Value Homophily Benefits Cooperation but Motivates Employing Incorrect Social Information

Paul Rauwolf[a,*], Dominic Mitchell[a], Joanna J. Bryson[a]

[a]*Department of Computer Science, University of Bath, Bath, BA27AY, UK*

Abstract

Individuals often judge others based on third-party gossip, rather than their own experience, despite the fact that gossip is error-prone. Rather than judging others on their merits, even when such knowledge is free, we judge based on the opinions of third parties. Here we seek to understand this observation in the context of the evolution of cooperation. If individuals are being judged on noisy social reputations rather than on merit, then agents might exploit this, eroding the sustainability of cooperation. We employ a version of the Prisoner's Dilemma, the Donation game, which has been used to simulate the evolution of cooperation through indirect reciprocity. First, we validate the proposition that adding homophily (the propensity to interact with others of similar beliefs) into a society increases the sustainability of cooperation. However, this creates an evolutionary conflict between the accurate signalling of ingroup status versus the veridical report of the behaviour of other agents. We find that conditions exist where signalling ingroup status outweighs honesty as the best method to ultimately spread cooperation.

Keywords: indirect reciprocity, cooperation, gossip, homophily, self-deception

1. Introduction

It is frequently argued that the key advantage which drives the evolution of social learning compared to individual learning is that it provides more or better information at a lower cost. An individual that can benefit from what others know can draw knowledge from a wider range of experience at lower personal risk than one limited to their own immediate life events (Boyd and Richerson, 1985; Fernández-Juricic and Kacelnik, 2004; King and Cowlishaw, 2007; Magurran and Higham, 1988; Rendell et al., 2010). What happens when an individual discovers that the socially-received information is false? If correctness is the paramount concern, we might expect that false socially-learned information would be replaced by a more reliable source such as a first-hand experience.

*Corresponding author

Email addresses: p.rauwolf@bath.ac.uk (Paul Rauwolf), d.mitchell@bath.ac.uk (Dominic Mitchell), J.J.Bryson@cs.bath.ac.uk (Joanna J. Bryson)

There is mounting evidence that humans do not do this. Sommerfeld et al. (2008) tested the circumstances under which a participant would donate money to another individual. In each round, participants were paired and one person (the donor) was offered the opportunity to donate to the other (the recipient). Each donor was given either: *a*) the directly-observed history of the receiver's tendency to donate when the receiver had been a donor; or, *b*) the gossip-spread reputation of the receiver from third parties. Significantly more variation in the tendency to donate was explained by individuals' use of reputation compared to their use of direct observation. Furthermore, Lorenz et al. (2011) showed that individuals edit their answers to questions based on other people's responses, though this often makes the average response of the group less correct. Even compared to other species of primates, humans continue to persist with inaccurate social views longer (Whiten et al., 2009).

The null hypothesis is that the above behaviours are maladaptive exceptions to what is typically an adaptive heuristic. Social learning could be the best strategy despite a high incidence of error when the full cost of accruing accurate information, including time, is taken into account (Mitchell et al., in prep; Bryson, 2009). Further, researchers have proposed multiple heuristics by which humans bias their search for the most useful socially-acquired information. Conformity bias — acting with the majority (Henrich and Boyd, 1998), prestige bias — imitating the most prestigious (Henrich and Gil-White, 2001), pay-off bias — imitating the most successful (Mesoudi, 2011) are examples. Additionally, although social information transmission may introduce error, so may individual learning. Thus, in the rare situations where correct direct observation is easily attainable (e.g. Sommerfeld et al. 2008), individuals may employ noisy social information instead of correct directly observed information, because typically direct observation is expensive or similarly error prone.

These explanations argue for error prone social learning as the 'least-worst' option, and that the human tendency to employ social information in contexts where it is not useful is merely a local exception to a generally adaptive heuristic. However, the underlying assumption is that the utility of information (whether gossip or asocial) rests upon the accuracy of the information. Here we propose an alternative explanation for ignoring accurate personal experience in favour of social information. If social information comes with social prescriptions as to the employment of that information, then the factors influencing one's decision to utilize the information may extend beyond accuracy alone.

We demonstrate that ignoring veridical personal experience can facilitate the cooperative exchange of information more generally. In particular, the mechanisms that generally facilitate cooperation can create a dilemma between two levels of information: *a*) information about the transmitter, and *b*) information to be transmitted. We begin with a model of society where cooperation is regulated via reputation. Agents decide whether to donate to other agents and the reputation of the agent is spread throughout the population. We show that when homophily (the tendency to act with others who share similar beliefs) is added to this model, the robustness of cooperation is increased against error in communication. However, as a consequence, it becomes adaptive to employ incorrect social information even when an individual agent has access to correct information. In conditions where the pay-offs for group unanimity outweigh the costs of acting based on inaccurate information, there is selective pressure for norm-

following.

Our examination employs both computer simulations and formal analysis and proceeds as follows. First, we briefly introduce the literature on homophily and the evolution of cooperation. Next, we model the Donation game to examine the effects of error on the evolution of cooperation. The Donation game has been utilized as an existence proof for the evolution of cooperation in highly mobile societies (Nowak and Sigmund, 2005). It can be described as a specific instantiation of the Prisoner's Dilemma (Suzuki and Kimura, 2013; Masuda, 2012; Uchida and Sigmund, 2010) and continues to be used for studying cooperation both theoretically (Tanabe et al., 2013; Masuda, 2012; Hilbe et al., 2013; Stewart and Plotkin, 2013; Uchida and Sasaki, 2013; Marshall, 2011, 2009; Nakamura and Masuda, 2012) as well as experimentally (Angerer et al., 2014; Sommerfeld et al., 2008). We confirm that the Donation game and the spreading of reputation can be used to sustain cooperation (Panchanathan and Boyd, 2003; Nowak and Sigmund, 2005). This result is employed as a baseline for measuring cooperation.

Next, we analyze the effects of value homophily (the propensity to interact with those who share your beliefs) on cooperation. We find that as interactions become biased toward shared beliefs, cooperation becomes increasingly robust to error. Finally, we allow individuals to discover in isolation whether the social information they have received is incorrect. We test the consequences of acting on this information. We find that in homophilous societies, agents employing correct information are invaded by agents communicating known error. This demonstrates that honest signalling about own-group membership can outweigh the importance of honest signalling about others' behaviour. We discuss some of the consequences of the results for the literature on self-deception.

2. Model & Context

2.1. The Problem of Cooperation

In order to explore these issues, we need a context which meets certain requirements. First, the agents must learn valuable information socially. Second, that information must be subject to error. And finally, individuals must possess the ability to overrule what they socially learn, but in doing so breach a social norm. For our model, we implement a version of the Prisoner's Dilemma, called the Donation game (Marshall, 2011). This game has been used to show that cooperation can be established in a society when individuals exchange social information about the reputations of others (Nowak and Sigmund, 1998, 2005). We will give more details of the model in the following section, but first we review the problem of cooperation.

A cooperative society is defined as one in which individuals benefit from the collective absence of defection (Axelrod and Hamilton, 1981). However, it is often the case that for any individual member, defection is advantageous when others are cooperative. Several mechanisms have been hypothesised to overcome this problem of defection, notably reciprocal altruism (Trivers, 1971). In reciprocal altruism, an agent behaves prosocially with another so that the other will reciprocate at some later date. However, mobile societies, such as human ones, are often seen as vulnerable to free-riders (Enquist and Leimar, 1993 though see Schonmann et al., 2013). Individuals might

Action	Reputation
Cooperate	G
Defect	B

Table 1: First order norm. An agent receives a good reputation for performing the cooperate action, regardless of the reputation of the recipient. An agent receives a bad reputation for not cooperating.

	Good	Bad
Cooperate	G	B
Defect	B	G

Table 2: Judging norm. An agent receives a good reputation for cooperating with good, or defecting against bad agents. This second order norm can enforce cooperation but creates scope for descriptive/normative conflict. See further the main text.

defect opportunistically and move on before the consequences of their behaviour can catch up with them. In such cases, a different mechanism may be required to explain cooperation.

Indirect reciprocity (Nowak and Sigmund, 1998) solves this problem as an agent behaves prosocially with another because it is likely to subsequently receive a benefit from a different agent. This can be achieved when individuals observe each other, judge behaviour according to a norm, and pass on the resulting reputation via social transmission. Defectors can no longer free-ride, however mobile they are, so long as for every interaction they are likely to be preceded by their reputation and suffer a cost.

It should be clear that accuracy of information can be measured: for example, how closely an individual's reputation matches their actual behaviour. But to test the hypothesis described above, we need information to have further normative implications. We therefore need to distinguish between first and second order judgements.

2.2. First and Second Order Norms.

Previous studies on how the spread of reputation can establish cooperation employed a first order norm (see Table 1). If an agent cooperates, it receives a good reputation. Reputationally-aware agents cooperate with good agents, or defect against bad. But employing first-order norms introduces a threat to a cooperative society from an unexpected source: those who do good indiscriminately (Nowak and Sigmund, 1998). In a cooperative society, indiscriminate cooperators may invade the population via neutral drift. Whilst indiscriminate cooperators do not directly introduce defection, since they do not pay heed to the reputation of others, a society comprised of indiscriminate cooperators is unable to punish defectors via exclusion, and thus is likely to be invaded by them. Therefore, it is not sufficient for cooperation to exploit a reputation for doing good or bad — a first order judgement. Instead, cooperation also requires reputational impact for doing good to the good and bad to the bad — a second order judgement (see Table 2, Ohtsuki and Iwasa, 2004; Panchanathan and Boyd, 2003).

However, the move to a second order norm cannot be made without introducing the risk of conflict between actions motivated by social versus asocial knowledge. When an individual interacts with another who is socially negatively reputed, but discovers in isolation that this partner's reputation is unwarranted, what is the individual's best strategy? If they respond 'honestly,' that is in accordance with their descriptive knowledge about their partner, the actor risks putting itself in breach of the norm their peers employ.

2.3. Descriptive versus Normative

We term as *descriptive* aspects of knowledge where the utility depends on its accuracy with respect to the target of its description. Our hypothesis, tested here, is that there may be instances where the utility of the information is not solely constituted by its accuracy. To illustrate how social information may come with implications beyond the solely descriptive, consider the following thought experiment. One friend advises another against a certain restaurant, but despite the advice the latter goes there to dine. We ask the question: regardless of the eventual quality of the restaurant, does going against the advice of one's peer have implications?

When an individual accepts information from others, this hypothesised aspect of knowledge which goes beyond the descriptive, we term *normative*. The distinction between normative and descriptive becomes salient when the individual possesses accurate personal experience which contradicts the socially-held view. In this case, if the utility of the information is solely descriptive, the best strategy is obvious: the individual benefits from overruling the inferior source of information with the more reliable one. But if the social information comes with implications beyond the descriptive, the choice is not so simple. The benefit gained from prioritizing accuracy of information may be outweighed by the costs incurred in terms of reputation. In this case, individuals may gain by employing information they know to be false if they prioritize a normative response to gossip (such as signalling ingroup membership) over a purely descriptive one. In this paper we explore the consequences of agents making such a choice.

2.4. Value Homophily

The likelihood of being faced with this dilemma not only depends on the amount of error in gossip, but also the topology of the social network through which the gossip transmits. Various studies have considered the implications of social network structure on cooperation (Ohtsuki et al., 2006; Santos et al., 2008; Taylor et al., 2007). However, we explore the consequences of varying the probability of an agent interacting with another who shares the same socially inaccurate information. To do this, we use the concept of value homophily.

