more widely adopted than direct reciprocity, as represented by the T4T strategy.

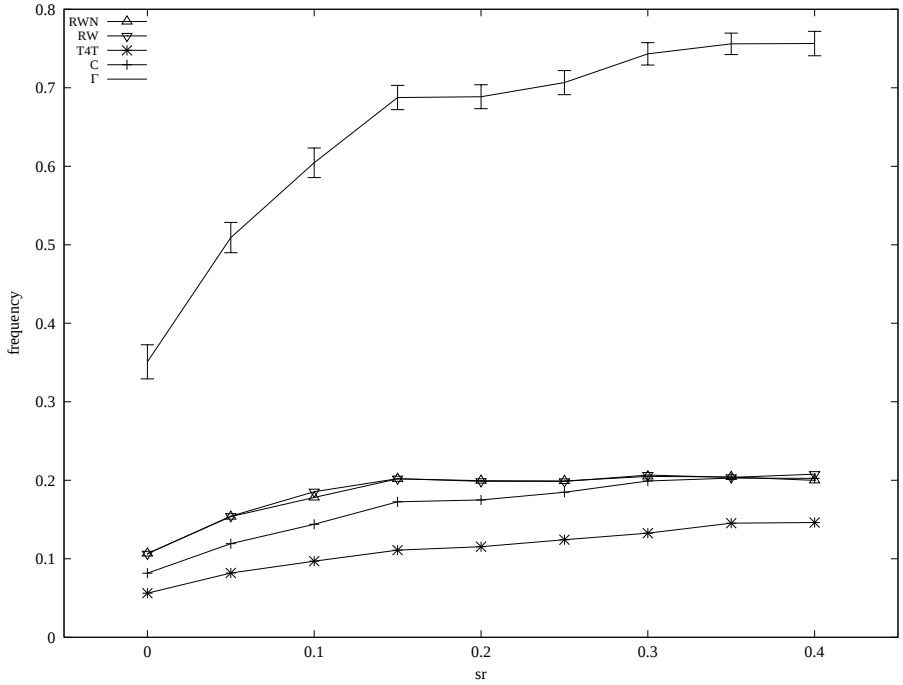
However, this situation is reversed when we introduce state into the learning algorithm by allowing

Figure 3: Mean reputation as function of the multiplier m and the proportion of the population configured as strong reciprocators sr.



These results were produced using 10^3 runs of the simulation for each position on the grid. The value of the multiplier m was varied over the range $[1.5, 3.0]$ with increments of 0.5 and the proportion of strong reciprocators over the range $[0, 0.4]$ with increments of $5 \times 10^{-}2$. All other parameters were treated as random variates with the distributions specified in Table 1.

Figure 4: Cooperation (Γ) against proportion of strong reciprocators (sr) — stateless treatment



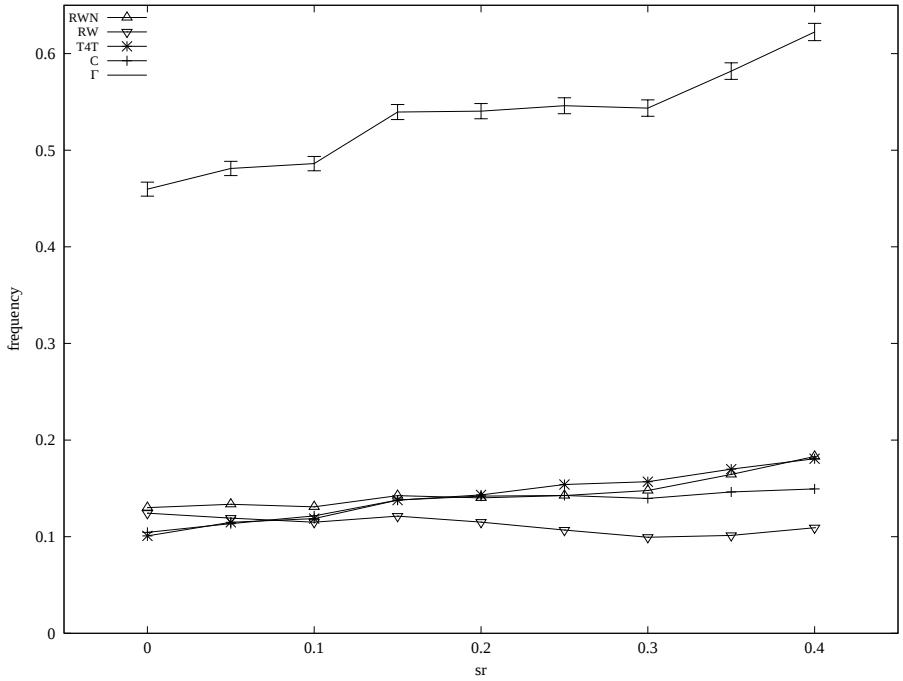
The above graph shows the mean frequency of every cooperative strategy in the learning fraction of the population as the number of strong reciprocators sr is increased under the stateless learning treatment when the multiplier is held constant at $m = 3$. The error bars show the confidence intervals for $p = 0.05$.

