PNAS

Supporting Information for

Evolution of norms for judging social behavior

Taylor A. Kessinger, Corina E. Tarnita, and Joshua B. Plotkin

Joshua B. Plotkin
E-mail: jplotkin@sas.upenn.edu

This PDF file includes:

Supporting text
Figs. S1 to S8
Tables S1 to S6
SI References

Contents

Supporting Information Text

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,
$$P^{GC} = (1 - u_x)(1 - u_a) + u_x u_a = \epsilon.$$

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
$$\begin{aligned} P^{BC} &= (1 - u_x)[p(1 - u_a) + (1 - p)u_a] + u_x[q(1 - u_a) + (1 - q)u_a] \\ &= (1 - u_x)[p - 2pu_a + u_a] + u_x[q - 2qu_a + u_a] \\ &= p - 2pu_a + u_a - pu_x + 2pu_x u_a - u_x u_a + qu_x - 2qu_x u_a + u_x u_a \\ &= p(1 - 2u_a - u_x + 2u_a u_x) + q(u_x - 2u_x u_a) + u_a \\ &= p(\epsilon - u_a) + q(1 - \epsilon - u_a) + u_a. \end{aligned} \quad [1]$$

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

$$P^{BD} = q(1-u_a) + (1-q)u_a = q(1-2u_a) + u_a. \qquad [2]$$

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

1. when $(p,q) = (0,1)$ (*Stern Judging*), Eq. [1] becomes $1-\epsilon$ and Eq. [2] becomes $1-u_a$.
2. when $(p,q) = (1,1)$ (*Simple Standing*), Eq. [1] becomes $1-u_a$ and Eq. [2] becomes $1-u_a$.
3. when $(p,q) = (1,0)$ (*Scoring*), Eq. [1] becomes ϵ and Eq. [2] becomes u_a.
4. when $(p,q) = (0,0)$ (*Shunning*), Eq. [1] becomes u_a and Eq. [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 S1.

observer view of recipient	good	good	bad	bad
donor intent	cooperate	defect	cooperate	defect
good reputation probability	P^{GC}	P^{GD}	P^{BC}	P^{BD}
general expression	ϵ	u_a	$p(\epsilon-u_a)+q(1-\epsilon-u_a)+u_a$	$q(1-2u_a)+u_a$
Shunning $(p=0, q=0)$	ϵ	u_a	u_a	u_a
Stern Judging $(p=0, q=1)$	ϵ	u_a	$1-\epsilon$	$1-u_a$
Scoring $(p=1, q=0)$	ϵ	u_a	ϵ	u_a
Simple Standing $(p=1, q=1)$	ϵ	u_a	$1-u_a$	$1-u_a$

Table S1. 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. Average reputations.

Before continuing, we define several types of “average” reputations in terms of the strategy- and group-specific reputations $g^s_{I,J}$. We begin with

$$g_{I,J} = \sum_s f^s_I g^s_{I,J},$$

which is the average reputation of group I in the eyes of group J, or equivalently the probability that group J considers a randomly chosen individual in group I to be good. We also define

$$g_{\bullet,J} = \sum_L \nu_L g_{L,J},$$

which is group J’s average view of the entire population, or the probability that group J considers a randomly chosen individual in the whole population to be good. We continue by defining averages over the population’s group structure to obtain the average reputation of individuals following strategy $s \in \{X, Y, Z\}$:

$$g^s = \sum_I \sum_J \nu_I \nu_J g^s_{I,J}. \qquad [3]$$

And we finally define the “grand” population-wide average reputation

$$g = \sum_I \sum_J \nu_I \nu_J \sum_s f^s_I g^s_{I,J} = \sum_I \sum_J \nu_I \nu_J g_{I,J}, \qquad [4]$$

which is also the probability that one randomly chosen individual considers another randomly chosen individual to be good.

1.2. 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 strategic type 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 $g_{\bullet,J}$); the I member intends to cooperate, which yields a good reputation with probability P^{GC}.

2. with someone group J sees as bad (probability $1 - g_{\bullet,J}$); the I member intends to cooperate, which yields a good reputation with probability P^{BC}.

We thus have

$$g_{I,J}^{X} = g_{\bullet,J}P^{GC} + (1 - g_{\bullet,J})P^{BC} = g_{\bullet,J}(P^{GC} - P^{BC}) + P^{BC}.$$

Similar reasoning for defectors yields

$$g_{I,J}^{Y} = g_{\bullet,J}P^{GD} + (1 - g_{\bullet,J})P^{BD} = g_{\bullet,J}(P^{GD} - P^{BD}) + P^{BD}.$$

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 $g_{\bullet,I}$); the I discriminator intends to cooperate, which yields a good reputation with probability P^{GC}.

2. with someone group I sees as bad (probability $1 - g_{\bullet,I}$); the I discriminator intends to defect, 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 s (probability $\nu_L f_L^s$) whom I sees as good (probability $g_{L,I}^s$) and whom J also sees as good (probability $g_{L,J}^s$); the I discriminator intends to cooperate, which yields a good reputation with probability P^{GC}.

2. with someone in an arbitrary group L following strategy s (probability $\nu_L f_L^s$) whom I sees as bad (probability $1 - g_{L,I}^s$) but whom J sees as good (probability $g_{L,J}^s$); the I discriminator intends to defect, which yields a good reputation with probability P^{GD}.

3. with someone in an arbitrary group L following strategy s (probability $\nu_L f_L^s$) whom I sees as good (probability $g_{L,I}^s$) but whom J sees as bad (probability $1 - g_{L,J}^s$); the I discriminator intends to cooperate, which yields a good reputation with probability P^{BC}.

4. with someone in an arbitrary group L following strategy s (probability $\nu_L f_L^s$) whom I sees as bad (probability $1 - g_{L,I}^s$) and whom J also sees as bad (probability $1 - g_{L,J}^s$); the I discriminator intends to defect, which yields a good reputation with probability P^{BD}.

Defining

$$G_{I,J} = \sum_L \nu_L \sum_s f_L^s g_{L,I}^s g_{L,J}^s,$$

we can sum over all groups and strategy combinations to obtain

$$\sum_L \nu_L \sum_s f_L^s g_{L,I}^s g_{L,J}^s = G_{I,J},$$

$$\sum_L \nu_L \sum_s f_L^s (1 - g_{L,I}^s) g_{L,J}^s = g_{\bullet,J} - G_{I,J},$$

$$\sum_L \nu_L \sum_s f_L^s g_{L,I}^s (1 - g_{L,J}^s) = g_{\bullet,I} - G_{I,J},$$

$$\sum_L \nu_L \sum_s f_L^s (1 - g_{L,I}^s)(1 - g_{L,J}^s) = 1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}.$$

Thus,

$$\begin{aligned}
g^Z_{I,J} &= \delta_{I,J}\big[g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})P^{BD}\big] \\
&\quad + (1-\delta_{I,J})\big[G_{I,J}P^{GC} + (g_{\bullet,J} - G_{I,J})P^{GD} + (g_{\bullet,I} - G_{I,J})P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J})P^{BD}\big] \\
&= \delta_{I,J}\big[g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})P^{BD}\big] \\
&\quad + (1-\delta_{I,J})\big[G_{I,J}(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g_{\bullet,J}(P^{GD} - P^{BD}) + g_{\bullet,I}(P^{BC} - P^{BD}) + P^{BD}\big].
\end{aligned}$$

1.3. Special case: *Scoring*. Under *Scoring* ($p = 1, q = 0$), we have

$$\begin{aligned}
P^{GC} &= P^{BC} = \epsilon, \\
P^{GD} &= P^{BD} = u_a.
\end{aligned}$$

In this case, the $I \neq J$ term of $g^Z_{I,J}$ becomes

$$\begin{aligned}
&G_{I,J}(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g_{\bullet,J}(P^{GD} - P^{BD}) + g_{\bullet,I}(P^{BC} - P^{BD}) + P^{BD} \\
&= g_{\bullet,I}(P^{BC} - P^{BD}) + P^{BD} \\
&= g_{\bullet,I}\epsilon + (1 - g_{\bullet,I})u_a.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
g^X_{I,J} &= g_{\bullet,J}\epsilon + (1-g_{\bullet,J})\epsilon = \epsilon, \\
g^Y_{I,J} &= g_{\bullet,J}u_a + (1-g_{\bullet,J})u_a = u_a, \\
g^Z_{I,J} &= \delta_{I,J}\big[g_{\bullet,J}\epsilon + (1-g_{\bullet,J})u_a\big] + (1-\delta_{I,J})\big[g_{\bullet,I}\epsilon + (1-g_{\bullet,I})u_a\big] \\
&= \delta_{I,J}\big[g_{\bullet,I}\epsilon + (1-g_{\bullet,I})u_a\big] + (1-\delta_{I,J})\big[g_{\bullet,I}\epsilon + (1-g_{\bullet,I})u_a\big] \\
&= g_{\bullet,I}\epsilon + (1-g_{\bullet,I})u_a.
\end{aligned}$$

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

$$\begin{aligned}
g_{\bullet,I} &= \sum_L \nu_L g_{L,I} = \sum_L \nu_L \sum_s f^s_L g^s_{L,I} \\
&= \sum_L \nu_L\big(f^X_L\epsilon + f^Y_L u_a + f^Z_L[g_{\bullet,I}\epsilon + (1 - g_{\bullet,I}u_a]\big) \\
&= \epsilon\sum_L \nu_L f^X_L + u_a\sum_L \nu_L f^Y_L + g_{\bullet,I}\epsilon\sum_L \nu_L f^Z_L + (1-g_{\bullet,I})u_a\sum_L \nu_L f^Z_L \\
\therefore g_{\bullet,I} &= \frac{\epsilon\sum_L \nu_L f^X_L + u_a\sum_L \nu_L(f^Y_L + f^Z_L)}{1 - \sum_L \nu_L f^Z_L(\epsilon - u_a)},
\end{aligned}$$

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

$$\begin{aligned}
g_{\bullet,I} &= \frac{\epsilon\sum_L \nu_L f^X_L + u_a\sum_L \nu_L(f^Y_L + f^Z_L)}{1 - \sum_L \nu_L f^Z_L(\epsilon - u_a)} \\
&= \frac{\epsilon f^X\sum_L \nu_L + (f^Y + f^Z)u_a\sum_L \nu_L}{1 - f^Z\sum_L \nu_L(\epsilon - u_a)} \\
&= \frac{\epsilon f^X + u_a(f^Y + f^Z)}{1 - f^Z(\epsilon - u_a)},
\end{aligned}$$

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.