Value homophily is the tendency to associate with individuals who share similar beliefs and values (McPherson et al., 2001). Computational models have argued for the evolutionary feasibility of homophilous behaviour biases (Fu et al., 2012) and that the propensity for homophily can lead to local cultural convergence with disparate global beliefs (Axelrod, 1997).

Furthermore, value homophily has been prevalently diagnosed within human society. Curry and Dunbar (2013) showed that shared hobbies, moral beliefs, and a sense of

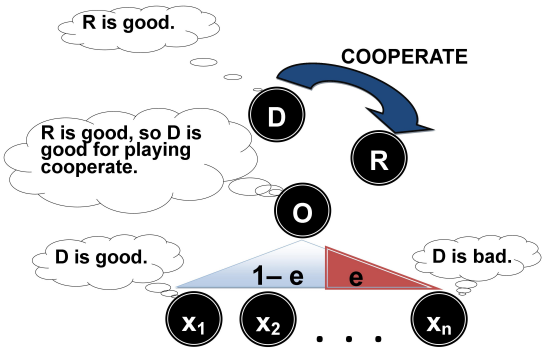


Figure 1: During a round of the Donation game the donor (D) either cooperates (pays a cost c) with the receiver (R) (gains a benefit b), or defects (no cost paid). The donor's behaviour is judged by an observer (O) based on a norm, and its reputation is altered accordingly. The observer then spreads the donor's newly formed reputation to the rest of the agents ($x_1 ... x_n$). However, the reputation may be spread with some amount of error.

humour are correlated to frequency of communication between friends. Humans more frequently interact with others who share values in dating (Fiore and Donath, 2005), drug use (Kandel, 1978), and several other self-reported personality traits (Adamic et al., 2003). Humans expect like-minded individuals to be more intelligent, moral, and knowledgeable of current events (Byrne, 1961). Ross et al. (2013) studied the differences in a folk-tale across geographic and cultural topographies. They found that "folktales from the same culture found 100 km apart are, on average, as similar as folktales found 10 km apart in different cultures" (Ross et al., 2013, p. 6). By adjusting the levels of value homophily, we consider the advantages of employing accurate personal information over inaccurate social information.

3. Model 1: Cooperation via Indirect Reciprocity

3.1. Simulation

Here we demonstrate the evolution of cooperation using the Donation game. This experiment replicates what is already well-known; reputation can sustain cooperation even against error in communication (Panchanathan and Boyd, 2003). We use this as a baseline for judging the sustainability of cooperation in the subsequent experiments. As such, we have chosen parameters in line with previous instantiations of the Donation game so that subsequent experiments are comparable with the existing literature.

In a round of the Donation game an individual has the opportunity to pay a cost to give some other agent a benefit. Initially, there are two players, and the *donor* decides to cooperate or defect with the *receiver*. If cooperation is selected, the donor pays a cost c to give a benefit b to the receiver. If the donor defects, the pay-off is zero for both players. As a result, the donating agent may garner a positive (or negative) reputation and be aided (or not) by someone else at some later date.

Symbol	Parameter	Value	Notes
n	population size	100	number of agents.
r	rounds	10,000	Donation games per generation.
g	generations	500	# of evolutionary iterations.
b	benefit	[1 - 6]	incremented by 0.5
c	cost	1	cost for cooperating. kept constant.
u	mutation rate	0.01	chance of strategy mutation.
h	homophilly	0, 0.5, 1	see main text.
e	error	0–1	see main text.
v	veracity	0 or 1	see main text.

Table 3: Table of free parameters, values used in present figures, and a sensitivity report including range of values tested.

We use a three-player version of the Donation game. Here, an additional player (the *observer*) is permitted to monitor the interaction (see Figure 1). The observer then spreads its reputational judgement of the donor. This article assumes an observer in all interactions, and that the observer judges the reputation of the donor via the 'Judging' social norm, explained below. Other norms are explored in the appendix.

An observer alters its belief in the donor's reputation via the Judging norm (see Table 2). When observers employ this norm, cooperation can be sustained via indirect reciprocity (Ohtsuki and Iwasa, 2006, 2004; Uchida and Sasaki, 2013; Uchida and Sigmund, 2010). If the donor elects to cooperate and the observer believes the recipient's reputation to be good, the observer will assign the donor a good reputation. Conversely, if the donor cooperates, but the observer believes the receiver to be bad, then the observer marks the donor as bad. The reverse is true if the donor defects. The observer then shares its new reputational view of the donor with the rest of the population, and they update their beliefs accordingly, though sometimes errors occur in communication (see Figure 1).

As per Nowak and Sigmund (1998), three strategies are considered: *always defect* (ALLD), *always cooperate* (ALLC), and *discriminate* (DISC). As donors, discriminating (DISC) agents cooperate or defect based upon their belief in the reputation of the receiver. If the donor believes the agent to be good, it will cooperate, otherwise it will defect. The other two strategies do not consider the recipient's reputation, but always cooperate or defect, as per their namesake.

Error (e) is limited to the range [0...1] and represents the percent of the population which receives the wrong reputational adjustment from the observer. If $e = 0$, then the observer convinces the entire population of the donor's new reputation. Otherwise, some fraction of the populace receives the wrong reputation. For instance, if $e = 0.2$, then 20% of the population inaccurately updates their reputation for that donor (see Figure 1). The individuals which receive the erroneous reputation update are selected at random during each interaction. Thus, one agent might possess inaccurate information for Agent A, but not Agent B.

We assume that there is a population of $N = 100$ individuals each with a single triallelic loci representing strategy and a list (of size $N - 1$) representing their *belief* about the reputation of every other individual. Each element of the list exists in one

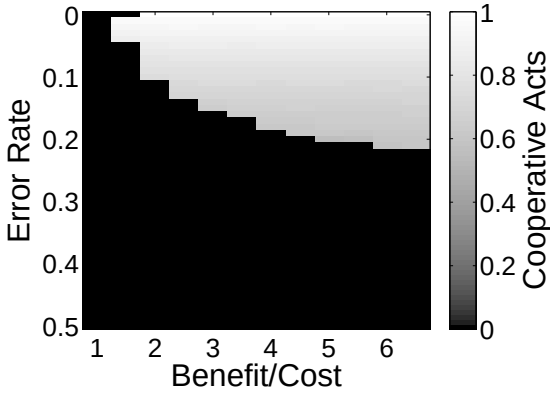


Figure 2: Average fraction of Donation games which were cooperative over final 50 of 500 generations. A parameter sweep of the error rate and benefit/cost ratio illustrates that the evolutionary sustainability of cooperation depends on the interplay of both parameters. See Table 3 for complete details of parameters

of two states, which we call *good* and *bad*. In order to test the sustainability of a cooperative society, initially the population is comprised solely of DISC agents. We then analyze whether cooperation can be maintained in the face of invading ALLDs, which may invade indirectly via vulnerable ALLCs.

During each round (r), an instance of the Donation game occurs with three randomly selected agents taking on the role of recipient, donor, and observer. Based on its strategy, the donor will choose whether or not to donate to the receiver. If the donor cooperates, it will pay a personal cost c to give a benefit of b to the receiver. Based on the donor's behaviour and the observer's belief of the reputation of the receiver, a fraction of the population $1 - e$, will change their reputational belief of the donor to sync with the observer's new belief. The rest of the population (e), will erroneously update their belief to be the opposite of the observer's.

These interactions are repeated for $r = 10,000$ rounds, constituting a single generation. At the end of a generation, the total pay-off (i.e. the sum of all the costs paid and benefits received of each individual during the generation) is calculated. This represents an individual i's fitness (p_i), which is used to calculate the proportion of the various strategies in the subsequent generation. To do this, we employ the evolutionary selection algorithm, fitness proportionate selection (Goldberg and Deb, 1991). In fitness proportionate selection, each member of the new generation is determined by selecting an individual from the previous generation, with the chance of selecting individual i being $p_i/(\sum_{j=1}^{n} p_j)$, the payoff of individual i divided by the total pay-off of the population. There is then a small chance for mutation $u = 0.01$, where an agent's strategy morphs to one of the three available strategies. The experiments runs for $g = 500$ generations (see Table 3 for parameters and sensitivity).

Results

Figure 2 illustrates the results, displaying the average fraction of cooperative actions per generation over the final fifty generations. The figure displays the fraction of cooperative acts through a parameter sweep of error rate (e) and benefit/cost (b/c) ratio. Clearly, there exists a threshold of discontinuity. In the white area of the curve the vast majority of actions are cooperative, whilst in the dark area cooperation destabilizes and ALLDs invade. The outlier at $b = 1.5$ and $e = 0$ illustrates that occasionally, with low error rates, ALLCs invade and are subsequently vulnerable to ALLDs (Fishman, 2003; Ohtsuki and Iwasa, 2004). Generally, the figure illustrates that as the benefit/cost ratio increases, a cooperative society can support increased error in reputational dissemination. However, regardless the benefit to cost ratio, cooperation seems to be limited by an error rate of approximately 0.2.

4. Model 2: Homophily

4.1. Simulation

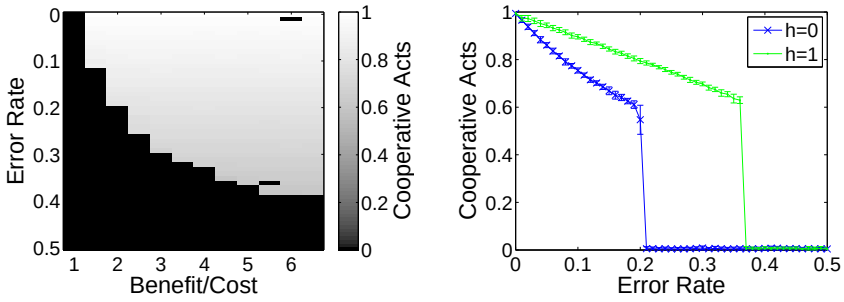
Model 1 shows that when agents randomly interact, selection for cooperation is limited by error in communication. However, in human societies, random interaction is uncommon; a correlation is frequently found between the beliefs of individuals and their propensity to interact (Adamic et al., 2003; Curry and Dunbar, 2013; Fiore and Donath, 2005; Kandel, 1978; McPherson et al., 2001). In Model 2 we use both computer simulations and formal analysis to examine the effects of this phenomenon, called *value homophily*.

To analyze the effects of value homophily on the evolution of pro-social behaviour, we define it as the correlation between informational or belief relatedness and the propensity to interact. Thus, we seek to juxtapose a randomly interacting society (low value homophily) with a society where individuals tend to interact with like-minded others (high value homophily).

We presume that there is some abstracted cultural variable affecting the distribution of reputation. Namely, an agent will more frequently interact with another agent that shares its reputational assessment of a given recipient compared to a random interaction. Thus, we added a new variable, homophily, $h = [0...1]$.

Rather than defining a specific social structure or network, we operationalised homophily as the probability of a donor interacting with an observer who agrees with the donor's belief in the reputation of the recipient. For each interaction the donor, recipient, and observer are initially chosen at random. However, if the donor and the observer disagree on the recipient's reputation, then h represents the probability of replacing the observer with an agent who agrees with the donor's assessment of the recipient's reputation (if such an agent exists). As an example, if $h = 0.2$, then if the donor and observer disagree, there is a 20% chance a different observer will be chosen. This observer will be drawn from the subset of agents that share the donor's belief with respect to the specific recipient.

