Indirect reciprocity in populations with group structure

Taylor A. Kessinger and Joshua B. Plotkin

1Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
2Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

April 25, 2022

Abstract

Indirect reciprocity, whereby individuals cooperate with others of good standing, is a powerful and parsimonious mechanism to sustain a high level of cooperation. Studies of indirect reciprocity typically assume either that each individual forms their own private assessment of the reputations of others or, conversely, that everyone shares a consensus view of the reputations in the population. Here we develop a mathematical framework for analyzing intermediate cases between strictly private and fully public information. We introduce group structure in information availability: the population is stratified into disjoint groups, and reputation information is shared only within each group. Even when social interactions and strategic imitations are well-mixed across the entire population, the partitioning of reputation information into disjoint groups weakens the fitness advantage of cooperative individuals, because it makes it less likely that interacting players will agree about each other’s reputations. In the limit of a single group, or in the limit of a large number of groups, we recover the classical cases of fully public, or completely private, information. We analyze the effect of group structure on the stability of behavioral strategies, and we identify critical values for the number of groups required to support cooperation, under four different social norms of reputation assessment.

1 Model

We consider a population of $N$ individuals who play a series of one-shot pairwise donation games (Rapoport et al., 1965) with each other. Every round, each individual plays the donation game twice with everyone else, once as a donor and once as a recipient. A donor may either cooperate, paying a cost $c$ to convey a benefit $b > c$ to the recipient, or defect, paying no cost and conveying no benefit. Each donor chooses an action based on their behavioral strategy: cooperators (denoted ALLC or $X$) always cooperate, defectors (ALLD, $Y$) always defect, and discriminators (DISC, $Z$) cooperate with those they consider to have good reputations and defect with those they consider to have bad reputations. (We do not consider the “reverse discriminator” strategy in this model.) Following the round of all pairwise game interactions, the players update their views of each others’ reputations, and they update their strategies according to payoff-based imitation, as described below.

1.1 Reputations

Each player belongs to one of $K$ distinct and disjoint groups, which comprise fractions $\nu_1, \nu_2, \ldots, \nu_K$ of the total population. An individual’s group membership determines their view of the reputations
of the other players: each group has a shared, consensus view of the reputation of every player in the population, but different groups may have different views of individuals’ reputations. This characterizes a situation where individuals transmit information about reputations to other members of their group via rapid gossip (Sommerfeld et al., 2007), or, alternatively, each group has its own “institution” (Radzilavicius et al., 2021) that broadcasts reputation assessments to its group.

Each round, everyone plays the donation game with everyone else, and then each group updates their (consensus) view of the reputation of each individual in the population, as follows. For a given individual, the group considers a single random interaction from that round in which that individual acted as a donor. Depending on the donor’s action, the group’s view of the recipient’s reputation, and a rule known as a social norm, the donor is assigned a new reputation by the group. We consider a generalized social norm in which:

1. cooperating with an individual with a good reputation is considered good.
2. defecting with an individual with a good reputation is considered bad.
3. cooperating with an individual with a bad reputation is considered good with probability \( p \).
4. defecting with an individual with a bad reputation is considered good with probability \( q \).

The social norm is thus parameterized by two probabilities, \( p \) and \( q \). When \((p, q) = (0, 1)\), for example, we recover the stern judging norm, which requires that a donor interacting with a recipient of bad standing must defect to earn a good standing. Setting \((p, q) = (0, 0), (1, 0),\) or \((1, 1)\) yields the other standard social norms shunning, scoring, and simple standing respectively.

1.2 Errors

We consider two possible sources of errors: error in social interaction and in reputation assessment. First, an individual who intends to cooperate may accidentally defect, which we call an execution error; this occurs with probability \( u_x \). (Individuals who intend to defect can never accidentally cooperate.) Second, an observer may erroneously assign an individual the wrong reputation, which we call an assessment error; this occurs with probability \( u_a \). We also define the related parameter

\[
\epsilon = (1 - u_x)(1 - u_a) + u_x u_a,
\]

which quantifies the chance that an individual who intends to cooperate with someone with a good reputation successfully does so and is correctly assigned a good reputation (first term) or accidentally defects but is erroneously assigned a good reputation nonetheless (second term).

Given the social norm and these error rates, we can characterize how a donor is assessed in terms of four probabilities:

- \( P_{GC} \), the chance that a donor who cooperates with a good recipient will be assigned a good reputation.
- \( P_{GD} \), the chance that a donor who defects with a good recipient will be assigned a good reputation.
- \( P_{BC} \), the chance that a donor who defects with a bad recipient will be assigned a good reputation.
- \( P_{BD} \), the probability that a donor who defects with a bad recipient will be assigned a good reputation.
For an arbitrary social norm \((p, q)\) and error rates \(u_a\) and \(u_x\), we can derive general expressions for these four probabilities that characterize reputation assessment (see Supplementary Information):

\[
\begin{align*}
P_{GC} &= \epsilon \\
P_{GD} &= u_a \\
P_{BC} &= p(\epsilon - u_a) + q(1 - \epsilon - u_a) + u_a \\
P_{BD} &= q(1 - 2u_a) + u_a.
\end{align*}
\]

### 1.3 Mean-field reputation dynamics

In the limit of large population size, we neglect fluctuations and consider an individual’s expected reputation over many rounds of play, prior to any strategic changes in the population. Let \(g_{i,J}^{I,J}\) be the probability that an individual with strategy \(i\) in group \(I\) has a good reputation in the eyes of an individual in group \(J\). (The first superscript index denotes “who”, the donor; the second index denotes “in whose eyes”, the observer.) Furthermore, let \(f_i^I\) be the frequency of individuals in group \(I\) who have strategy \(i\), so that \(\sum_i f_i^I = 1\) (not \(\nu_I\)). We define

\[
g_{i,J}^{I,J} = \sum_{i \in \{X, Y, Z\}} f_i^I g_{i,J}^{I,J},
\]

which represents the expected fraction of individuals in group \(I\) who are seen as good from the point of view of someone in group \(J\). Note that the summation index \(i\) in this expression, and all other such expressions below, denotes a sum over strategic types, namely \(i \in \{X, Y, Z\}\). We further define

\[
g_{\bullet,J}^{I,J} = \sum_{L=1}^{K} \nu_L g_{L,J}^{I,J},
\]

which represents the fraction of individuals in the whole population whom an individual in group \(J\) sees as good. Here, and elsewhere in this document, capital letter summation indices (such as \(L\), \(J\), or \(I\)) denote a sum over all groups \(\in \{1, 2, \ldots, K\}\).

In the SI, we show that the reputations associated with different strategic types satisfy

\[
\begin{align*}
g_{X,J}^{I,J} &= g_{\bullet,J}^{I,J} P_{GC} + (1 - g_{\bullet,J}^{I,J}) P_{BC}, \\
g_{Y,J}^{I,J} &= g_{\bullet,J}^{I,J} P_{GD} + (1 - g_{\bullet,J}^{I,J}) P_{BD}, \\
g_{Z,J}^{I,J} &= \delta_{I,J} \left[ g_{\bullet,J}^{I,J} P_{GC} + (1 - g_{\bullet,J}^{I,J}) P_{BD} \right] \\
&\quad + (1 - \delta_{I,J}) \left[ g_{\alpha,J}^{I,J} P_{GC} + (g_{\bullet,J}^{I,J} - g_{\alpha,J}^{I,J}) P_{GD} + (g_{\bullet,J}^{I,J} - g_{\alpha,J}^{I,J}) P_{BC} + (1 - g_{\bullet,J}^{I,J} - g_{\alpha,J}^{I,J}) P_{BD} \right],
\end{align*}
\]

where the term \(g_{\alpha,J}^{I,J}\) is defined as

\[
g_{\alpha,J}^{I,J} = \sum_{L=1}^{K} \nu_L \sum_{i \in \{X, Y, Z\}} f_i^I g_{i,J}^{I,J} g_{L,I}^{L,J},
\]

which reflects the probability that distinct groups \(I \neq J\) agree that a randomly chosen individual in the population has a good reputation.
1.4 Payoffs

Payoffs accrue to each individual based on their gameplay in pairwise interactions. That is, an individual acquires a payoff $b$ for each interaction either with a cooperator ($X$) or with a discriminator ($Z$) who sees them as good. A cooperator pays cost $c$ in each interaction, and a discriminator pays cost $c$ in each interaction with someone whom they see as good. Thus, the average payoff for each of the three strategic types in an arbitrary group $I$ is

$$
\Pi_X^I = (1 - u_x) \left[ b \sum_{J=1}^{K} \nu_J (f_X^J + f_Z^J g_X^{I,J}) - c \right] 
$$

$$
\Pi_Y^I = (1 - u_x) \left[ b \sum_{J=1}^{K} \nu_J (f_X^I + f_Z^J g_Y^{I,J}) \right] 
$$

$$
\Pi_Z^I = (1 - u_x) \left[ b \sum_{J=1}^{K} \nu_J (f_X^I + f_Z^J g_Z^{I,J}) - cg^{I,J} \right].
$$

Note that these payoffs are averaged over all pairwise interactions, i.e., they are already divided by the population size $N$.

1.5 Strategy updating

Each round, after all pairwise games have occurred and all reputations have been updated, a randomly chosen individual considers updating their strategy. They compare their payoff, averaged over all games in which they have played, to that of another random individual in the population. If the focal individual has strategy $i$ and is in group $I$, and their comparison partner has strategy $j$ and is in group $J$, they copy their partner’s strategy with a probability given by the Fermi function

$$
\phi(\Pi_i^I, \Pi_j^J) = \frac{1}{1 + \exp \left[ w(\Pi_i^I - \Pi_j^J) \right]}.
$$

Here, $w$ is a parameter known as the strength of selection (Traulsen et al., 2007, 2010).

In the limit of small $w$ and large population size $N \rightarrow \infty$, the process of gameplay, reputation assessment, and strategic updating can be described by a deterministic replicator equation (Hofbauer and Sigmund, 1998) after an appropriate re-scaling of time (see derivation in Supplementary Information). The form of the resulting replicator equation depends upon how individuals choose comparison partners for strategic updates. We present two possibilities here:

1. disjoint copying, in which an individual may choose any member of their in-group, but not their out-group, as a comparison partner.

2. well-mixed copying, in which an individual is equally likely to choose any other individual as a comparison partner.

In the SI, we consider a more general model in which individuals choose members of their in-group with probability $1 - m$ and choose a random member of the population (irrespective of group identity) with probability $m$, and we show that this model reduces to the ones above in the limits $m \rightarrow 0$ and $m \rightarrow 1$, respectively.
1.5.1 Disjoint copying

If an individual in group $I$ chooses only other individuals in group $I$ as potential comparison partners, then the frequency of strategy $i$ in group $I$ changes over time according to the following replicator equation:

$$
\dot{f}_i = f_i (\Pi_i - \bar{\Pi}), \text{ with } \\
\bar{\Pi} = \sum_i f_i \Pi_i.
$$

(9)

Note that, in this case, even though strategic imitation occurs only within each group, game-play and payoff accumulation occur among all members of the population, so strategy frequencies are not independent across groups.