1.4. The "staying" norm. Under the "staying" norm (1), individuals do not change their opinions of a donor who interacts with a bad recipient at all, irrespective of the donor's action. In our notation, this is tantamount to replacing P^{BC} and P^{BD} with $g^{s}_{I,J}$, where s is the strategy whose reputation is being assessed, since, if the recipient has a bad reputation, the donor's reputation is unchanged. Reputations are thus given by

$$\begin{aligned}
g^{X}_{I,J} &= g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})g^{X}_{I,J}\\
g^{Y}_{I,J} &= g_{\bullet,J}P^{BC} + (1-g_{\bullet,J})g^{Y}_{I,J}\\
g^{Z}_{I,J} &= \delta_{I,J}\big[g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})g^{Z}_{I,J}\big]\\
&\quad + (1-\delta_{I,J})\big[G_{I,J}P^{GC} + (g_{\bullet,J}-G_{I,J})P^{GD} + (g_{\bullet,I}-G_{I,J})g^{Z}_{I,J} + (1-g_{\bullet,J}-g_{\bullet,I}+G_{I,J})g^{Z}_{I,J}\big]\\
&= \delta_{I,J}\big[g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})g^{Z}_{I,J}\big]\\
&\quad + (1-\delta_{I,J})\big[G_{I,J}P^{GC} + (g_{\bullet,J}-G_{I,J})P^{GD} + (1-g_{\bullet,J})g^{Z}_{I,J}\big]\\
&= \delta_{I,J}\big[g_{\bullet,J}P^{GC} + (1-g_{\bullet,J})g^{Z}_{I,J}\big]\\
&\quad + (1-\delta_{I,J})\big[G_{I,J}(P^{GC}-P^{GD}) + g_{\bullet,J}(P^{GD}-g^{Z}_{I,J}) + g_{\bullet,I}(P^{BC}-g^{Z}_{I,J}) + g^{Z}_{I,J}\big].
\end{aligned}$$

In equilibrium, this yields

$$\begin{aligned}
g^{X}_{I,J} = g^{Z}_{I,J}|_{I=J} &= P^{GC},\\
g^{Y}_{I,J} &= P^{GD},
\end{aligned}$$

but the out-group version of $g^{Z}_{I,J}$ can still be very complicated.

2. Invasibility of discriminators in a single group

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.

2.1. Invasibility by defectors. Let $f = f^{Z}$. Defectors resist invasion by discriminators provided

$$\begin{aligned}
\partial_f \dot{f}^{Y}|_{f=0} &< 0\\
\partial_f[f^{Y}(\Pi^{Y}-\bar{\Pi})]|_{f=0} &< 0\\
\partial_f[(1-f)(\Pi^{Y}-(1-f)\Pi^{Y}-f\Pi^{Z})]|_{f=0} &< 0\\
\partial_f[(1-f)(\Pi^{Y}-(1-f)\Pi^{Y}-f\Pi^{Z})]|_{f=0} &< 0\\
\partial_f[(f-f^2)(\Pi^{Z}-\Pi^{Y})]|_{f=0} &< 0\\
[(1-2f)(\Pi^{Z}-\Pi^{Y}) + (f-f^2)\partial_f(\Pi^{Z}-\Pi^{Y})]|_{f=0} &< 0\\
\Pi^{Z}|_{f=0} &< \Pi^{Y}|_{f=0}\\
(bfg^{Z}-cg)|_{f=0} &< bfg^{Y}|_{f=0}\\
\therefore -cg^{Y} &< 0.
\end{aligned}$$

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

$$\begin{aligned}
\partial_f \dot{f}^Y|_{f=1} &< 0 \\
\partial_f [f^Y(\Pi^Y - \bar{\Pi})]|_{f=1} &< 0 \\
\partial_f [(1-f)(\Pi^Y - (1-f)\Pi^Y - f\Pi^Z)]|_{f=1} &< 0 \\
\partial_f [(1-f)(\Pi^Y - (1-f)\Pi^Y - f\Pi^Z)]|_{f=1} &< 0 \\
\partial_f [(f-f^2)(\Pi^Z - \Pi^Y)]|_{f=1} &< 0 \\
[(1-2f)(\Pi^Z - \Pi^Y) + (f-f^2)\partial_f(\Pi^Z - \Pi^Y)]|_{f=1} &< 0 \\
\Pi^Z|_{f=1} &> \Pi^Y|_{f=1} \\
(bfg^Z - cg)|_{f=1} &> bfg^Y|_{f=1} \\
(bg^Z - cg^Z)|_{f=1} &> bg^Y|_{f=1} \\
b(g^Z - g^Y)|_{f=1} &> cg^Z|_{f=1} \\
bg(P^{GC} - P^{GD}) &> cg \\
\therefore \frac{b}{c} &> \frac{1}{P^{GC} - P^{GD}} = \frac{1}{\epsilon - u_a}.
\end{aligned}$$

This can also be written in terms of g: discriminators resist invasion by defectors provided

$$\begin{aligned}
\Pi^Z|_{f^Z=1} &> \Pi^Y|_{f^Z=1} \\
bg^Z|_{f^Z=1} - cg &> bg^Y|_{f^Z=1} \\
(b-c)g &> b[gP^{GD} + (1-g)P^{BD}] \\
g(b - c - b[P^{GD} - P^{BD}] &> bP^{BD} \\
g(b[1 - P^{GD} + P^{BD}] - c) &> bP^{BD} \\
\therefore g &> \frac{P^{BD}}{1 - P^{GD} + P^{BD} - c/b}.
\end{aligned} \tag{5}$$

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)

$$\begin{aligned}
\dot{f} &= 0 \\
f(\Pi^Z - \bar{\Pi}) &= 0 \\
\Pi^Z - f\Pi^Z - (1-f)\Pi^Y &= 0 \\
\Pi^Z &= \Pi^Y \\
bfg^Z - cg &= bfg^Y \\
bf(gP^{GC} + [1-g]P^{BD}) - cg &= bf(gP^{GD} + [1-g]P^{BD}) \\
bfg(P^{GC} - P^{GD}) &= cg \\
\therefore f &= \frac{c}{b}\frac{1}{P^{GC} - P^{GD}} = \frac{c}{b}\frac{1}{\epsilon - u_a}.
\end{aligned}$$

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

$$f^Z(g^Z - g^Y) > c/b. \tag{6}$$

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.

2.2. Invasibility by cooperators. Finally, we consider conditions under which *cooperators* can invade a population of discriminators. We proceed by reasoning similar to Eq. [5], 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 $f = f^Z$)

$$\begin{aligned}
\Pi^X|_{f=1} &> \Pi^Z|_{f=1} \\
(bfg^X - c)|_{f=1} &> (bfg^Z - cg)|_{f=1} \\
b(g^X - g) &> c(1-g) \\
b(gP^{GC} + (1-g)P^{BC} - g) &> c(1-g) \\
b(g[P^{GC} - P^{BC} - 1] + P^{BC}) &> c(1-g) \\
g(b[P^{GC} - P^{BC} - 1] + c) &> c - bP^{BC} \\
&\therefore \begin{cases} g > \dfrac{c - bP^{BC}}{b(P^{GC} - P^{BC} - 1) + c} & \textit{Shunning, Stern Judging}, \\ g < \dfrac{c - bP^{BC}}{b(P^{GC} - P^{BC} - 1) + c} & \textit{Scoring, Simple Standing}. \end{cases}
\end{aligned}$$

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*!

3. Multiple groups with well-mixed strategic imitation

When individuals choose their comparison partners from the entire population at random, irrespective of group identity, all strategic frequencies f_I^s rapidly equilibrate to a common value f^s; we show this in SI Section 8.4. We refer to this scenario as "well-mixed" strategic imitation. Under this scenario, the only interesting dynamical quantities are the "total" (group-averaged) strategy fitnesses Π^s, viz.:

$$\dot{f}^s = f^s\left[\left(\sum_I \nu_I \Pi_I^s\right) - \bar{\Pi}\right] = f^s[\Pi^s - \bar{\Pi}].$$

When strategic imitation is well-mixed, we can simplify Eq. [4] and write

$$\begin{aligned}
g &= \sum_I \sum_J \nu_I \nu_J \sum_s f_I^s g_{I,J}^s \\
&= \sum_I \sum_J \nu_I \nu_J \sum_s f^s g_{I,J}^s \\
&= \sum_s f^s \sum_I \sum_J \nu_I \nu_J g_{I,J}^s \\
&= \sum_s f^s g^s.
\end{aligned}$$

We continue by determining the group-averaged fitnesses. By summing the fitnesses over all groups, we obtain

$$\begin{aligned}
\Pi^Z &= \sum_I \nu_I \Pi_I^Z = \sum_I \Big\{ \nu_I (1-u_x) \Big[b \sum_J \nu_J (f_J^X + f_J^Z g_{I,J}^Z) - c g_{\bullet,I} \Big] \Big\} \\
&= (1-u_x) \sum_I \Big\{ \nu_I \Big[b \sum_J \nu_J (f^X + f^Z g_{I,J}^Z) - c g_{\bullet,I} \Big] \Big\} \\
&= (1-u_x) \Big[b(f^X + f^Z \sum_I \sum_J \nu_I \nu_J g_{I,J}^Z) - c \sum_I \nu_I g_{\bullet,I} \Big] \\
&= (1-u_x) \Big[b(f^X + f^Z \sum_I \sum_J \nu_I \nu_J g_{I,J}^Z) - c \sum_I \sum_J \nu_I \nu_J g_{J,I} \Big] \\
&= (1-u_x) \big[b(f^X + f^Z g^Z) - cg \big] \text{, and likewise} \\
\Pi^X &= \sum_I \nu_I \Pi_I^X = (1-u_x) \big[b(f^X + f^Z g^X) - c \big], \\
\Pi^Y &= \sum_I \nu_I \Pi_I^Y = (1-u_x) \big[b(f^X + f^Z g^Y) \big].
\end{aligned} \tag*{[7]}$$

The group-averaged strategic reputations are given by

$$\begin{aligned}
g^X &= \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J g_{I,J}^X \\
&= \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BC} \big] \\
&= \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J g_{\bullet,J} P^{GC} + \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J (1 - g_{\bullet,J}) P^{BC} \\
&= \sum_{I=1}^K \nu_I g P^{GC} + \sum_{I=1}^K \nu_I (1-g) P^{BC} \\
&= g P^{GC} + (1-g) P^{BC} \text{, and likewise} \\
g^Y &= g P^{GD} + (1-g) P^{BD}.
\end{aligned}$$