Importantly, it should be noted that the propagation of error (e) is unchanged. Error is introduced entirely by observers misinforming e percent of the population. The only difference between Models 1 and 2 is via the selection criterion of the observer. Thus,



(a) Average fraction of cooperative acts.

(b) Comparing $h = 1$ with $h = 0$.

Figure 3: **(a)** Depicts the average fraction of Donation games which were cooperative over the final 50 generations. Homophily is 100% ($h = 1$). **(b)** Comparison of the fraction of cooperative acts with $h = 1$ and $h = 0$. Benefit/cost is held static at 5. Error rate varies [0,0.5] in increments of 0.01

when the donor is randomly selected, the probability that the donor's belief concerning the recipient's reputation is erroneous remains unchanged from the first model.

We define homophily as congruency between the donor and observer. However, our choice is unique compared to recent work on indirect reciprocity involving in-group/outgroup dynamics. In much of the literature, it is the donor and recipient, rather than the donor and observer, that belong to the same group (Nakamura and Masuda, 2012; Masuda, 2012; Jusup et al., 2014; Matsuo et al., 2014).

What is the definition of ingroup in the context of indirect reciprocity? The observer is at least as likely to share the donor's social network as the recipient is. As indirect reciprocity is founded upon the assumption that agents do not meet twice, consider an example with mobile agents. In Village Xan, Alex is well-reputed, but in Village Yar, Alex is ill-reputed. A highly-mobile agent visits Xan, hears the reputation of Alex, and then meets Alex. Since the mobile agent is visiting Xan, when it meets Alex, it has a higher probability of being observed by someone from Xan, rather than from Yar. Thus, the mobile agent will likely be observed and judged by the people of Xan. After the interaction, the mobile agent moves on to other villages and is judged by their inhabitants. In this sense, an agent does not have a group with which it interacts, but when it interacts with an agent, it is observed by the group who informed the wandering agent of the recipient's reputation. We return to this point in the discussion and test the consequences of defining homophily differently.

Results

Figure 3 illustrates the consequences of homophily on the sustainability of cooperation. If one juxtaposes Figure 3(a) — which shows the maximum degree of homophily ($h = 1$) — with Figure 2, where $h = 0$, it is clear that given unanimity between the donor and the observer, stable cooperation can tolerate greater amounts of error. In Figure 2 cooperation fails at an error rate of approximately $e = 0.2$. With homophily, Figure 3(a), cooperation can sustain an error rate of around $e = 0.4$. Furthermore, this result is reproducible in a variety of environments. In Appendix A we show that this

result extends to other social norms (e.g. Standing and Shunning). In Appendix B the result is duplicated when reputation is not binary, but continuous.

Figure 3(b) compares a subsection of the results underlying Figures 2 and 3(a), where the benefit/cost ratio is 5. The graph shows the average fraction of cooperative acts for both full homophily ($h = 1$) and randomly selected observers ($h = 0$). For a given error rate, the number of cooperative acts is increased when the donors and observers agree upon the recipient's reputation, regardless of the accuracy of the reputation. When $h = 0$ cooperation fails at $e \approx 0.2$, while unanimity in reputational assessment allows stable cooperation to survive twice the error rate ($e \approx 0.4$). Furthermore, not only does homophily sustain cooperation against more error, when cooperation is sustained for either society, a homophilous society also performs more cooperative acts. For example, at $e = 0.1$ cooperation is stable for both homophilous and non-homophilous societies, however more cooperative acts take place when $h = 1$.

4.2. Simplified Analytical Model

While simulations demonstrate difficult to elucidate consequences of a theory, it cannot speak to the underlying mechanism. Here analytically model the repercussions of homophily on cooperation. We make a few simplifying assumptions compared to the simulation. First, we presume that an agent will first be selected as a donor and then as a recipient. Second, we only consider two strategies, DISC and ALLD. With the Judging norm and error in communication, there is a relatively low danger of ALLC agents drifting into the population sufficiently to risk a subsequent ALLD invasion (Takahashi and Mashima, 2006; Fishman, 2003; Ohtsuki and Iwasa, 2004). Consequentially, we focus on when ALLDs can invade a society of DISCs, and how a homophilic society guards against this. Finally, we do not employ evolution, rather we only test when rare ALLDs perform better than a population comprised almost entirely from DISCs.

We presume an infinite population where Donation games are played for a finite, but extended period of time. We define P_r and P_d as the probabilities of interacting with a DISC (i.e. reciprocater) and an ALLD, respectively. After each game, w is the probability that another round of the Donation game will take place. Similar to Panchanathan (2011), we presume that the proportion of reputations goes to equilibrium by the second round. Given this, the fitness of an ALLD (π_d) may be expressed as:

$$\pi_d = bP_r + \frac{w}{1-w}(bP_rG_d) \tag{1}$$

In the first round, each agent is considered good, so an ALLD agent will receive a benefit it is meets a DISC agent (bP_r). In subsequent rounds, the ALLD agent will receive a benefit if it meets a DISC agent, and the DISC agent believes the ALLD agent is good. G_d represents the probability that another agent considers the ALLD agent good.

The fitness of a DISC agent (π_r) may be expessed by equation 2. In the first round, the agent will receive a benefit if it interacts with another DISC. Additionally, it will always pay the cost of cooperating as each agent begins with a good reputation ($bP_r - c$). In subsequent rounds, it will receive a benefit if the agent interacts with another DISC, and that agent believes the DISC to be good (G_r). It will pay the cost of cooperating if it interacts with any agent who it believes to be good.

$$\pi_r = bP_r - c + \frac{w}{1-w}[bP_rG_r - c(P_rG_r + P_dG_d)] \tag{2}$$

In our simulation we test whether a population of DISCs is stalwart against invasion. In the analytical model, this would be true as long as the average fitness of DISC agents is greater than the average fitness of ALLD agents:

$$\pi_r - \pi_d > 0 \tag{3}$$

Furthermore, in the invasion scenario, we presume that initially $P_r \approx 1$ and $P_d \approx 0$. Presuming play continues for an extended period ($w \approx 1$), we expect a population of DISCs to be stable against invasion if (see Appendix D for details):

$$G_r > \frac{bG_d}{b-c} \tag{4}$$

Equation 4 shows that the stability of a DISC population is dependant on the likelihood that agents of a strategy are well-reputed. The probability of having a good reputation not only depends on a donor's actions, but on the error rate of reputational dissemination, and how likely agents with different reputational beliefs are to interact. This is how homophily alters the interactions of a population.

In a randomly interacting population ($h = 0$), the probability of a DISC agent possessing a good reputation is:

$$G_r^{h=0} = (1-e)[(1-e)^2 + e^2] + e[(1-e)e]^2 \tag{5}$$

Given the Judging social norm, a donor DISC agent earns a good reputation if the observer and it either: *a*) agree on the reputation of the recipient, and the donor's reputation is passed on correctly, or *b*) disagree on the reputation of the recipient, but the donor's reputation is passed on erroneously.

In the first case, there are two ways the donor and observer can agree on the reputation of the recipient. Either, both possess correct information about the recipient $(1-e)^2$, or they both possess incorrect social information, e^2. The good reputation is then disseminated to $(1-e)$ fraction of the population. In the second case, the donor and observer disagree with probability $[e(1-e)]^2$, and the DISC agent receives a bad reputation. However, e fraction of the population are erroneously told that the donor is good.

In a population with homophilic interactions (h), the probability of a DISC agent possessing a good reputation is altered. Equation 5 becomes:

$$G_r = h(1-e) + (1-h)G_r^{h=0} \tag{6}$$

If $h = 0$ then Equation 6 = Equation 5, since each Donation game is played with randomly selected agents. In a fully homophilous society $h = 1$, the donor and the observer will always agree, therefore the DISC agent will always earn a good reputation. However, because there is error in transmission, only $1 - e$ fraction of the population will believe the donor is good. Thus, $G_r = 1 - e$.

The probability of an ALLD agent receiving a good reputation (G_d) is:

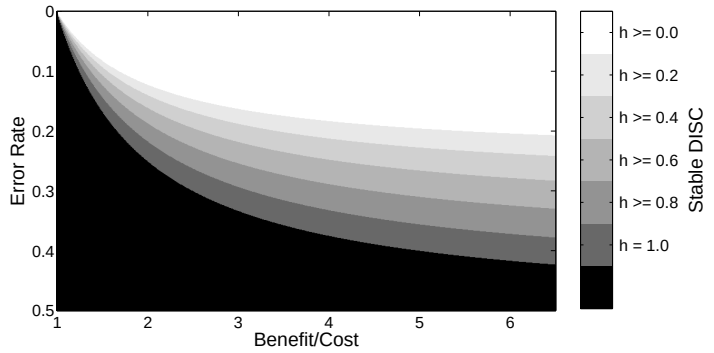


Figure 4: Plotting Equation 4 for differing values of homophily. The **black** area shows where ALLDs are predicted to invade. The **grey** areas represent where, for certain values of homophily, DISCs are stable against invasion. Lighter greys are subsets, of darker greys. For example, the area of DISC stability for $h = 0.0$ is a subset of $h = 0.2$. As the frequency of homophilous interaction increase, DISCs can avoid invasion despite greater frequency of errors in communication.

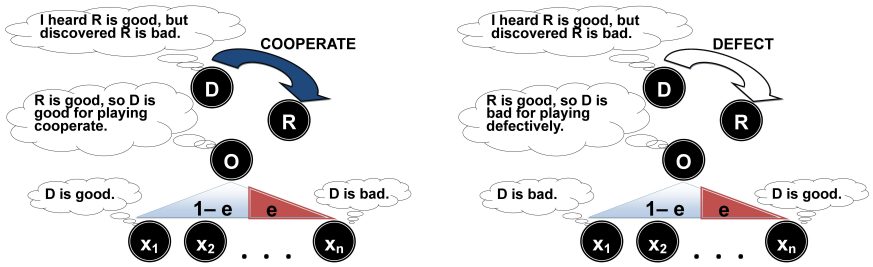
$$\begin{aligned} G_d &= (1-e)[P_r(1-G_r) + P_d(1-G_d)] + e[P_r(G_r) + P_d(G_d)] \\ G_d &= (1-e)(1-G_r) + eG_r \end{aligned} \tag{7}$$

An ALLD always defects, so it will receive a good reputation if the observer believes the recipient is bad. If the recipient is a DISC agent, then the observer will believe the agent is bad with probability $(1 - G_r)$. Similarly, if the recipient is an ALLD, then the observer will hold it in negative repute with probability $(1 - G_d)$. The observer will then spread this good reputation to $(1 - e)$ of the population. Furthermore, if the observer believes the agent is good, it will give the ALLD donor a bad reputation, but e fraction of the population will erroneously consider the agent good. However, since the assumption for invasion is that $P_r \approx 1$ and $P_d \approx 0$, the equation simplifies to $G_d = (1 - e)(1 - G_r) + eG_r$.

The probability an agent will consider an ALLD agent good (G_d) does not change based on homophily. In a homophilous interaction a donor and recipient are chosen at random. The observer is then selected to agree with the donor, but the donor's opinion of the recipient is selected from the same probability as the observer's opinion of the recipient. However, while Equation 7 remains the same for any value of homophily, the calculation of G_r changes, so indirectly the value of G_d is altered.