1.5.2 Well-mixed copying

If an individual in group $I$ is equally likely to choose anyone in the population as a comparison partner, then differences in strategy frequencies between groups do not persist: they converge to a value $f_i$ that is common to all groups, as we show in the SI. We have the resulting replicator equation for the frequencies of strategic types over time:

$$
\dot{f}_i = f_i \sum_J (\nu_J \Pi_i^J - \bar{\Pi}), \text{ with } \\
\bar{\Pi} = \sum_J \nu_J \sum_i f_i \Pi_i^J.
$$

(10)

Because the strategy frequencies cannot vary by group, the quantity that ultimately determines the change in the frequency of each strategy is the group-averaged fitness

$$
\Pi_i = \sum_J \nu_J \Pi_i^J.
$$

(11)

As we show in the SI, this formulation allows us to study the time-evolution and stability of strategies in terms of the average reputations,

$$
\bar{g}_i = \sum_I \sum_J \nu_I \nu_J \bar{g}_{i\,J},
$$

(12)

which represents the probability that a randomly chosen member of the population considers a random individual following strategy $i$ to have a good reputation. By averaging over groups and leveraging the fact that strategy frequencies do not differ by group, we can remove the fitness-dependence on an individual’s group identity $I$. Equation 7 thus simplifies to

$$
\Pi_X = (1 - u_x) \left[ b(f_X + f_Z \bar{g}_X) - c \right] \\
\Pi_Y = (1 - u_x) \left[ b(f_X + f_Z \bar{g}_Y) \right] \\
\Pi_Z = (1 - u_x) \left[ b(f_X + f_Z \bar{g}_Z) - c \bar{g} \right].
$$

(13)

We further show in the SI that the average reputations of cooperators and defectors, $\bar{g}_X$ and $\bar{g}_Y$, can be expressed straightforwardly as

$$
\bar{g}_X = \bar{g}_P + (1 - \bar{g})P_{BC} \\
\bar{g}_Y = \bar{g}_P + (1 - \bar{g})P_{BD}.
$$

(14)

whereas the form of $\bar{g}_Z$ depends sensitively on the group structure of the population.
2 Results

2.1 Population structure weakens the incentive to cooperate

When populations are separated into groups with distinct reputational judgments, it is generally more difficult to achieve a high level of cooperation. To demonstrate this, we first recapitulate the behavior of a population with just one group, which has been studied extensively in prior work (Hilbe et al., 2018; Radzvilavicius et al., 2019, 2021; Santos et al., 2016, 2018; Pacheco et al., 2006; Ohtsuki and Iwasa, 2004, 2006; Ohtsuki et al., 2009; Schmid et al., 2021a,b). Then we compare those results to a population with $K > 1$ distinct groups, each with their own independent reputation information.

2.1.1 Cooperation in a well-mixed population

In a well-mixed population ($K = 1$) with public sharing of reputations, there are two stable equilibria: $f_Y = 1$ (a population consisting entirely of defectors) and $f_Z = 1$ (a population consisting entirely of discriminators). In the SI, we show that an all-defector population can never be invaded, and an all-discriminator population can resist invasion by defectors provided

$$\frac{b}{c} > \frac{1}{P_{GC} - P_{GD}} = \frac{1}{\epsilon - u_a}. \quad (15)$$

For small error rates, this critical benefit-to-cost ratio that guarantees stability of the all-discriminator population is a little greater than 1. There is also an unstable equilibrium consisting of a mixture of $Y$ and $Z$ (defectors and discriminators), located at

$$f_Z^* = \frac{c}{b} \frac{1}{P_{GC} - P_{BD}} = \frac{c}{b} \frac{1}{\epsilon - u_a}$$

$$f_Y^* = 1 - f_Z^*$$

$$f_X^* = 0. \quad (16)$$

As a result, when only discriminators and defectors are present, and discriminators are at a frequency higher than $f_Z^*$, they are destined to fix in the population. Because discriminators attempt to cooperate with everyone they consider to have good reputations, the cooperation rate is given by $(1 - u_x)g$, where $g$, the proportion of the population considered good, satisfies $gZ|_{f_Z=1} = gP_{GC} + (1 - g)P_{BD}$. Solving for $g$ gives

$$g = \frac{P_{BD}}{1 - P_{GC} + P_{BD}} = \frac{q(1 - 2u_a) + u_a}{1 - \epsilon + q(1 - 2u_a) + u_a} = \begin{cases} \frac{u_a}{1 - \epsilon + u_a} & \text{shunning, scoring}, \\ \frac{1 - u_a}{2 - \epsilon + u_a} & \text{stern judging, simple standing}. \end{cases}\quad (17)$$

Under stern judging and simple standing, this value of $g$ is close to 1 (for small error rates), meaning that most of the population is considered good, once the population consists of all discriminators. For example, with $u_a = u_x = .02$, the resulting $g$ is roughly 0.96, and so roughly 96% of the population will be cooperating at the all-discriminator stable equilibrium. Indirect reciprocity with public information thus provides a powerful and reasonable mechanism to not only produce a high level of cooperation but also protect cooperative individuals from the temptation to become defectors.
2.1.2 Cooperation in a population with multiple groups

In models of indirect reciprocity, discriminators tend to enjoy a substantial fitness advantage when information about reputations flows freely, for example by rapid dissemination of reputation information (Sommerfeld et al., 2007) or by adherence to a centralized institution that broadcasts assessments (Radzvilavicius et al., 2021). This generates a high level of agreement among discriminators, which means they are likely to reward each other’s good behavior by cooperating. And so the benefit of mutual cooperation disproportionately accrues to those who are willing to cooperate.

However, even when social interactions occur across the entire well-mixed population, the free flow of reputation information can be disrupted if the population is stratified into groups with potentially different views about reputations. Such a partitioning is expected to temper the advantage of discriminators, who may not agree about the reputations of their interaction partners and are thus more likely to punish each other by defecting.

We demonstrate the effects of population structure on reputation information by solving equation 10 across a varying number of groups $K$ of equal size. In this analysis we allow individuals to copy strategies by payoff comparison with anyone in the population (well-mixed copying), even though information about reputations may vary across groups. The resulting strategy dynamics are shown in the upper panels of figure 1, for a representative set of typical parameters $(b = 2, c = 1, u_x = u_a = 0.02)$.

Under the shunning social norm, the fitness difference between discriminators and defectors is generally small. We have $\bar{g}_Y = u_a = 0.02$, and $\bar{g}_Z$ starts at around 0.34 (for $K = 1$) but rapidly drops for higher $K$. Discriminators thus cooperate in fairly few interactions and receive little fitness benefit from other discriminators. Cooperators are rapidly driven to low frequency due to the cost of cooperation, and subsequent competition between discriminators and defectors occurs close to the $Y - Z$ edge of the strategy simplex. The all-discriminator state is a stable equilibrium for $K$ sufficiently small, but it quickly becomes unstable as $K$ increases.

Under stern judging, the location of the unstable equilibrium along the $Y - Z$ edge moves significantly toward the all-defect vertex at $f_Z = 1$ as $K$ increases. When $K = 2$, the cooperation rate at the all-discriminator stable equilibrium drops from 0.96 to 0.72. When $K \geq 3$, it drops further to 0.65, and the all-discriminator equilibrium ceases to be stable altogether. This instability arises because, as $K$ increases, it is less likely that discriminators will interact with discriminators who share their views of the rest of the population, so the slice of phase space in which discriminators are expected to dominate shrinks.

Finally, under simple standing, the all-$Z$ equilibrium is always stable against invasion by defectors, but the unstable equilibrium along the $Y - Z$ edge shifts slightly toward the all-$Z$ vertex with increasing $K$. In the SI we consider the ability of cooperators to invade the all-discriminator equilibrium under simple standing: we show that cooperators can do so if $\bar{g}_Z$ falls below a fairly high threshold. Accordingly, for sufficiently large $K$, the all-discriminator equilibrium is stable against invasion by defectors but not by cooperators, and there exists a stable equilibrium consisting of a mix of cooperators and discriminators (seen in Figure 1 for $K \geq 5$).

We show in the SI that the dynamics under scoring do not depend on the number or relative size of groups, so we do not present results for scoring here.

We can also analyze the effects of multiple groups on the rate of cooperation at the all-discriminator vertex. Discriminators can resist invasion by defectors only when their fitness exceeds the fitness of a rare defector mutant near the $f_Z = 1$ vertex, i.e., when $(b - c)\bar{g} |_{f_Z=1} > b \bar{g}_Y |_{f_Z=1}$. This condition
Figure 1: Strategy frequency dynamics for various numbers of equally-sized groups, $K$, under shunning, stern judging, and simple standing, in populations where strategies are freely copied across group lines. As $K$ increases, the dynamics rapidly approach those of a model with private assessment (Radzvilavicius et al., 2021) (fourth row). Open circles indicate unstable equilibria; filled circles indicate stable equilibria. In all panels, $b = 2$, $c = 1$, $u_a = u_x = 0.02$. 
In Figure 2 we plot the value of $\bar{g}_{fZ} = 1$ and demonstrate that, as $K$ increases, the cooperation rate in a population of all discriminators decreases below this threshold under shunning and stern judging, whereas it remains above this threshold for simple standing. And so a sufficiently large number of separate groups, with access to disjoint sets of reputation judgments, destabilizes cooperation under two of the norms we consider, but it does not threaten cooperation under simple standing.

### 2.1.3 Many groups

As the number of groups $K$ approaches infinity, we recover the reputation dynamics for discriminators in a population of private assessors (see Supplementary Information):

$$g_Z = g_2 P_{GC} + (g - g_2)(P_{GD} + P_{BC}) + (1 - 2g + g_2)P_{BD},$$

$$g_2 = \sum_i f_i g_i^2.$$  \hfill (19)

These expressions are identical to those derived in Radzvilavicius et al. (2019) in the case of no empathy: the three terms of $g_Z$ correspond (respectively) to the donor and observer agreeing that the recipient is good, disagreeing about the recipient’s reputation, and agreeing that the recipient is bad. This result makes intuitive sense because, in this limit of infinitely many information groups, each individual in the population effectively has an independent view from all other individuals – which is equivalent to individuals with private information about reputations.

Figure 1 also reflects these results. We see that the average reputation of the all-discriminator population, $\bar{g}_{fZ}$, rapidly approaches the private-assessment limit as the number of groups $K$ increases. Under simple standing, the asymptotic private-assessment limit still exceeds the reputation
value required for discriminators to resist invasion by defectors. This is why, even under private assessment, simple standing allows discriminators to persist in a sizable region of parameter space: there is a stable equilibrium that consists of a mixture of discriminators and cooperators.

2.1.4 Groups of different size

We also consider a scenario in which a fraction \( \nu \) of the population belongs to one large group and the remaining \( K - 1 \) groups each comprise fractions \((1 - \nu)/(K - 1)\) of the population. In the SI, we show that, as \( K \) approaches infinity, this case reduces to a model with a mix of adherents to a single institution (in the group of size \( \nu \)) and private assessors (in the remaining groups), which has been previously studied (Radzvilavicius et al., 2021).