The form of g^Z will vary depending on the specific scenario, but we can obtain a couple of general relations. First, note that Eq. [5] becomes

$$\begin{aligned}
\Pi^Z|_{f^Z=1} &> \Pi^Y|_{f^Z=1} \\
bg^Z|_{f^Z=1} - cg &> bg^Y|_{f^Z=1} \\
(b-c)g &> b\big[g P^{GD} + (1-g) P^{BD} \big] \\
g(b - c - b[P^{GD} - P^{BD}]) &> bP^{BD} \\
g(b[1 - P^{GD} + P^{BD}] - c) &> bP^{BD} \\
\therefore g &> \frac{P^{BD}}{1 - P^{GD} + P^{BD} - c/b},
\end{aligned} \tag*{[8]}$$

and Eq. [6] becomes

$$f^Z (g^Z - g^Y) > c/b.$$

That is, under well-mixed strategic imitation, the conditions for discriminators to resist invasion by defectors and to increase in frequency over time can be written in terms of average reputations g^s, though the value of those reputations will vary depending on the group structure. An equivalent way to write Eq. [8] is

$$\begin{aligned}
\frac{b}{c} &> \frac{g}{g - g^Y} \\
\therefore \frac{b}{c} &> \frac{g}{g(1 + P^{BD} - P^{GD}) - P^{BD}}.
\end{aligned}$$

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 g^Z:

$$
\begin{aligned}
g^Z &= \sum_{I=1}^{K}\sum_{J=1}^{K} \nu_I \nu_J g_{I,J}^Z \\
&= \sum_{I=1}^{K}\sum_{J=1}^{K} \nu_I \nu_J \Big(\delta_{I,J} \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \\
&\quad + (1 - \delta_{I,J}) \big[G_{I,J} P^{GC} + (g_{\bullet,J} - G_{I,J}) P^{GD} + (g_{\bullet,I} - G_{I,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big] \Big) \\
&= \sum_{J=1}^{K} (\nu_J)^2 \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \\
&\quad + \sum_{\substack{I=1 \\ I \neq J}}^{K} \sum_{J=1}^{K} \nu_I \nu_J \big[G_{I,J} P^{GC} + (g_{\bullet,J} - G_{I,J}) P^{GD} + (g_{\bullet,I} - G_{I,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big].
\end{aligned}
$$

The first term simplifies to

$$
\begin{aligned}
\sum_{J=1}^{K} (\nu_J)^2 \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] &= \frac{1}{K} \sum_{I=1}^{K} \nu_J \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \\
&= \frac{1}{K} \big[g P^{GC} + (1 - g) P^{BD} \big].
\end{aligned}
$$

The second becomes

$$
\begin{aligned}
&\sum_{\substack{I=1 \\ I \neq J}}^{K} \sum_{J=1}^{K} \nu_I \nu_J \big[G_{I,J} P^{GC} + (G_{I,J} - g_{\bullet,J}) P^{GD} + (G_{I,J} - g_{\bullet,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big] \\
&= \sum_{\substack{I=1 \\ I \neq J}}^{K} \sum_{J=1}^{K} \nu_I \nu_J \big[G_{I,J} (P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g_{\bullet,I} (P^{GD} - P^{BD}) + g_{\bullet,J} (P^{BC} - P^{BD}) + P^{BD} \big].
\end{aligned}
$$

Because all the groups are the same size and strategy frequencies are identical across groups, the values of $g_{I,J}^s$ can only vary depending on whether $I = J$ or not. Define $g_{\text{in}}^s = g_{I,I}^s$ and $g_{\text{out}}^s = g_{I,J}^s \big|_{I \neq J}$. We exploit this symmetry to obtain

$$
\begin{aligned}
g_{\bullet,I} &= \sum_{L}^{K} \nu_L g_{L,I} \\
&= \frac{1}{K} \sum_{L}^{K} g_{L,I} \\
&= \frac{1}{K} g_{\text{in}} + \frac{K-1}{K} g_{\text{out}}, \\
g_{\bullet,J} &= \frac{1}{K} g_{\text{in}} + \frac{K-1}{K} g_{\text{out}}
\end{aligned}
$$

and

$$\begin{aligned}
G_{I,J} &= \sum_L \nu_L \sum_s f_L^s g_{L,I}^s g_{L,J}^s \\
&= \frac{1}{K} \sum_L \sum_s f_L^s g_{L,I}^s g_{L,J}^s \\
&= \frac{1}{K} \sum_L \sum_s f^s g_{L,I}^s g_{L,J}^s \\
&= \frac{1}{K} \sum_s f^s g_{\text{in}}^s g_{\text{out}}^s + \frac{1}{K} \sum_s f^s g_{\text{out}}^s g_{\text{in}}^s + \frac{K-2}{K} \sum_s f^s g_{\text{out}}^s g_{\text{out}}^s \\
&= \frac{2}{K} \sum_s f^s g_{\text{in}}^s g_{\text{out}}^s + \frac{K-2}{K} \sum_s f^s (g_{\text{out}}^s)^2.
\end{aligned}$$

Thus

$$\begin{aligned}
g^Z &= \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J g_{I,J}^Z \\
&= \frac{1}{K}[g P^{GC} + (1-g) P^{BD}] \\
&\quad + \frac{K-1}{K}\Big[\Big(\frac{2}{K} \sum_s f^s g_{\text{in}}^s g_{\text{out}}^s + \frac{K-2}{K} \sum_s f^s (g_{\text{out}}^s)^2\Big)(P^{GC} - P^{GD} - P^{BC} + P^{BD}) \\
&\quad + \Big(\frac{1}{K} g_{\text{in}} + \frac{K-1}{K} g_{\text{out}}\Big)(P^{GD} + P^{BC} - 2P^{BD}) + P^{BD}\Big].
\end{aligned} \tag{9}$$

Observe that setting $f^Z = 1$ in Eq. [9] yields exactly the system of equations one would need to solve in order to obtain g in a population of equally sized groups, viz.:

$$\begin{aligned}
g &= \frac{g_{\text{in}} + (K-1) g_{\text{out}}}{K}, \\
g_{\text{in}} &= g P^{GC} + (1-g) P^{BD} = g(P^{GC} - P^{BD}) + P^{BD}, \\
g_{\text{out}} &= \Big(\frac{2}{K} g_{\text{in}} g_{\text{out}} + \frac{K-2}{K}(g_{\text{out}})^2\Big)(P^{GC} - P^{GD} - P^{BC} + P^{BD}), \\
&\quad + \Big(\frac{1}{K} g_{\text{in}} + \frac{K-1}{K} g_{\text{out}}\Big)(P^{GD} + P^{BC} - 2P^{BD}) + P^{BD}.
\end{aligned}$$

For example, choosing *Stern Judging* as the norm and setting $u_x = 0$ yields $g_{\text{in}} = 1 - u_a$, $g_{\text{out}} = 1/2$, consistent with the reputation expressions from (2).

3.2. Limit of many groups: private reputations. As K approaches infinity, the contribution of g_{in}^s to the total average reputation of s, g^s, tends to zero, so that the entirety of g^s is due to the g_{out}^s terms. We thus have that

$$\lim_{K\to\infty} g^Z = \sum_s f^s (g^s)^2 (P^{GC} - P^{GD} - P^{BC} + P^{BD}) + 2g(P^{GD} + P^{BD} - 2P^{BD}) + P^{BD}.$$

Defining

$$\begin{aligned}
g_\star &= \sum_s f^s (g^s)^2, \\
d_\star &= \sum_s f^s g^s (1 - g^s) = g - g_\star, \\
b_\star &= \sum_s f^s (1 - g^s)^2 = 1 - 2g + g_\star
\end{aligned}$$

allows us to rewrite this as

$$\lim_{K\to\infty} g^Z = g_\star P^{GC} + d_\star (P^{GD} + P^{BC}) + b_\star P^{BD}.$$

This is the bottom term of Eq. 5 from (3) with empathy parameter $E = 0$, corresponding to fully private reputation assessment. 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 mean reputation in a population of discriminators is given by a solution to

$$0 = g^2(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g(P^{GD} + P^{BC} - 2P^{BD} - 1) + P^{BD}$$

$$\therefore g = \begin{cases} \dfrac{1 - \sqrt{1 - 4(\epsilon - u_a)u_a}}{2(\epsilon - u_a)}, & \textit{Shunning}, \\ \dfrac{1}{2}, & \textit{Stern Judging}, \\ \dfrac{u_a}{1 - \epsilon + u_a}, & \textit{Scoring}, \\ \dfrac{1 - u_a - \sqrt{(1 - u_a)(1 - \epsilon)}}{\epsilon - u_a}, & \textit{Simple Standing}. \end{cases}$$

We have picked out the solutions that are viable for $1 > u_x > 0$ and $1 > u_a > 0$. The result for *Stern Judging* was previously obtained by (4), which also showed that, in the presence of errors under private assessment, $g^s = 1/2$ for any strategy s, irrespective of the population's strategic composition.

