The effects of reputation on inequality in network cooperation games

Milena Tsvetkova*

Department of Methodology, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom, ORCID ID 0000-0002-3552-108X

Keywords: reputation, cooperation, inequality, networks, experiments, agent-based model

Abstract

In the last several decades, ample evidence from across evolutionary biology, behavioural economics, and econophysics has solidified our knowledge that reputation can promote cooperation across different contexts and environments. Higher levels of cooperation entail higher final payoffs on average but how are these payoffs distributed among individuals? This study investigates how public and objective reputational information affects payoff inequality in repeated social dilemma interactions in large groups. We consider two aspects of inequality: excessive dispersion of final payoffs and diminished correspondence between final payoff and cooperative behaviour. We use a simple heuristics-based agent model to demonstrate that reputational information does not always increase the dispersion of final payoffs in strategically updated networks, and actually decreases it in randomly rewired networks. More importantly, reputational information almost always improves the correspondence between final payoffs and cooperative behaviour. We analyse empirical data from nine experiments of the repeated Trust, Helping, Prisoner's Dilemma, and Public Good games in networks of ten or more individuals to provide partial support for the predictions. Our research suggests that reputational information not only improves cooperation but may also reduce inequality.

Introduction

We, humans, have the distinctive ability to form complex social interaction and communication networks. These networks allow us to share information and cooperate with each other, achieving more collectively than any single individual could by themselves. In fact, scientists argue that our ability to gather and disseminate information about others is one of the reasons why we trust and help each other: that is, reputational information promotes cooperation.

Reputation is the information about someone's past behaviour that we obtain from direct observation, centralized institutions (reputation systems), or via social networks (gossip). In social dilemma situations where cooperation is individually undesirable but collectively beneficial, reputation allows cooperation to emerge via indirect reciprocity or reputation-based partner choice (1). Indirect reciprocity is the tendency to help individuals who have been helpful towards others, while reputation-based partner choice is the tendency to select helpful partners and avoid unhelpful ones. Indirect reciprocity is evolutionarily advantageous and thus, deeply ingrained in us (2, 3). Due to indirect reciprocity, individuals realize that their decisions affect their reputation, which in turn affects how others treat them in future interactions, and hence they become more likely to cooperate (4–6). Due to reputation-based partner choice, individuals increase their cooperation so that they are more likely to be chosen as partners and benefit from future interactions (1).

By fostering cooperation, reputation increases individuals' final payoffs on average and improves collective wealth. However, how are the payoffs distributed among different individuals? And do they reward cooperators more than defectors? These are essentially questions about inequality. The problem of inequality is one of the most fundamental problems modern organizations and societies face. Inequality decreases individual productivity and job satisfaction at the workplace, lowers organizational performance (7, 8), and more generally, negatively impacts happiness (9, 10) and health (11). Thus, a more comprehensive understanding of the effects of reputation systems on inequality would allow us to evaluate any potential tradeoffs and design more efficacious and sustainable organizations, institutions, online marketplaces, and social media communities.

This study takes a major step in this direction by investigating the effect of reputational information on inequality in repeated social dilemmas in networks with different dynamics. We investigate inequality in

*Author for correspondence (m.tsvetkova@lse.ac.uk).

terms of the *dispersion* of final payoffs and the *correspondence* of final payoffs to cooperative behaviour. Looking through the lens of the co-evolution of reputations, social networks, and cooperation (12), we compare situations where individuals are randomly matched with new partners, such as speed dating events, round-robin tournaments, anonymous chat rooms, or rotating teams for school assignments and work projects, with situations where individuals can choose with whom they interact, such as online markets or self-assembled project teams. We argue that, when the networks are dictated by random matching, the higher levels of cooperation that reputation produces lower the dispersion of payoffs. However, when the networks are updated strategically, reputation fosters the clustering of cooperators and exclusion of defectors and thus may increase the dispersion of payoffs, yet reduce inequality in terms of better correspondence between payoffs and cooperative behaviour. We use an agent-based model to visualize the group-level expectations and data from nine network cooperation experiments to provide empirical evidence for them.

Theoretical background

In non-cooperation settings, reputation has long been understood as a prominent mechanism for inequality. If we use a person's past behaviour and performance to predict their future actions and achievements, reputation can produce a positive feedback loop whereby past accomplishments and resources turn into new accomplishments and resources. The process can cause initially small or accidental differences in behaviour and performance to compound and amplify over time. As a result, individual outcomes become less predictable and more extremely distributed. Merton studied this process in the context of academic success and famously labelled it "the Mathew effect" (13). The process is also known as "the rich get richer", increasing returns, and cumulative advantage (14, 15).

Reputation has been shown to create inequality in cooperation settings too. Hackel and Zaki (16) demonstrate that reputation serves to propagate existing inequalities in cooperation experiments. They find that participants tend to misattribute good reputation to those who have an arbitrary resource advantage that allows them to give more to others. As others rely on this reputational information to make investment decisions, they end up reproducing and reinforcing the inequalities in a different setting. Further, Frey and van de Rijt (17) provide evidence for reputation-driven cumulative advantage in experimental buyer-seller markets. The authors find that sellers who cooperate early get disproportionately rewarded because buyers are more likely to choose sellers of good repute. This implies that initial differences in cooperative behaviour could have long-lasting effects that get reinforced and exaggerated over time. As a result, social groups where information about past behaviour is available will have higher inequality than groups without reputation tracking.

Evaluating inequality in the context of cooperation, however, is a complex problem. Cooperative behaviour is collectively beneficial and hence, valued and desirable. Importantly, cooperative behaviour is not necessarily inherent but conditional on others. If we disassociate individuals from the cooperative behaviour they exhibit, then we can measure inequality with the dispersion of final payoffs. In essence, if we consider differences in cooperative behaviours emergent, and hence, somewhat arbitrary, then a higher payoff difference between the wealthiest and the poorest would define higher inequality. On the other hand, however, if we hold individuals accountable for their cooperative behaviour, then inequality will depend on the extent to which payoffs correspond to cooperation. If we assume defectors willingly chose to defect, then situations where they end up better off compared to cooperators would be more unequal. This distinction between *dispersion* and *correspondence* parallels the contrast between dispersion and rank reversal that Freda Lynn and colleagues delineate in the context of status hierarchies (18). The underlying idea is that if we start with a population that is heterogeneous on a valued characteristic or behaviour, then inequality can take two forms: outcomes that are more dispersed than the underlying heterogeneity, and outcomes that correspond less to the valued characteristic or behaviour.

Furthermore, in cooperation settings, reputation can have complex effects on inequality. On the one hand, we know that reputation tends to increase the level of cooperation. Starting from the initial rounds of interaction, individuals cooperate more due to forward-looking behaviour: aware that their reputation will affect others' behaviour towards them in the future, they immediately start investing in it (4, 19). Over the course of interacting, individuals further learn from experience: they realize that being a reputable partner gives you an advantage and consequently, switch to cooperating (20). Since more uniform behaviour entails more similar individual outcomes, the higher levels of cooperation that reputation brings can lower the dispersion of final payoffs. Multiple studies confirm that reputation promotes cooperation in simple dyadic interactions (21–23) but there are also studies that fail to find the expected effect (24).

Higher levels of cooperation will entail lower payoff dispersion if there is no segregation between cooperators and defectors. However, if reputation allows cooperators to find other cooperators and exclude defectors, the differences in payoffs between the two groups can widen. Simultaneously, the correspondence between payoffs and cooperative behaviour would improve. Indeed, previous research shows that individuals are

more likely to select partners who have reputation as cooperators (17), as well as more likely to cooperate with them (23, 25). At the group level, this implies that reputation will lead to networks with higher clustering by cooperativeness and higher degree centrality for cooperators. In fact, network experiments where payoff depends on the number of partners show the emergence of cooperative hubs as cooperative individuals attract higher number of connections (17, 19). The evidence for clustering by cooperativeness, however, is less convincing. Melamed et al. (26) find no effect from reputation on clustering as partner choice alone can induce near uniform cooperation. Gallo and Yan (19) discover that reputational information needs to be combined with knowledge of the network in order for cooperators to cluster.

One prior study that uses data from nine network cooperation experiments to investigate the effects of reputational information on the dispersion of final payoffs reports statistically significant effects in both the positive and negative directions, as well as near-zero effects (27). Here, we extend this work in a couple of ways. First, we present a more nuanced view of inequality by investigating both the dispersion of payoffs and the correspondence between payoffs and cooperative behaviour. Second, we pinpoint one specific factor that moderates the effect of reputational information on inequality: the dynamics of the underlying interaction network. We compare the effects of reputation under homogenous mixing in randomly rewired networks to strategic partner selection in dynamic networks.