2.2 Strategic imitation restricted by population structure

In the preceding analyses, we have assumed that individuals freely copy strategies across groups (well-mixed copying), so that the only impact of population structure was to partition reputation information into distinct groups. In this section we consider a model in which, in addition, strategic imitation occurs only within groups, disallowing imitation between groups (disjoint copying). Even when strategic updates are constricted in this manner, game-play interactions between groups mean that the strategic composition of one group may shape the composition in another group. Here we consider the case of \( K = 2 \) groups, disallowing strategic imitation across groups. This model requires that we keep track of the strategy frequencies in each of the groups separately.

We study the behavior of strategies in group 1, when group 2 is exogenously fixed for either DISC or ALLD. Figure 3 shows the dynamics that arise in these cases, for one choice of parameter values: \( \nu_1 = \nu_2 = 1/2, b = 2, c = 1, \) and \( u_a = u_x = 0.02. \)

In group 1, the three corners of the simplex (populations fixed for ALLC, ALLD, or DISC) are all equilibria. In the SI, we obtain some useful general results in this setting. First, we show that, under disjoint copying, when group 2 is fixed for defectors, the all-defector equilibrium in group 1 remains stable. However, when group 2 is fixed for discriminators, the all-defector equilibrium in group 1 can become unstable. We further show that, under disjoint copying, when group 2 is fixed for defectors, the all-discriminator equilibrium in group 1 is stable against invasion by defectors provided

\[
\frac{b}{c} > \frac{1}{\nu_1(P_{GC} - P_{GD})} = \frac{1}{\nu_1(\epsilon - u_a)}.
\]  

(20)

(This reduces to the one-group case in the limit \( \nu_1 \to 1. \) In the example shown in Figure 3, we have \( b = 2 \) and \( c = 1. \) The critical \( b/c \) value is slightly greater than 2, meaning that the all-discriminator equilibrium just barely fails to be stable (bottom row). When groups are partitioned in this manner and copying is disjoint, discriminators in one group “waste” effort on defectors in the other group: the other group contains no discriminators, so they can only accrue a payoff due to discriminators in their own group. This wasted effort manifests as a lower average payoff for discriminators, which increases the temptation to defect.

When group 2 is fixed for discriminators, however, this can help group 1 discriminators resist invasion by defectors (middle row of Figure 3). How much help is provided by group 2 depends sensitively on the social norm, specifically how likely discriminators in one group are to look kindly upon discriminators with different reputational views. Under shunning, any interaction with an individual with a bad reputation yields a bad reputation; under simple standing, any such interaction yields a good reputation; and stern judging is intermediate between the two. This helps explain the
behavior of equilibria along the $Y - Z$ edge of the simplex seen in Figure 3. Under shunning, the all-$Z$ equilibrium is unstable; under stern judging and simple standing, it is stable, and there exists an unstable mixed $Y - Z$ equilibrium, corresponding to a slice of phase space that is drawn to the all-$Z$ stable equilibrium. This slice of phase space is larger under simple standing than under stern judging, as expected.

When strategies are freely copied across groups (well-mixed copying), strategy frequencies equilibrate quickly, and their dynamics can be understood in terms of group-averaged reputations. What we have shown here is that even in the polar opposite copying scenario (disjoint copying), the fact that individuals freely interact across group lines causes their dynamics to be linked. The general tendency is that discriminators in one group make it easier for discriminators to proliferate in the other group, whether by making the other group’s discriminators stable against invasion or even by making defectors vulnerable to invasion by discriminators. Conversely, defectors in one group can render another group more vulnerable to invasion by defectors. And so even without direct strategy copying, gameplay between disjoint groups can cause their strategy compositions to resemble each other.
Figure 3: Strategy frequency dynamics for one of two equally-sized groups ($\nu_1 = \nu_2 = 1/2$) under shunning, stern judging, and simple standing. In all panels, the dynamics of strategy frequencies in group 1 are shown. The top row corresponds to well-mixed copying; the bottom two rows correspond to disjoint copying, with group 2 fixed for either DISC or ALLD. Under stern judging of simple standing, fixing group 2 for DISC increases the basin of attraction for cooperation in group 1, whereas fixing group 2 for ALLD reduces the cooperative basin in group 1. In all panels, $b = 2$, $c = 1$, $u_a = u_x = 0.02$. 
References

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, 115(48):12241–12246.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of theoretical biology*, 231(1):107–120.

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

Ohtsuki, H., Iwasa, Y., and Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, 457(7225):79.

Pacheco, J. M., Santos, F. C., and Chalub, F. A. C. (2006). Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS computational biology*, 2(12):e178.

Radzvilavicius, A. L., Kessinger, T. A., and Plotkin, J. B. (2021). Adherence to public institutions that foster cooperation. *Nature Communications*, 12(1):3567.

Radzvilavicius, A. L., Stewart, A. J., and Plotkin, J. B. (2019). Evolution of empathetic moral evaluation. *eLife*, 8:e44269.

Rapoport, A., Chammah, A. M., and Orwant, C. J. (1965). *Prisoner’s dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press.

Santos, F. P., Pacheco, J. M., and Santos, F. C. (2016). Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Scientific reports*, 6:37517.

Santos, F. P., Santos, F. C., and Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695):242.

Schmid, L., Chatterjee, K., Hilbe, C., and Nowak, M. A. (2021a). A unified framework of direct and indirect reciprocity. *Nature Human Behaviour*, 5(10):1292–1302.

Schmid, L., Shati, P., Hilbe, C., and Chatterjee, K. (2021b). The evolution of indirect reciprocity under action and assessment generosity. *Scientific Reports*, 11(1):17443.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the national academy of sciences*, 104(44):17435–17440.

Traulsen, A., Pacheco, J. M., and Nowak, M. A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of theoretical biology*, 246(3):522–529.

Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M. (2010). Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences*, 107(7):2962–2966.
1. Generalized social norms

In this section, we derive the expressions for $P_{GC}$, $P_{GD}$, $P_{BC}$ and $P_{BD}$ used in the main text.

We begin by generalizing the “big four” social norms, which occur as special cases of a two-parameter family of norms. We suppose that cooperation with a bad individual yields a good reputation with probability $p$ (barring errors) and defecting with a bad individual yields a good reputation with probability $q$ (again barring errors). We recover stern judging, simple standing, scoring, and shunning with $(p, q) = (0, 1), (1, 1), (1, 0), (0, 0)$, respectively.

In the presence of errors of execution and assessment, an individual can obtain a good reputation in the following ways. They may be observed:

1. interacting with an individual with a good reputation and intending to cooperate.
   (a) With probability $1 - u_x$, they successfully cooperate. With probability $1 - u_a$, they are successfully assigned a good reputation.
   (b) With probability $u_x$, they accidentally defect. With probability $u_a$, they are accidentally assigned a good reputation.

   Thus,
   $$P_{GC} = (1 - u_x)(1 - u_a) + u_xu_a = \epsilon.$$

2. interacting with an individual with a good reputation and intending to defect. This is always considered a bad action, so such an individual can only achieve a good reputation on accident. Thus,
   $$P_{GD} = u_a.$$

3. interacting with an individual with a bad reputation and intending to cooperate.
   (a) With probability $1 - u_x$, they successfully cooperate. With probability $p$, this is considered a “good reputation” action. With probability $1 - u_a$, they are successfully assigned a good reputation.
(b) With probability $1 - u_x$, they successfully cooperate. With probability $1 - p$, this is considered a “bad reputation” action. With probability $u_a$, they are accidentally assigned a good reputation.

(c) With probability $u_x$, they accidentally defect. With probability $q$, this is considered a “good reputation” action. With probability $1 - u_a$, they are successfully assigned a good reputation.

(d) With probability $u_x$, they accidentally defect. With probability $1 - q$, this is considered a “bad reputation” action. With probability $u_a$, they are accidentally assigned a good reputation.

Thus, the total probability is

$$P_{BC} = (1 - u_x)[p(1 - u_a) + (1 - p)u_a] + u_x[q(1 - u_a) + (1 - q)u_a]$$
$$= (1 - u_x)[p - 2pu_a + u_a] + u_x[q - 2qu_a + u_a]$$
$$= p - 2pu_a + u_a - pu_x + 2pu_xu_a - u_xu_a + qu_x - 2qu_xu_a + u_xu_a$$
$$= p(1 - 2u_a - u_x + 2u_a u_x) + qu_x - 2u_x u_a + u_a$$
$$= p(\epsilon - u_a) + q(1 - \epsilon - u_a) + u_a. \quad (1)$$

(4) interacting with an individual with a bad reputation and intending to defect.

(a) With probability $q$, this is considered a “good reputation” action. They defect and successfully obtain a good reputation with probability $1 - u_a$.

(b) With probability $1 - q$, this is considered a “bad reputation” action. They defect and accidentally obtain a good reputation with probability $u_a$.

Thus, the total probability is

$$P_{BD} = q(1 - u_a) + (1 - q)u_a = q(1 - 2u_a) + u_a. \quad (2)$$

We recover the traditional four social norms in the following limits:

(1) when $(p, q) = (0, 1)$ (stern judging), equation 1 becomes $1 - \epsilon$ and equation 2 becomes $1 - u_a$.

(2) when $(p, q) = (1, 1)$ (simple standing), equation 1 becomes $1 - u_a$ and equation 2 becomes $1 - u_a$.

(3) when $(p, q) = (1, 0)$ (scoring), equation 1 becomes $\epsilon$ and equation 2 becomes $u_a$.

(4) when $(p, q) = (0, 0)$ (shunning), equation 1 becomes $u_a$ and equation 2 becomes $u_a$.

The values of $P_{BC}, P_{GC}, P_{GD}$, and $P_{BD}$ for these four norms are summarized in SI Table 1.

| observer view of recipient donor intent | good cooperate $P_{GC}$ | good defect $P_{GD}$ | bad cooperate $P_{BC}$ | bad defect $P_{BD}$ |
|---------------------------------------|-------------------------|---------------------|------------------------|---------------------|
| general expression                    | $\epsilon$              | $u_a$               | $p(\epsilon - u_a) + q(1 - \epsilon - u_a) + u_a$ | $q(1 - 2u_a) + u_a$ |
| shunning $(p = 0, q = 0)$             | $\epsilon$              | $u_a$               | $u_a$                  | $u_a$               |
| stern judging $(p = 0, q = 1)$        | $\epsilon$              | $u_a$               | $1 - \epsilon$        | $1 - u_a$           |
| scoring $(p = 1, q = 0)$              | $\epsilon$              | $u_a$               | $\epsilon$            | $u_a$               |
| simple standing $(p = 1, q = 1)$      | $\epsilon$              | $u_a$               | $1 - u_a$             | $1 - u_a$           |

**Table 1.** Probability that an observer will assign a donor a good reputation based on the donor’s action and the observer’s view of the recipient, under various social norms. Here, $\epsilon = (1 - u_a)(1 - u_x) + u_a u_x$ is the probability that an individual who intends to cooperate with a recipient who has a good reputation is ultimately themselves assigned a good reputation. They may either successfully cooperate and be correctly assigned a good reputation (first term) or accidentally defect and be wrongly assigned a good reputation (second term).
1.1. Reputation dynamics for cooperators, defectors, and discriminators. Given the expressions for $P_{BC}, P_{GC}, P_{GD},$ and $P_{BD}$ derived above, we now consider what portion of each types of strategy will be assigned a good reputation, and by whom.