3.3. One large group and many small groups. Without loss of generality, suppose that group 1 has size ν and the remaining $K - 1$ groups each have size $(1 - \nu)/(K - 1)$. Starting with Eq. [7], we can unpack g^Z. We have

$$\begin{aligned}
g^Z &= \sum_I \sum_J \nu_I \nu_J g^Z_{I,J} \\
&= \sum_{I=1}^K \sum_{J=1}^K \nu_I \nu_J \Big(\delta_{I,J} \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \\
&\quad + (1 - \delta_{I,J}) \big[G_{I,J} P^{GC} + (g_{\bullet,J} - G_{I,J}) P^{GD} + (g_{\bullet,I} - G_{I,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big] \Big) \\
&= \nu_2 \big[g_{\bullet,1} P^{GC} + (1 - g_{\bullet,1}) P^{BD} \big] \\
&\quad + \nu \frac{1 - \nu}{K - 1} \sum_{I=2}^K \big[G_{I,1} P^{GC} + (g_{\bullet,1} - G_{I,1}) P^{GD} + (g_{\bullet,I} - G_{I,1}) P^{BC} + (1 - g_{\bullet,1} - g_{\bullet,I} + G_{I,1}) P^{BD} \big] \\
&\quad + \nu \frac{1 - \nu}{K - 1} \sum_{J=2}^K \big[G_{1,J} P^{GC} + (g_{\bullet,J} - G_{1,J}) P^{GD} + (g_{\bullet,1} - G_{1,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,1} + G_{1,J}) P^{BD} \big] \\
&\quad + \left(\frac{1 - \nu}{K - 1} \right)^2 \Bigg(\sum_{J=2}^K \big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \\
&\quad + \sum_{\substack{I=2 \\ I \neq J}}^K \sum_{J=2}^K \big[G_{I,J} P^{GC} + (g_{\bullet,J} - G_{I,J}) P^{GD} + (g_{\bullet,I} - G_{I,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big] \Bigg) \\
&= \nu_2 \big[g_{\bullet,1} P^{GC} + (1 - g_{\bullet,1}) P^{BD} \big] \\
&\quad + \nu(1 - \nu) \big[G_{I,1} P^{GC} + (g_{\bullet,1} - G_{I,1}) P^{GD} + (g_{\bullet,I} - G_{I,1}) P^{BC} + (1 - g_{\bullet,1} - g_{\bullet,I} + G_{I,1}) P^{BD} \big] \Big|_{I \neq 1} \\
&\quad + \nu(1 - \nu) \big[G_{1,J} P^{GC} + (g_{\bullet,J} - G_{1,J}) P^{GD} + (g_{\bullet,1} - G_{1,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,1} + G_{1,J}) P^{BD} \big] \Big|_{J \neq 1} \\
&\quad + \frac{(1 - \nu)^2}{K - 1} \Big(\big[g_{\bullet,J} P^{GC} + (1 - g_{\bullet,J}) P^{BD} \big] \Big) \Big|_{J \neq 1} \\
&\quad + \left(\frac{(1 - \nu)(K - 2)}{K - 1} \right)^2 \big[G_{I,J} P^{GC} + (g_{\bullet,J} - G_{I,J}) P^{GD} \\
&\qquad + (g_{\bullet,I} - G_{I,J}) P^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P^{BD} \big] \Big|_{I \neq J \neq 1}.
\end{aligned}$$

As $K \to \infty$, the $I = J$ elements drop out. What remains is a special case of equation 12 of (5), in which part of the population consists of adherents to a single institution of reputation assessment and the remainder consists of private assessors.

4. Switching group membership

We now consider a variant of our model in which individuals can switch gossip group identity depending on their difference in fitness (with probability given by a Fermi function, as in the rest of our analysis). We assume $K = 2$ groups, both of which are fixed for strategy Z.

4.1. Same norm and payoffs. When both groups follow the same social norm and payoffs are group-independent, we have

$$\begin{aligned}\dot{\nu}_1 &= \nu_1(\Pi_1^Z - \bar{\Pi}),\\ \dot{\nu}_2 &= \nu_2(\Pi_2^Z - \bar{\Pi}), \text{with}\\ \bar{\Pi} &= \nu_1\Pi_1^Z + \nu_2\Pi_2^Z,\end{aligned}$$

as we show in SI Section 8.2. We expect ν_1 to grow if

$$\begin{aligned}\dot{\nu}_1 &> 0\\ \therefore \nu_1(\Pi_1^Z - \bar{\Pi}) &> 0\\ \therefore \nu_1(\Pi_1^Z - \nu_1\Pi_1^Z - (1-\nu_1)\Pi_2^Z) &> 0\\ \therefore (\nu_1 - (\nu_1)^2)(\Pi_1^Z - \Pi_2^Z) &> 0\\ \therefore \Pi_1^Z &> \Pi_2^Z\\ \therefore b(\nu_1 g_{1,1}^Z + \nu_2 g_{1,2}^Z) - c g_{\bullet,1} &> b(\nu_1 g_{2,1}^Z + \nu_2 g_{2,2}^Z) - c g_{\bullet,2}\\ \therefore b(\nu_1 g_{1,1}^Z + \nu_2 g_{1,2}^Z) - c(\nu_1 g_{1,1}^Z + \nu_2 g_{2,1}^Z) &> b(\nu_1 g_{2,1}^Z + \nu_2 g_{2,2}^Z) - c(\nu_1 g_{1,2}^Z + \nu_2 g_{2,2}^Z)\\ \therefore \nu_1(b[g_{1,1}^Z - g_{2,1}^Z] - c[g_{1,1}^Z - g_{1,2}^Z]) &> \nu_2(b[g_{2,2}^Z - g_{1,2}^Z] - c[g_{2,2}^Z - g_{2,1}^Z])\\ \therefore \frac{\nu_1}{\nu_2} &> \frac{b(g_{2,2}^Z - g_{1,2}^Z) - c(g_{2,2}^Z - g_{2,1}^Z)}{b(g_{1,1}^Z - g_{2,1}^Z) - c(g_{1,1}^Z - g_{1,2}^Z)}.\end{aligned} \quad [10]$$

When both groups follow the same social norm and are of the same size, there cannot be any difference between $g_{1,1}^Z$ and $g_{2,2}^Z$, nor between $g_{1,2}^Z$ and $g_{2,1}^Z$. The last line of Eq. [10] thus simplifies to 1, which at least suggests $\nu_1 = \nu_2 = 1/2$ is significant. We can show that ν_1 grows when it is greater than $1/2$ (i.e., the critical value ν_1^* equals $1/2$) by explicitly solving for the ratio ν_1/ν_2:

$$\begin{aligned}\frac{\nu_1}{\nu_2} &> \frac{b(g_{2,2}^Z - g_{1,2}^Z) - c(g_{2,2}^Z - g_{2,1}^Z)}{b(g_{1,1}^Z - g_{2,1}^Z) - c(g_{1,1}^Z - g_{1,2}^Z)}\\ &> \frac{b(1 + \nu_1(P^{BD} - P^{GC}) + \nu_2(P^{BC} - P^{GD})) - c(1 - \nu_1(P^{GC} + P^{GD} - P^{BC} - P^{BD}))}{b(1 + \nu_2(P^{BD} - P^{GC}) + \nu_1(P^{BC} - P^{GD})) - c(1 - \nu_2(P^{GC} + P^{GD} - P^{BC} - P^{BD}))}\\ \therefore \frac{\nu_1}{1-\nu_1} &> \frac{b(1 + \nu_1(P^{BD} - P^{GC}) + (1-\nu_1)(P^{BC} - P^{GD})) - c(1 - \nu_1(P^{GC} + P^{GD} - P^{BC} - P^{BD}))}{b(1 + (1-\nu_1)(P^{BD} - P^{GC}) + \nu_1(P^{BC} - P^{GD})) - c(1 - (1-\nu_1)(P^{GC} + P^{GD} - P^{BC} - P^{BD}))}.\end{aligned}$$

We can collect powers of ν_1 to obtain

$$\begin{aligned}&(\nu_1)^2(b[P^{GC} - P^{GD} + P^{BC} - P^{BD}] - c[P^{GC} + P^{GD} - P^{BC} - P^{BC}])\\ &\quad + \nu_1(b[1 - P^{GC} + P^{BD}] - c[1 - P^{GC} - P^{GD} + P^{BC} + P^{BD}])\\ &> (\nu_1)^2(b[P^{BC} - P^{BD} + P^{GC} - P^{GD}] - c[P^{GC} + P^{GD} - P^{BC} - P^{BD}])\\ &\quad + \nu_1(b[2P^{GD} - P^{GC} + -2P^{BC} + P^{BD} - 1] - c[P^{BC} + P^{BD} - P^{GC} - P^{GD} - 1])\\ &\quad + b(1 + P^{BC} - P^{GD}) - c.\end{aligned}$$

The quadratic terms cancel, leaving

$$\begin{aligned}\therefore \nu_1(b[2 + 2P^{BC} - 2P^{GD}] - 2c) &> b(1 + P^{BC} - P^{GD}) - c\\ \therefore \nu_1 &> 1/2.\end{aligned}$$

4.2. Different norms and payoffs. We now allow social norms and payoffs to differ between groups. Suppose group J uses a social norm with reputation probabilities $P_J^{GC}, P_J^{GD}, P_J^{BC}, P_J^{BD}$ when assessing others' reputations, and suppose an individual in group I who cooperates with an individual in group J conveys a benefit $b_{I,J}$ but pays a cost $c_{I,J}$. (A natural application of variable benefits and costs is to allow in-group and out-group interactions to differ, so that $b_{I,J} = \delta_{I,J} b_{\text{in}} + (1-\delta_{I,J}) b_{\text{out}}$ and $c_{I,J} = \delta_{I,J} c_{\text{in}} + (1-\delta_{I,J}) c_{\text{out}}$.) For completeness, we present payoffs for all three strategic types:

$$\begin{aligned}
\Pi_I^X &= (1-u_x)\Big\{\sum_J \nu_J \big[b_{I,J}(f_J^X + f_J^Z g_{I,J}^X)\big] - c_{I,J}\Big\} \\
\Pi_I^Y &= (1-u_x)\Big\{\sum_J \nu_J \big[b_{I,J}(f_J^X + f_J^Z g_{I,J}^Y)\big]\Big\} \qquad [11] \\
\Pi_I^Z &= (1-u_x)\Big\{\sum_J \nu_J \big[b_{I,J}(f_J^X + f_J^Z g_{I,J}^Z)\big] - c_{I,J} g_{\bullet,I}\Big\}.
\end{aligned}$$

Group-dependent social norms mean that reputation equations change:

$$\begin{aligned}
g_{I,J}^X &= g_{\bullet,J} P_J^{GC} + (1-g_{\bullet,J}) P_J^{BC} \\
g_{I,J}^Y &= g_{\bullet,J} P_J^{GD} + (1-g_{\bullet,J}) P_J^{BD} \\
g_{I,J}^Z &= \delta_{I,J}\big[g_{\bullet,J} P_J^{GC} + (1-g_{\bullet,J}) P_J^{BD}\big] \\
&\quad + (1-\delta_{I,J})\big[G_{I,J} P_J^{GC} + (g_{\bullet,J} - G_{I,J}) P_J^{GD} + (g_{\bullet,I} - G_{I,J}) P_J^{BC} + (1 - g_{\bullet,J} - g_{\bullet,I} + G_{I,J}) P_J^{BD}\big], \text{with} \\
g_{I,J} &= \sum_s f_I^s g_{I,J}^s, \qquad [12] \\
g_{\bullet,J} &= \sum_L \nu_L g_{L,J}, \\
G_{I,J} &= \sum_L \nu_L \sum_s f^s g_{L,I}^s g_{L,J}^s.
\end{aligned}$$

In the main text, we do not vary the $b_{I,J}$ and $c_{I,J}$, but we simultaneously solve Eqs. [11] and [12] for cases where I and J follow different social norms and both populations are fixed for Z. The dynamics are at most bistable, with ν_1 shrinking unless it is above a critical frequency ν_1^*. For *Stern Judging*, this critical frequency is almost always less than $1/2$, meaning groups that follow *Stern Judging* are likely to grow over a larger region of phase space than any of the second-order norms we consider.

within-group reputations ($g_{1,1}$)

	SJ	SS	SC	SH
SJ	0.97	0.96	0.96	0.97
SS	0.96	0.96	0.96	0.96
SC	0.73	0.80	0.34	0.09
SH	0.10	0.10	0.06	0.06

between-group reputations ($g_{1,2}$)

	SJ	SS	SC	SH
SJ	0.47	0.75	0.65	0.36
SS	0.83	0.90	0.82	0.38
SC	0.78	0.86	0.34	0.06
SH	0.07	0.07	0.02	0.02

Table S2. In-group and out-group reputations for $K = 2$ equally sized groups. The social norm used in group 1 is indicated at the top of each column; the norm used in group 2 is indicated at the left of each row. Darker colors denote more strongly positive opinions. When group 1 follows *Stern Judging*, it typically has a high view of itself but a somewhat low view of group 2, so that its members will often cooperate with each other and are less likely to engage in un-reciprocated cooperation with the opposing group. No other norm satisfies both of these conditions, which is why *Stern Judging* tends to outcompete other social norms across a wide range of costs and benefits c and b (see Main Text Fig. 1). Error rates are $u_a = u_x = 0.02$.