Expectations

We use an agent-based model to demonstrate how the effects of reputational information on payoff dispersion and correspondence may differ for different network dynamics. We rely on a generic and simple model that assumes agents follow fixed behavioural strategies and heuristic-based rules to respond to others' actions and reputational information. The model we use intentionally avoids problematizing the evolution of cooperation (in contrast to most prior work) and instead takes it for granted in order to focus on how cooperation affects the distribution of payoffs. Our goal is to demonstrate the theoretical complexity of the phenomenon, rather than provide quantitative predictions for the empirical analyses.

The agents in the model play an N-person Prisoner's Dilemma game in which they choose between cooperating (C) and defecting (D) with payoffs for mutual cooperation $CC = 5$, mutual defection $DD = 2$, defecting with a cooperating partner $DC = 8$, and cooperating with a defecting partner $CD = 0$. Following empirical research on behavioural proclivities in the general population (28, 29), we assume that agents belong to three different fixed-strategy types: p_D fraction are persistent defectors, p_A are persistent altruists, and $p_R = 1 - p_A - p_D$ conditional cooperators. Altruists always cooperate, defectors never do, and conditional cooperators initially cooperate with some probability, after which they reciprocate by cooperating if at least a certain fraction of their interaction partners cooperated in the last period. While p_A and p_D define the minimum and maximum possible levels of cooperation, respectively, p_R determines the proportion of responsive and interdependent agents, who exhibit more realistic variation in behaviour and drive the system's emergent properties.

Agents interact in a network of size N and average node degree m. To compare the effect of network dynamics, we study two modes of network updating. For the randomly rewired networks, agents are placed in a new network every period. For the strategically updated networks, every period each agent is given the opportunity to replace one of their defecting neighbours with someone else. An existing link gets deleted if one of the two agents drops it, but for a new link to appear, both agents need to desire it. To facilitate mutual nominations but avoid skewed degree distributions, we assume that agents can nominate multiple new potential partners, as long as the number of actual partners does not exceed $2m$.

To avoid stochastically unstable outcomes, we assume a small probability for error $\varepsilon = 0.005$ such that the agent executes an action that is opposite to the one they originally chose, and another small probability for error $\gamma = 0.005$ such that the agent does not update their network even if they have decided to.

We study the outcomes in networks with random rewiring and strategic updating when there is no reputational information and when reputation is provided as the average action over the last r periods. When reputational information is available, forward-looking individuals are more likely to cooperate when they are aware of the negative consequences from reputation as a non-cooperator (4, 19, 26). We implement this by specifying that conditional cooperators initially cooperate with probability $\theta_0 + a_0 r$, and then cooperate if at least $\theta_C - a_C r$ of their neighbours cooperated over the last r periods, where θ_0 and θ_C are the behavioural thresholds without reputation and a_0 and a_C define the strength of the reputation effects ($0 < \theta_0, \theta_C, a_0, a_C < 1$). Reputational information also comes in play when agents select new partners (17, 19, 26). Without reputational information, agents pick new partners randomly from those with whom they are not yet linked. Otherwise, they pick only among those who have cooperated in at least θ_C of the past r periods.

For the results we report here, we fix the initial tendency to cooperate $\theta_0 = p_A$, that is, we assume that the willingness to cooperate on first encounter is equal to the probability of encountering an unconditional cooperator. Additionally, we fix the subsequent cooperation threshold $\theta_C = 0.5$ and the reputation effects $a_0 = 0.1$ and $a_C = 0.05$. Generally, higher θ_0, p_A, a_0, a_C and lower θ_C increase the rate at which cooperation emerges and the equilibrium level of cooperation. The specific values were chosen because they produce sufficient variation in the level of cooperation and replicate the empirically validated expectation that reputational information and strategic network updating independently increase the level of cooperation (Fig. S1A, S2A, S3A). Further, we assume that agents are initially embedded in a random network, where an interaction link between any pair of nodes is formed with a small fixed probability; this produces initial networks with low clustering and Poisson degree distributions, where we fix the network size $N = 100$ and average node degree $m = 2$. We also assume that payoffs are averaged over all interactions in a period and, hence, do not depend on the number of interaction partners. The model allows to modify the assumptions about the initial network structure, average node degree m (see Fig.S1), payoff function (Fig.S2), and reputational information r (Fig.S3), but the outcomes are qualitatively similar.

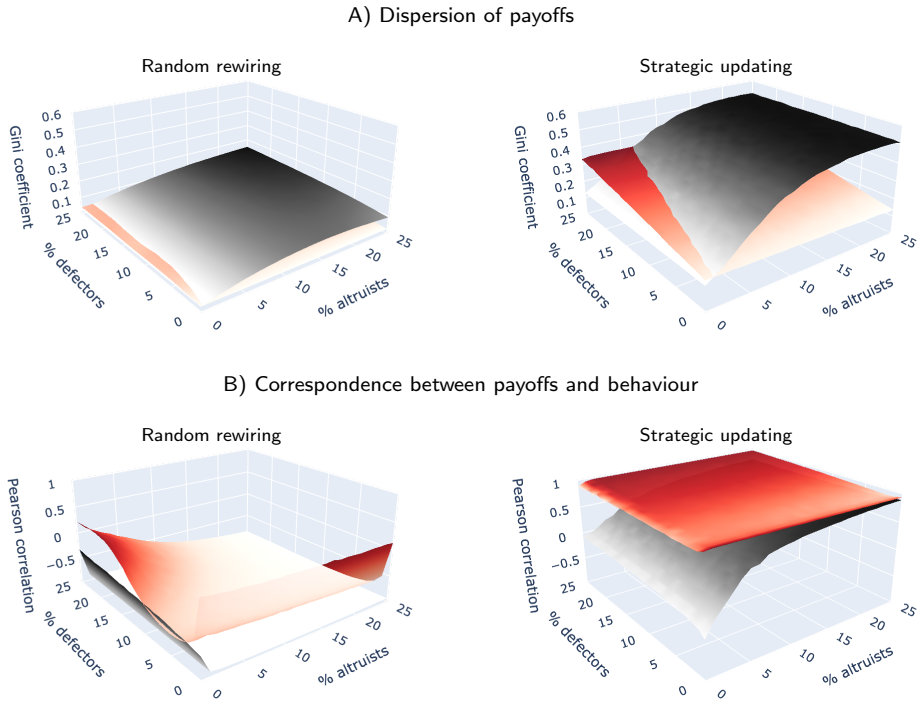

Figure 1. A) Agent-based models reveal that reputation (shown in red) generally decreases the dispersion of payoffs in randomly rewired networks but might increase it in networks with strategic updating when the percentage of steady defectors is sufficiently high. B) Reputation almost always results in better correspondence between payoffs and cooperative behaviour. The plots show (A) the mean Gini coefficient of accumulated payoffs and (B) the mean Pearson correlation between proportion of periods when cooperating and final payoff when no reputational information is available ($r = 0$; grey) and when agents know everyone's actions in the previous period ($r = 1$; red). The means are calculated over 1000 runs in networks of $N = 100$ agents who start interaction in a random network with $m = 2$ partners on average and play for $T = 100$ periods. The simulations vary the percent of steady defectors (x-axis) and steady altruists (y-axis), with the rest being conditional cooperators.

Although the model is simple enough to be solved analytically, we focus here on visualizing our theoretical intuition and demonstrating the variability of outcomes expected for a range of cooperation outcomes due to changes in the proportion of persistent altruists, persistent defectors, and conditional cooperators. We measure inequality in terms of dispersion with the Gini coefficient of the payoffs accumulated after $T = 100$ periods. The Gini coefficient is 0 when all individuals have equal payoffs and 1 when only a single individual has a non-zero payoff. We measure inequality in terms of correspondence with the Pearson correlation between the proportion of periods in which the agent cooperated and the agent's final payoff. A Pearson correlation of 1 means that higher cooperation is always proportionally rewarded with higher payoffs, while a Pearson correlation of –1 indicates perfect inverse proportionality.

The model comparing outcomes for $r = 0$ and $r = 1$ confirms our intuition (Fig. 1). The results indicate that reputational information decreases the dispersion of payoffs in randomly rewired networks but could increase it in strategically updated networks when persistent defectors are numerous (Fig. 1A). In randomly rewired networks, the Gini coefficient is always low but reputation brings it even lower because conditional cooperators can make agents pay back for last-period's defection. In strategically updated networks, the Gini increases with higher proportions of persistent altruists in the absence of reputation but with higher proportions of persistent defectors when reputational information is available. The reason is that reputation allows the complete isolation of defectors, while its lack does not prevent the exploitation of altruists.