A cooperator in group $I$ can be assigned a good reputation in the eyes of group $J$ in two ways. $J$ can observe the $I$ cooperator’s interaction:

1. with someone group $J$ sees as good (probability $g_{*(*)}$); the $I$ member cooperates, which yields a good reputation with probability $P_{GC}$.
2. with someone group $J$ sees as good (probability $1 - g_{*(*)}$); the $I$ member cooperates, which yields a good reputation with probability $P_{BD}$.

We thus have

$$g_{*(*)}^{I,J} = g_{*}^{I,J}P_{GC} + (1 - g_{*}^{I,J})P_{BC}.$$  

Similar reasoning for defectors yields

$$g_{*(*)}^{I,J} = g_{*}^{I,J}P_{GD} + (1 - g_{*}^{I,J})P_{BD}.$$  

Discriminators vary their behavior according to the reputation of the recipient, but discriminators in different groups are not guaranteed to have the same views of each recipient’s reputation. Thus, discriminators will be viewed differently by their in-group versus their out-group. A discriminator in group $I$ can gain a good reputation in the eyes of group $I$ (their in-group) in two ways. $I$ can observe the $I$ discriminator’s interaction:

1. with someone group $I$ sees as good (probability $g_{*(*)}$); the $I$ discriminator cooperates, which yields a good reputation with probability $P_{GC}$.
2. with someone group $I$ sees as bad (probability $1 - g_{*(*)}$); the $I$ discriminator defects, which yields a good reputation with probability $P_{BD}$.

A discriminator in group $I$ can gain a good reputation in the eyes of group $J \neq I$ (their out-group) in four ways. $J$ can observe the $I$ discriminator’s interaction:

1. with someone in an arbitrary group $L$ following strategy $i$ (probability $\nu_{L}f_{i}^{L,J}$) whom $I$ sees as good (probability $g_{i}^{L,J}$) and whom $J$ also sees as good (probability $g_{i}^{L,J}$); the $I$ discriminator cooperates, which yields a good reputation with probability $P_{GC}$.
2. with someone in an arbitrary group $L$ following strategy $i$ (probability $\nu_{L}f_{i}^{L,J}$) whom $I$ sees as bad (probability $1 - g_{i}^{L,J}$) but whom $J$ sees as good (probability $g_{i}^{L,J}$); the $I$ discriminator defects, which yields a good reputation with probability $P_{GD}$.
3. with someone in an arbitrary group $L$ following strategy $i$ (probability $\nu_{L}f_{i}^{L,J}$) whom $I$ sees as good (probability $g_{i}^{L,J}$) but whom $J$ sees as bad (probability $1 - g_{i}^{L,J}$); the $I$ discriminator cooperates, which yields a good reputation with probability $P_{BC}$.
4. with someone in an arbitrary group $L$ following strategy $i$ (probability $\nu_{L}f_{i}^{L,J}$) whom $I$ sees as bad (probability $1 - g_{i}^{L,J}$) and whom $J$ also sees as bad (probability $1 - g_{i}^{L,J}$); the $I$ discriminator defects, which yields a good reputation with probability $P_{BD}$.

We can sum over all groups and strategy combinations to obtain

$$g_{\alpha}^{L,J} = \sum_{L} \nu_{L} \sum_{i} f_{i}g_{i}^{L,J},$$

$$g_{\beta}^{L,J} = \sum_{L} \nu_{L} \sum_{i} f_{i}(1 - g_{i}^{L,J})g_{i}^{L,J} = g_{*}^{I,J} - g_{\alpha}^{L,J},$$

$$g_{\gamma}^{L,J} = \sum_{L} \nu_{L} \sum_{i} f_{i}g_{i}^{L,J}(1 - g_{i}^{L,J}) = g_{*}^{I,J} - g_{\delta}^{L,J},$$

$$g_{\delta}^{L,J} = \sum_{L} \nu_{L} \sum_{i} f_{i}(1 - g_{i}^{L,J})(1 - g_{i}^{L,J}) = 1 - g_{*}^{I,J} - g_{*}^{I,J} + g_{\alpha}^{L,J}.$$
Thus,
\[
g_{Z}^{I,J} = \delta_{I,J} \left[ g^{*,J} P_{GC} + (1 - g^{*,J}) P_{BD} \right] \\
+ (1 - \delta_{I,J}) \left[ g_{\alpha}^{I,J} P_{GC} + g_{\beta}^{I,J} P_{GD} + g_{\gamma}^{I,J} P_{BC} + g_{\delta}^{I,J} P_{BD} \right] \\
= \delta_{I,J} \left[ g^{*,J} P_{GC} + (1 - g^{*,J}) P_{BD} \right] \\
+ (1 - \delta_{I,J}) \left[ g_{\alpha}^{I,J} (P_{GC} - P_{GD} - P_{BC} + P_{BD}) + g^{*,J} (P_{GD} - P_{BD}) + g^{*,J} (P_{BC} - P_{BD}) + P_{BD} \right].
\]

(3)

1.2. Special case: scoring \((p = 1, q = 0)\). Under scoring, we have

\[
P_{GC} = P_{BC} = \epsilon, \\
P_{GD} = P_{BD} = u_{a}.
\]

In this case,

\[
g_{\alpha}^{I,J} P_{GC} + g_{\beta}^{I,J} P_{GD} + g_{\gamma}^{I,J} P_{BC} + g_{\delta}^{I,J} P_{BD} \\
= g^{I,J} (P_{GC} - P_{GD} - P_{BC} + P_{BD}) + g^{*,J} (P_{GD} - P_{BD}) + g^{*,J} (P_{BC} - P_{BD}) + P_{BD} \\
= g^{*,J} (P_{BC} - P_{BD}) + P_{BD} \\
= g^{*,J} \epsilon + (1 - g^{*,J}) u_{a}.
\]

We likewise have

\[
g_{X}^{I,J} = g^{*,J} \epsilon + (1 - g^{*,J}) \epsilon = \epsilon, \\
g_{Y}^{I,J} = g^{*,J} u_{a} + (1 - g^{*,J}) u_{a} = u_{a}, \\
g_{Z}^{I,J} = \delta_{I,J} \left[ g^{*,J} \epsilon + (1 - g^{*,J}) u_{a} \right] + (1 - \delta_{I,J}) \left[ g^{*,J} \epsilon + (1 - g^{*,J}) u_{a} \right] \\
= \delta_{I,J} \left[ g^{*,J} \epsilon + (1 - g^{*,J}) u_{a} \right] + (1 - \delta_{I,J}) \left[ g^{*,J} \epsilon + (1 - g^{*,J}) u_{a} \right] \\
= g^{*,J} \epsilon + g^{*,J} u_{a}.
\]

The last line implies that \(J\)’s opinion of \(I\) discriminators depends solely on whom \(I\) sees as good, not whom \(J\) sees as good. This is reasonable; scoring is a first-order norm, in which any cooperation is considered good and any defection is considered bad, meaning that an \(I\) discriminator will be considered good as a result of their interactions with those \(I\) sees as good (with whom they therefore cooperate). Likewise, an \(I\) discriminator will be considered good as a result of their interactions with those \(I\) sees as bad (with whom they therefore defect). One may note that

\[
g^{*,I} = \sum_{L} \nu_{L} g^{L,I} \\
= \sum_{L} \nu_{L} \left( f^{L}_{X} (g^{*,J} \epsilon + (1 - g^{*,J}) u_{a}) + f^{L}_{Y} (g^{*,J} \epsilon + (1 - g^{*,J}) u_{a}) \right) \\
= \epsilon \sum_{L} \nu_{L} f^{L}_{X} + u_{a} \sum_{L} \nu_{L} f^{L}_{Y} + g^{*,J} \epsilon \sum_{L} \nu_{L} f^{L}_{Z} + (1 - g^{*,J}) u_{a} \sum_{L} \nu_{L} f^{L}_{Z} \\
\therefore g^{*,I} = \frac{\epsilon \sum_{L} \nu_{L} f^{L}_{X} + u_{a} \sum_{L} \nu_{L} f^{L}_{Y} + f^{L}_{Z}}{1 - \sum_{L} \nu_{L} f^{L}_{Z} (\epsilon - u_{a})},
\]

which is independent of \(I\). In this way, under scoring, the reputation of discriminators does not depend on their group identity. Moreover, if there is no difference in strategy frequency among
groups, we have
\[
g_{\bullet \ell} = \sum_{L} \nu_{L} \left( f_{X}^{L} + f_{Z}^{L} \right) \epsilon \sum_{L} \nu_{L}(\epsilon - u_{a})
\]
\[
= \frac{\sum_{L} \nu_{L} f_{X}^{L} + u_{a} \sum_{L} \nu_{L}(f_{Y}^{L} + f_{Z}^{L})}{1 - f_{Z} \sum_{L} \nu_{L}(\epsilon - u_{a})}
\]
which is independent of the number of groups and their relative sizes. Thus, under scoring, if strategy frequencies are equal among groups, imposing a group structure on the population does not affect reputations at all, and hence it does not affect the strategy dynamics.

2. Single-group competition between defectors and discriminators

When \( K = 1 \), there are two stable equilibria: a population consisting entirely of defectors (\( Y \)) and a population consisting entirely of discriminators (\( Z \)). Here, we consider the circumstances under which these equilibria are stable against invasion. Let \( f = f_{Z} \). Defectors resist invasion by discriminators provided

\[
\partial_{f} f_{Y} | f = 0 < 0
\]
\[
\partial_{f} f_{Y}(\Pi_{Y} - \Pi_{X}) | f = 0 < 0
\]
\[
\partial_{f} [(1 - f)(\Pi_{Y} - (1 - f)\Pi_{Y} - f\Pi_{Z})] | f = 0 < 0
\]
\[
\partial_{f} [(1 - f)(\Pi_{Y} - (1 - f)\Pi_{Y} - f\Pi_{Z})] | f = 0 < 0
\]
\[
\partial_{f} [(f - f^{2})(\Pi_{Z} - \Pi_{Y})] | f = 0 < 0
\]
\[
[(1 - 2f)(\Pi_{Z} - \Pi_{Y}) + (f - f^{2})\partial_{f}(\Pi_{Z} - \Pi_{Y})] | f = 0 < 0
\]
\[
\Pi_{Z} | f = 0 < \Pi_{Y} | f = 0
\]
\[
(bf_{Z} - cg) | f = 0 < bf_{Y} | f = 0
\]
\[
-cg_{Y} < 0.
\]