The sole exception is when *Stern Judging* competes against *Shunning* and b is sufficiently small (e.g., $b = 2$), in which case $\nu_1^* > 1/2$. In SI figure S1, we show that this case is distinctive because, as the *Shunning* group expands, the population becomes invasible by *defectors*. This means that, in a model with both group and strategy evolution (i.e., $0 < \tau < 1$), it cannot be guaranteed that the population would continue to consist of discriminators; a lucky defector mutant that invades at the right time might take over the entire population.

4.3. Co-evolution of strategies and norms. Up until now, we have assumed either that strategy frequencies evolve *or* group sizes evolve, with the other one being fixed. In this section, we relax this assumption and allow both group sizes

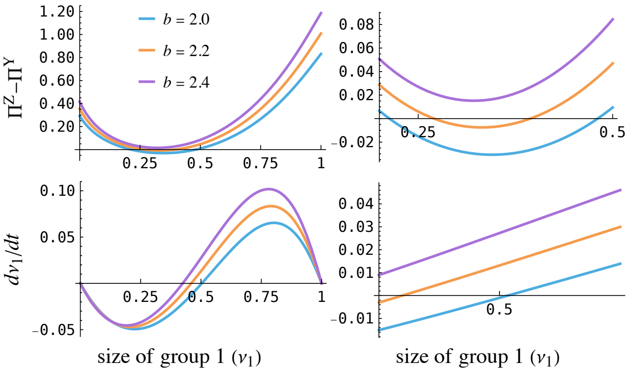

Fig. S1. Competition between *Stern Judging* (group 1) and *Shunning* (group 2) in $K = 2$ groups. For low values of b we have $\nu_1^* > 1/2$, so that *Shunning* (not *Stern Judging*) can take over the population even when starting from a minority. In this regime, however, as the *Shunning* group grows, the population passes through a regime where it becomes vulnerable to invasion by pure defectors (top two plots). Increasing b to the point where this instability no longer occurs is sufficient to push ν_1^* below $1/2$, so that *Stern Judging* will take over the population when starting from a minority. And so, in summary, in all regimes where the population resists invasion by defectors, *Stern Judging* out-competes *Shunning*, even when starting in the minority.

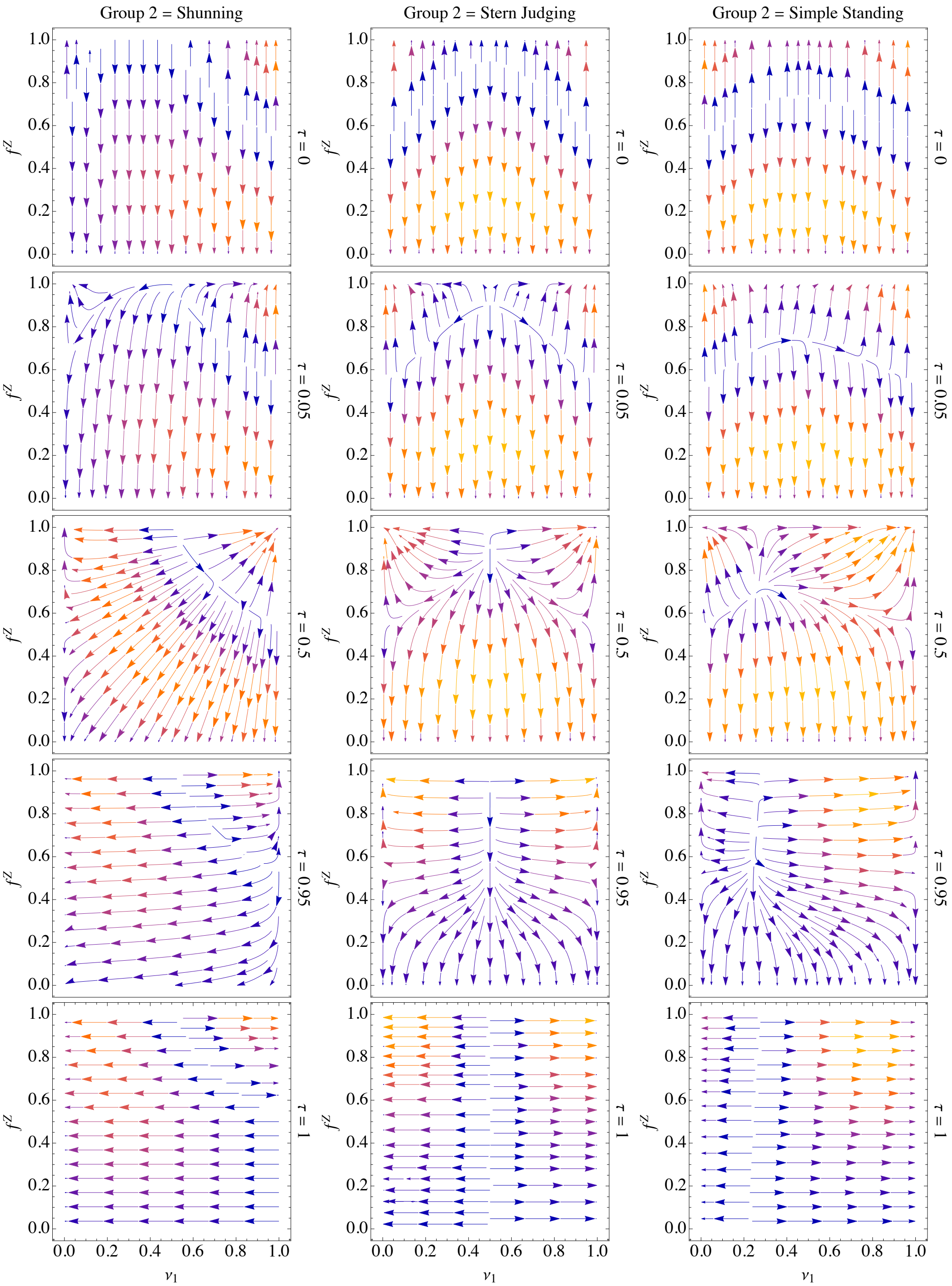

Fig. S2. Coevolution of strategy frequencies and group sizes under well-mixed copying, in $K = 2$ groups. Group 1 follows the Stern Judging norm, and group 2's norm is indicated in the column headings at the top. Each plot shows the vector field of replicator dynamics, with the x-axis indicating the frequency of group 1 versus group 2, and the y-axis indicating the frequency of DISC versus ALLD. The top two rows ($\tau = 0$ and 0.05) correspond to slow dynamics of group imitation relative to strategic imitation; and the bottom rows correspond to rapid dynamics of group imitation relative to strategic imitation. In all plots, $b = 2$, $c = 1$, $u_a = u_x = 0.02$.

and strategy frequencies to co-evolve. In this model, when an individual is chosen to update, they decide to update their group identity with probability τ or their behavioral strategy with probability $1-\tau$; we continue to assume well-mixed copying. As we show in SI section 8.6, group sizes and strategy frequencies are then governed by the system of differential equations

$$\dot{\nu}_I = \nu_I \tau(\Pi_I - \bar{\Pi}),$$
$$\dot{f}^s = f^s(1-\tau)(\Pi^s - \bar{\Pi}).$$

We present numerical solutions to these equations as vector fields in SI Figure S2, showing the co-evolution of group sizes and strategy frequencies for *Stern Judging* versus other social norms, for a range of different values of τ. When $\tau = 0$, we recover the limit of fixed group sizes (i.e., only behavioral strategies evolve); when $\tau = 1$, behavioral strategies are fixed and only group sizes evolve.

The results are consistent with our prior findings in the limits of pure strategy competition ($\tau = 0$) or pure norm competition ($\tau = 1$). The analysis also extends some of our results for $\tau = 1$ to cases with intermediate τ: in particular, even when strategies and norms co-evolve, there remains a large basin of attraction towards *Stern-Judging* discriminators in competition with *Simple Standing* and defectors. And we also find qualitatively new phenomena: there is a large basin of attraction towards defectors when *Stern Judging* competes with *Shunning* while both norms and strategies evolve at similar rates.

In the case of competition between Simple Standing and Stern Judging (Figure S2, right column), the top edge of each vector field ($f^Z = 1$) recapitulates the result that Stern Judging can out-compete Simple Standing even when starting from a minority (consistent with main text Figure 1). The same result is also seen in the bottom row of Figure S2, when only group membership evolves ($\tau = 1$)–and indeed it holds regardless of the frequency of defectors. Moreover, there is a substantial basin of attraction towards Stern Judging and DISC, including from initial conditions where Stern Judging starts in the minority, even when strategies and group membership co-evolve ($0 < \tau < 1$).

In the case of two competing groups that each follow Stern Judging (Figure S2, middle column), we see that a population undergoing strategic evolution is more vulnerable to eventual domination by defectors when the groups are equally sized (top row, $\tau = 0$), a result that is consistent with our analysis of pure strategic evolution in the main text (main text Figure 3, middle column). This result continues to hold when strategies and group membership co-evolve ($\tau > 0$).