Although reputational information can increase the dispersion of payoffs in strategically updated networks under certain conditions, it almost always provides better correspondence between payoffs and cooperative behaviour, regardless of the network dynamics (Fig. 1B). With strategic updating, reputation guarantees that defectors are excluded and thus suffer, while cooperators prosper. Under random rewiring, even though reputation rarely produces positive rewards to cooperation, it still reduces the rewards for defection by making payoffs less negatively correlated with cooperative behaviour. In short, reputational information combined with the strategic choice of partners could increase the dispersion of payoffs but this is because persistent defectors deservingly lose out.

Data and methods

We provide empirical evidence for the theoretical expectations with data from nine network cooperation experiments (Table 1). These experiments were originally designed to study the effect of reputational information on the emergence of cooperation and the published articles associated with them do not report on inequality. We identified these studies after conducting online searches on the Scopus database, the Google Scholar search engine, and the Cooperation Databank (30). Specifically, we searched for network cooperation experiments that involve the repeated play of a social dilemma game in networks of at least 10 and that manipulate the information available on other participants' past actions. We only considered accurate and objective information, excluding studies with gossip or subjective ratings. We identified ten studies and after contacting the corresponding authors, obtained data for nine of them (with (19) missing).

Table 1. Summary of the experimental data.

Experim.	Game: payoffs	N_i	N_n	N	T	Network	m	**Updating**	**Reputation**
BOLT04	TG: 50, 70, 35	144	9	16	30	pair	1	random	0, all
BOLT05a	HG: -0.25, 1.25	96	6	16	14	pair	1	random	0, 1, 1+1
BOLT05b	HG: -0.75, 1.25	96	6	16	14	pair	1	random	0, 1, 1+1
SEIN06	HG: -150, 250	112	8	14	>90	pair	1	random	1, 6
STAH13	PD: 80, 10, 90, 20	92	8	22, 24	24, 39	pair	1	random	0, 1
BAYE16a	PG: $100 - c_i + 0.8 \sum c_j$	224	12	16–20	24	pair	1	random	0, last
CUES15	PD: 7, 0, 10, 0	243	22	17–25	25	(cycle)	(4)	strategic	0, 1, 3, 5
BAYE16b	PG: $100 - c_i + 0.8 \sum c_j$	224	12	16–20	24	pair	1	disincent. strategic	0, last
BAYE16c	PG: $100 - c_i + 0.8 \sum c_j$	224	12	16–20	24	pair	1	incent. strategic	0, last
KAME17a	PG: $10 - c_i + 0.65 \sum c_j$	120	12	10	40	pair	1	strategic	0, part, avg
KAME17b	PG: $10 - c_i + 0.85 \sum c_j$	130	13	10	40	pair	1	strategic	0, part, avg
HARR18a	tPD: 50, -50, 100, 0	334	20	12–24	12	(random)	(4)	part strategic	0, avg
HARR18b	tPD: 50, -50, 100, 0	334	20	12–24	12	(random)	(4)	strategic	0, avg
MELA18a	PD: 50, -50, 100, 0	810	15	19–28	16	(random)	(4)	strategic	0, loc, avg
MELA18b	PD: 50, -50, 100, 0	810	15	19–28	16	(clustered)	(4)	strategic	0, loc, avg
MELA18c	tPD: 50, -50, 100, 0	810	15	19–28	16	(random)	(4)	strategic	0, loc, avg
MELA18d	tPD: 50, -50, 100, 0	810	15	19–28	16	(clustered)	(4)	strategic	0, loc, avg
MELA18e	PD: 50, -50, 100, 0	472	16	19–28	16	(random)	(4)	slow strategic	0, loc, avg
MELA18f	tPD: 50, -50, 100, 0	472	14	19–28	16	(random)	(4)	slow strategic	0, loc, avg

The games played in the experiments are Prisoner's Dilemma (PD), targeted Prisoner's Dilemma (tPD), Public Good (PG), Helping (HG), and Trust (TG) game. Payoffs are shown for CC, CD, DC, DD for PD and CC, CD, D for TG. For HG, the payoff numbers correspond to *cost*, *benefit* for the gift and for PG, c_i (c_j) is the amount contributed by the player (all players) to the public good. N_i = number of unique participants, N_n = number of networks, N = network size, T = number of periods. Network refers to network structure, m—the number of interaction partners, and updating—how the network is updated. Values in brackets refer to the network in the first period only. Reputation refers to reputation tracking with the number indicating the number of previous periods over which partner's actions are observed; all = observe all actions in previous periods; avg = observe average behavior over all previous periods; last = observe average behavior over previous four-period-long phase; part = observe average behavior over 50% of previous periods; loc = observe average behavior for partners' partners only; 1+1 = observe partner's action in last period, as well as the action of the partner's partner from the period before that.

Our choice to re-analyse existing data, instead of developing bespoke experiments, has some advantages, as well as disadvantages. On the positive side, we insure ourselves against Type I errors, or falsely reporting positive effects, because none of the studies could have been cherry-picked to report significant results with respect to the dependent variables. In addition, the repurposed data allow us to affirm the robustness of our findings across different experimental setups, incentives, and participant pools; this will be prohibitively expensive to do from scratch. Still, on the negative side, we risk Type II errors, i.e., failing to find true positives, because the experiments were not calibrated to produce the highest variation in the outcome of interest, nor scaled up sufficiently for group-level analyses. We can only test the predictions qualitatively, not quantitatively, because the model and experiments vary in more than one aspect. Finally, we can only do an indirect test of the predictions because only one experiment crosses reputational information with random rewiring and strategic updating. Consequently, we analyse the effect of reputation separately in randomly rewired and strategically updated networks.

The experiments use one of four common social dilemma games: the Trust (TG), Helping (HG), Prisoner's Dilemma (PD), or Public Good (PG) game. The TG is most intuitively understood in the context of buyer-seller relationships, where the buyer (trustor) decides whether to send a sum of money to the seller (trustee), and if so, the seller gets to choose whether to ship the purchased item, i.e., send something of value back, or not. Honoured trust (*CC*) is mutually beneficial but abusing trust (*CD*) is tempting for the trustee. In the HG, the player decides whether to give a gift of a cost c to another player who will benefit b, where giving is collectively beneficial ($b > c$) but not guaranteed to be reciprocated. In the PG, players invest in a common pot and then share the multiplied investments equally. In the PD, players choose either to cooperate or defect, where mutual cooperation (*CC*) is collectively beneficial but individually unprofitable because defecting (*DC* or *DD*) is always the better choice, regardless of what the other player does.

For random rewiring, we analyse six different network cooperation experiments from five studies (21–23, 31, 32). The experiments randomly re-match pairs of participants at regular intervals in networks of 14-24. For strategic updating, we analyse 13 experimental setups from five studies (20, 25, 26, 32, 33). Typically, the interaction starts in a network with a certain structure and density and throughout the game, participants are given the opportunity to drop some of their current interaction partners and select new ones. In the experiments using the PD, participants play the dyadic game with each of their partners; thus, each participant's payoff depends on the number of partners, as well as their actions. In some of these experiments the participant chooses one action and plays it against all their partners, while in others, they can choose a different action against each partner (shown as tPD, or targeted PD, in Table 1). The one experiment that manipulates both reputational information and network updating (32) uses a dyadic PG where partners are updated (randomly assigned or strategically chosen in the different treatments) every four periods.

The experiments manipulate the amount of information available about other players' past actions beyond knowledge of what one's partners did in the previous round. In the control condition, no such information is available and in (25), even information about what one's own partners individually chose in the previous period is missing. To provide reputational information, some of the experiments reveal to participants what their current or potential partners did in the past r periods, where $r = 1, 3, 5, 6$, or all previous periods. Other experiments show the average rate of cooperation/contribution over all previous periods (avg), over half of previous periods selected at random (part avg), or over the last four-period game phase (last avg). One study shows the average rate of cooperation over all previous periods but only towards one's partners' partners (loc avg). Another study includes information on the action of the partner's partner to which one's partner responds (abbreviated as 1+1).

For the analyses, we measure final payoff with the number of in-game monetary units the player accumulated at the end of the game. As in the model, we use the Gini coefficient to measure the dispersion of final payoffs. The Gini coefficient is particularly suited in this case because it is invariant to scale, which helps us compare outcomes across experiments with very different incentive structures (34). Analogously, we use the Pearson correlation coefficient between the rate of cooperation and final payoff to measure the correspondence between behaviour and rewards.

To test the effects of reputational information on inequality, we compare the Gini and Pearson correlation coefficients in the conditions with reputational information to the baseline condition without. We use the Mann-Whitney test to assess the differences in each control-treatment pair. This is a non-parametric test that does not assume a normal distribution for the residuals. It essentially checks against the null hypothesis that a randomly selected value from one condition would be equally likely to be less than or greater than a randomly selected value from the other condition.