Results

Figure 4 illustrates where the model predicts DISC stability, $G_r > \frac{bG_d}{b-c}$. A parameter sweep of benefits (b) and error rates (e) are tested. The black area defines where $G_r \leq \frac{bG_d}{b-c}$, and thus, where we always predict an ALLD invasion. The grey areas depict parameter locations where $G_r > \frac{bG_d}{b-c}$ for certain values of h, and thus, where a DISC population is stable.



(a) DISC (b) VDISC

Figure 5: Difference between veridical (VDISC) and socially-biased (DISC) agents. The donor (D) has been socially informed that the recipient (R) is good, but has discovered that this information is inaccurate. **(a)** The DISC donor acts on normative information and is judged to be good. **(b)** The VDISC donor acts on descriptive information, despite negative normative consequences.

Figure 4 is in line with the simulation results. The whitest curve in Figure 4 ($h = 0$) is comparable to Figure 2, and the darkest grey (that is not black, $h >= 1$) is comparable to Figure 3(a). Despite the exclusion of ALLCs and the lack of evolution, the analytic model predicts similar DISC stability.

The simulations showed that cooperation is more robust to error in homophilous societies. Here we demonstrate the reason why. Homophily increases the ratio of good DISCs over good ALLDs (i.e. G_r/G_d). When the benefit (b), cost (c), and error rate (e) are held constant, larger values of h increase the robustness of a cooperative society.

5. Model 3: Incorrect Social Information v. Correct Personal Information

5.1. Simulation

Lastly, we introduce the conflict between descriptive and normative information. Thus far, we have assumed that reputational errors were spread randomly, without intention. However, while reputation is spread, an individual may realize that its normative (socially-acquired) reputational belief concerning another agent is in error. In such circumstances, is it in the agent's best interest to update its reputational belief and act on its accurate, descriptive information? Or, is it better off continuing to act upon the inaccurate, socially-informed information?

To model this we added a new binary variable, *veracity*, v. If $v = 0$, then a donating agent will always employ its socially received reputational belief when selecting whether to cooperate or defect against a recipient. In Models 1 and 2, v was implicitly zero. Donors always acted based on their socially-acquired belief concerning the recipient's reputation. If $v = 1$, then the donating agent will access and employ correct, descriptive information when interacting with a recipient. This simulates the direct experience of the agent. To do this, we give the donor access to the reputation the recipient would have garnered if there was no error ($e = 0$) — the recipient's warranted reputation. The recipient can then be judged on its merits rather than on error-prone

gossip. While offering the recipient's warranted reputation without a cost is not realistic (though see Fetchenhauer and Dunning, 2010 and Sigmund, 2009), we test this as the extreme limit case, to understand the impact of having such information freely available. If it is not in an agent's best interest to employ correct information even when it is free, then the agent would not pay a cost to attain the information.

We now add a fourth agent strategy, *veridical DISC (VDISC)*. VDISC agents behave like DISC agents, except that when a VDISC agent is a donor, and the agent's reputational belief of the receiver is in error, then the VDISC behaves based on its descriptive knowledge rather than its socially received normative knowledge.

Figure 5 illustrates the difference between VDISC and DISC strategies. In both Figure 5(a) and 5(b), the donor has been socially informed that the recipient is good. Furthermore, the donor has learned, or is able to observe, that this socially-acquired information is in error. The discriminating (DISC) agent (Figure 5(a)), despite this new information, continues to employ the normative (social) information. The veridical discriminating (VDISC) agent (Figure 5(b)) acts on the accurate, descriptive information.

To analyze the effects of veracity (v), we test whether VDISC agents can be invaded by DISCs. As such, the initial population is comprised entirely of VDISC agents. Mutation now offers the possibility of four strategies entering the population, $S_i \in \{ALLD, ALLC, DISC, VDISC\}$. We observed the consequences across three values of homophily, $h \in \{0, 0.5, 1\}$.

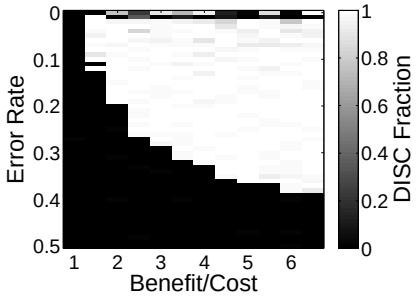
In conditions where $h > 0$, finding a homophilous observer precedes the donor's discovery of the recipient's warranted reputation (see Experiment 2). Therefore both donor and observer hold the normative belief before the donor gains access to the warranted reputation. We therefore explore the question — is it worth re-enforcing good and punishing bad behaviour at the cost of violating an ingroup norm?

Results

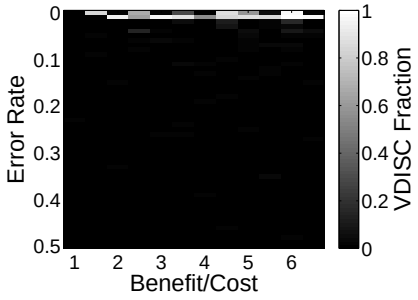
Figure 6 depicts the fraction of the population consisting of DISCs and VDISCs after 500 generations for varying values of homophily. Since at generation 0 the entire population was comprised of VDISCs, the figure illustrates that DISCs invade when $h = 1$ or $h = 0.5$. This result employed the Judging norm but also holds for the Standing and Shunning norm, but not Image Scoring (see Appendix A.2). Figures 6(a) and 6(b) demonstrate the consequences of full homophily ($h = 1$). Agents acting on correct information are invaded by the those acting upon socially garnered information. When $h = 0.5$, the DISC strategy still invades the VDISC agents (Figures 6(c) and 6(d)). Finally, Figures 6(e) and 6(f) show that in an environment where the interactions are completely independent from the beliefs of a given recipient ($h = 0$), VDISCs withstand invasion by DISCs, provided there is any error to be corrected and that ALLDs do not dominate both.

5.2. VDISC Stability: Simplified Analytical Model

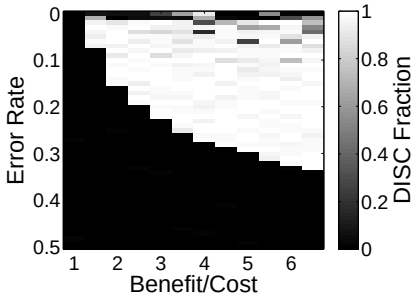
Next, we formally analyze the stability of VDISCs. Again, for reasons of tractability, we do not consider ALLCs and ALLDs, focusing on the interaction between the two strategies that exploit reputation. Most assumptions remain unchanged from Section 4.2. $w \approx 1$ is the chance of playing another round. Each agent starts with a good



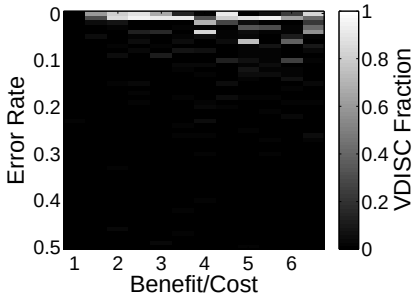
(a) Fraction of DISC agents when $h = 1$.



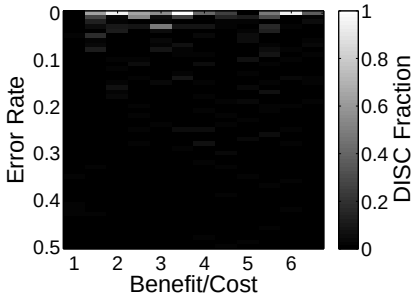
(b) Fraction of VDISC agents when $h = 1$.



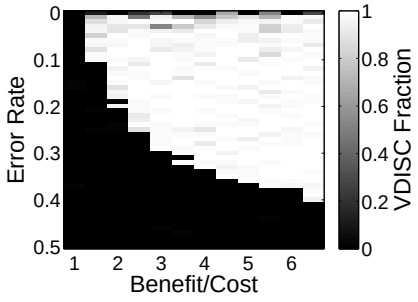
(c) Fraction of DISC agents when $h = 0.5$.



(d) Fraction of VDISC agents when $h = 0.5$.



(e) Fraction of DISC agents when $h = 0$.



(f) Fraction of VDISC agents when $h = 0$.

Figure 6: Fraction of population composed of a strategy after 500 generations. The initial population is solely comprised of VDISCS.

reputation, and we presume the percentage of well-reputed agents equilibrate by round 2. The population begins with mostly VDISCs and few DISCs, such that the probability of interacting with a VDISC is approximately one ($P_v \approx 1$), and the chance of meeting a DISC is approximately 0 ($P_r \approx 0$).

A DISC agent's fitness (π_r) can be define as:

$$\pi_r = b - c + \frac{w}{1-w}[b(P_r G_r + P_v G_r) - c(P_r G_r + P_v G_v)] \tag{8}$$

In the first round each agent is positively reputed, so a DISC both receives a benefit and pays the cost of cooperation ($b-c$). In subsequent rounds, a DISC agent receives a benefit if it interacts with a DISC or VDISC who considers it good, $b(P_r G_r + P_v G_r)$. G_v and G_r are the chance an agent will consider a VDISC and DISC to be good, respectively. Furthermore, the DISC agent pays a cost if it interacts with any agent it considers good $c(P_r G_r + P_v G_v)$.

A VDISC agent's fitness (π_v) is:

$$\pi_v = b - c + \frac{w}{1-w}[b(P_r G_v + bP_v G_v) - c(P_r G_r + P_v G_v)] \tag{9}$$

Similar to the DISC, in the first round a VDISC will both receive a benefit and cooperate ($b - c$). In subsequent rounds, it will receive a benefit if it meets an agent who considers it good. It will pay a cost and cooperate if it meets an agent it believes to be good. To test whether VDISCs can repel an invasion of DISCs, we analyze when the fitness of a VDISC agent is greater than that of the DISC:

$$\pi_v - \pi_r > 0 \tag{10}$$

Since we presume $P_v \approx 1$, $P_r \approx 0$, and $w \approx 1$, Equation 10 simplifies (see Appendix D.2 for details) to:

$$G_v > G_r \tag{11}$$

If the chance that a VDISC is considered good is greater than that of a DISC, then the population of VDISC is stable. Next, we show that both G_v and G_r are functions of error rate (e) and homophily (h). We then analyze the interplay between both on the stability of a VDISC population.