Since \( c \) and \( g_{Y} \) are both positive, this condition always obtains: discriminators can never invade a population of defectors. Likewise, discriminators resist invasion by defectors provided

\[
\partial_{f} f_{Y} | f = 1 < 0
\]
\[
\partial_{f} f_{Y}(\Pi_{Y} - \Pi_{X}) | f = 1 < 0
\]
\[
\partial_{f} [(1 - f)(\Pi_{Y} - (1 - f)\Pi_{Y} - f\Pi_{Z})] | f = 1 < 0
\]
\[
\partial_{f} [(1 - f)(\Pi_{Y} - (1 - f)\Pi_{Y} - f\Pi_{Z})] | f = 1 < 0
\]
\[
\partial_{f} [(f - f^{2})(\Pi_{Z} - \Pi_{Y})] | f = 1 < 0
\]
\[
[(1 - 2f)(\Pi_{Z} - \Pi_{Y}) + (f - f^{2})\partial_{f}(\Pi_{Z} - \Pi_{Y})] | f = 1 < 0
\]
\[
\Pi_{Z} | f = 1 > \Pi_{Y} | f = 1
\]
\[
(bf_{Z} - cg) | f = 1 > bf_{Y} | f = 1
\]
\[
(bg_{Z} - cg_{Z}) | f = 1 > bg_{Y} | f = 1
\]
\[
b(g_{Z} - g_{Y}) | f = 1 > cg_{Z} | f = 1
\]
\[
b(P_{GC} - P_{GD}) > cg
\]
\[
\therefore \frac{b}{c} > \frac{1}{P_{GC} - P_{GD}} = \frac{1}{\epsilon - u_{a}}.
\]
This can also be written in terms of \( g \): discriminators resist invasion by defectors provided

\[
\Pi_Z|_{f_Z=1} > \Pi_Y|_{f_Z=1}
\]
\[
bfg_Z|_{f_Z=1} - cg > bfg_Y|_{f_Z=1}
\]
\[
(b-c)g > b[gP_{GD} + (1-g)P_{BD}]
\]
\[
g(b-c - b[P_{GD} - P_{BD}] > bP_{BD}
\]
\[
g(b[1-P_{GD} + P_{BD}] - c) > bP_{BD}
\]
\[
\therefore g > \frac{P_{BD}}{1-P_{GD} + P_{BD} - c/b}.
\]

We do not need to flip the inequality because \( b > c \) and because \( 1 + P_{BD} - P_{GD} \) is guaranteed to be greater than or equal to 1 for every social norm we consider. Finally, there is a third equilibrium between the two which, by the mean value theorem, is unstable, at (letting \( f = f_Z \) again)

\[
\dot{f} = 0
\]
\[
f(\Pi_Z - \Pi) = 0
\]
\[
\Pi_Z - f\Pi_Z - (1-f)\Pi_Y = 0
\]
\[
\Pi_Z = \Pi_Y
\]
\[
bfg_Z - cg = bfg_Y
\]
\[
bf(gP_{GC} + [1-g]P_{BD}) - cg = bf(gP_{GD} + [1-g]P_{BD})
\]
\[
bf(gP_{GC} - P_{GD}) = cg
\]
\[
\therefore f = \frac{c}{b} \frac{1}{P_{GC} - P_{GD}} = \frac{c}{b} \frac{1}{\epsilon - u_a}.
\]

An equivalent way to express this is that discriminators rise in frequency provided

\[
f_Z(g_Z - g_Y) > c/b. \tag{5}
\]

If there is no value of \( f_Z \) for which this is true, then discriminators do not rise in frequency: if it is not true for \( f_Z = 1 \) even when the inequality is relaxed, discriminators cannot resist invasion by defectors.

Finally, we note that there are conditions under which cooperators can invade a population of discriminators. We proceed by reasoning similar to equation 4, noting that the stability of an equilibrium against invasion is determined by evaluating the fitnesses of the resident and the invader at that equilibrium. Cooperators can invade discriminators when (letting \( f = f_Z \))

\[
\Pi_X|_{f=1} > \Pi_Z|_{f=1}
\]
\[
(bfg_X - c)|_{f=1} > (bfg_Z - cg)|_{f=1}
\]
\[
b(g_X - g) > c(1-g)
\]
\[
b(gP_{GC} + (1-g)P_{BC} - g) > c(1-g)
\]
\[
b(gP_{GC} - P_{BC} - 1) + P_{BC}) > c(1-g)
\]
\[
g(b[1-P_{GC} - P_{BC} - 1] + P_{BC}) > c - bP_{BC}
\]
\[
\therefore \begin{cases} 
 g > \frac{c - bP_{BC}}{b(P_{GC} - P_{BC} - 1) + c} & \text{shunning, stern judging,} \\
 g < \frac{c - bP_{BC}}{b(P_{GC} - P_{BC} - 1) + c} & \text{scoring, simple standing.}
\end{cases}
\]

For small error rates, this condition is never satisfied under shunning or stern judging (the right hand side is generally greater than 1), but it can be met under scoring and simple standing. With \( u_x = u_a = .02 \) and \( b = 2, c = 1 \), the cutoff is about 0.92 for both simple standing and scoring; for \( b = 5, c = 1 \), the cutoff is about 0.95. This means that if discriminators do not view each other
as having sufficiently good reputations, they become vulnerable to invasion by cooperators! This explains the emergence of a mixed ALLC-DISC equilibrium under these norms.

3. **Dynamics of Multiple Groups with Well-Mixed Copying**

Under the assumption of well-mixed copying, the only interesting dynamical quantities are the “total” (group-averaged) strategy fitnesses $\Pi_i$, viz.:

$$\dot{f}_i = f_i \left[ \left( \sum_l \nu_l \Pi_i^l \right) - \bar{\Pi} \right] = f_i \left[ \Pi_i - \bar{\Pi} \right].$$

By summing over all groups, we obtain

$$\Pi_Z = \sum_l \nu_l \Pi_Z^l = \sum_l \left\{ \nu_l (1 - u_x) \left[ b \sum_j \nu_j (f_X^l + f_Y^l g_{Z}^{l,j}) - c g^{\bullet \cdot l} \right] \right\}$$

$$= (1 - u_x) \sum_l \left\{ \nu_l \left[ b \sum_j \nu_j (f_X^l + f_Y^l g_{Z}^{l,j}) - c g^{\bullet \cdot l} \right] \right\}$$

$$= (1 - u_x) \left[ b(f_X + f_Z \sum_l \sum_j \nu_l \nu_j g_{Z}^{l,j}) - c \sum_l \nu_l g^{\bullet \cdot l} \right]$$

$$= (1 - u_x) \left[ b(f_X + f_Z \bar{g}_Z) - c \bar{g} \right],$$

and likewise

$$\Pi_X = \sum_l \nu_l \Pi_X^l = (1 - u_x) \left[ b(f_X + f_Z \bar{g}_X) - c \right],$$

$$\Pi_Y = \sum_l \nu_l \Pi_Y^l = (1 - u_x) \left[ b(f_X + f_Z \bar{g}_Y) \right].$$

Here, $\bar{g}_i$ is the probability that a randomly chosen individual of type $i$ is seen as good by a randomly chosen observer in the population; and likewise $\bar{g}$ is the probability that a randomly chosen individual is seen as good by a randomly chosen observer (by the linearity of averages we have $\bar{g} = f_X \bar{g}_X + f_Y \bar{g}_Y + f_Z \bar{g}_Z$). In general we will have

$$\bar{g}_X = \sum_{l=1}^K \sum_{J=1}^K \nu_l \nu_J g_{X}^{l,J}$$

$$= \sum_{l=1}^K \sum_{J=1}^K \nu_l \nu_J [g^{\bullet \cdot J} P_{GC} + (1 - g^{\bullet \cdot J}) P_{BC}]$$

$$= \sum_{l=1}^K \sum_{J=1}^K \nu_l \nu_J g^{\bullet \cdot J} P_{GC} + \sum_{l=1}^K \sum_{J=1}^K \nu_l \nu_J (1 - g^{\bullet \cdot J}) P_{BC}$$

$$= \sum_{l=1}^K \nu_l \bar{g}_l P_{GC} + \sum_{l=1}^K \nu_l (1 - \bar{g}) P_{BC}$$

$$= \bar{g}_l P_{GC} + (1 - \bar{g}) P_{BC},$$

and likewise

$$\bar{g}_Y = \bar{g}_l P_{GD} + (1 - \bar{g}) P_{BD}.$$
The form of $\bar{g}_Z$ will vary depending on the specific scenario, but we can obtain a couple of general relations. First, note that equation 4 becomes

$$\Pi_{Z|f_Z=1} > \Pi_{Y|f_Z=1}$$
$$b\bar{g}_Z|_{f_Z=1} - cg > b\bar{g}_Y|_{f_Z=1}$$
$$(b - c)\bar{g} > b[g_{GD} + (1 - \bar{g})P_{BD}]$$
$$\bar{g}(b - c - b[P_{GD} - P_{BD}] > bP_{BD}$$
$$\bar{g}(b[1 - P_{GD} + P_{BD}] - c) > bP_{BD}$$
$$\therefore \bar{g} > \frac{P_{BD}}{1 - P_{GD} + P_{BD} - c/b},$$

and equation 5 becomes

$$f_Z(\bar{g}_Z - \bar{g}_Y) > c/b.$$ 

That is, under well-mixed copying, the conditions for discriminators to resist invasion by defectors and to increase in frequency over time can be written in terms of average reputations $\bar{g}_i$, though the value of those reputations will vary depending on the group structure.

3.1. **Groups of identical size.** When all $K$ groups have the same size $1/K$ and strategies spread via well-mixed copying, we can solve for $\bar{g}_Z$:

$$\bar{g}_Z = \sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J g_{Z}^{I,J}$$
$$= \sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J \left(\delta_{I,J} [g_{\star}^{*J}P_{GC} + (1 - g_{\star}^{*J})P_{BD}] + (1 - \delta_{I,J}) \left[ g_{\alpha}^{I,J}P_{GC} + g_{\beta}^{I,J}P_{GD} + g_{\gamma}^{I,J}P_{BC} + g_{\delta}^{I,J}P_{BD} \right] \right)$$
$$= \sum_{J=1}^{K} \nu_J^2 [g_{\star}^{*J}P_{GC} + (1 - g_{\star}^{*J})P_{BD}] + \sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J \left( g_{\alpha}^{I,J}P_{GC} + g_{\beta}^{I,J}P_{GD} + g_{\gamma}^{I,J}P_{BC} + g_{\delta}^{I,J}P_{BD} \right).$$

The first term simplifies to

$$\sum_{J=1}^{K} \nu_J^2 [g_{\star}^{*J}P_{GC} + (1 - g_{\star}^{*J})P_{BD}] = \frac{1}{K} \sum_{I=1}^{K} \nu_I [g_{\star}^{*J}P_{GC} + (1 - g_{\star}^{*J})P_{BD}]$$
$$= \frac{1}{K} \left[ g_{P_{GC}} + (1 - \bar{g})P_{BD} \right].$$