Since the experiments were not designed to test group-level hypotheses, we do not always have large enough effect sizes and statistical power to provide evidence at the level of a single experiment. Hence, we additionally conduct a meta-analysis across all control-treatment pairs using the sign test (35). We first count the number of positive and negative effects, regardless of whether they are statistically significant. We then

conduct the Binomial test, testing against the null hypothesis that there is no effect in reality and thus negative and positive effects are equally likely to occur by chance. This approach is somewhat limited because it does not take into account the amount of evidence: neither the effect magnitudes nor the sample sizes. However, our goal is not to estimate an effect size but test a causal relationship. We note that effect sizes are not very meaningful in controlled social experiments as they are highly sensitive to aspects of the experimental design such as the framing of the decision situation, the monetary incentives, the experience of the participant pool, and experimenter demand effects (36).

Results

We first test the prediction that reputational information lowers the dispersion of payoffs in randomly rewired networks. Using Mann-Whitney tests to compare the Gini coefficients for payoffs accumulated at the end of interactions between the control and treatment conditions, we find statistically significant evidence for this prediction in BOLT04, SEIN06, and STAH13 (Fig. 2A). Overall, the effect direction is as predicted in seven out of the eight control-treatment pairs, yielding a sign test result that is significant at the 0.05-level (1-sided $p = 0.035$).

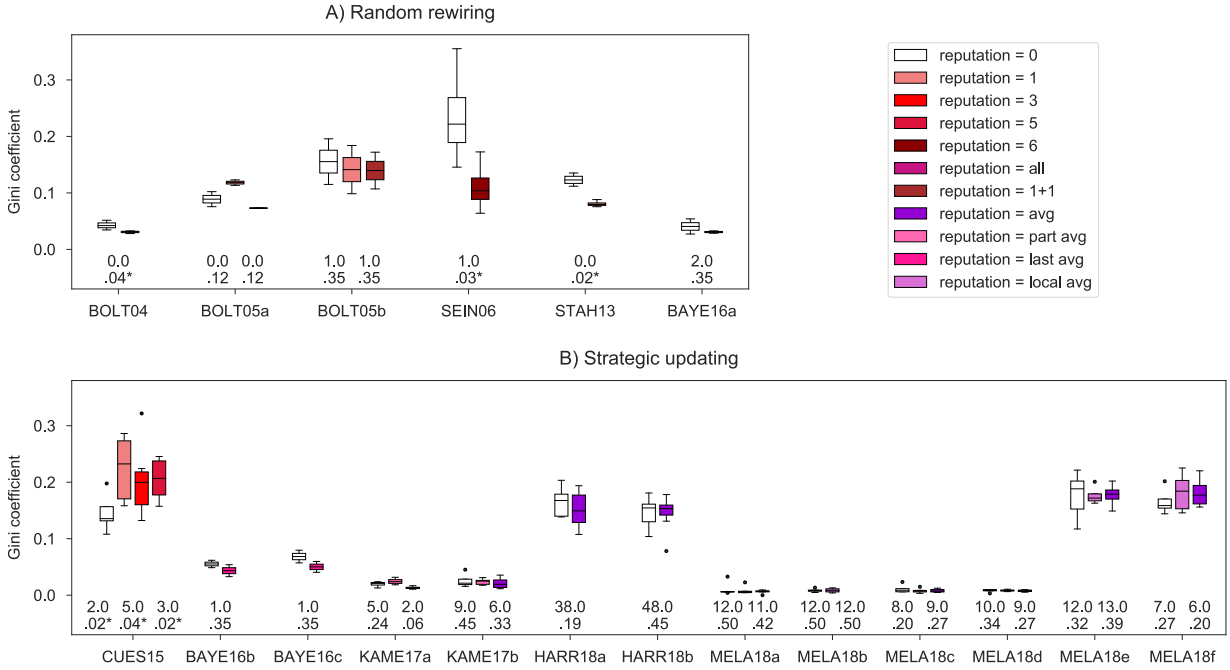

Figure 2. A) The empirical analyses confirm that reputational information decreases the dispersion of final payoffs in randomly rewired networks. B) As predicted, in networks with strategic updating, reputational information could increase the dispersion of final payoffs, as in CUES15, but this does not occur in most cases. The figure shows boxplots of the Gini coefficient for final payoffs for each experimental condition and results from Mann-Whitney tests comparing each condition with reputation to the control condition (reputation = 0) in each experiment (Mann-Whitney U on top and 2-sided p-value on bottom, with asterisk if $p < 0.05$). Description of the experimental setups and treatment conditions can be found in Table 1.

Next, we investigate whether reputational information could increase the dispersion of payoffs in strategically updated networks. We find statistically significant supporting evidence only in CUES15 (Fig. 2B). Out of the 23 control-treatment pairs, only ten have a positive effect from reputation on inequality, resulting in 2-sided $p = 0.678$ for the sign test. It is worth noting that this result is skewed by eight of the pairs from MELA18 where groups achieve nearly universal cooperation within a couple of periods and consequently, end up with little variation in final payoffs (Fig.S4). Nevertheless, even after excluding these outliers, we are left with results in both directions (8 positive out of 13, 2-sided $p = 0.999$). In sum, although possible, reputation is unlikely to produce higher inequality in strategically updated networks. This finding corroborates previous research showing that reputation does not contribute much to the clustering and proliferation of cooperators beyond what the possibility to exclude defecting partners can already do (26). The positive effect we find in CUES15 is likely contingent on two design decisions in the experiment: 1) the assumption that memory is a source of reputation, such that in the no reputation condition, individuals do not know even their own partners' actions, and 2) the payoff matrix with $CD = DD$, which means that cooperators do not lose by interacting with defectors. Because of these two assumptions, exclusion does not occur when reputation = 0, evidenced by the fact that the networks have much higher density then (25).

Regarding the expected positive effect of reputational information on the correspondence between payoffs and cooperative behaviour, we find supporting evidence that is statistically significant in BOLT04, SEIN06, STAH13, CUES15, and KAME17a (Fig.3). The overall results, however, are not statistically significant for randomly rewired networks (6/8, 1-sided $p = 0.145$) and significant for strategically updated networks only if we remove the eight outliers from MELA18 (12/15, 1-sided $p = 0.018$). Figures S5 and S6 show in more detail how reputational information increases both the level of cooperation and the rewards for cooperating in each experiment.

Discussion

This study aimed to expand our understanding of the effects of reputation systems in social groups. Complementing prior research, which focuses on cooperation and collective welfare, we investigated how reputational information affects inequality. We studied inequality in terms of both the dispersion and fairness of rewards: we analysed how the final payoffs of the poorest differ from those of the richest but also, how the final payoffs correspond to cooperative behaviour. Overall, our aim was to answer the question of whether reputation systems pose a tradeoff between efficiency and equality: Do the higher levels of cooperation and higher collective wealth imply undesirable side effects that have been overlooked? We argued and demonstrated with an agent-based model that reputation may increase the dispersion of payoffs under some conditions in strategically updated networks, but almost always decreases it in randomly rewired networks; it also provides better rewards to cooperative behaviour regardless of the dynamics of the interaction network.

We hypothesized and found partial empirical evidence that reputational information decreases inequality in interaction situations where partners are periodically randomly re-matched. Markets driven by one-off transactions that employ some form of randomization in allocating sellers to buyers exemplify such situations. Ride-hailing services such as Uber and Lyft present a good case in point, as they match drivers with customers based on time availability, which introduces a component of randomness. Networks that are periodically randomly rewired can also occur in schools and work environments, when teachers and managers randomly allocate students or employees to team projects. In this kind of networks, reputation increases the level of cooperation but also the rewards for cooperating, and this diminishes the dispersion of final payoffs. In other words, in randomly rewired networks, reputation works even better than we thought: it benefits both efficiency and equality.