agents to condition their choice of strategy on reputation scores. Fig. 5 shows the same cross section of $\bar{\Gamma}$ under the stateful learning treatment. Levels of direct reciprocity and network-discounted indirect reciprocity are similar, both of which are more prevalent than the non-networked indirect reciprocity strategy (RW). In this treatment, overall levels of cooperation are less responsive to changes in sr. On the one hand there is greater cooperation in the absence of strong reciprocators, but on the other hand as strong reciprocators are introduced there is relatively little improvement (relative to the stateless treatment) in the final level of cooperation.

Additionally, the difference in frequency between the networked and non-networked strategies is sensitive to the value of the multiplier parameter. Fig. 6 shows the difference in frequency with which the RW (non-networked) and RWN (networked) are executed by the learning algorithm in steady-state when the proportion of strong reciprocators is held constant at $sr = 0.4$. The difference in steady-state adoption rates is more pronounced for intermediate values of the multiplier and disappears altogether when the multiplier is low.

Thus our results indicate that the source of information regarding the trustworthiness of other agents does indeed matter, as per Bolton et al.'s empirical study [2], in that both direct and indirect reciprocity come into play. Thus direct information about the history of interaction together with global reputation

Figure 5: Cooperation (Γ) against proportion of strong reciprocators (sr) — stateful treatment.



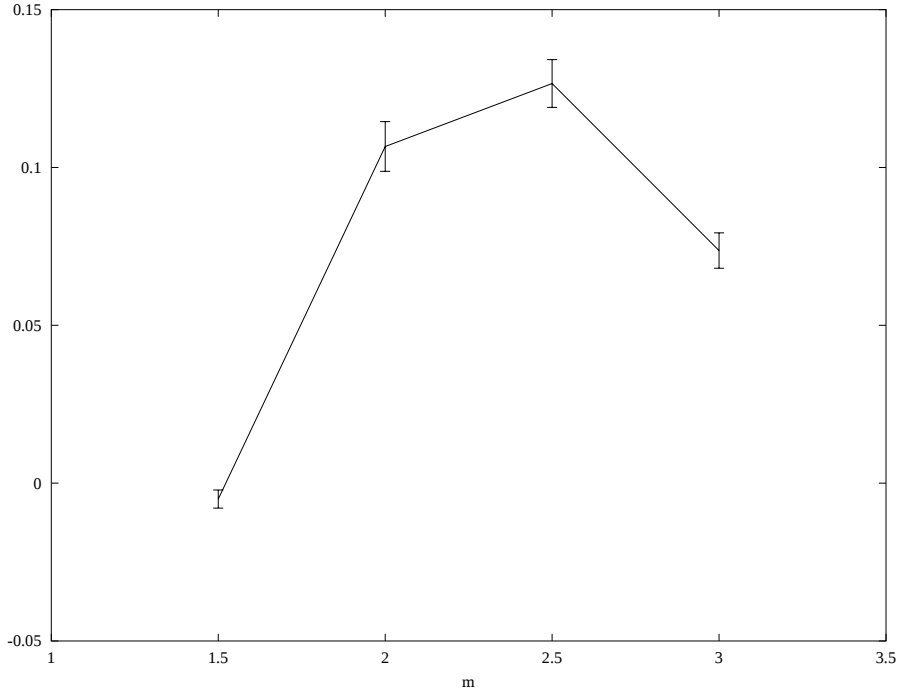
The above graph shows the mean frequency of each cooperative strategy in the learning fraction of the population once learning has reached a steady-state, against the number of strong reciprocators sr. The error bars show the confidence intervals for $p = 0.05$.