In the case of competition between Shunning and Stern Judging (Figure S2, left column), the results for discriminators ($f^Z = 1$) are consistent with SI Figure S1; specifically, when the Shunning group begins to overtake the Stern Judging group, the population becomes vulnerable to invasion by defectors. When strategies and group membership evolve at similar rates (middle row), there is a small basin of attraction towards Stern Judging and discriminators, a small basin towards Shunning and discriminators, and a large basin towards ALLD. And when only group membership evolves ($\tau = 1$, bottom row), Shunning will often overtake Stern Judging when discriminators predominate, and Shunning will always do so when defectors predominate.

4.4. Switching costs. Hitherto, we have assumed that switching group membership does not incur any kind of fitness penalty. In the real world, there may be social barriers or initiation costs associated with joining a new group. Such costs could be "immediate", thus affecting the perceived fitness change associated with switching group membership, or they could be transient, requiring time to overcome. In this section, we consider several different "switching cost" models and show that they do not affect our main results.

We first consider the possibility that individuals regard it as costly to switch groups; in effect, when they compare themselves with an individual in a different group, they consider the out-group individual to have their fitness lowered by a value α. This means the Fermi function becomes

$$\phi(\Pi_J, \Pi_I) = \frac{1}{1 + \exp[\beta(\Pi_J - \Pi_I - \alpha)]}$$

for individuals in groups I and J. In SI section 8.7, we show that this has no effect on the dynamics, because the total rate of change of group I, $\dot{\nu}_I$, is given by the difference in the rates at which J individuals switch to I and at which I individuals switch to J. In the weak selection limit, the switching cost penalizes both of these rates equally.

We then consider a more sophisticated switching cost model, in which individuals who are "new" to a group have their fitness penalized by α until some amount of time has elapsed; they transition out of the "new" state (with fitness penalty α) and into the "old" state (no fitness penalty) at rate σ. We limit ourselves to the case of two groups, so that

$$\Pi_1^{\text{new}} = \Pi_1^{\text{old}} - \alpha,$$
$$\Pi_2^{\text{new}} = \Pi_2^{\text{old}} - \alpha;$$

the expressions for Π_1^{old} and Π_2^{old} are the standard expressions for discriminator fitnesses.

We assume, further, that individuals who are "new" to their group may still switch back to their former group without penalty and be regarded as "old" in that group. We use ν_1^{new}, ν_1^{old}, ν_2^{new}, and ν_2^{old} to denote the fraction of the population that is "new" in group 1, "old" in group 1, and so on.

In SI Section 8.8, we exploit separation of timescales to show that the fraction of "new" individuals in each group, ρ_1 and ρ_2, rapidly equilibrates to a value

$$\rho_1 = \frac{1-\nu_1}{1+\sigma}, \rho_2 = \frac{\nu_1}{1+\sigma};$$

the dynamics then follow

$$\dot{\nu}_1 = \nu_1(\bar{\Pi}_1 - \bar{\Pi}), \dot{\nu}_2 = \nu_2(\bar{\Pi}_2 - \bar{\Pi}),$$

in which

$$\begin{aligned}\bar{\Pi}_1 &= (\nu_1^{\text{new}}\Pi_1^{\text{new}} + \nu_1^{\text{old}}\Pi_1^{\text{old}})/\nu_1,\\ \bar{\Pi}_2 &= (\nu_2^{\text{new}}\Pi_2^{\text{new}} + \nu_2^{\text{old}}\Pi_2^{\text{old}})/\nu_2.\end{aligned}$$

Note that, while ρ_1 and ρ_2 feature no explicit dependence on the switching cost α, the cost appears in $\bar{\Pi}_1$ and $\bar{\Pi}_2$. We then explore the effects of changing parameter values on $\dot{\nu}_1$. It is worth noting that

$$\begin{aligned}\dot{\nu}_1 &= \nu_1(\bar{\Pi}_1 - \bar{\Pi})\\ &= \nu_1([\nu_1^{\text{new}}\Pi_1^{\text{new}} + \nu_1^{\text{old}}\Pi_1^{\text{old}}]/\nu_1 - \nu_1^{\text{new}}\Pi_1^{\text{new}} - \nu_1^{\text{old}}\Pi_1^{\text{old}} - \nu_2^{\text{new}}\Pi_2^{\text{new}} - \nu_2^{\text{old}}\Pi_2^{\text{old}})\\ &= \nu_1(\rho_1\Pi_1^{\text{new}} + (1-\rho_1)\Pi_1^{\text{old}} - \nu_1[\rho_1\Pi_1^{\text{new}} + (1-\rho_1)\Pi_1^{\text{old}}] - \nu_2[\rho_2\Pi_2^{\text{new}} + (1-\rho_2)\Pi_2^{\text{old}}])\\ &= \nu_1([1-\nu_1][\rho_1\Pi_1^{\text{new}} + (1-\rho_1)\Pi_1^{\text{old}}] - [1-\nu_1][\rho_2\Pi_2^{\text{new}} + (1-\rho_2)\Pi_2^{\text{old}}])\\ &= \nu_1(1-\nu_1)[\rho_1\Pi_1^{\text{new}} + (1-\rho_1)\Pi_1^{\text{old}} - \rho_2\Pi_2^{\text{new}} - (1-\rho_2)\Pi_2^{\text{old}}]\\ &= \nu_1(1-\nu_1)[\Pi_1^{\text{old}} + \rho_1(\Pi_1^{\text{new}} - \Pi_1^{\text{old}}) - \Pi_2^{\text{old}} - \rho_2(\Pi_2^{\text{new}} - \Pi_2^{\text{old}})]\\ &= \nu_1(1-\nu_1)[\Pi_1^{\text{old}} - \rho_1\alpha - \Pi_2^{\text{old}} + \rho_2\alpha].\end{aligned}$$

This will have zeros at $\nu_1 = 0$, $\nu_1 = 1$, and $\Pi_1^{\text{old}} - \rho_1\alpha = \Pi_2^{\text{old}} - \rho_2\alpha$.

We can then wonder whether there is ever a value of ν_1 such that $\Pi_1^{\text{old}} > \Pi_2^{\text{old}}$ (i.e., $\dot{\nu}_1$ would be *positive* in the "null" case, with no switching cost) but $\Pi_1^{\text{old}} - \rho_1\alpha < \Pi_2^{\text{old}} - \rho_2\alpha$ (i.e., $\dot{\nu}_1$ becomes negative once the switching cost is imposed). Using the separation of timescales argument above, we obtain

$$\begin{aligned}\Pi_1^{\text{old}} - \frac{1-\nu_1}{1+\sigma}\alpha &< \Pi_2^{\text{old}} - \frac{\nu_1}{1+\sigma}\alpha\\ \therefore \Pi_1^{\text{old}} - \Pi_2^{\text{old}} &< \frac{1-\nu_1}{1+\sigma}\alpha - \frac{\nu_1}{1+\sigma}\alpha\\ \therefore \Pi_1^{\text{old}} - \Pi_2^{\text{old}} &< \frac{1-2\nu_1}{1+\sigma}\alpha\\ \therefore \frac{1+\sigma}{\alpha}(\Pi_1^{\text{old}} - \Pi_2^{\text{old}}) &< 1 - 2\nu_1\\ \therefore \frac{1}{2} - \frac{1+\sigma}{2\alpha}(\Pi_1^{\text{old}} - \Pi_2^{\text{old}}) &> \nu_1.\end{aligned}$$

Since the subtrahend on the left hand side is positive, this implies a switching cost can *only* cause $\dot{\nu}_1$ to switch from positive to negative for $\nu_1 < 1/2$. If ν_1^*, the critical size below which ν_1 shrinks and above which it grows, is less than $1/2$, then ν_1 is positive everywhere on the interval $(\nu_1^*, 1/2]$. The above argument demonstrates that it is only possible to switch ν_1 from positive to negative in the interval $(\nu_1^*, 1/2)$. By the mean value theorem, imposing a switching cost therefore cannot move the non-trivial zero of $\dot{\nu}_1$ from a value less than $1/2$ to a value greater than or equal to $1/2$. A similar argument shows that, if $\nu_1^* = 1/2$, it will still be $1/2$ after a switching cost is imposed; likewise, if $\nu_1^* > 1/2$, it will still be $1/2$ with a switching cost.

We verify this result in SI Figure S3. Note that if either the switching cost is small (α near 0), or if the rate of establishment in a new group is fast (σ large), then the dynamics are quantitatively similar to a model without any switching cost whatsoever. Nonetheless, regardless of the switching cost (α) or the length of time an individual must pay the cost after switching groups ($1/\sigma$), which one of two competing norms is "stronger" (i.e., which norm can win even when starting from a minority of the population) remains unchanged.

In summary, while a switching cost can quantitatively affect competition between norms, it *cannot* affect which norm is "stronger" in the sense of being able to overtake the population starting from a minority.

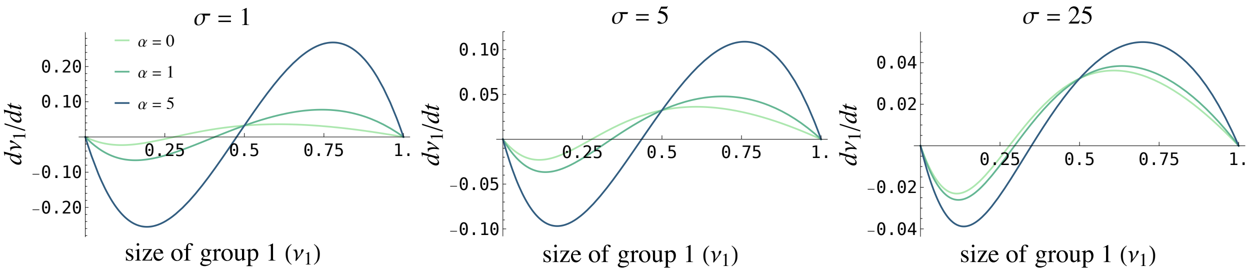

Fig. S3. Norm competition with a switching cost. In all plots, $u_a = u_x = 0.02$, $b = 2$, $c = 1$, group 1 follows Stern Judging, and group 2 follows Simple Standing. The rate σ at which individuals "establish" in their new groups (and thus are absolved of the switching cost) is indicated along the top of each plot. Each plot shows curves for three different values of the switching cost, including the case $\alpha = 0$ which reduces to the model without costs. When the switching cost α is high and the establishment rate σ is low, the critical value ν_1^* above which group 1 grows increases. But regardless of the cost α or rate σ, ν_1^* never exceeds $1/2$, meaning that Stern Judging is always the stronger norm – consistent with the general result we derive analytically.