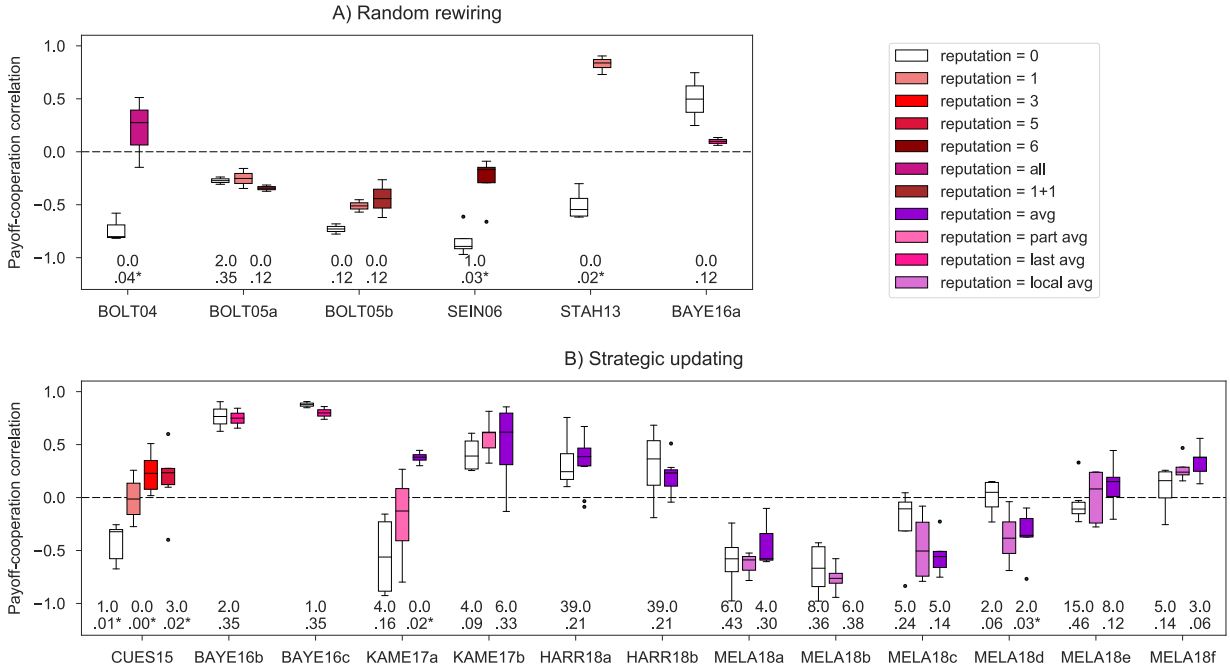

Figure 3. Reputational information generally improves the correspondence of payoffs to cooperation. The overall effect, however, is only significant in strategically updated networks (B) if we exclude MELAa-d, where the level of cooperation is > 95%. The figure shows boxplots of the Pearson correlation between final payoffs and individual cooperation, the latter defined as the proportion of periods in which the participant chose to cooperate. The text shows results from Mann-Whitney tests comparing each condition with reputation to the control condition (reputation = 0) in each experiment (Mann-Whitney U on top and 2-sided p-value on bottom, with asterisk if $p < 0.05$).

In interaction situations where individuals strategically choose their partners, we predicted that reputational information increases the dispersion of payoffs only under restricted conditions and indeed, we found that this is not common empirically. More importantly, just as in randomly rewired networks, in networks with strategic updating, reputation increases the rewards to cooperators and any increase in the dispersion of payoffs occurs at the expense of defectors, who are excluded and isolated. Yet, in practice, as (26) argue, strategic updating can already induce segregation between cooperators and defectors, and reputation does not contribute additionally. Thus, even in situations where partners are strategically selected, reputation systems appear to improve equality and fairness.

Our study extends and qualifies previous research showing that reputation systems could produce arbitrary and enduring inequalities in social dilemma situations (16, 17). Our fist contribution is to introduce, in addition to the dispersion of payoffs, another dimension to consider when analysing inequality in cooperation settings – the *correspondence* between payoffs and cooperative behaviour. This dimension is important to the extent to which we attribute cooperative behaviour to structural constraints and incentives, rather than individual agency. If cooperation is enabled by an arbitrary resource advantage (16) or constrained by the lack of opportunities to act (17), correspondence is less meaningful. When everyone faces the same decision situation, however, our sense of justice and fairness tells us that rewards should correspond to actions and this becomes an important consideration for inequality.

Our second contribution to existing research was to account for the multiplicity of equilibria. We accomplished this by presenting a generic model that considers different combinations of individual strategies and analysing experiments with different incentives and setups. This approach does not allow us to explain differences in the results between the model and any specific experiment, or between any pair of experiments, due to the fact that the model and the experiments differ on more than one dimension. Nevertheless, the approach allowed us to theoretically isolate the main driving factors behind the phenomenon we study – the emergent level of behaviour, the variability in individual behaviour, and the assortativity between individuals with similar behaviour – and to test the frequency and robustness of the predictions under variable empirical settings. Our work brings attention to the danger of drawing strong conclusions from a single model or experiment. Specifically, although prior research offers empirical evidence for the intuitive expectation that strategic partner selection increases inequality in terms of payoff dispersion (17), we revealed that this outcome is in fact rare for most network cooperation settings.

We acknowledge that the results presented here are not conclusive and that further research is needed. The biggest limitation of our study is that the empirical results regarding strategically updated networks are based only on symmetrical games, where both parties choose partners and face the same decision situation. What is more, the networks were degree-constrained, such that most individuals had a similar number of interaction partners. For example, this is the case when students team up for a course project. In contrast, buyer-seller markets involve asymmetric relations and preferential attachment, and research of such settings suggests that reputation systems could produce enduring inequalities (17, 37, 38). Our study could be extended with a bespoke experiment that investigates the effects of reputational information on inequality in strategically updated networks for both symmetric and asymmetric social dilemma games, systematically varying the maximum number of possible partners. This would test whether the negative effects of reputation on inequality are contingent on preferential attachment and exclusion, which prevent the opportunity to modify behaviour.

Overall, our present finding that reputation systems rarely worsen inequality in social dilemma situations in degree-constrained networks and, in fact, improve it under most conditions and in terms of rewarding cooperative behaviour suggests that it is possible to benefit from reputation systems without undesirable side effects. If reputation is used to encourage defectors to cooperate, rather than punish and isolate them, then everyone will be better off. We could achieve this by, for instance, implementing smart search algorithms in online markets and more deliberate team-formation strategies in schools and organizations.

Acknowledgments

This research was made possible through the generous support of the Volkswagen Foundation (https://www.volkswagenstiftung.de/) under Grant Ref. 92 173. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The author would like to thank Ralph-Christopher Bayer, Gary Bolton, David Melamed, Louis Putterman, Anxo Sánchez, Arthur Schram, and Dale Stahl for sharing data.

References

1. G. Roberts, *et al.*, The benefits of being seen to help others: indirect reciprocity and reputation-based partner choice. *This Issue*.

2. M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).

3. C. Wedekind, M. Milinski, Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).

4. M. Milinski, D. Semmann, H.-J. Krambeck, Reputation helps solve the 'tragedy of the commons.' *Nature* **415**, 424–426 (2002).

5. F. Fu, C. Hauert, M. A. Nowak, L. Wang, Reputation-based partner choice promotes cooperation in social networks. *Phys. Rev. E* **78**, 026117 (2008).

6. J. Wu, D. Balliet, P. A. M. V. Lange, Reputation, gossip, and human cooperation. *Soc. Personal. Psychol. Compass* **10**, 350–364 (2016).

7. K. Stainback, D. Tomaskovic-Devey, S. Skaggs, Organizational Approaches to Inequality: Inertia, Relative Power, and Environments. *Annu. Rev. Sociol.* **36**, 225–247 (2010).

8. H. Bapuji, G. Ertug, J. D. Shaw, Organizations and societal economic inequality: A review and way forward. *Acad. Manag. Ann.* (2019) https:/doi.org/10.5465/annals.2018.0029 (October 27, 2019).

9. A. Alesina, R. Di Tella, R. MacCulloch, Inequality and happiness: Are Europeans and Americans different? *J. Public Econ.* **88**, 2009–2042 (2004).

10. E. F. P. Luttmer, *et al.*, Neighbors as negatives: relative earnings and well-being. *Q. J. Econ.* **120**, 963–1002 (2005).

11. R. Wilkinson, K. Pickett, *The Spirit Level: Why Equality is Better for Everyone* (Penguin UK, 2010).

12. K. Takács, *et al.*, Networks of reliable reputations and cooperation: A review. *This Issue*.

13. R. K. Merton, The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis* **79**, 606–623 (1988).

14. T. A. DiPrete, G. M. Eirich, Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annu. Rev. Sociol.* **32**, 271–297 (2006).

15. A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley, Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc. Natl. Acad. Sci.* **108**, 18–23 (2011).

16. L. M. Hackel, J. Zaki, Propagation of economic inequality through reciprocity and reputation. *Psychol. Sci.*, 956797617741720 (2018).

17. V. Frey, A. van de Rijt, Arbitrary inequality in reputation systems. *Sci. Rep.* **6**, 38304 (2016).

18. F. B. Lynn, J. M. Podolny, L. Tao, A sociological (de)construction of the relationship between status and quality. *Am. J. Sociol.* **115**, 755–804 (2009).

19. E. Gallo, C. Yan, The effects of reputational and social knowledge on cooperation. *Proc. Natl. Acad. Sci.* **112**, 3647–3652 (2015).

20. K. Kamei, L. Putterman, Play it again: Partner choice, reputation building and learning from finitely repeated dilemma games. *Econ. J.* **127**, 1069–1095 (2017).