both inform agents' decisions. This result is also corroborated by our earlier game-theoretic modelling [20].

Moreover, the source of *reputation information itself* matters in that agents have a higher propensity to adopt the RWN strategy than the RW strategy. However, the source of reputation information plays an important role only when the benefits from cooperation are relatively greater than the costs, and this effect diminishes as the benefits grow larger. Intuitively, this follows the same reasoning that explains greater cooperation as the multiplier is increased: for small multipliers cooperation is difficult to sustain in general and the choice of which cooperative strategy to adopt is less relevant. On the other hand, for higher multipliers cooperation is easier to sustain in general, and similarly the particular choice of which form of cooperation to adopt is less critical. It is only for intermediate values of the multiplier that the game becomes subtle, and the exact mechanism for transmitting reputation becomes important. Within this regime the emergent social network plays an important role, and it becomes important to pay attention not only to the reputation of potential partners but also their social proximity — we will return to this discussion in Section 6.

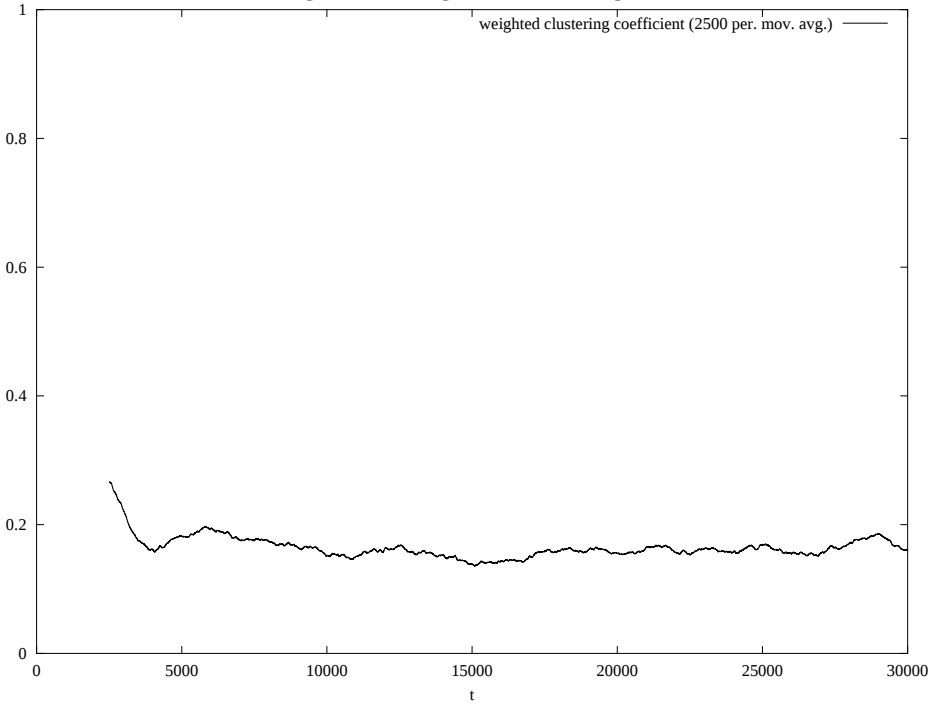
Thus the network relationships that arise from the interactions between agents play an important causal role in determining those very same interactions, resulting in emergent networks which co-evolve with the choice of agents' strategies. We are especially interested in how these emergent networks

Figure 6: Difference between RWN and RW frequencies in steady-state for $sr = 0.4$



The above graph shows the difference between the frequency with which the networked reputation (RWN) and non-networked reputation (RW) strategies are adopted in steady-state plotted against the multiplier m when the proportion of strong reciprocators is held constant at $sr = 0.4$. The error bars show the confidence intervals for $p = 0.05$. For intermediate values of the multiplier there is a pronounced difference between the frequency with which network-discounted reputation (RWN) is chosen over the global variant of this strategy (RW).