In a society with random interactions ($h = 0$), the fraction of good VDISC agents is:

$$G_v^{h=0} = [P_v + P_r(1-e)](1-e) + (P_r(e))^2 \tag{12}$$

A VDISC, like a DISC agent, receives a good reputation if it agrees with the observer on the reputation of the donor. Since the VDISC donor always switches to the correct information, it will agree with the observer with probability $1 - e$. Furthermore, when the VDISC becomes a recipient, this good reputation is accurately used by $[P_v + P_r(1-e)]$. Only $(1-e)$ fraction of the population hears the correct social information, however, all VDISCs will employ the correct reputation. Lastly, if the observer is in error, then the VDISC and observer will disagree, garnering the VDISC

a bad reputation. However, when the VDISC becomes a recipient, it will receive a benefit if the donor is a DISC who is in error of the original error. Since we presume $P_v \approx 1$ and $P_r \approx 0$, the equation simplifies to:

$$G_v^{h=0} = 1 - e \tag{13}$$

Adding homophily, the chance that a donor agent who interacts with a VDISC recipient will consider the recipient good, is:

$$G_v = h(1-e)[P_v + P_r(1-e)] + (1-h)G_v^{h=0} \tag{14}$$

In a homophilous interaction a VDISC donor will be matched with an observer that heard the same social information regarding the reputation of the recipient. Either, both the donor and recipient heard the correct social information with probability $1 - e$, or both heard incorrect information with probability e. However, the VDISC donor then discovers the correct reputation of the recipient, and employs that. Thus, it will only agree with the recipient with probability $1 - e$. This reputation will then be employed by a donor with probability $[P_v + P_r(1-e)]$. The correct, good reputation will always be employed by a VDISC, and will be employed by a DISC if it heard the correct information through gossip, $P_r(1 - e)$. Presuming $P_v \approx 1$ and $P_r \approx 0$, the equation simplifies to:

$$G_v = h(1-e) + (1-h)(1-e) \tag{15}$$

For a DISC agent in a randomly interacting population ($h = 0$), the probability of possessing a good reputation is:

$$G_r^{h=0} = [P_v + P_r(1-e)][(1-e)^2 + e^2] + P_r e[(1-e)e]^2 \tag{16}$$

A donating DISC will be treated as if it has a good reputation if the observer agrees with it regarding the reputation of the recipient $[(1-e)^2 + e^2]$, and then, as a recipient, the agent interacts with a VDISC donor (who will act on the correct information), or a DISC donor who heard correct gossip ($[P_v + P_r(1 - e)]$). Additionally, it will be considered good if it interacts with an observer who disagrees $[(1 - e)e]^2$, but as a recipient meets a DISC donor in error $(P_r e)$. In the invasion case, this simplifies to:

$$G_r^{h=0} = (1-e)^2 + e^2 \tag{17}$$

Adding homophily:

$$G_r = h[P_v + P_r(1-e)] + (1-h)G_r^{h=0} \tag{18}$$

If $h = 0$ then Equation 18 = Equation 16, since each Donation game is played with randomly selected agents. In a fully homophilous society $h = 1$, the donor and the observer will always agree, therefore the DISC agent will always earn a good reputation. However, because there is error in transmission, only $P_v + P_r(1 - e)$ fraction of the population will believe the donor is good. Again, presuming $P_v \approx 1$ and $P_r \approx 0$, and substituting in $G_r^{h=0}$, the fraction of good DISC agents is:

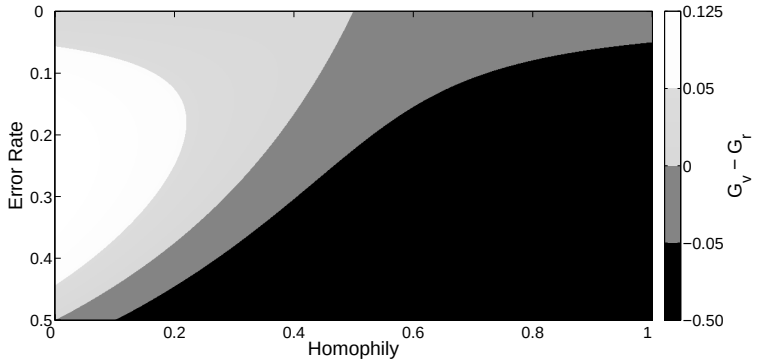


Figure 7: Plot of $G_v - G_r$ over a parameter sweep of error rate (e) and homophily (h). The difference is in the range $0.125 \geq G_v - G_r \geq -0.5$. A positive difference represents the phase space where a population of VDISCs is stable against invasion from DISCs. A negative difference suggests invasion. **White** represents strong selection for VDISCs. **Light Grey** is weaker selection for VDISCs. **Black** represents strong selection for DISCs. **Dark Grey** illustrates weaker selection for DISCs. The delineation between light and dark grey marks where VDISC and DISC fitness are equal.

$$G_r = h + (1-h)[(1-e)^2 + e^2] \tag{19}$$

According to Equation 11, a VDISC population is stable against a DISC invasion if Equation 15 - Equation 19 > 0. Substituting, we get:

$$h(1-e) + (1-h)(1-e) - (h + (1-h)[(1-e)^2 + e^2]) > 0 \tag{20}$$

Results

Graphing the equation, Figure 7 depicts the phase space where VDISCs are stable against DISCs. The Figure represents Equation 11, reduced to a function of e and h, by Equation 20 and is plotted over a parameter sweep of error rate (e) and homophily (h). A positive difference represents where a population of VDISCs is stable against invasion from DISCs. A negative difference suggests invasion.

The curve between the light grey and dark grey regions represents where both strategies perform equally. The result explains the simulation result in Figure 6, where VDISCs are stable against invasion without homophilic interactions ($h = 0$), but are invaded when $h = 0.5$ or $h = 1$. As homophily increases, VDISCs require higher fidelity communication in order to repel invasion. However, when $h \geq 0.5$, any error in communication favors DISC agents. Furthermore, a larger difference implies stronger selection. Since selection pressure is weak for low values of error, this explains why, in Figure 6, DISCs and VDISCs co-inhabit for low values of error.

It should be noted that our simple analytical model only compares the fitness of VDISCs to DISCs when the population is almost entirely comprised of VDISCs. The simulation offers a more complicated, evolutionary solution. It takes into account the repercussions of DISCs increasing in frequency, as well as interacting with ALLD and ALLC agents.

6. Discussion

We have presented agent-based and formal models indicating that contexts exist where selection can work against the employment of veridical information, despite generally supporting cooperation. We started from a baseline showing results that are becoming increasingly well-known. First, cooperation is adaptive provided that the overall benefits outweigh the costs. Second, many kinds of social structures, including value homophily, extend the range of cost/benefit ratios over which cooperation can flourish or even fixate. Unjustified reputations can threaten cooperation, and previous studies have sought methods for removing false reputations from a population (Nakamaru and Kawata, 2004). The present article tested the percentage of false reputations which threatens cooperation, and found that stratifying the interaction propensity of those with correct and incorrect information aids in sustaining a cooperative society.

As value homophily increases, not only does the robustness of a cooperative society increase, but, paradoxically, so does selective pressure for employing error-prone social information — even when correct information is freely available. As normative pressures rise, cooperation is enhanced, but there is a consequent pressure for conformity, meaning that the relative utility of descriptive information decreases. Where there is no value homophily, descriptive (veridical) information about the strategies of others holds the advantage. But where value homophily is deployed, conformity to norms is more adaptive than acting on the truth, at least in the context of these simple agents.

6.1. Gossip

Gossip is widespread in humans (Dessalles, 2007) and yet is also widely disparaged. Gossips were burnt in medieval Europe (Emler, 1990), and women, in particular, have been repressed to avoid it (Funder, 1995; Gilmore, 1978). Intuitively, this is understandable: gossip is often false, conspiratorial, unverifiable and malicious. Yet evidence suggests that gossip is pervasive in its influence on human behaviour (Ellwardt et al., 2012). Further, empirical research suggests that much gossip focusses on the true and positive (Ellwardt et al., 2012; Herrmann et al., 2013; Ingram and Bering, 2010; Sommerfeld et al., 2008), and theoretical results, including those reported here, show that it can be beneficial for spreading information that allows individuals to choose who to interact with (Giardini and Conte, 2012; Mitchell et al., in prep; Nakamura and Masuda, 2011; Traag et al., 2011). Our results suggest one way in which the seeming contradiction between gossip's negative capacities and its pervasive presence might be resolved. Behaviour which is occasionally unfair to individuals who are the object of incorrect gossip nevertheless delivers an overall benefit.

Giardini and Conte (2012) posit that while gossip does not necessarily transmit accurate descriptive information, it likely imparts what others believe is the descriptive information — deemed meta-evaluation. In a human experiment of the Donation game, Sommerfeld et al. (2008) demonstrated that gossip (i.e. the recipient's reputation) is a better predictor of a donor's decision to cooperate or defect than descriptive information (i.e. the history of the recipient's actions). They posit that individuals might "adjust their own behaviour in order to not depart from the public opinion of their local group; they do not want to stand out" (Sommerfeld et al., 2008). Our model offers an

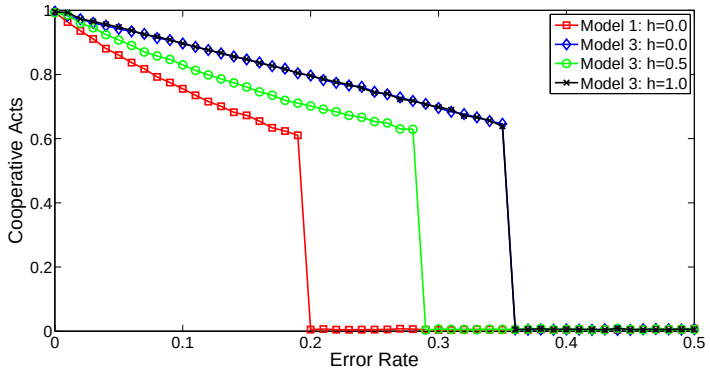


Figure 8: Comparing the consequences of Experiment 3 with cooperation. Each line represents the fraction of cooperative actions when benefit/cost was held at 4.5. Interestingly, employing the veridical strategy (using descriptive information) is an excellent strategy in the absence of homophily, performing equally well as the conformity driven strategy in the fully homophilous society. This result assumes that perfect access to the veridical truth is available, though similarly homophily of 1 assumes perfect consensus.

explanation of such a behaviour, namely that meta-evaluation *is* descriptive information and should be considered since the normative pressures embedded in gossip may outweigh the utility of reputational descriptive information.

6.2. Cooperation

Experiment 3 shows that DISCs invade VDISCs at high levels of homophily, but what are the effects of the two strategies on cooperation? Figure 8 depicts the fraction of cooperative actions for a fixed cost/benefit ratio (4.5). When $h = 0$, VDISC agents sustain cooperation against error just as effectively as DISC agents do when $h = 1$. This is despite the fact that in both conditions, VDISCs exploit descriptive (veridical) information, while DISCs use error-prone gossip.

These results might suggest that rather than generating robust cooperation via homophily (which breeds agents that ignore correct information), a society might better consist of randomly interacting agents that employ correct information. Whilst true in theory, in practice correct information comes at a high cost. For our model, we offered accurate information without cost, and even so DISCs invaded in an unrealistically extreme context. Cost-free descriptive information is not feasible in the natural world. In contrast, there is significant real world evidence for value homophily (Adamic et al., 2003; Curry and Dunbar, 2013; Fiore and Donath, 2005; Kandel, 1978; McPherson et al., 2001). Our results contribute to mounting evidence for the utility of homophily in generating cooperative behaviour, and more generally for exploiting and developing novel adaptive traits (Dieckmann and Doebeli, 1999; Panhuis et al., 2001; Powers et al., 2011).

Furthermore, while personally gathered information may be less noisy than socially garnered knowledge, it is certainly not free from error. In our model, personal information was always veridical, and thus agents employing personal information were given an unrealistic advantage. Again, this was employed as the limiting case. If this

constraint was lifted and error in personally gathered information was included in the model, this would reduce the fitness of the descriptively-biased agents. In terms of cooperation, we have shown two strategies that are equally effective in theory, but only one of which is plausible in practice.

6.3. Ingroup and Indirect Reciprocity

Humans tend to treat individuals within and without of their group distinctly (Brewer, 1979). Recently, indirect reciprocity research has begun to study ingroup biases (Nakamura and Masuda, 2012; Masuda, 2012; Jusup et al., 2014; Matsuo et al., 2014). There are two major differences between our model and the current literature. First, the existing models presume predefined group structure. In our model, groups emerge from shared beliefs, which are in turn determined by random errors in transmission. Second, as mentioned earlier, ingroup dynamics are defined by the likelihood that donors and recipients are within the same group, rather than donors and observers. We discuss the consequences of each of these in turn.