21. G. E. Bolton, E. Katok, A. Ockenfels, How effective are electronic reputation mechanisms? An experimental investigation. *Manag. Sci.* **50**, 1587–1602 (2004).

22. G. E. Bolton, E. Katok, A. Ockenfels, Cooperation among strangers with limited information about reputation. *J. Public Econ.* **89**, 1457–1468 (2005).

23. I. Seinen, A. Schram, Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.* **50**, 581–602 (2006).

24. R. Corten, S. Rosenkranz, V. Buskens, K. S. Cook, Reputation effects in social networks do not promote cooperation: An experimental test of the Raub & Weesie model. *PLOS ONE* **11**, e0155703 (2016).

25. J. A. Cuesta, C. Gracia-Lázaro, A. Ferrer, Y. Moreno, A. Sánchez, Reputation drives cooperative behaviour and network formation in human groups. *Sci. Rep.* **5**, 7843 (2015).

26. D. Melamed, A. Harrell, B. Simpson, Cooperation, clustering, and assortative mixing in dynamic networks. *Proc. Natl. Acad. Sci.*, 201715357 (2018).

27. M. Tsvetkova, C. Wagner, A. Mao, The emergence of inequality in social groups: Network structure and institutions affect the distribution of earnings in cooperation games. *PLOS ONE* **13**, e0200965 (2018).

28. U. Fischbacher, S. Gächter, E. Fehr, Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).

29. R. Kurzban, D. Houser, Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proc. Natl. Acad. Sci.* **102**, 1803–1807 (2005).

30. G. Spadaro, *et al.*, The Cooperation Databank (2020) https:/doi.org/10.31234/osf.io/rveh3 (May 12, 2021).

31. D. O. Stahl, An experimental test of the efficacy of a simple reputation mechanism to solve social dilemmas. *J. Econ. Behav. Organ.* **94**, 116–124 (2013).

32. R.-C. Bayer, Cooperation in partnerships: The role of breakups and reputation. *J. Institutional Theor. Econ. JITE* **172**, 615–638 (2016).

33. A. Harrell, D. Melamed, B. Simpson, The strength of dynamic ties: The ability to alter some ties promotes cooperation in those that cannot be altered. *Sci. Adv.* **4**, eaau9109 (2018).

34. P. D. Allison, Measures of inequality. *Am. Sociol. Rev.* **43**, 865–880 (1978).

35. Borenstein, *Introduction to Meta-Analysis*, 1 edition (John Wiley & Sons, 2009).

36. D. J. Zizzo, Experimenter demand effects in economic experiments. *Exp. Econ.* **13**, 75–98 (2010).

37. A. van de Rijt, S. M. Kang, M. Restivo, A. Patil, Field experiments of success-breeds-success dynamics. *Proc. Natl. Acad. Sci.* **111**, 6934–6939 (2014).

38. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).

Supplementary material for:

The effects of reputation on inequality in network cooperation games

Milena Tsvetkova

Department of Methodology, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom, ORCID ID 0000-0002-3552-108X

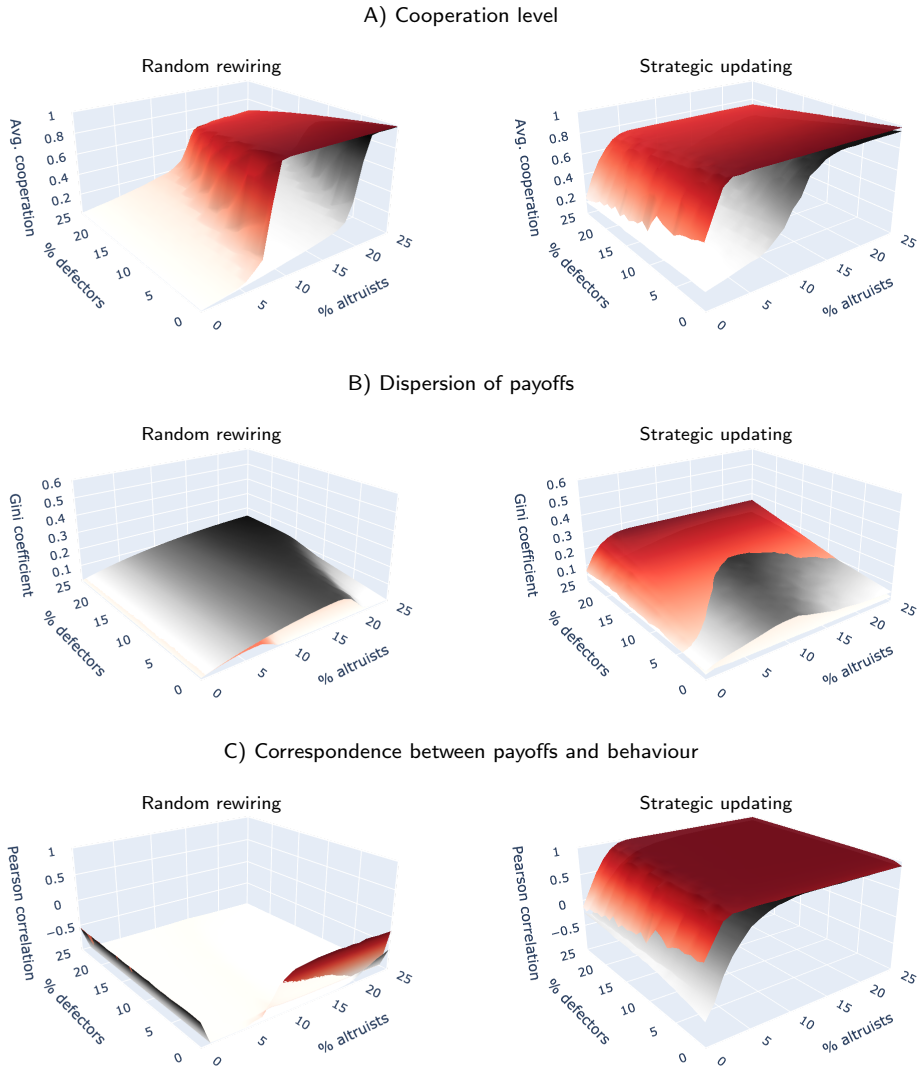

Figure S1. (A) The level of cooperation, (B) the dispersion of final payoffs, and (C) the correspondence between payoffs and cooperative behaviour in the agent-based model where networks are denser. The values are means calculated over 1000 runs in networks of 100 agents who have average node degree $m = 5$ and interact for 100 periods. Gray shows results without reputational information ($r = 0$), red – results when agents know everyone's actions in the previous period ($r = 1$). The simulations varied the percent of steady defectors (x-axis) and steady altruists (y-axis), with the rest being conditional cooperators.