The second becomes

$$\sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J \left( g_{\alpha}^{I,J}P_{GC} + g_{\beta}^{I,J}P_{GD} + g_{\gamma}^{I,J}P_{BC} + g_{\delta}^{I,J}P_{BD} \right).$$

$$= \sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J \left[ g_{\alpha}^{I,J}(P_{GC} - P_{GD} - P_{BC} + P_{BD}) + g_{\star}^{I,J}(P_{GD} - P_{BD}) + g_{\star}^{I,J}(P_{BC} - P_{BD}) + P_{BD} \right].$$
Because all the groups are the same size and strategy frequencies are identical across groups, the values of $g_i^{L,I}$ can only vary depending on whether $I = J$ or not. We exploit this symmetry to obtain

$$g^{•,i} = \sum_{L} \nu_L g_{L,i}^{L,I}$$

$$= \frac{1}{K} \sum_{L} g_{L,i}^{L,I} = \frac{1}{K} g_{L,i}^{I,I} + \frac{K - 1}{K} g_{L,i}^{L,I} \bigg|_{L \neq I},$$

$$g^{•,j} = \frac{1}{K} g_{L,j}^{I,I} + \frac{K - 1}{K} g_{L,j}^{L,I} \bigg|_{L \neq I}$$

and

$$g_{\alpha}^{I,I} = \sum_{L} \nu_L \sum_{i} f_i g_i^{L,I} L_{L,i}^{L,I}$$

$$= \frac{1}{K} \sum_{L} \sum_{i} f_i g_i^{L,I} L_{L,i}^{L,I}$$

$$= \frac{1}{K} \sum_{L} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I}$$

$$= \frac{1}{K} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I} \bigg|_{L \neq I} + \frac{1}{K} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I} \bigg|_{L \neq J} + \frac{K - 2}{K} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I} \bigg|_{L \neq I \neq J}$$

$$= \frac{2}{K} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I} \bigg|_{L \neq I} + \frac{K - 2}{K} \sum_{i} f_i (g_i^{I,I})^2 \bigg|_{L \neq I}^{\cdot}.$$  

Thus

$$\bar{g}_Z = \sum_{I=1}^{K} \sum_{J=1}^{K} \nu_I \nu_J g_{\bar{Z}}^{I,J}$$

$$= \frac{1}{K} [\bar{g} P_{GC} + (1 - \bar{g}) P_{BD}] + \frac{K - 1}{K} \left[ \left( \frac{2}{K} \sum_{i} f_i g_i^{I,I} L_{L,i}^{I,I} \bigg|_{L \neq I} + \frac{K - 2}{K} \sum_{i} f_i (g_i^{I,I})^2 \bigg|_{L \neq I} \right) (P_{GC} - P_{GD} - P_{BC} + P_{BD}) \right]$$

$$+ \left( \frac{1}{K} g_{L,i}^{I,I} + \frac{K - 1}{K} g_{L,i}^{L,I} \bigg|_{L \neq I} \right) (P_{GD} + P_{BC} - 2P_{BD}) + P_{BD} \right].$$

As $K$ approaches infinity, the contribution of $g_i^{I,I}$ to the total average reputation of $i$, $\bar{g}_i$, tends to zero, so that the entirety of $\bar{g}_i$ is due to the $g_i^{L,I} \bigg|_{L \neq I}$ terms. We thus have that

$$\lim_{K \to \infty} \bar{g}_Z = \sum_{i} f_i \bar{g}_i^2 (P_{GC} - P_{GD} - P_{BC} + P_{BD}) + 2\bar{g}(P_{GD} + P_{BD} - 2P_{BD}) + P_{BD}.$$
Defining

\[ g_2 = \sum_i f_i \bar{g}_i^2, \]
\[ d_2 = \sum_i f_i \bar{g}_i (1 - \bar{g}_i) = g - g_2, \]
\[ b_2 = \sum_i f_i (1 - \bar{g}_i)^2 = 1 - 2g + g_2 \]

allows us to rewrite this as

\[
\lim_{K \to \infty} \bar{g}_Z = g_2 P_{GC} + d_2 (P_{GD} + P_{BC}) + b_2 P_{BD}.
\]

This is the bottom term of equation 5 from Radzvilavicius et al. (2019) with empathy parameter \( E = 0 \). This result confirms that, when the number of groups goes to infinity, our model with separate groups is identical to everyone in the population following independent or private reputation assessment. In this limit, the reputation in a population of discriminators is given by a solution to

\[
0 = g^2 (P_{GC} - P_{GD} - P_{BC} + P_{BD}) + g (P_{GD} + P_{BC} - 2P_{BD} - 1) + P_{BD}
\]

\[ \therefore g = \begin{cases} 
\frac{1 - \sqrt{1 - 4(\epsilon - u_a)u_a}}{2(\epsilon - u_a)}, & \text{shunning,} \\
\frac{1}{2}, & \text{stern judging,} \\
\frac{u_a}{1 - \epsilon + u_a}, & \text{scoring,} \\
\frac{1 - u_a - \sqrt{(1 - u_a)(1 - \epsilon)}}{\epsilon - u_a}, & \text{simple standing.}
\end{cases} \]

We have picked out the solutions that are viable for \( 1 > u_x > 0 \) and \( 1 > u_a > 0 \).

3.2. **One large group, other smaller groups of equal size.** Without loss of generality, suppose that one group has size \( \nu \) and the remaining \( K - 1 \) groups each have size \( (1 - \nu)/(K - 1) \). Let group 1 be the largest group.
Starting with equation 7, we can unpack what $\bar{g}_Z$ looks like. We have

$$\bar{g}_Z = \sum_{i} \sum_{j} \nu_i \nu_j g_{Z}^{i,j}$$

$$= \sum_{i} \sum_{j} \nu_i \nu_j \left( \delta_{i,j} \left[ g_{\star} \cdot_{j} P_{GC} + (1 - g_{\star} \cdot_{j}) P_{BD} \right] + (1 - \delta_{i,j}) \left[ g_{\alpha}^{l,j} P_{GC} + g_{\beta}^{l,j} P_{GD} + g_{\gamma}^{l,j} P_{BC} + g_{\delta}^{l,j} P_{BD} \right] \right),$$

$$= \nu^2 \left[ g_{\star}^{1,1} P_{GC} + (1 - g_{\star}^{1,1}) P_{BD} \right] + \nu \frac{1 - \nu}{K - 1} \sum_{i=2}^{K} \left[ g_{\alpha}^{l,1} P_{GC} + g_{\beta}^{l,1} P_{GD} + g_{\gamma}^{l,1} P_{BC} + g_{\delta}^{l,1} P_{BD} \right]$$

$$+ \nu \frac{1 - \nu}{K - 1} \sum_{j=2}^{K} \left[ g_{\alpha}^{1,l,j} P_{GC} + g_{\beta}^{1,l,j} P_{GD} + g_{\gamma}^{1,l,j} P_{BC} + g_{\delta}^{1,l,j} P_{BD} \right]$$

$$+ \left( \frac{1 - \nu}{K - 1} \right)^2 \left( \sum_{j=2}^{K} \left[ g_{\star}^{1,l,j} P_{GC} + (1 - g_{\star}^{1,l,j}) P_{BD} \right] + \sum_{i=2}^{K} \sum_{j \neq i}^{K} \left[ g_{\alpha}^{1,l,i} P_{GC} + g_{\beta}^{1,l,i} P_{GD} + g_{\gamma}^{1,l,i} P_{BC} + g_{\delta}^{1,l,i} P_{BD} \right] \right)$$

$$= \nu^2 \left[ g_{\star}^{1,1} P_{GC} + (1 - g_{\star}^{1,1}) P_{BD} \right] + \nu(1 - \nu) \left[ g_{\alpha}^{l,1} P_{GC} + g_{\beta}^{l,1} P_{GD} + g_{\gamma}^{l,1} P_{BC} + g_{\delta}^{l,1} P_{BD} \right]_{I \neq 1}$$

$$+ \nu(1 - \nu) \left[ g_{\alpha}^{1,l,j} P_{GC} + g_{\beta}^{1,l,j} P_{GD} + g_{\gamma}^{1,l,j} P_{BC} + g_{\delta}^{1,l,j} P_{BD} \right]_{j \neq 1}$$

$$+ \left( \frac{1 - \nu}{K - 1} \right)^2 \left[ g_{\alpha}^{1,l,i} P_{GC} + g_{\beta}^{1,l,i} P_{GD} + g_{\gamma}^{1,l,i} P_{BC} + g_{\delta}^{1,l,i} P_{BD} \right]_{I \neq j \neq 1}.$$

As $K \to \infty$, the diagonal elements of the last term drop out. What remains is a special case of equation 12 of Radzvilavicius et al. (2021), in which part of the population consists of adherents to a single institution of reputation assessment and the remainder consists of private assessors.

4. Dynamics of Multiple Groups with Disjoint Copying

When strategies cannot be copied between groups, we independently track the fitnesses of types within each group. We zero in on the case of two groups:

$$\Pi^1_X = (1 - u_x) \left[ b \left( \nu_1 [ f_X^1 + f_{ZG}^{1,1} ] + \nu_2 [ f_X^2 + f_{ZG}^{1,2} ] \right) \right] - c$$

$$\Pi^1_Y = (1 - u_x) \left[ b \left( \nu_1 [ f_Y^1 + f_{ZG}^{1,1} ] + \nu_2 [ f_Y^2 + f_{ZG}^{1,2} ] \right) \right]$$

$$\Pi^1_Z = (1 - u_x) \left[ b \left( \nu_1 [ f_Z^1 + f_{ZG}^{1,1} ] + \nu_2 [ f_Z^2 + f_{ZG}^{1,2} ] \right) \right] - c g_{\star}^{1,1}$$

$$\Pi^2_X = (1 - u_x) \left[ b \left( \nu_1 [ f_X^1 + f_{ZG}^{2,1} ] + \nu_2 [ f_X^2 + f_{ZG}^{2,2} ] \right) \right] - c$$

$$\Pi^2_Y = (1 - u_x) \left[ b \left( \nu_1 [ f_Y^1 + f_{ZG}^{2,1} ] + \nu_2 [ f_Y^2 + f_{ZG}^{2,2} ] \right) \right]$$

$$\Pi^2_Z = (1 - u_x) \left[ b \left( \nu_1 [ f_Z^1 + f_{ZG}^{2,1} ] + \nu_2 [ f_Z^2 + f_{ZG}^{2,2} ] \right) \right] - c g_{\star}^{2,2}.$$

For any social norm, with a single group and public reputations, there are two stable equilibria, consisting respectively of DISC and ALLD. We consider here what happens to these equilibria in one group when the other is fixed either for DISC or for ALLD. When these are the only two types
present, their fitnesses are given by

\[
\Pi_1^Y = (1 - u_x) \left[ b(\nu_1 f_2^1 g_Y^{1,1} + \nu_2 f_2^2 g_Y^{1,2}) \right]
\]
\[
\Pi_1^Z = (1 - u_x) \left[ b(\nu_1 f_2^1 g_Z^{1,1} + \nu_2 f_2^2 g_Z^{1,2}) - cg^{*1} \right]
\]
\[
\Pi_2^Y = (1 - u_x) \left[ b(\nu_1 f_2^1 g_Y^{2,1} + \nu_2 f_2^2 g_Y^{2,2}) \right]
\]
\[
\Pi_2^Z = (1 - u_x) \left[ b(\nu_1 f_2^1 g_Z^{2,1} + \nu_2 f_2^2 g_Z^{2,2}) - cg^{*2} \right].
\]

4.0.1. Both groups fixed for ALLD. Let \( f = f_Z^2 \) be the frequency of \( Z \) in group 2. For DISC to invade ALLD in group 2, we would require

\[
(\partial f \dot{f})|_{f=0} > 0
\]
\[
\Pi_2^Z|_{f^1_Y=1, f^2_Y=1} > \Pi_2^Y|_{f^1_Y=1, f^2_Y=1} - cg^{*2} > 0.
\]

Since \( c \) and \( g^{*2} \) are both positive numbers, \( Z \) cannot invade.