Figure 7: Weighted clustering coefficient as a time series



The global properties of the networks arising from our model are relatively stable; this graph shows the moving average of the weighted global clustering coefficient as a time series for a typical simulation run. In contrast to the degree rankings of individual agents which are highly volatile (see Fig. 8), the average global network properties do not change significantly over time.

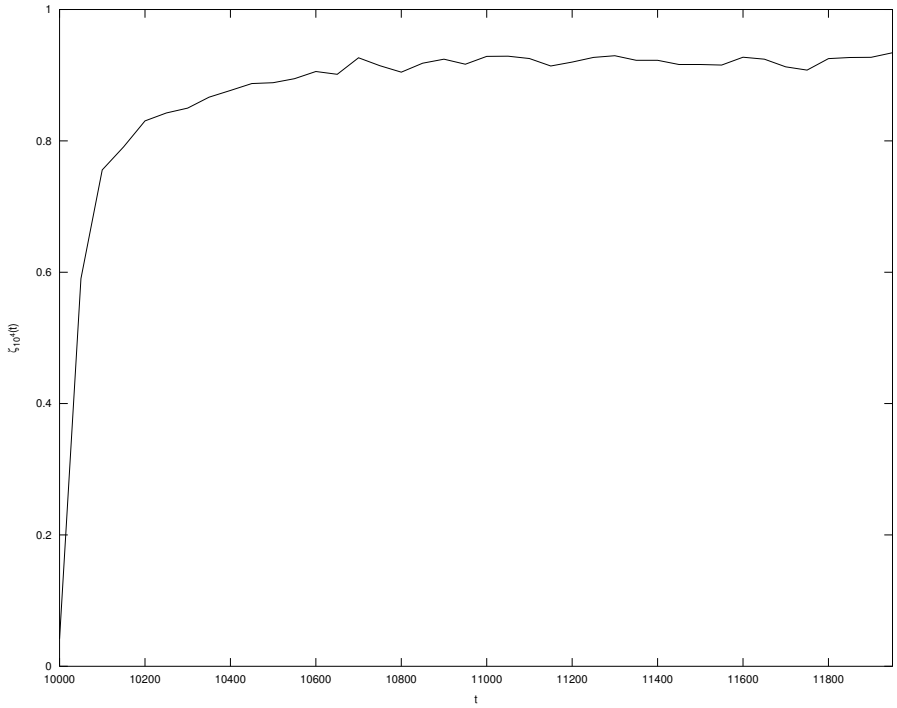
evolve over time. As in [12] we examine both individual and network-level degree metrics as time series. The former is measured by the correlation of the ranking of agents' degree values over time using a measure based on Spearman's rank correlation coefficient ρ [27]. The network dissimilarity coefficient $\zeta_{t_0}(t) = 1-\rho^2$ tracks the proportion of the variance in the weighted degree rankings at time t that cannot be explained by the rankings sampled at time t_0.

Fig. 8 shows the dissimilarity coefficient $\zeta_{10^4}(t)$ as a time series. This shows how agents' degree rankings evolve over time once the population has reached a steady-state; that is, we start from the weighted degree distribution of the network at $t = 10^4$ and compute the rank correlation with the weighted degree values at successive time periods. Despite the fact that the population as a whole is in a steady-state during this time period, as illustrated in Fig. 2, the individual-level network properties continue to evolve[1].

We find that the global network properties are relatively stable when compared to individual properties; this is in accordance with the empirical study of Kossinets and Watts [12]. Fig. 7 shows the weighted clustering coefficient ([25]) as a time a series for a typical run of the simulation. After a brief period of significant change at the beginning of the simulation the network converges to a steady-state behaviour

[1]This is most easily visualised as video — see `http://www.youtube.com/watch?v=rltW6CGj1WQ` for an illustration of a typical run.