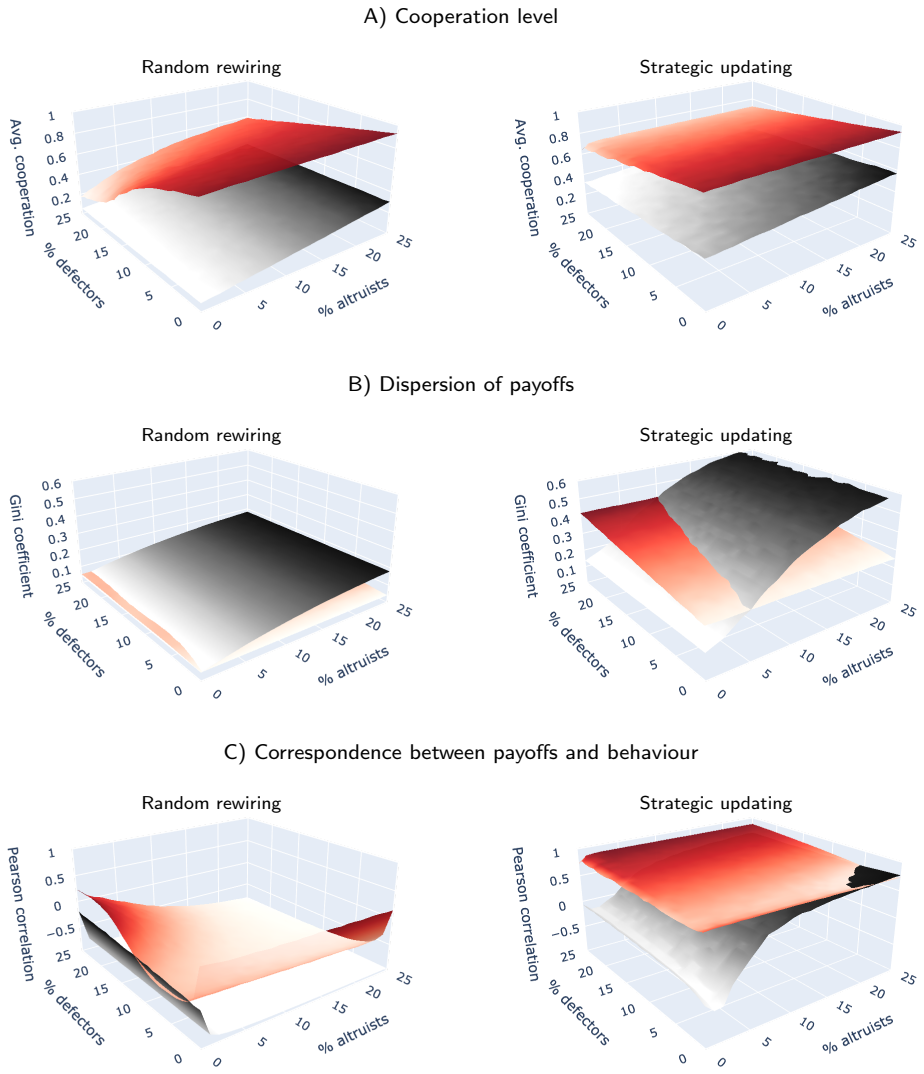

Figure S2. (A) The level of cooperation, (B) the dispersion of final payoffs, and (C) the correspondence between payoffs and cooperative behaviour in the agent-based model where payoffs are summed over all interactions in a period (rather than averaged). The values are means calculated over 1000 runs in networks of 100 agents who have average node degree $m = 2$ and interact for 100 periods. Gray shows results without reputational information ($r = 0$), red – results when agents know everyone's actions in the previous period ($r = 1$). The simulations varied the percent of steady defectors (x-axis) and steady altruists (y-axis), with the rest being conditional cooperators.

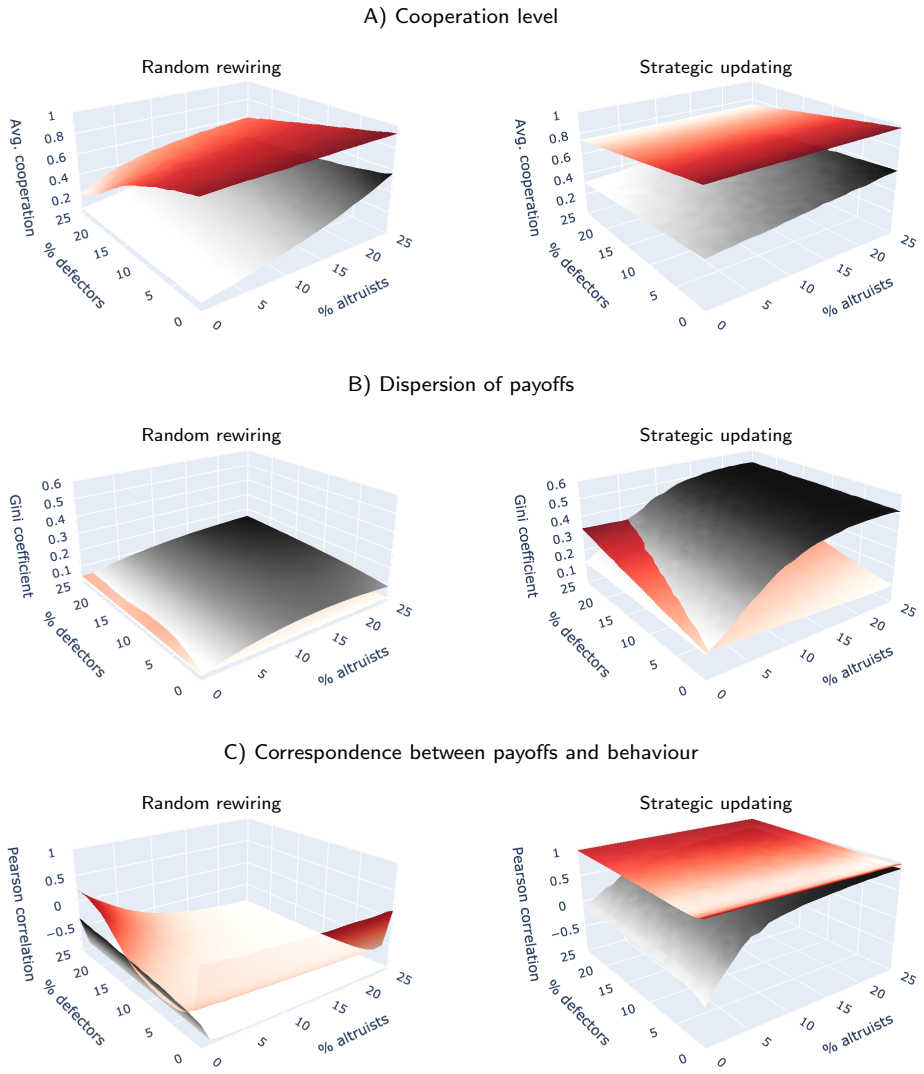

Figure S3. (A) The level of cooperation, (B) the dispersion of final payoffs, and (C) the correspondence between payoffs and cooperative behaviour in the agent-based model where reputational information covers more periods. The values are means calculated over 1000 runs in networks of 100 agents who have average node degree $m = 2$ and interact for 100 periods. Gray shows results without reputational information ($r = 0$), red – results when agents know everyone's actions in the previous three periods ($r = 3$). The simulations varied the percent of steady defectors (x-axis) and steady altruists (y-axis), with the rest being conditional cooperators. The left-hand plots are identical to the left-hand plots in Figure 1 since they are generated with the same model parameters.

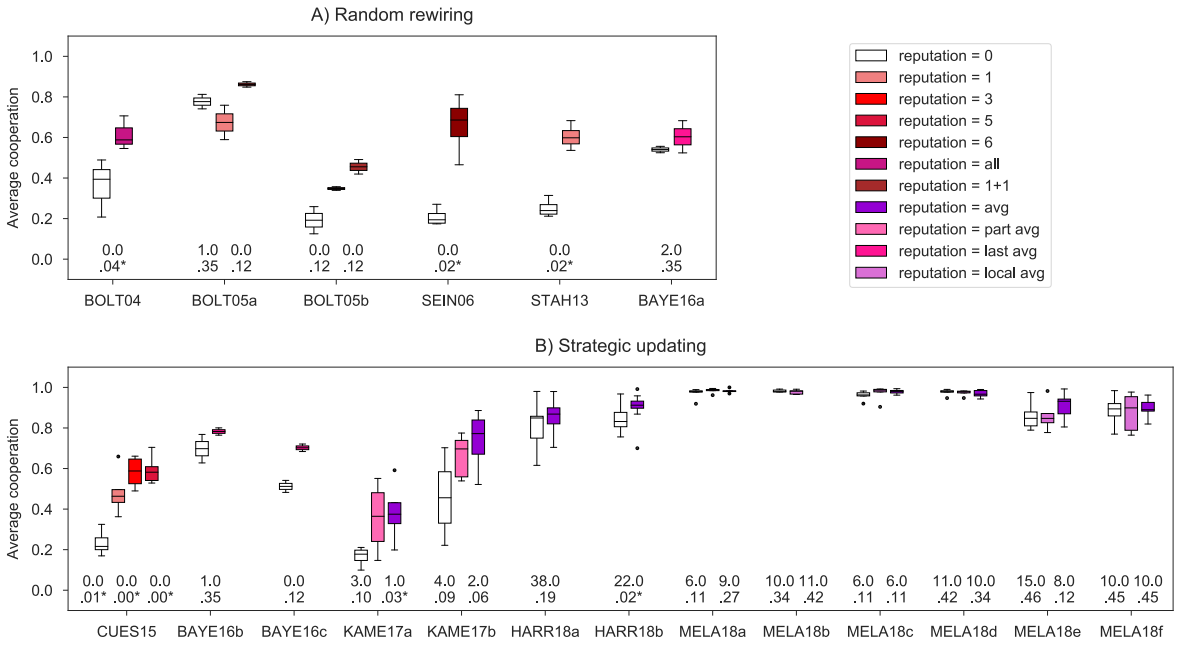

Figure S4. Group-level analyses replicate previously reported individual-level findings that reputational information increases cooperation: (A) 1-sided $p = 0.035$ from sign test for random rewiring; (B) 1-sided $p = 0.005$ from sign test for strategic updating. The effects of reputation are less pronounced for networks with strategic updating compared to networks with random rewiring because, there, cooperation could be nearly universal even without reputation. The figure shows boxplots for each experimental condition and results from Mann-Whitney tests comparing each condition with reputation to the control condition (reputation = 0) in each experiment (Mann-Whitney U on top and 2-sided p-value on bottom, with asterisk if $p < 0.05$). Description of the experimental setups and treatment conditions can be found in Table 1.