4.0.2. Both groups fixed for DISC. Assume now that both groups are fixed for DISC. For ALLD to invade DISC in group 2, we would require

\[
\Pi_2^2|_{f^2_Z=1, f^2_Z=1} > \Pi_2^2|_{f^2_Z=1, f^2_Z=1} - cg^{*2} > 0
\]
\[
b \left( \nu_1 g_Y^{2,1} + \nu_2 g_Y^{2,2} \right)|_{f^2_Y=1, f^2_Y=1} > b \left( \nu_1 g_Z^{2,1} + \nu_2 g_Z^{2,2} \right)|_{f^2_Y=1, f^2_Y=1} - cg^{*2}|_{f^2_Y=1, f^2_Y=1}
\]
\[
b \frac{c}{g^{*2}} < \frac{\nu_1 (g_Z^{2,1} - g_Y^{2,1}) + \nu_2 (g_Z^{2,2} - g_Y^{2,2})}{g_Y^{2,1} \nu_1 (P_{GC} - P_{GD} - P_{BC} + P_{BD}) + g^{*2} \nu_1 (P_{BC} - P_{BD}) + \nu_2 (P_{GC} - P_{GD})}
\]

This is less stringent than the standard condition \( b/c < 1/(P_{GC} - P_{GD}) = 1/(\epsilon - u_a) \) for one group. Discriminators in group 1 contribute much more weakly to the fitness of discriminators in group 2 and thus offer limited protection against invasion by defectors.

4.0.3. One group fixed for ALLD, other for DISC. Suppose that group 1 is fixed for DISC and 2 is fixed for ALLD. We now investigate whether ALLD can invade 1 and DISC can invade 2, respectively.
In the first case, let $f$ be the frequency of ALLD in group 1 (that is, $f_1^\downarrow$). ALLD can invade $1$ provided

\[
(\partial_f \hat{f})|_{f=0} > 0
\]

\[
\therefore (\partial_f[f(\Pi_Y^1 - \Pi^1)])|_{f=0} > 0
\]

\[
\therefore (\partial_f[(f - f^2)\Pi_Y^1 - (f - f^2)^2\Pi_Y^2])|_{f=0} > 0
\]

\[
\therefore (\partial_f[(f - f^2)\Pi_Y^1 - (f - f^2)^2\Pi_Y^2])|_{f=0} > 0
\]

\[
\therefore ([1 - 2f]\Pi_Y^1 - \Pi_Y^2)|_{f=0} > 0
\]

\[
\therefore \Pi_Y^1|_{f=0} > \Pi_Y^2|_{f=0}
\]

\[
\therefore b(\nu_1 f_Z^1 g_Y^1 + \nu_2 f_Z^2 g_Y^2)|_{f_Z^1=0, f_Z^2=1} > [b(\nu_1 f_Z^1 g_Y^1 + \nu_2 f_Z^2 g_Y^2)|_{f_Z^1=0, f_Z^2=1} - c g^{*1}]|_{f_Z^1=0, f_Z^2=1}
\]

\[
\therefore b\nu_1 (g_Y^1 - g_Y^2)|_{f_Z^1=0, f_Z^2=1} < c g^{*1}
\]

\[
\therefore \frac{b}{c} < g^{*1} / \nu_1 (g_Y^1 - g_Y^2)
\]

\[
\therefore \frac{b}{c} < \nu_1(\nu_1^1 P_G + (1 - g^{*1})P_B - g^{*1} P_G - (1 - g^{*1})P_B)
\]

\[
\therefore \frac{b}{c} < \nu_1 g^{*1}(P_B - P_G)
\]

\[
\therefore \frac{b}{c} < \nu_1(\frac{1}{P_B} - \frac{1}{P_G}) = \frac{1}{\nu_1(\epsilon - u_a)}
\]

Letting $\nu_1 \to 1$ allows us to recover the one-group condition, $b/c < 1/(P_B - P_G) = 1/(\epsilon - u_a)$. Since $\nu_1 < 1$, this is generally less strict than the one-group condition: the fact that the second group consists entirely of defectors makes it more difficult for the first group to resist invasion by defectors.

We now consider the second case, i.e., whether DISC can invade 2, which is fixed for ALLD. Let $f$ now be the frequency of DISC in group 2 (that is, $f_Z^2$). DISC being able to invade requires

\[
(\partial_f \hat{f})|_{f=0} > 0
\]

\[
\therefore \Pi_Y^2|_{f=0} > \Pi_Y^2|_{f=0}
\]

\[
\therefore b \left(\nu_1 f_Z^1 g_Y^1 + \nu_2 f_Z^2 g_Y^2 - c g^{*2}\right)|_{f_Z^1=1, f_Z^2=0} > b \left(\nu_1 f_Z^1 g_Y^1 + \nu_2 f_Z^2 g_Y^2\right)|_{f_Z^1=1, f_Z^2=0}
\]

\[
\therefore b\nu_1 (g_Y^1 - g_Y^2)|_{f_Z^1=1, f_Z^2=0} < c g^{*1}
\]

\[
\therefore \frac{b}{c} < g^{*1} / \nu_1 (g_Y^1 - g_Y^2)
\]

\[
\therefore \frac{b}{c} < \nu_1 \left[ \frac{1}{g_0^1(\epsilon - u_a)} (P_B - P_G - P_B + P_G) + g^{*2}(P_B - P_G) \right]
\]

This is distinct from the single-group case, in which DISC ($Z$) can never invade ALLD ($Y$) (which corresponds to $\nu_1 \to 0$, blowing up the denominator). In this scenario, discriminators in group 1 can help discriminators in group 2 rise in frequency, even though they are not guaranteed to have good opinions of discriminators in group 2.
5. Derivation of Replicator Equation Under Different Copying Models

Here we explicitly derive the replicator dynamics for various group-wise strategy copying scenarios.

5.1. One group. Consider first the case of a single group. We have the following events to take into account:

(1) Increase. A type \( j (\neq i) \) individual is chosen to update with probability \( f_j \). With probability \( f_i \), the compared individual is type \( i \). The update happens with probability \( \phi(\Pi_j, \Pi_i) = 1/(1 + \exp[w(\Pi_j - \Pi_i)]) \). The frequency of type \( i \) increases by \( 1/N \).

(2) Decrease. A type \( i \) individual is chosen to update with probability \( f_i \). With probability \( f_j \), the compared individual is type \( j (\neq i) \). The update happens with probability \( \phi(\Pi_i, \Pi_j) \). The frequency of type \( i \) decreases by \( 1/N \).

Thus,

\[
\mathbb{E}[\Delta f_i] = \frac{1}{N} f_i \sum_j \left( f_j f_i \phi(\Pi_j, \Pi_i) - f_i \sum_j f_j \phi(\Pi_i, \Pi_j) \right) = \frac{1}{N} f_i \left( \sum_j f_j \left[ \phi(\Pi_j, \Pi_i) - \phi(\Pi_i, \Pi_j) \right] \right).
\]

Note that

\[
\phi(\Pi_j, \Pi_i) = \frac{1}{1 + \exp[w(\Pi_j - \Pi_i)]} \approx \frac{1}{1 + \exp[w(\Pi_j - \Pi_i)]} \bigg|_{w=0} + w \left( \frac{d}{dw} \frac{1}{1 + \exp[w(\Pi_j - \Pi_i)]} \right) \bigg|_{w=0} + \mathcal{O}(w^2)
\]

\[
= \frac{1}{2} + w \frac{\Pi_i - \Pi_j}{4} + \mathcal{O}(w^2).
\]

Hence

\[
\phi(\Pi_j, \Pi_i) - \phi(\Pi_i, \Pi_j) \approx w \frac{\Pi_i - \Pi_j}{2} + \mathcal{O}(w^2).
\]

We therefore have

\[
\mathbb{E}[\Delta f_i] = \frac{1}{N} f_i \sum_j \left[ \frac{w}{2} f_j (\Pi_i - \Pi_j) + \mathcal{O}(w^2) \right]
\]

\[
\approx \frac{w}{2N} f_i \sum_j f_j (\Pi_i - \Pi_j)
\]

\[
= \frac{w}{2N} f_i (\Pi_i \sum_j f_j - \sum_j f_j \Pi_j)
\]

\[
= \frac{w}{2N} f_i (\Pi_i - \bar{\Pi}).
\]

This is what ultimately justifies the use of the replicator equation under pairwise comparison. Rescaling time so that, on average, one update event occurs per time step yields

\[
\dot{f}_i = f_i (\Pi_i - \bar{\Pi}).
\]
5.2. **Multiple groups, copying only from one’s in-group (“disjoint copying”).** When there is more than one group ($K > 1$), the preceding analysis holds, except that we must specify that an individual with strategy $i$ in group $I$ can only copy from another individual in group $I$ (their in-group). We thus obtain

$$j_i^t = f_i(\Pi_i^t - \Pi^t), \text{with}$$

$$\Pi^t = \sum_i f_i^t \Pi_i^t.$$  

(9)

5.3. **Multiple groups with well-mixed strategic copying.** We now derive the analogous case for multiple groups ($K > 1$) with “well-mixed copying”, i.e., individuals do not distinguish between their in-group and out-group when deciding whom to compare their fitness against and potentially imitate. Let $\nu_I$ be the frequency of group $I$, and let $n_I^t = N\nu_I f_i^t$ be the absolute number of individuals of type $I$ following strategy $i$. The following events may occur.

1. **Increase.** A type $j$ individual in group $I$ is chosen to update with probability $\nu_I f_j^t$. With probability $\nu_I f_j^t$, the compared individual is type $i$, $J$, with $J \in \{1 \ldots K\}$ (i.e., $J$ can take on the same value as $I$). The update happens with probability $\phi(\Pi_j^t, \Pi_i^t)$. $n_i^t$ increases by 1.

2. **Decrease.** A type $i$ individual in group $I$ is chosen to update with probability $\nu_I f_i^t$. With probability $\nu_I f_i^t$, the compared individual is type $j (\neq i)$, $J$, with $J \in \{1 \ldots K\}$ (i.e., $J$ can take on the same value as $I$). The update happens with probability $\phi(\Pi_i^t, \Pi_j^t)$. $n_i^t$ decreases by 1.