4.5. Private reputations. We briefly consider what happens when two groups each adhere to *private* reputation assessment but follow different norms. We assume both groups are fixed for discriminators. In that case,

$$\begin{aligned} g_{I,J} &= \sum_L \nu_L \left(g_{I,L} g_{J,L} P_J^{GC} + g_{I,L}[1 - g_{J,L}] P_J^{BC} + [1 - g_{I,L}] g_{J,L} P_J^{GD} + [1 - g_{I,L}][1 - g_{J,L}] P_J^{BD} \right) \\ &= G_{I,J} P_J^{GC} + (g_{\bullet,I} - G_{I,J}) P_J^{BC} + (g_{\bullet,J} - G_{I,J}) P_J^{GD} + (1 - g_{\bullet,I} - g_{\bullet,J} + G_{I,J}) P_J^{BD}. \end{aligned}$$

An important difference between this expression and $g^Z_{I,J}$ in Eq. [12] is that individuals *in the same group* are not guaranteed to share reputational views of anyone else in the population, meaning that the $I = J$ term does not have a different form. Thus, for two groups, we have

$$\begin{aligned} g_{1,1} &= G_{1,1} P_1^{GC} + (g_{\bullet,1} - G_{1,1})(P_1^{GD} + P_1^{BC}) + (1 - 2g_{\bullet,1} + G_{1,1}) P_1^{BD} \\ g_{1,2} &= G_{1,2} P_2^{GC} + (g_{\bullet,1} - G_{1,2}) P_2^{BC} + (g_{\bullet,2} - G_{1,2}) P_2^{GD} + (1 - g_{\bullet,1} - g_{\bullet,2} + G_{1,2}) P_2^{BD} \\ g_{2,1} &= G_{2,1} P_1^{GC} + (g_{\bullet,2} - G_{2,1}) P_1^{BC} + (g_{\bullet,1} - G_{2,1}) P_1^{GD} + (1 - g_{\bullet,2} - g_{\bullet,1} + G_{2,1}) P_1^{BD} \\ g_{2,2} &= G_{2,2} P_2^{GC} + (g_{\bullet,2} - G_{2,2})(P_2^{GD} + P_2^{BC}) + (1 - 2g_{\bullet,2} + G_{2,2}) P_1^{BD}. \end{aligned}$$

When $\nu_1 \to 1$ (i.e., there is only one group), this reduces to

$$g = g_\star P^{GC} + (g - g_\star)(P^{GD} + P^{BC}) + (1 - 2g + g_\star) P^{BD},$$

which is the standard expression for one group with private reputations. When information about reputations is shared within a group, individuals can benefit both by choosing a more socially beneficial norm (one that minimizes their risk of unreciprocated cooperation) and by aligning themselves with the reputational assessments of a larger group. When both groups rely solely on private assessments, the second advantage is weakened, because merely following the same norm as a given group is not sufficient to ensure a good reputation in the eyes of that group: this is especially true of *Stern Judging* and *Shunning*, which are relatively intolerant of disagreement.

We consider competition between social norms under private assessment in SI figure S4, by allowing individuals to switch group identity and, thus, which norm they use in assessing others. A murkier picture emerges than under group-wise public assessment. *Stern Judging* and *Shunning* are both capable of "beating" other norms, in the sense of growing in size even when their group is less than half the population, under certain circumstances. However, the fact that they are themselves vulnerable to invasion by defectors (whereas *Simple Standing* is not) means it is difficult to draw a general conclusion about the "strength" of these norms. *Simple Standing* emerges as a "strong" norm that can generally outcompete other norms, especially for high values of b, and is itself capable of fomenting a high level of cooperation under private assessment.

5. Insular social interactions

We now consider the possibility that, instead of interacting equally with everyone in the population, individuals have different interaction rates with their in-group versus their out-group. With probability $\omega_{I,J}$, a possible interaction between individuals in groups I and J happens (for simplicity we assume $\omega_{I,J} = \omega_{J,I}$). We average fitnesses over all interactions that actually happen; for an individual in group I, this will be given by

$$\mathcal{M}_I = \sum_L \nu_L \omega_{I,L},$$

and the total number of interactions an individual in group I engages in with someone in group J will be given by $\nu_J \omega_{I,J}$ (times N, which we divide out).

5.1. Reputations and fitnesses. Before writing down fitnesses, it is instructive to consider how reputations change. We need to consider the fact that interactions that do not happen cannot be observed and therefore cannot factor into updating someone's reputation. We thus assume that an observer is equally likely to observe any of the donor's actions *that actually happened*, which means that when they consider a random interaction of a group I individual, it is with an individual in arbitrary group L with probability $\nu_L \omega_{I,L} / \mathcal{M}_I$.

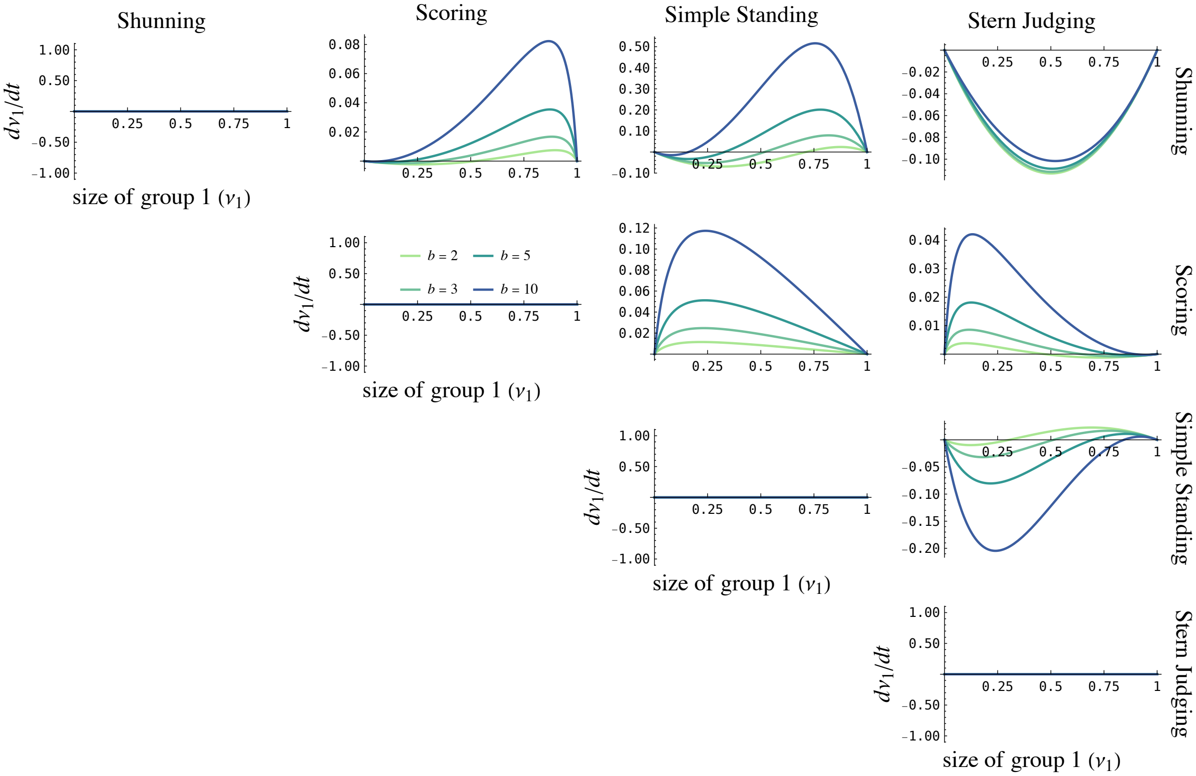

Fig. S4. Group size dynamics for $K = 2$ groups and varying values of the benefit b, with *private* (individually held) reputations rather than public reputations shared among group members. The norm used in group 1 is along the top: the norm used in group 2, along the left. In all plots, $c = 1$, $u_a = u_x = 0.02$. Values of b are as inset in the *Scoring-Scoring* figure.

5.1.1. Cooperator reputation. 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 in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group J sees as good (probability $g_{L,J}$); the I group member cooperates, which yields a good reputation with probability P_J^{GC}.

2. with someone in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group J sees as bad (probability $1 - g_{L,J}$); the I group member cooperates, which yields a good reputation with probability P_J^{BC}.

If we define

$$\gamma_{I,J} = \frac{1}{\mathcal{M}_I}\sum_L \nu_L\omega_{L,I}g_{L,J},$$

the average reputation (in J's eyes) of the component of the population I interacted with, then

$$g_{I,J}^X = \gamma_{I,J}P_J^{GC} + (1-\gamma_{I,J})P_J^{BC} = \gamma_{I,J}(P^{GC} - P^{BC}) + P^{BC}.$$

This is different from the classical case of uniform population-wide interaction, because how many interactions, and with whom, a member of I engages in now depends on I. That is, $\gamma_{I,J}$ could be thought of as $g_{\bullet,J}$, corrected for the fact that I no longer interacts uniformly with the entire population: J can only judge I based on whom I actually interacted with. Setting all $\omega_{I,J} = 1$ yields $\gamma_{I,J} = g_{\bullet,J}$, $\mathcal{M}_I = 1$, and $\gamma_{I,J} = G_{I,J}$.

5.1.2. Defector reputation. A defector in group I can be assigned a good reputation in the eyes of group J in two ways. J can observe the I defector's interaction:

1. with someone in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group J sees as good (probability $g_{L,J}$); the I group member defects, which yields a good reputation with probability P_J^{GD}.

2. with someone in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group J sees as bad (probability $1 - g_{L,J}$); the I group member defects, which yields a good reputation with probability P_J^{BD}.

Thus,

$$g_{I,J}^Y = \gamma_{I,J}P_J^{GD} + (1-\gamma_{I,J})P_J^{BD} = \gamma_{I,J}(P^{GD} - P^{BD}) + P^{BD}. \qquad [13]$$

5.1.3. Discriminator reputation. 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 in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group I sees as good (probability $g_{L,I}$); the I discriminator cooperates, which yields a good reputation with probability P_J^{GC}.

2. with someone in arbitrary group L (probability $\nu_L\omega_{L,I}/\mathcal{M}_I$) whom group I sees as bad (probability $1 - g_{L,I}$); the I discriminator defects, which yields a good reputation with probability P_J^{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 s (probability $\nu_L\omega_{L,I}f_L^s/\mathcal{M}_I$) whom I sees as good (probability $g_{L,I}^s$) and whom J also sees as good (probability $g_{L,J}^s$); the I discriminator cooperates, which yields a good reputation with probability P^{GC}.