Figure 8: Mean dissimilarity coefficient



The plot above shows the mean dissimilarity coefficient $\zeta_{10^4}(t)$ between the network at time t and the network at $t = 10^4$ averaged across 100 runs of the simulation with parameters drawn from the distributions shown in Table 1. The ζ values are sampled at intervals of 100. Despite the fact that the population has already reached a steady-state during this time period (see Fig. 2), the network continues to evolve at the level of individual agents.

in which the moving average of the weighted clustering coefficient hovers around a central value with only very small fluctuations away from the mean.

6 Discussion

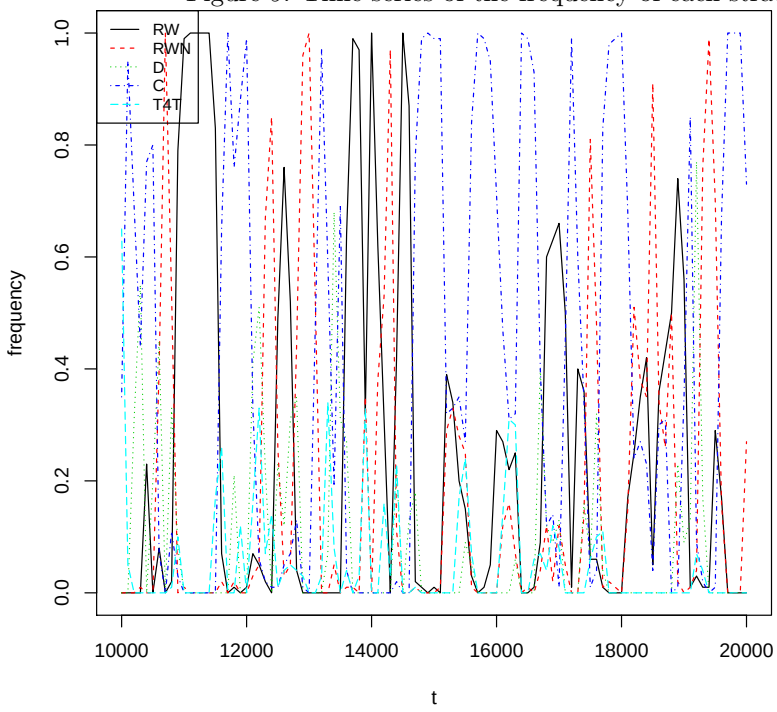
It is a salient feature of our model that it gives rise to networks which are simultaneously dynamic at the individual level but stable at the global level in line with empirical observations of large-scale human social networks [12]. This is particularly intriguing given that there is nothing to prevent the weighted clustering coefficient from changing over time; the network structure is driven by the choice of strategies made by the agents and different combinations of strategies can give rise to a diverse range of network structures with varying global metrics.

It is also striking that the average frequency with which each strategy is globally adopted also reaches a steady-state equilibrium (Fig. 2), and indeed this is what explains the stability of global properties of the network. In line with many other models of cooperation, agents attempt to learn a *pure* strategy and only randomise over other strategies in so far as they explore them randomly as determined by the parameter ϵ. When we analyse the behaviour of each individual agent, we see that the global equilibrium illustrated in Fig. 2 is in fact the result of highly dynamic behaviour at the level of individual agents who are switching between pure strategies in response to a changing environment. Although the population-level frequencies form a learning-theoretic equilibrium in which the proportions with which each strategy are adopted remain stable, this is the result of punctuated switching between pure strategies at the level of individual agents, as can be seen from Fig. 9 which shows how the propensity to play each strategy changes over time for a single agent.

Moreover, provided that the recency parameter α is sufficiently high, this phenomena persists even when we switch to an alternative learning model which allows agents to learn a policy which randomises over strategies; we observe similar behaviour when agents use softmax action selection as defined by Equation 13.