Thus

$$\mathbb{E}[\Delta n_i^t] = \mathbb{P}(\Delta n_i^t = 1) - \mathbb{P}(\Delta n_i^t = -1)$$

$$= \nu_I \sum_j f_j^t \sum_j \nu_I f_j^t \phi(\Pi_j^t, \Pi_i^t) - \nu_I f_i^t \sum_j \nu_I f_j^t \phi(\Pi_i^t, \Pi_j^t)$$

$$\approx \nu_I \sum_j f_j^t \sum_j \nu_I f_j^t \left(\frac{1}{2} + w \frac{\Pi_i^t - \Pi_j^t}{4}\right) - \nu_I f_i^t \sum_j \nu_I f_j^t \left(\frac{1}{2} + w \frac{\Pi_j^t - \Pi_i^t}{4}\right)$$

$$= \nu_I \frac{1}{2} \left[\sum_j f_j^t \sum_j \nu_I f_j^t - f_i^t \sum_j \nu_I f_j^t\right]$$

$$+ \nu_I \frac{w}{4} \sum_j f_j^t \left[\sum_j \nu_I f_j^t (\Pi_i^t - \Pi_j^t) - f_i^t \sum_j \nu_I f_j^t (\Pi_j^t - \Pi_i^t)\right]$$

$$= \nu_I \frac{1}{2} \left[\sum_j \nu_I \left( f_j^t \sum_j f_j^t - f_i^t \sum_j f_j^t\right)\right]$$

$$+ \nu_I \frac{w}{4} \left[\sum_j \nu_I \left( f_j^t f_j^t (\Pi_i^t - \Pi_j^t) - f_i^t f_j^t (\Pi_j^t - \Pi_i^t)\right)\right].$$

$$= \nu_I \frac{1}{2} \left[\sum_j \nu_I \left( f_j^t - f_i^t\right)\right]$$

$$+ \nu_I \frac{w}{4} \left[\sum_j \nu_I \left( f_j^t (\Pi_i^t - \sum_j f_j^t \Pi_j^t) + f_i^t (\Pi_i^t - \sum_j f_j^t \Pi_j^t)\right)\right].$$
Recalling that \( n_I^t = N \nu_I f_I^t \), and dropping the 1/2 prefactor, we have

\[
\dot{f}_I^t \propto \sum_J \nu_J \left[ f_J^t (\Pi_J^t - \sum_J f_J^t \Pi_J^t) + f_I^t (\Pi_I^t - \sum_J f_J^t \Pi_J^t) \right].
\] (10)

The proportionality constant will depend on how we rescale time. Note that the first term does not have a \( w \) prefactor and roughly corresponds to “neutral” mixing between the two groups. This means that that term will dominate, and thus we expect \( f_I^t \) to equilibrate rapidly to a value common to all \( I \). If we mandate this, the only dynamical quantity becomes \( f_i = \sum_I \nu_I f_i^t \), so we have

\[
\dot{f}_i = \sum_I \nu_I f_i^t
\]

\[
\propto \sum_I \nu_I \left[ \sum_J \nu_J \left( f_J^t (\Pi_J^t - \sum_J f_J^t \Pi_J^t) + f_I^t (\Pi_I^t - \sum_J f_J^t \Pi_J^t) \right) \right]
\]

\[
= \sum_J \nu_J f_i \Pi_J^t - f_i \sum_I \nu_I \sum_J f_J \Pi_J^t + \sum_I \nu_I f_i \Pi_I^t - f_i \sum_J \nu_J \sum_J f_J \Pi_J^t
\]

\[
= f_i \left( \sum_J \nu_J \Pi_J^t + \sum_I \nu_I \Pi_I^t - \sum_J f_J \sum_J \nu_J \Pi_J^t - \sum_J f_J \sum_J \nu_J \Pi_J^t \right)
\]

\[
\propto f_i \sum_J \nu_J (\Pi_J^t - \Pi^t).
\]

Rescaling time allows us to write this as an equality:

\[
\dot{f}_i = f_i \sum_J \nu_J (\Pi_J^t - \Pi^t)
\]

\[
= f_i \left( \sum_J \nu_J \Pi_J^t - \Pi^t \right) = f_i (\Pi_i - \bar{\Pi}), \text{ with}
\]

\[
\Pi_i = \sum_L \nu_L \Pi_L^t,
\]

\[
\bar{\Pi} = \sum_L \nu_L \sum_i f_i \Pi_i^t = \sum_L \nu_L \bar{\Pi}^t = \sum_i f_i \Pi_i.
\] (11)

5.4. Multiple groups, in-group favored (“partially-mixed copying”). We have seen that if individuals freely copy across group lines, strategy frequencies change much faster due to mixing than due to selection. We now consider the possibility of partially, but not completely, restricting partner choice for strategy imitation. Let \( m \) (for “imitation” or, equivalently, for “mixing”) be the weight that an individual assigns to the opposite group when deciding whom to imitate, so that \( m = 0 \) corresponds to no mixing and \( m = 1 \) corresponds to full mixing. An individual in group \( I \) thus chooses an individual in their own group with probability \( 1 - m \) and chooses an individual in a random group \( J \) (which could be \( I \)) with probability \( \nu_J m \). For \( n_I^t \), the following events are possible.

1. **Increase**. A type \( j \) individual in group \( I \) is chosen to update with probability \( \nu_j f_j^t \). With probability \( (1 - m) f_I^t \), the compared individual is type \( i, I \), and with probability \( \nu_J m f_j^t \), the compared individual is type \( i, J \) (\( J \) can be \( I \)). The update happens with probability \( \phi(\Pi_J^t, \Pi_I^t) \) (for \( i, I \)) or \( \phi(\Pi_I^t, \Pi_J^t) \) (for \( i, J \)). In either case, \( n_I^t \) increases by 1.

2. **Decrease**. A type \( i \) individual in group \( I \) is chosen to update with probability \( \nu_i f_i^t \). With probability \( (1 - m) f_j^t \), the compared individual is type \( j(\neq i), I \), and with probability
\( \nu_{jm} f_j^I \), the compared individual is type \( j(\neq i) \), \( J \) (\( J \) can be \( I \)). The update happens with probability \( \phi(\Pi_i^I, \Pi_j^J) \) (for \( j, I \)) or \( \phi(\Pi_i^I, \Pi_j^I) \) (for \( j, J \)). In either case, \( n_i^I \) decreases by 1.

We thus have

\[
\mathbb{E} [\Delta n_i^I] = P(\Delta n_i^I = 1) - P(\Delta n_i^I = -1)
\]

\[
= (1 - m) \left[ \nu_I \sum_j f_j^I f_i^I \phi(\Pi_j^I, \Pi_i^I) - \nu_I \sum_j f_j^I \phi(\Pi_i^I, \Pi_j^I) \right]
\]

\[
\quad + m \left[ \nu_I \sum_j f_j^I \nu_J f_i^I \phi(\Pi_j^I, \Pi_i^I) - \nu_I \sum_j \nu_J f_j^I \phi(\Pi_i^I, \Pi_j^I) \right]
\]

\[
\approx (1 - m) \left[ \nu_I \sum_j f_j^I f_i^I \left( \frac{1}{2} + w \frac{\Pi_j^I - \Pi_i^I}{4} \right) - \nu_I f_i^I \sum_j f_j^I \left( \frac{1}{2} + w \frac{\Pi_j^I - \Pi_i^I}{4} \right) \right]
\]

\[
\quad + m \left[ \nu_I \sum_j f_j^I \nu_J f_i^I \left( \frac{1}{2} + w \frac{\Pi_j^I - \Pi_i^I}{4} \right) - \nu_I f_i^I \sum_j \nu_J f_j^I \left( \frac{1}{2} + w \frac{\Pi_j^I - \Pi_i^I}{4} \right) \right]
\]

\[
= m \nu_I \frac{1}{2} \left[ \sum_j f_j^I \nu_J f_i^I - f_i^I \sum_j \nu_J f_j^I \right]
\]

\[
\quad + (1 - m) \nu_I \frac{w}{2} \sum_j f_j^I f_i^I (\Pi_j^I - \Pi_j^I)
\]

\[
\quad + m \nu_I \frac{w}{4} \sum_j \nu_J f_i^I (\Pi_j^I - \Pi_j^I) - f_i^I \sum_j \nu_J f_j^I (\Pi_j^I - \Pi_j^I)
\]

\[
= m \nu_I \frac{1}{2} \left[ \sum_j \nu_J \left( f_j^I \sum_j f_j^I - f_i^I \sum_j f_j^I \right) \right]
\]

\[
\quad + (1 - m) \nu_I \frac{w}{2} f_i^I \left( \Pi_i^I - \sum_j f_j^I \Pi_j^I \right)
\]

\[
\quad + m \nu_I \frac{w}{4} \left[ \sum_j \nu_J \sum_j (f_j^I f_i^I (\Pi_i^I - \Pi_j^I) - f_i^I f_j^I (\Pi_j^I - \Pi_i^I)) \right].
\]

\[
= m \nu_I \frac{1}{2} \left[ \sum_j \nu_J \left( f_j^I - f_i^I \right) \right]
\]

\[
\quad + (1 - m) \nu_I \frac{w}{2} f_i^I \left( \Pi_i^I - \Pi_i^I \right)
\]

\[
\quad + m \nu_I \frac{w}{4} \left[ \sum_j \nu_J \left( f_i^I (\Pi_i^I - \Pi_i^I) + f_i^I (\Pi_i^I - \Pi_i^I) \right) \right].
\]

(12)
Recalling that \( n_i^I = N \nu_i f_i^I \), the replicator dynamics are given by
\[
\dot{f}_i^I \propto m \frac{1}{2} \left[ \sum_J \nu_J (f_J^I - f_i^I) \right] \\
+ (1 - m) \frac{w}{2} f_i^I (\Pi_i^I - \bar{\Pi}_i^I) \\
+ m \frac{w}{4} \left[ \sum_J \nu_J (f_J^I (\Pi_i^J - \bar{\Pi}_i^J) + f_i^I (\Pi_i^I - \bar{\Pi}_i^J)) \right] \\
\propto m \frac{w}{w} \left[ \sum_J \nu_J (f_J^I - f_i^I) \right] \\
+ (1 - m) f_i^I (\Pi_i^I - \bar{\Pi}_i^I) \\
+ \frac{m}{2} \left[ \sum_J \nu_J (f_J^I (\Pi_i^J - \bar{\Pi}_i^J) + f_i^I (\Pi_i^I - \bar{\Pi}_i^J)) \right]
\]

As usual, the \( \propto \) can be converted into = by rescaling time. In each equation, the first term (proportional to \( m/w \)) sets the rate of between-group “neutral” mixing, the second corresponds to within-group selection, and the third corresponds to between-group selection. Note that setting \( m = 0 \) yields equation 9 and setting \( m = 1 \) yields equation 11, subject to rescaling.

**References**

Radzvilavicius, A. L., Kessinger, T. A., and Plotkin, J. B. (2021). Adherence to public institutions that foster cooperation. *Nature Communications*, 12(1):3567.

Radzvilavicius, A. L., Stewart, A. J., and Plotkin, J. B. (2019). Evolution of empathetic moral evaluation. *eLife*, 8:e44269.