Nakamura and Masuda (2012) presume group structure, where each group of agents agrees upon the reputations of other agents. They show that this and the Judging norm are sufficient for the development of strong ingroup favoritism. Similar to our model, individuals in a group always agree on the reputation of agents, though they may be in error. So, when group members interact (even if the agreed upon reputation is in error), DISC agents receive a good reputation. This leads to ingroup members possessing better reputations for each other, relative to outgroup agents, deemed ingroup favoritism.

The present model does not include rigid group structure, per se. Shared belief does increase the probability of interaction. However, in our model there is no specific subset of agents with which one agents shares all beliefs. This more agile form of affiliation still affords the cooperative advantages of homophily, and the results may be more applicable to contemporary urban societies.

The second difference between our model and most related literature, is that ordinarily ingroup behaviour is defined by the relationship between the donor and recipient. In our model, the recipient is always chosen randomly, and it is the donor and observer who agree. This model better accounts for most observed ingroup behavior, such as social exclusion or charity, and also allows for interactions to occur between groups. In Appendix C we analyze the repercussions to a society if homophily is defined by agreement between *a*) the donor and recipient; or, *b*) the observer and recipient. In both cases, the redefined homophily neither benefits nor hinders cooperation. Given these differences, we believe it would be interesting to see the consequences to Nakamura and Masuda (2012)'s model if an individual discovered the group reputation was wrong and attempted to switch to the correct reputation.

6.4. Self-Deception

Our model may seem to describe a dire world, where truth is to be avoided. However, this would be an over-interpretation. First, truth is deeply valuable — the conflict comes when an agent must choose either to signal honestly about its knowledge, or another's, but not both. Second, our model is agnostic on the intentionality of the

agent. What matters is the behaviour (i.e. aligning with the normative information), not whether the agent intends to deceive. As such, this research may have implications for the research on self-deception. von Hippel and Trivers (2011) argue that intentional deception is cognitively demanding, and that self-deception might evolve to mitigate that cost. Two naturally observed mechanisms which could enable such self-deception are (i.) confirmation bias and (ii.) groupthink.

Confirmation bias is a phenomenon whereby individuals bias their information gathering to retain the fidelity of their existing beliefs (Nickerson, 1998; Jonas et al., 2001; Schulz-Hardt et al., 2000). If, as in our model, it is not beneficial to employ correct information, an agent may attempt not to learn the truth, and thus never be confronted with the conflict which possessing correct information generates. Second, if it is easier to deceive another if one does not know one is deceiving the other (Trivers, 1991), then it may be better for an agent not to have explicit access to some knowledge — that is, it should not be able to act on that knowledge, even if it is still acquiring it 'unconsciously' (Bryson, 2009). This is similar to the phenomenon of groupthink, where being in a group biases the way individuals process information (Janis, 1971; Turner and Pratkanis, 1998). In group contexts, individuals might not be aware of the information they have or the opinions they would have developed if uninfluenced by the group (Janis, 1972; Bénabou, 2013). While deceiving others can be advantageous, the cost of detection is likely high. If accurate information performs worse than inaccurate, and if deception comes at a cost, then equilibrium may rest at a point where the agent would be willing to pay some cost to not acquire the information in the first place (i.e. confirmation bias), or to not be able to access it in some behavioural contexts (i.e. groupthink).

6.5. Parsimonious Causes for Ignoring Correct Information

Finally, the correlation between this simple model and the natural world is clearly speculative. However, the Donation game has been used as an existence proof for the evolution of cooperation via the spread of reputation (Nowak and Sigmund, 1998). We show that taking this simple model and adding one additional constraint (i.e. value homophily) leads to a phenomenon seen in the natural world. More variance on whether an individual will donate to another is explained by using error-prone reputation rather than an accurate record of a recipient's previous actions (Sommerfeld et al., 2008). At the very least, we believe this model increases insight into the variables required to give rise to a phenomenon witnessed in the natural world.

7. Conclusion

We make two claims regarding the role of homophily and misinformation in employing the Donation game to sustain societal level cooperation. First, increasing homophily facilitates cooperation. Even when information is communicated with a high probability of error, social information is better able to sustain cooperation. Second, if a society employs homophily, then when faced with a choice between being accurate in gossip or correctly signalling group membership, it can be beneficial to cooperation overall to favour the latter. Therefore, our overall conclusion is that when social norms

(a) Standing

	Good	Bad
Coop	G	G
Defect	B	G

(b) Shunning

	Good	Bad
Coop	G	B
Defect	B	B

(c) Image Scoring

	Good	Bad
Coop	G	G
Defect	B	B

Table A.4: Definition of tested social norms. Each column represents whether the observer believes the recipient is good or bad. The first row is the reputation the observer will impart on the donor, if the donor cooperates. The second, is the reputational repercussions if the donor defects.

achieve cooperation there are contexts (e.g. homophilous societies) where persisting with inaccurate normatively-acquired information despite knowledge of its descriptive falsity is advantageous and adaptive.

8. Acknowledgements

We would like to acknowledge Marina De Vos and Daniel Taylor for support and suggestions. Paul Rauwolf is supported by a University of Bath studentship.

Appendix A. Social Norms

Appendix A.1. Model 1 & 2: Homophily

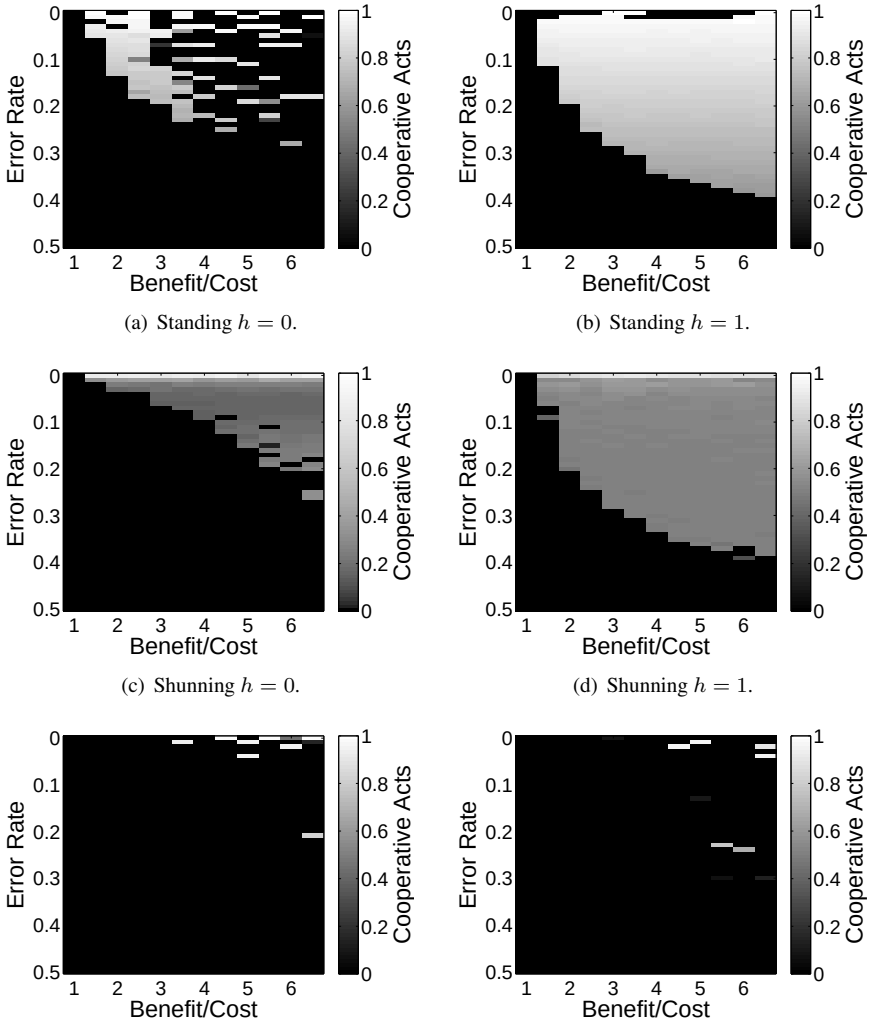
Thus far we have shown that the Judging norm (see Table 2) and homophily can increase the robustness of cooperation in the face of erroneous communication. Here we check whether this result holds for other social norms. We employ three norms, shown in Table A.4.

With the Standing norm (Table 4(a)), if a donor cooperates with the recipient, it always receives a good reputation. The only situation where an agent becomes ill-reputed is if it refuses to cooperate with a good recipient. When agents always share the same belief about an given agent, even if unjustified, the Standing norm has been shown to maintain cooperation (Ohtsuki and Iwasa, 2004). However, Takahashi and Mashima (2006) showed a Standing society is susceptible to invasion if error is disseminated as in our model, where agents can have differing opinions regarding the same agent.

The Shunning norm (Table 4(b)) is the only norm other than the Judging norm which sustains cooperation when agents can hold different beliefs about the same agent (Takahashi and Mashima, 2006). With Shunning, an agent can only receive a good reputation if it cooperates with a good recipient. Finally, with Image Scoring (Table 4(c)) an agent receives a good reputation for cooperating, and a bad reputation for defecting. As a first order norm, it is well-known to fail to maintain cooperation (see Section 2.2).

Results

Figure A.9 depicts the fraction of cooperative actions performed utilizing three different norms for two values of homophily. In a society using the Standing norm without homophily (Figure 9(a)) cooperation is vulnerable because ALLCs can perform better than DISCs and are subsequently vulnerable to invasion by ALLDs (see Takahashi



(a) Standing $h = 0$.

(b) Standing $h = 1$.

(c) Shunning $h = 0$.

(d) Shunning $h = 1$.

(e) Image Scoring $h = 0$.

(f) Image Scoring $h = 1$.

Figure A.9: Fraction of cooperative acts averaged over last 50 generations. The figures illustrate the consequences of adding homophily to three social norms.

and Mashima (2006) for a discussion). Interestingly, in a society with homophilous interactions (Figure 9(b)) cooperation is aided and more robust against invasion.

Figure 9(c) illustrates the cooperative propensity of the Shunning norm without homophily. The Shunning norm has been shown to remain stable when $e = 0.025$ (Takahashi and Mashima, 2006). Here, we extend this research and show that, for certain levels of benefit (b), cooperation can remain stable in the face of erroneous communication. Furthermore, homophily extends this range, stabilizing cooperation despite lower values for b, and higher error rates. Finally, we reproduce the well known result that Image Scoring cannot sustain cooperation (Figure 9(e)). Furthermore, homophily does not aid in maintaining cooperation (Figure 9(f)).

In conclusion, Figure A.9 illustrates that the results from Section 4 extend to additional social norms. Homophily increases cooperation for Judging and Shunning societies, which are the two norms stable in populations where agents may disagree on the reputation of an individual (Takahashi and Mashima, 2006). Furthermore, homophily increases the stability of an unstable norm in Standing.

Appendix A.2. Model 3: VDISC Stability

Here we test whether DISCs invade a population of VDISCs with social norms other then Judging. We use the three norms shown in Table A.4 and described in more detail in Appendix A.1. As in Section 5, each population starts entirely comprised of VDISC agents. ALLD, ALLC, and DISC agents can enter the population through mutation. We test whether DISC agents invade.