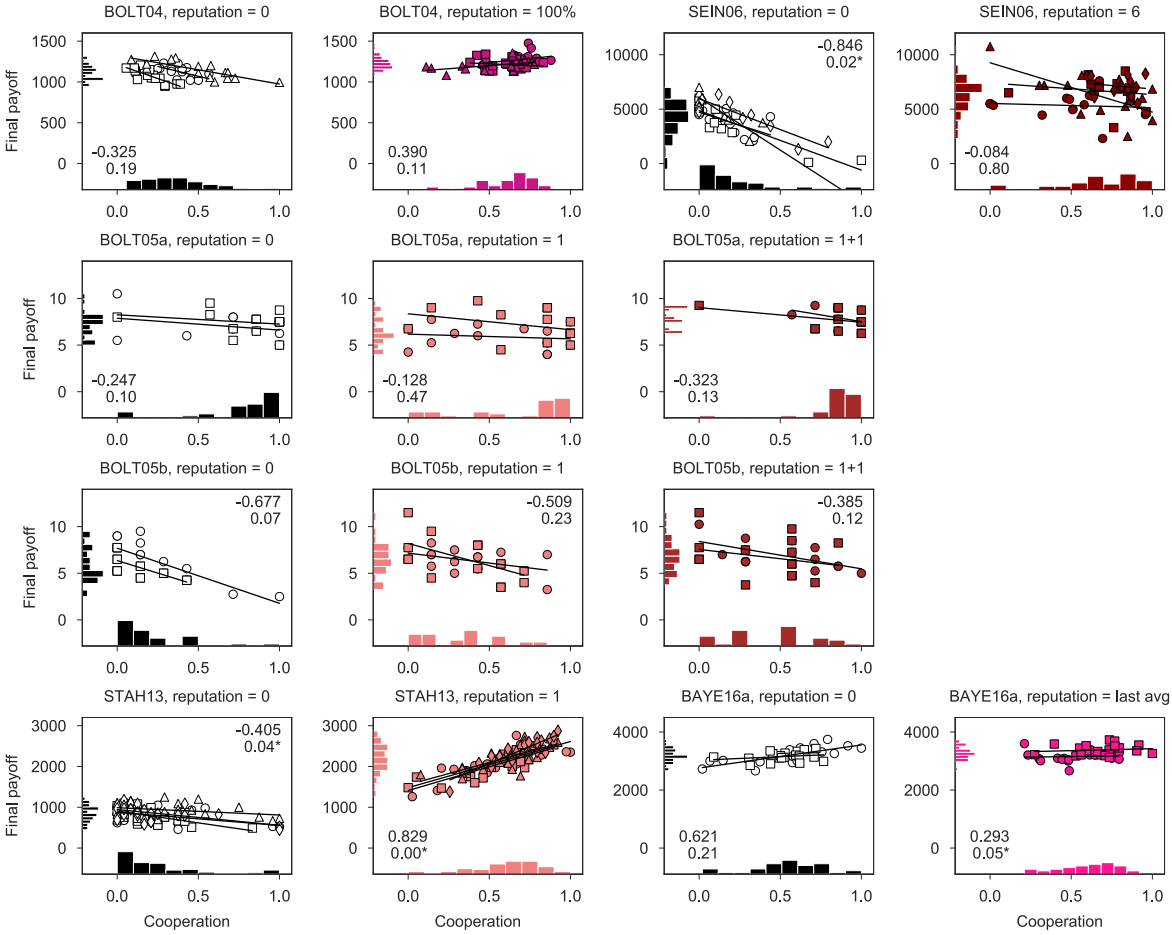

Figure S5. Empirical data from experiments on networks with random rewiring reveal that reputational information increases both the incidence of cooperation and the rewards for cooperating, skewing the distribution of final payoffs to higher values; as a result, the dispersion of final payoffs decreases, while the correspondence between payoffs and cooperative behaviour improves. The figure shows scatter plots for individuals' final payoffs by cooperation, as well as the distribution of final payoffs and the distribution of individual cooperation. Individual cooperation is defined as the proportion of periods in which the participant chose to cooperate. Lines show best linear fit for each interaction group and numbers show the standardized coefficient (top) and 2-sided *p*-value (bottom) estimated in ordinary least-square regression models with standard errors clustered by group.

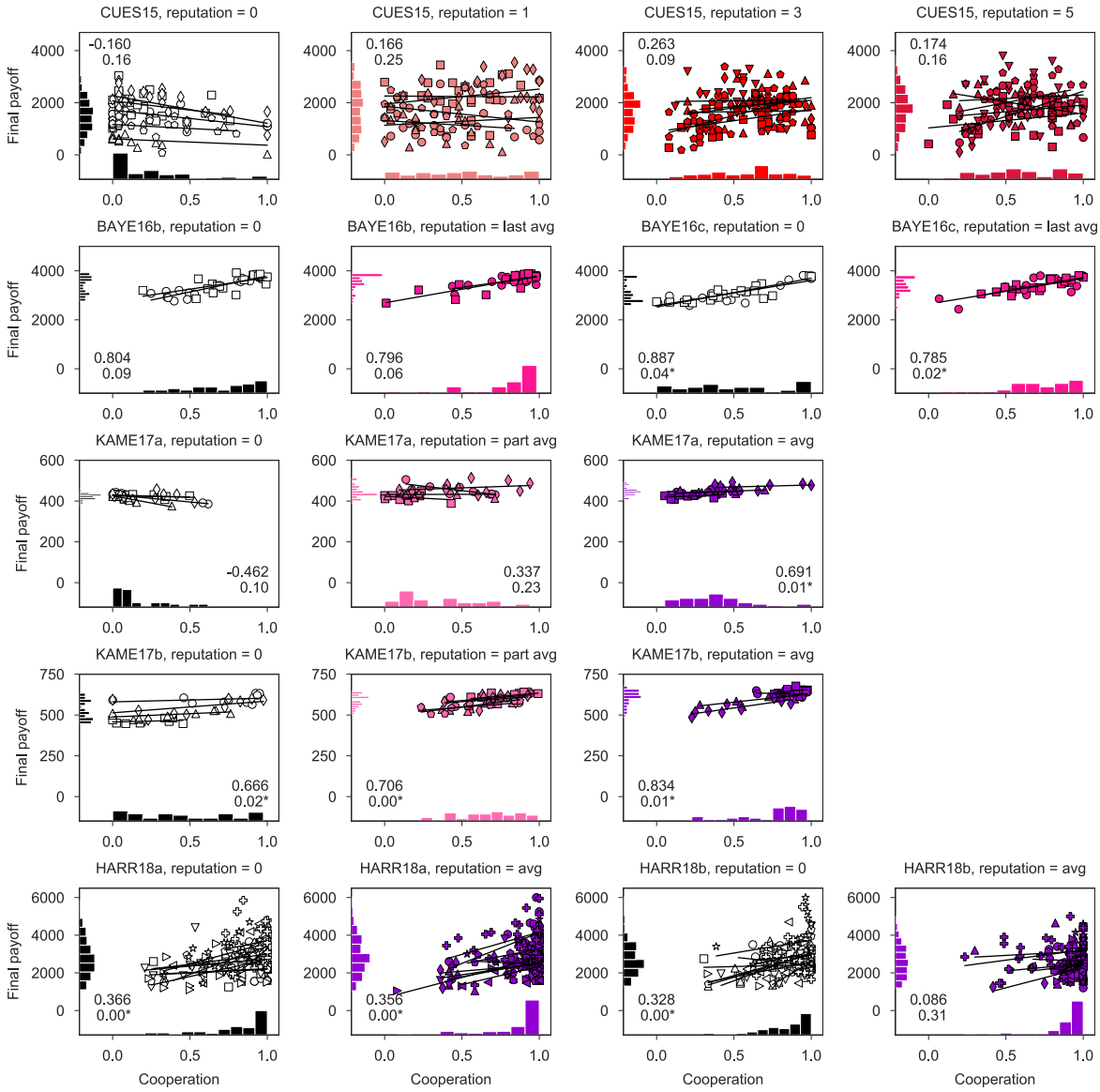

Figure S6. Empirical data from experiments on networks with strategic updating show that similarly to networks with random rewiring (Fig.S5), reputational information often increases both the incidence of cooperation and the rewards for cooperating. However, in contrast, the distribution of final payoffs does not change overall shape, with the exception of CUES15, where it becomes more stretched at the upper end. The figure shows scatter plots for individuals' final payoffs by cooperation, as well as the distribution of final payoffs and the distribution of individual cooperation. Individual cooperation is defined as the proportion of periods in which the participant chose to cooperate. Lines show best linear fit for each interaction group and numbers show the standardized coefficient (top) and 2-sided *p*-value (bottom) estimated in ordinary least-square regression models with standard errors clustered by group.