It is not possible to provide a full sensitivity analysis of parameter settings under this treatment because the appropriate value of the temperature parameter τ depends in a non-trivial way on the other parameters. Nevertheless, it serves to demonstrate that the dynamic network behaviour is not merely an artifact of the particular action-selection policy used in our model of learning. Rather, the dynamics are a direct result of learning using *recency*, in which more weight is placed on recent information. From the perspective of each individual agent, performing learning with a high recency weighting is entirely rational in a strategic environment populated by other agents who learn inductively; even if the true Nash strategy profiles were known to the agent there could be a benefit from playing non-equilibrium

Figure 9: Time series of the frequency of each strategy for a single agent



strategies against other inductive agents who are themselves not necessarily playing a Nash strategy. Moreover since other agents are dynamically adjusting their strategy through learning, it is rational to treat the other agents as a dynamic environment and deploy the standard techniques for dealing with such, viz. weighting more recent data.

Although there might be rational justifications for off-equilibrium play by individual agents, a priori we would not expect such deviations to be systematic: game-theoretic considerations suggest that agents who are currently adopting non-equilibrium strategies which are not best responses will eventually be exploited until they learn another strategy. Thus the existence of a game-theoretic equilibrium should create stability at the level of the population through negative feedback by driving out off-equilibrium behaviour, albeit with latency.

The other surprising phenomenon arising from our model is that agents learn to use the emergent social network to discount reputation information according to the social proximity of other agents (Fig. 6). Thus the source of information about the trustworthiness of other agents *does* matter, in line with empirical findings from experimental economics [2]. This feature is not built into our model since agents are free to choose between the network-discounted version of the reputation-weighted strategy (RWN) or the strategy which uses a simple global reputation score (RW) according to which version performs the best. Thus it is not a priori obvious why social proximity and reputation interact in this way.

A closer analysis reveals that the RWN strategy yields higher payoff than the RW strategy only in the presence of direct reciprocity (T4T) and the absence of unconditional cooperation (C). In such a scenario, agents adopting RWN are able to form cliques with each other and also with T4T. Agents within these cliques are able to selectively invest their endowment with each other, whilst also receiving low-level donations from the excluded agents which are below the threshold necessary for creating a social tie (parameter κ in eq. 8). These cliques can arise because RWN also embodies a form of direct reciprocity: the network distance metric used to discount reputation scores is undirected, and therefore RWN is more favourable to agents that have in turn favoured it. Thus RWN is able to exploit a niche in which it simultaneously benefits from both direct and indirect reciprocity.

7 Conclusion

In this paper we have introduced a model of cooperation which incorporates two distinct forms of reciprocity. Direct reciprocity uses a private source of information based on personal history of interaction with others. On the other hand, indirect reciprocity makes use of public information in the form of reputation. Whereas other studies have looked at each of these in isolation, in contrast our model allows for *interaction* between these two classes of strategy. Our key contributions have been a) to show that *both* forms of reciprocity play an important role with neither dominating the other; and moreover b) the interaction between these strategies gives rise to complex social networks which *emerge* bottom-up from the lower-level actions taken by agents. Moreover these networks evolve over time in a similar way to those observed in empirical studies [12]: global network metrics remain stable whilst individuals' degree rankings are highly dynamic.

This discrepancy between global and local network properties arises directly from the dynamics of learning. The global network properties are determined by the frequency with which each type of strategy is adopted in the population as a whole. This in turn determines the expected payoff to each strategy thus creating a static game-theoretic equilibrium which stabilises the global network structure. However, as is the case in nearly all non-trivial agent-based models, underlying this static global equilibrium is a highly dynamic switching between different strategies more reminiscent of "punctuated" equilibrium [5] as agents constantly adjust their strategies in response to a changing environment. Two key features of reinforcement learning are *recency* and *experimentation* and these result in off-equilibrium behaviour at the level of the individual agents in our model.

We conjecture that a similar process underlies the phenomena observed in the Kossinets and Watts study [12]. Proving this conjecture will entail controlled experiments with human subjects, which is the subject of our future work.

References

[1] R. Axelrod. *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration.* Princeton University Press, 1997.

[2] G. E. Bolton, E. Katok, and A. Ockenfels. How Effective are Electronic Reputation Mechanisms? An Empirical Investigation. *Management Science*, 50(11):1587—-1602, 2004.