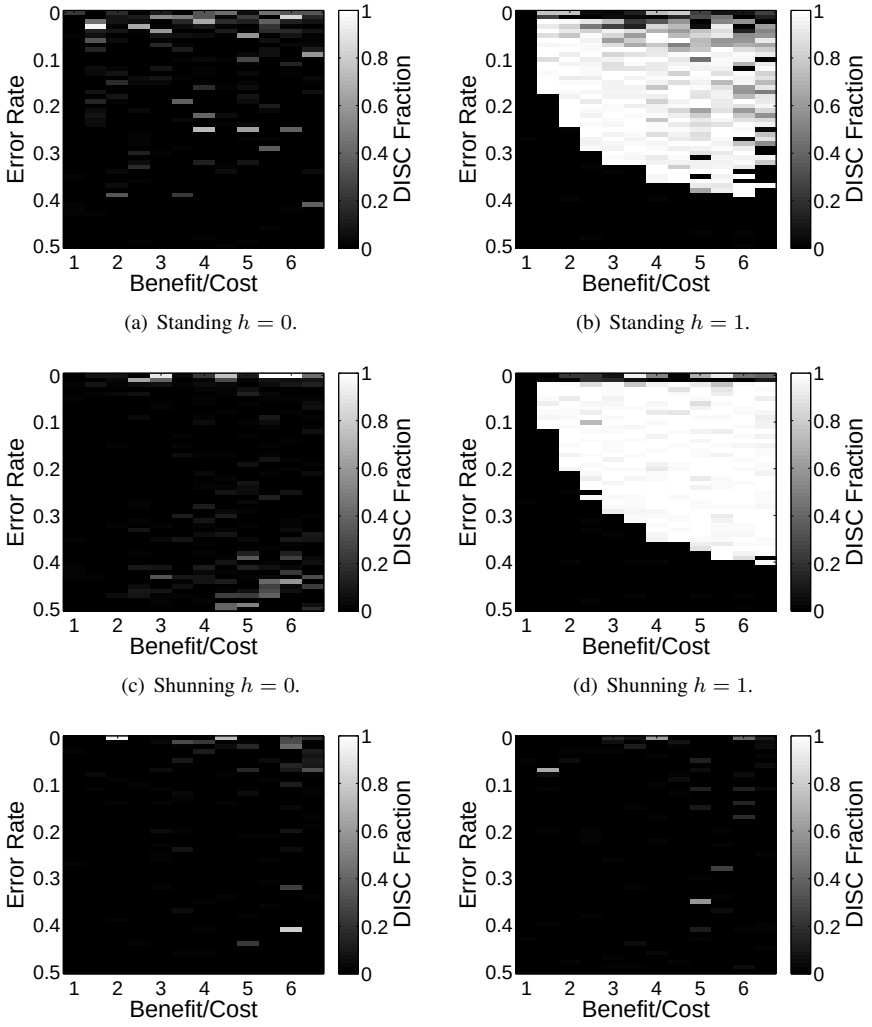
Results

Figure A.10 extends the results shown in Section 5 for the Standing, Shunning, and Image Scoring social norms. The graphs in A.10 depict the fraction of DISC agents after 500 generations. Without homophily DISC agents do not invade with Standing (Figure 10(a)), Shunning (Figure 10(c)), or Image Scoring (Figure 10(e)). However, when homophily is added, DISCs invade for both Standing (Figure 10(b)) and Shunning (Figure 10(d)). As Image Scoring does not sustain cooperation, DISCs do not invade because ALLDs invade, rather than because of VDISC stability — Figure 10(f). For both Standing and Shunning, homophily increases a cooperative society's robustness against error in communication (see Section 4 and Appendix A.1), but consequentially selects for agents who do not employ accurate reputational information over erroneous social information.

Appendix B. Model 1 & 2: Continuous Reputations

Up until now, an agent's reputation was based on its most recent action. If an agent had incurred a negative reputation many times in the past, it could alter its reputation with one action. Here, we test the consequences of homophily when there is some historical memory.

Each agent begins with a good reputation, $R = 1$. Depending on the judgement of the observer, the reputation of the donor moves ± 0.2. The error rate (e) affects the sign of the reputational movement. So, if the observer tells the population to update



(a) Standing $h = 0$.

(b) Standing $h = 1$.

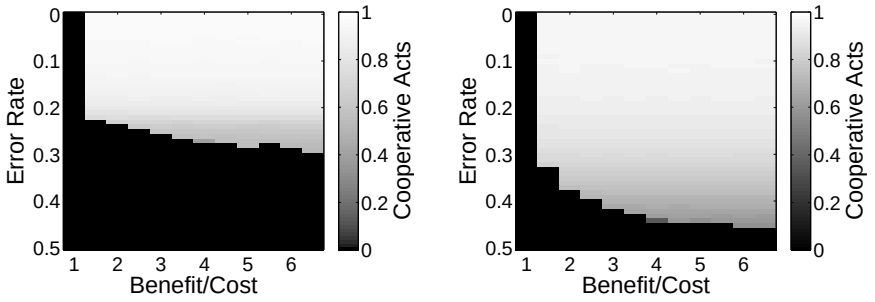
(c) Shunning $h = 0$.

(d) Shunning $h = 1$.

(e) Image Scoring $h = 0$.

(f) Image Scoring $h = 1$.

Figure A.10: Fraction of DISC agents after 500 generations. The figures illustrate whether DISC agents invade a population of VDISC using three social norms.



(a) $h = 0$ (b) $h = 1$

Figure B.11: Depicts the consequences of continuous reputations for different levels of homophily. The figures illustrate the average fraction of Donation games which were cooperative over the final 50 of 500 generations. **(a)** Homophily is 0% ($h = 0$). **(b)** Homophily is 100% ($h = 1$). With continuous reputations, homophily still increases robustness against error in communication.

its reputation of the donor by $+0.2$, e percent of the population will update it with -0.2. When judging the recipient, the threshold between good and bad is 0.5. Thus, if a DISC donor meets a recipient with a reputation greater than 0.5, it will cooperate, otherwise it will defect. Reputation is limited in the range $[0, 1]$, so if an agent receives a reputation higher than 1, it remains at 1. For this simulation, we used the Judging norm (see Table 2).

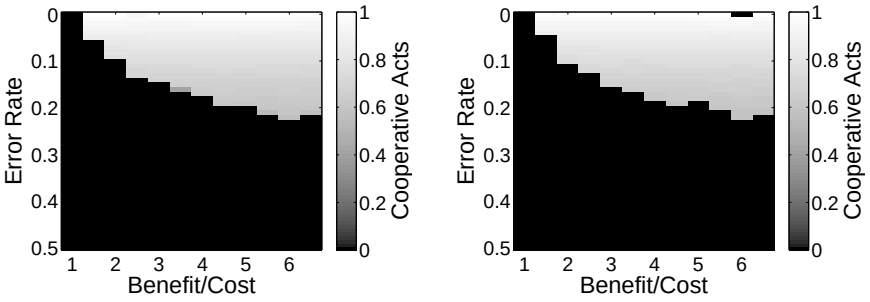
Results

Figure B.11 shows the result of homophily with continuous reputations. Clearly, a homophilous society ($h = 1$) is able to sustain cooperation in the face of increased error. Figure 11(a) illustrates that cooperation fails if $e > 0.3$, while Figure 11(b) demonstrates that a homophilous society can maintain cooperation for an error rate near 0.45.

Appendix C. Defining Homophily

In Section 4 we defined value homophily as the propensity for the donor and observer to agree on the reputation of the recipient. Here we test the repercussions of redefining it. We evaluate two other definitions, namely, value homophily is the propensity for *a*) a donor and recipient to share a belief regarding another; and, *b*) an observer and recipient to share a belief regarding another.

In both cases, initially, a donor, recipient, and observer are selected at random. In the first case, in a homophilic interaction, the donor and recipient share a reputation with another. As the observer is chosen randomly, we compare the donor's belief of the observer with the recipient's belief in the observer. If the two disagree, then h is the probability that a new recipient is selected which agrees with the donor on the reputation of the observer.



(a) Donor and Recipient agree

(b) Observer and Recipient agree

Figure C.12: The fraction of cooperative actions for two definitions of homophily. $h = 1$ for both, and the fraction of cooperative acts is average of the last 50 generations.

In the final instantiation of homophily, an observer and recipient must share a belief in the reputation of the donor. If they do not, then h is the probability that a new observer is selected which agrees with the recipient.

Results

Figure C.12 displays the number of cooperative acts when $h = 1$ for the two definitions of homophily. Recall that Figure 2 represents a baseline, where all actors are randomly selected ($h = 0$). Whether homophily is defined as unanimity between donor and recipient (Figure 12(a)) or recipient and observer (12(b)), homophily does not aid or hinder cooperation, relative to the baseline. This is in contrast to Figure 3(a), where homophily is defined as agreement between donor and observer.

Appendix D. Simplification

Appendix D.1. Simplification of DISC Stability

Here we simply the expression defining when a DISC population is stable against an ALLD invasion. We start with Equation 3:

$$\pi_r - \pi_d > 0 \tag{D.1}$$

Substituting Equations 1 and 2:

$$bP_r - c + \frac{w}{1-w}[bP_rG_r - c(P_rG_r + P_dG_d)] - [bP_r + \frac{w}{1-w}(bP_rG_d)] > 0 \tag{D.2}$$

Since we are testing ALLD invasion criteria, we presume that the proportion of DISCs is close to one, and the proportion of ALLDs is close to zero. In other words, $P_r \approx 1$ and $P_d \approx 0$. Replacing these:

$$b - c + \frac{w}{1-w}[bG_r - cG_r] - [b + \frac{w}{1-w}(bG_d)] > 0 \tag{D.3}$$

For simplicity, we replace $\alpha = \frac{w}{1-w}$.

$$b - c + \alpha[bG_r - cG_r] - [b + \alpha(bG_d)] > 0 \tag{D.4}$$

Dividing by α we get:

$$\frac{b-c}{\alpha} + [bG_r - cG_r] - \frac{b}{\alpha} - bG_d > 0 \tag{D.5}$$

Since $\frac{1}{\alpha} = \frac{1-w}{w}$ and $w \approx 1$, then $\frac{1}{\alpha} \approx 0$, the equation simplifies to:

$$\begin{aligned} bG_r - cG_r - bG_d &> 0 \\ G_r(b-c) - bG_d &> 0 \\ G_r &> \frac{bG_d}{b-c} \end{aligned} \tag{D.6}$$

Appendix D.2. Simplification of VDISC Stability

Here we simplify $\pi_v - \pi_r > 0$. Substituting Equations 8 and 9 we get:

$$\begin{aligned} b - c + \frac{w}{1-w}[b(P_rG_v + bP_vG_v) - c(P_rG_r + P_vG_v)] \\ - (b-c) - \frac{w}{1-w}[b(P_rG_r + P_vG_r) - c(P_rG_r + P_vG_v)] > 0 \end{aligned} \tag{D.7}$$

In the first round both the VDISC and DISC agent receive the same pay-off $(b-c)$. $(b-c) - (b-c) = 0$, so they are removed. Furthermore, for simplicity, we replace $\alpha = \frac{w}{1-w}$.

$$\begin{aligned} \alpha[b(P_rG_v + bP_vG_v) - c(P_rG_r + P_vG_v)] \\ - \alpha[[b(P_rG_r + P_vG_r) - c(P_rG_r + P_vG_v)] > 0 \end{aligned} \tag{D.8}$$

Since we are testing whether DISCs can invade, the probability of interacting with a VDISC is close to one ($P_v \approx 1$), and DISCs are rare ($P_r \approx 1$). Thus, we get:

$$\alpha[bG_v - cG_v] - \alpha[bG_r - cG_v] > 0 \tag{D.9}$$

Since $\alpha > 0$, dividing by α, simplifies the formula to:

$$bG_v - cG_v - bG_r + cG_v > 0 \tag{D.10}$$

$cG_v - cG_v = 0$. Since $b > 0$, dividing by b reduces the expression to:

$$G_v - G_r > 0 \tag{D.11}$$

References

Adamic, L., Buyukkokten, O., Adar, E., 2003. A social network caught in the web. First Monday 8 (6).

Angerer, S., Glätzle-Rützler, D., Lergetporer, P., Sutter, M., 2014. Donations, risk attitudes and time preferences: A study on altruism in primary school children. Tech. rep., Ifo Institute for Economic Research at the University of Munich.

Axelrod, R., 1997. The dissemination of culture a model with local convergence and global polarization. Journal of Conflict Resolution 41 (2), 203–226.

Axelrod, R., Hamilton, W. D., 27 March 1981. The evolution of cooperation. Science 211, 1390–1396.

Bénabou, R., 2013. Groupthink: Collective delusions in organizations and markets. The Review of Economic Studies 80 (2), 429–462.

Boyd, R., Richerson, P. J., 1985. Culture and the Evolutionary Process. The University of Chicago Press.

Brewer, M. B., 1979. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychological Bulletin 86 (2), 307–329.

Bryson, J. J., 2009. Representations underlying social learning and cultural evolution. Interaction Studies 10 (1), 77–100.

Byrne, D., 1961. Interpersonal attraction and attitude similarity. The Journal of Abnormal and Social Psychology 62 (3), 713–715.

Curry, O., Dunbar, R. I., 2013. Do birds of a feather flock together? Human Nature, 1–12.

Dessalles, J.-L., 2007. Why we talk: The evolutionary origins of language. Oxford University Press.

Dieckmann, U., Doebeli, M., 1999. On the origin of species by sympatric speciation. Nature 400 (6742), 354–357.

Ellwardt, L., Labianca, G. J., Wittek, R., 2012. Who are the objects of positive and negative gossip at work? A social network perspective on workplace gossip. Social Networks 34 (2), 193–205.

Emler, N., 1990. A social psychology of reputation. European Review of Social Psychology 1 (1), 171–193.

Enquist, M., Leimar, O., 1993. The evolution of cooperation in mobile organisms. Animal Behaviour 45, 747–757.

Fernández-Juricic, E., Kacelnik, A., 2004. Information transfer and gain in flocks: The effects of quality and quantity of social information at different neighbour distances. Behavioral Ecology and Sociobiology 55 (5), 502–511.

Fetchenhauer, D., Dunning, D., 2010. Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. Psychological Science 21 (2), 189–193.

Fiore, A. T., Donath, J. S., 2005. Homophily in online dating: When do you like someone like yourself? In: CHI '05 Extended Abstracts on Human Factors in Computing Systems. CHI EA '05. ACM, New York, NY, USA, pp. 1371–1374.

Fishman, M. A., 2003. Indirect reciprocity among imperfect individuals. Journal of Theoretical Biology 225 (3), 285–292.

Fu, F., Nowak, M. A., Christakis, N. A., Fowler, J. H., 2012. The evolution of homophily. Scientific Reports 2.

Funder, D. C., 1995. On the accuracy of personality judgment: A realistic approach. Psychological Review 102 (4), 652–670.

Giardini, F., Conte, R., 2012. Gossip for social control in natural and artificial societies. Simulation 88 (1), 18–32.

Gilmore, D., 1978. Varieties of gossip in a spanish rural community. Ethnology 17 (1), 89–99.

Goldberg, D. E., Deb, K., 1991. A comparative analysis of selection schemes used in genetic algorithms. In: Foundations of Genetic Algorithms. Morgan Kaufmann, pp. 69–93.

Henrich, J., Boyd, R., 1998. The evolution of conformist transmission and the emergence of between-group differences. Evolution and Human Behavior 19 (4), 215–241.

Henrich, J., Gil-White, F. J., 2001. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior 22 (3), 165–196.

Herrmann, E., Keupp, S., Hare, B., Vaish, A., Tomasello, M., 2013. Direct and indirect reputation formation in nonhuman great apes and human children. Journal of Comparative Psychology 127 (1), 63–75.

Hilbe, C., Nowak, M. A., Sigmund, K., 2013. Evolution of extortion in iterated prisoners dilemma games. Proceedings of the National Academy of Sciences 110 (17), 6913–6918.

Ingram, G. P. D., Bering, J. M., 2010. Children's tattling: The reporting of everyday norm violations in preschool settings. Child Development 81 (3), 945–957.

Janis, I. L., 1971. Groupthink. Psychology Today 5 (6), 43–46.

Janis, I. L., 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin.

Jonas, E., Schulz-Hardt, S., Frey, D., Thelen, N., 2001. Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. Journal of Personality and Social Psychology 80 (4), 557–571.

Jusup, M., Matsuo, T., Iwasa, Y., 05 2014. Barriers to cooperation aid ideological rigidity and threaten societal collapse. PLoS Comput Biol 10 (5), e1003618.

Kandel, D. B., 1978. Homophily, selection, and socialization in adolescent friendships. American Journal of Sociology, 427–436.

King, A. J., Cowlishaw, G., 2007. When to use social information: The advantage of large group size in individual decision making. Biology Letters 3 (2), 137–139.

Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D., 2011. How social influence can undermine the wisdom of crowd effect. Proceedings of the National Academy of Sciences 108 (22), 9020–9025.

Magurran, A., Higham, A., 1988. Information transfer across fish shoals under predator threat. Ethology 78 (2), 153–158.

Marshall, J. A., 2009. The donation game with roles played between relatives. Journal of Theoretical Biology 260 (3), 386–391.

Marshall, J. A., 2011. Ultimate causes and the evolution of altruism. Behavioral Ecology and Sociobiology 65 (3), 503–512.

Masuda, N., 2012. Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. Journal of Theoretical Biology 311 (0), 8–18.

Matsuo, T., Jusup, M., Iwasa, Y., 2014. The conflict of social norms may cause the collapse of cooperation: Indirect reciprocity with opposing attitudes towards in-group favoritism. Journal of Theoretical Biology 346 (0), 34–46.

McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 415–444.

Mesoudi, A., 2011. An experimental comparison of human social learning strategies: Payoff-biased social learning is adaptive but underused. Evolution and Human Behavior 32 (5), 334–342.

Mitchell, D., Bryson, J. J., Ingram, G. P., in prep. On the reliability of unreliable information: Gossip as cultural memory. Unpublished.

Nakamaru, M., Kawata, M., 2004. Evolution of rumours that discriminate lying defectors. Evolutionary Ecology Research 6 (2), 261–283.

Nakamura, M., Masuda, N., 07 2011. Indirect reciprocity under incomplete observation. PLoS Comput Biol 7 (7), e1002113.

Nakamura, M., Masuda, N., 2012. Groupwise information sharing promotes ingroup favoritism in indirect reciprocity. BMC Evolutionary Biology 12 (1), 213.

Nickerson, R. S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology 2 (2), 175–220.

Nowak, M. A., Sigmund, K., 11 June 1998 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Nowak, M. A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437 (7063), 1291–1298.

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

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? Reputation dynamics in indirect reciprocity. Journal of Theoretical Biology 231 (1), 107–120.

Ohtsuki, H., Iwasa, Y., 2006. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. Journal of Theoretical Biology 239 (4), 435–444.

Panchanathan, K., 2011. Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity. Journal of Theoretical Biology 277 (1), 48–54.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. Journal of Theoretical Biology 224 (1), 115–126.

Panhuis, T. M., Butlin, R., Zuk, M., Tregenza, T., 2001. Sexual selection and speciation. Trends in Ecology & Evolution 16 (7), 364–371.

Powers, S. T., Penn, A. S., Watson, R. A., 2011. The concurrent evolution of cooperation and the population structures that support it. Evolution 65 (6), 1527–1543.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L., Ghirlanda, S., Lillicrap, T., Laland, K. N., 2010. Why copy others? Insights from the social learning strategies tournament. Science 328 (5975), 208–213.

Ross, R. M., Greenhill, S. J., Atkinson, Q. D., 2013. Population structure and cultural geography of a folktale in europe. Proceedings of the Royal Society B: Biological Sciences 280 (1756).

Santos, F. C., Santos, M. D., Pacheco, J. M., 2008. Social diversity promotes the emergence of cooperation in public goods games. Nature 454 (7201), 213–216.

Schonmann, R. H., Vicente, R., Caticha, N., 08 2013. Altruism can proliferate through population viscosity despite high random gene flow. PLoS ONE 8 (8), e72043.

Schulz-Hardt, S., Frey, D., Lüthgens, C., Moscovici, S., 2000. Biased information search in group decision making. Journal of Personality and Social Psychology 78 (4), 655–669.

Sigmund, K., 2009. Sympathy and similarity: The evolutionary dynamics of cooperation. Proceedings of the National Academy of Sciences 106 (21), 8405–8406.

Sommerfeld, R. D., Krambeck, H.-J., Milinski, M., 2008. Multiple gossip statements and their effect on reputation and trustworthiness. Proceedings of the Royal Society B: Biological Sciences 275 (1650), 2529–2536.

Stewart, A. J., Plotkin, J. B., 2013. From extortion to generosity, evolution in the iterated prisoners dilemma. Proceedings of the National Academy of Sciences 110 (38), 15348–15353.

Suzuki, S., Kimura, H., 2013. Indirect reciprocity is sensitive to costs of information transfer. Scientific Reports 3.

Takahashi, N., Mashima, R., 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. Journal of Theoretical Biology 243 (3), 418–436.

Tanabe, S., Suzuki, H., Masuda, N., 2013. Indirect reciprocity with trinary reputations. Journal of Theoretical Biology 317, 338–347.

Taylor, P. D., Day, T., Wild, G., 2007. Evolution of cooperation in a finite homogeneous graph. Nature 447 (7143), 469–472.

Traag, V. A., Van Dooren, P., Nesterov, Y., 2011. Indirect reciprocity through gossiping can lead to cooperative clusters. In: Artificial Life (ALIFE), 2011 IEEE Symposium on. IEEE, pp. 154–161.

Trivers, R., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46 (1), 35–57.

Trivers, R., 1991. Deceit and self-deception: The relationship between communication and consciousness. In: Robinson, M., Tiger, T. (Eds.), Man and Beast Revisited. Smithsonian Press, pp. 175–191.

Turner, M. E., Pratkanis, A. R., 1998. Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. Organizational Behavior and Human Decision Processes 73 (2), 105–115.

Uchida, S., Sasaki, T., 2013. Effect of assessment error and private information on stern-judging in indirect reciprocity. Chaos, Solitons & Fractals 56, 175–180.

Uchida, S., Sigmund, K., 2010. The competition of assessment rules for indirect reciprocity. Journal of Theoretical Biology 263 (1), 13–19.

von Hippel, W., Trivers, R., 2 2011. The evolution and psychology of self-deception. Behavioral and Brain Sciences 34, 1–16.

Whiten, A., McGuigan, N., Marshall-Pescini, S., Hopper, L. M., 2009. Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1528), 2417–2428.