[3] S. Bowles and H. Gintis. The evolution of strong reciprocity: cooperation in heterogenous populations. *Theoretical Population Biology*, 65:17–28, Feb. 2004.

[4] R. Boyd and P. J. Richerson. The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, 132:337–356, 1988.

[5] N. Eldredge and S. J. Gould. Punctuated equilibria: an alternative to phyletic gradualism. In T. J. M. Schopf, editor, *Models in paleobiology*, pages 82–115. Cooper & Co., San Francisco, 1972.

[6] I. Erev and A. E. Roth. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88(4):848–881, 1998.

[7] E. Fehr and S. Gächter. Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14(3):159–181, 2000.

[8] A. Flache and M. W. Macy. Stochastic Collusion and the Power Law of Learning: A General Reinforcement Learning Model of Cooperation. *The Journal of Conflict Resolution*, 46(5):629–653, Oct. 2002.

[9] M. Granovetter. Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, 91(3):481–510, 1985.

[10] S. S. Izquierdo, L. R. Izquierdo, and N. M. Gotts. Reinforcement Learning Dynamics in Social Dilemmas. *Journal of Artificial Societies and Social Simulation*, 11(2):1, 2008.

[11] M. Kaisers and K. Tuyls. Frequency adjusted multi-agent Q-learning. In Van Der Hoek, Kamina, Lespérance, Luck, and Sen, editors, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 309–316. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[12] G. Kossinets and D. J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, Jan. 2006.

[13] M. Matsumoto and T. Nishimura. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, 1998.

[14] A. Nowak and K. Sigmund. Evolution of indirect reciprocity. *Nature*, 437:1291–1298, Oct. 2005.

[15] M. Nowak and K. Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 383:537–577, 1998.

[16] M. A. Nowak and K. Sigmund. The alternating prisoner's dilemma. *Journal of theoretical Biology*, 168:219–226, 1994.

[17] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, May 2006.

[18] S. Phelps. Binaries and source-code for reciprocity model. `http://jabm.sourceforge.net`. online; accessed 14/12/2011.

[19] S. Phelps. Applying dependency injection to agent-based modeling: the JABM toolkit. *ACM Transactions on Computer Simulation and Modeling*, 2011. (In submission).

[20] S. Phelps, G. Nevarez, and A. Howes. The effect of group size and frequency of encounter on the evolution of cooperation. In *LNCS, Volume 5778, ECAL 2009, Advances in Artificial Life: Darwin meets Von Neumann*, pages 37–44, Budapest, 2009. Springer.

[21] P. J. Richerson and R. Boyd. The evolution of human ultra-sociality. In I. Eibl-Eibisfeldt and F. Salter, editors, *Idealogy, Warfare and Indoctrinability*, pages 71–95. Berhan Books, 1998.

[22] G. Roberts and T. N. Sherratt. Development of cooperative relationships through increasing investment. *Nature*, 394:175–179, 1998.

[23] F. Santos, J. Rodrigues, and J. Pacheco. Graph topology plays a determinant role in the evolution of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 273(1582):51–55, Jan. 2006.

[24] F. C. Santos, J. M. Pacheco, and T. Lenaerts. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Comput Biol*, 2(10):1284–1291, 2006.

[25] J. Saramäki, M. Kivelä, J. P. Onnela, K. Kaski, and J. Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2), 2007.

[26] J. M. Smith and E. Szathmáry. *The Major Transitions In Evolution*. Oxford University Press, 1995.

[27] C. Spearman. The Proof and Measurement of Association between Two Things. *American Journal of Psychology*, 15(1):72–101, 1904.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[29] M. V. Vught, M. Roberts, and C. Hardy. Competitive altruism: Development of reputation-based cooperation in groups. In *Handbook of Evolutionary Psychology*, chapter 36. Oxford University Press, 2007.

[30] J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[31] C. Wedekind and M. Milinski. Human cooperation in the simultaneous and alternating prisoner's dilemma: Pavlov verses generous tit-for-tat. *Proceedings of the National Academy of Science USA*, 93:2686–2689, 1996.