2. with someone in an arbitrary group L following strategy s (probability $\nu_L\omega_{L,I}f_L^s/\mathcal{M}_I$) whom I sees as bad (probability $1 - g_{L,I}^s$) but whom J sees as good (probability $g_{L,J}^s$); the I discriminator defects, which yields a good reputation with probability P^{GD}.

3. with someone in an arbitrary group L following strategy s (probability $\nu_L\omega_{L,I}f_L^s/\mathcal{M}_I$) whom I sees as good (probability $g_{L,I}^s$) but whom J sees as bad (probability $1 - g_{L,J}^s$); the I discriminator cooperates, which yields a good reputation with probability P^{BC}.

4. with someone in an arbitrary group L following strategy s (probability $\nu_L\omega_{L,I}f_L^s/\mathcal{M}_I$) whom I sees as bad (probability $1 - g_{L,I}^s$) and whom J also sees as bad (probability $1 - g_{L,J}^s$); the I discriminator defects, which yields a good reputation with probability P^{BD}.

We can sum over all groups and strategy combinations to obtain

$$\begin{aligned}
\frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} \sum_s f_L^s g_{L,I}^s g_{L,J}^s &= \Gamma_{I,J}, \\
\frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} \sum_s f_L^s (1-g_{L,I}^s) g_{L,J}^s &= \gamma_{I,J} - \Gamma_{I,J}, \\
\frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} \sum_s f_L^s g_{L,I}^s (1-g_{L,J}^s) &= \gamma_{I,I} - \Gamma_{I,J}, \\
\frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} \sum_s f_L^s (1-g_{L,I}^s)(1-g_{L,J}^s) &= 1 - \gamma_{I,J} - \gamma_{I,I} + \Gamma_{I,J}.
\end{aligned}$$

Thus,

$$\begin{aligned}
g_{I,J}^Z &= \delta_{I,J}\left[\gamma_{I,J} P_J^{GC} + (1-\gamma_{I,J}) P_J^{BD}\right] \\
&\quad + (1-\delta_{I,J})\left[\Gamma_{I,J} P_J^{GC} + (\gamma_{I,J} - \Gamma_{I,J}) P_J^{GD} + (\gamma_{I,I} - \Gamma_{I,J}) P_J^{BC} + (1 - \gamma_{I,J} - \gamma_{I,I} + \Gamma_{I,J}) P_J^{BD}\right] \\
&= \delta_{I,J}\left[\gamma_{I,J} P_J^{GC} + (1-\gamma_{I,J}) P_J^{BD}\right] \\
&\quad + (1-\delta_{I,J})\left[\Gamma_{I,J}(P_J^{GC} - P_J^{GD} - P_J^{BC} + P_J^{BD}) + \gamma_{I,J}(P_J^{GD} - P_J^{BD}) + \gamma_{I,I}(P_J^{BC} - P_J^{BD}) + P_J^{BD}\right].
\end{aligned}$$

5.1.4. Group-averaged reputations. When individuals in different groups do not assuredly interact but rather interact only with probability $\omega_{I,J}$, Eq. [3] generalizes to

$$g^s = \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} g_{I,J}^s. \qquad [14]$$

Here, we have simply re-weighted $g_{I,J}^s$ by the probability $\omega_{I,J}$ that potential interactions actually occur; setting all $\omega_{I,J} = 1$ reduces to Eq. [3]. The population-averaged reputation is likewise given by

$$g = \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} g_{I,J},$$

which we can also write as

$$g = \sum_I \nu_I \gamma_{I,I}.$$

This expression reduces to Eq. [4] when all $\omega_{I,J} = 1$. Furthermore, we can define the related term

$$\hat{g} = \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} g_{J,I},$$

which is a weighted average of an arbitrary individual's view of the rest of the population. Note that $\hat{g} = g$ if all the ν_I are equal or all the $\omega_{I,J}$ are equal.

5.1.5. Fitnesses. We can now write down fitnesses. An individual in group I acquires a payoff $b_{I,J}$ for each group J interaction either with a cooperator or with a discriminator who sees them as good. In group I, a cooperator pays cost $c_{I,J}$ in each interaction they engage in, and a discriminator pays cost $c_{I,J}$ in each interaction with someone whom they see as good. An arbitrary individual in group I engages in $\mathcal{M}_I$ interactions. If the individual is a discriminator, then of these, $\nu_J \omega_{J,I}$ interactions will be with someone in group J, and the discriminator will regard them as good with probability $g_{J,I}$.

Finally, we average the payoffs differently. In the no-insularity case, we divide payoffs by all N interactions an individual engages in. With insularity, individuals engage in $\mathcal{M}_I$ interactions (times N), so we normalize by $\mathcal{M}_I$ to obtain their interaction-averaged payoff. Thus, the average payoff for each of the three strategic types in an arbitrary group I is

$$\begin{aligned}
\Pi_I^X &= \frac{1}{\mathcal{M}_I}(1-u_x)\left\{\sum_J \nu_J \omega_{I,J}\left[b_{I,J}(f_J^X + f_J^Z g_{I,J}^X) - c_{I,J}\right]\right\} \\
\Pi_I^Y &= \frac{1}{\mathcal{M}_I}(1-u_x)\left\{\sum_J \nu_J \omega_{I,J}\left[b_{I,J}(f_J^X + f_J^Z g_{I,J}^Y)\right]\right\} \\
\Pi_I^Z &= \frac{1}{\mathcal{M}_I}(1-u_x)\left\{\sum_J \nu_J \omega_{I,J}\left[b_{I,J}(f_J^X + f_J^Z g_{I,J}^Z) - c_{I,J} g_{J,I}\right]\right\}.
\end{aligned}$$

5.1.6. Hybrid strategies. The above treatment makes it trivial to write down reputations for strategies that distinguish explicitly between in-group and out-group, for example, cooperate with one's in-group and discriminate with one's out-group. These "hybrid" strategies were prominently featured in, e.g., (6). As an example, we present the strategy of discriminating with one's in-group and defecting with one's out-group, which we denote ZY. We have

$$\begin{aligned} g_{ZY}^{I,J} &= \delta_{I,J}\big[\gamma_{I,J}P_J^{GC} + (1-\gamma_{I,J})P_J^{BD}\big] \\ &\quad + (1-\delta_{I,J})\big[\gamma_{I,J}P_J^{GD} + (1-\gamma_{I,J})P_J^{BD}\big] \\ &= \delta_{I,J}\big[\gamma_{I,J}P_J^{GC} + (1-\gamma_{I,J})P_J^{BD}\big] \\ &\quad + (1-\delta_{I,J})\big[\gamma_{I,J}P_J^{GD} + (1-\gamma_{I,J})P_J^{BD}\big]. \end{aligned}$$

The fitness term is easily written down depending on the other strategies in the population; a ZY individual accrues a benefit from cooperators, from discriminators who see them as good, and from other ZY individuals *in the same group* who see them as good, and they pay the cost for any individuals *in the same group* whom their group sees as good.

5.1.7. Favoring of in-group interactions. Suppose we have

$$\omega_{I,J} = \delta_{I,J} + (1-\delta_{I,J})\omega,$$

i.e., individuals always interact with their in-group, but out-group interactions only happen with probability ω. Suppose, also, that benefits and costs do not vary by group. We can study this numerically: see main text figure 3.

When $\omega \to 0$, we have

$$\begin{aligned} \mathcal{M}_I &= \sum_L \nu_L \omega_{I,L} = \nu_I, \\ \gamma_{I,J} &= \frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} g_{L,J} = \frac{1}{\mathcal{M}_I}\nu_I g_{I,J} = g_{I,J}, \\ \Pi_I^X &= \frac{1}{\mathcal{M}_I}(1-u_x)\Big\{\sum_J \nu_J\omega_{I,J}\big[b(f_J^X + f_J^Z g_{I,J}^X) - c\big]\Big\} = (1-u_x)[b(f_I^X + f_I^Z g_{I,I}^X) - c], \\ \Pi_I^Y &= \frac{1}{\mathcal{M}_I}(1-u_x)\Big\{\sum_J \nu_J\omega_{I,J}\big[b(f_J^X + f_J^Z g_{I,J}^Y)\big]\Big\} = (1-u_x)[b(f_I^X + f_I^Z g_{I,I}^Y)], \\ \Pi_I^Z &= \frac{1}{\mathcal{M}_I}(1-u_x)\Big\{\sum_J \nu_J\omega_{I,J}\big[b(f_J^X + f_J^Z g_{I,J}^Z) - c g_{J,I}\big]\Big\} = (1-u_x)[b(f_I^X + f_I^Z g_{I,I}^Z) - c g_{I,I}]. \end{aligned}$$

Thus, when groups are completely insular, they only accrue payoffs from (and pay costs for) interactions with their own group: K groups effectively behave as K completely disjoint, independent populations.

5.2. Average fitnesses under well-mixed copying. In SI Section 3, we saw that it was possible to study the dynamics of strategy evolution solely by considering the group-averaged fitnesses and reputations. Here, we show that this is likewise possible when insularity is introduced. We have

$$\begin{aligned} \Pi^Z &= \sum_I \nu_I \Pi_I^Z \\ &= (1-u_x)\sum_I \nu_I \frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J}\big[b(f^X + f^Z g_{I,J}^Z) - c g_{J,I}\big] \\ &= (1-u_x)\Big\{b\sum_I \nu_I\frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J}(f^X + f^Z g_{I,J}^Z) - \sum_I \nu_I\frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J}g_{J,I}\Big\} \\ &= (1-u_x)\Big[bf^X\sum_I \nu_I\frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J} + bf^Z\sum_I \nu_I\frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J}g_{I,J}^Z - \sum_I \nu_I\frac{1}{\mathcal{M}_I}\sum_J \nu_J\omega_{I,J}g_{J,I}\Big] \\ &= (1-u_x)\Big[bf^X\sum_I \nu_I\frac{1}{\sum_J \nu_J\omega_{I,J}}\sum_J \nu_J\omega_{I,J} + bf^Z\sum_I \nu_I\frac{1}{\sum_J \nu_J\omega_{I,J}}\sum_J \nu_J\omega_{I,J}g_{I,J}^Z \\ &\quad - \sum_I \nu_I\frac{1}{\sum_J \nu_J\omega_{I,J}}\sum_J \nu_J\omega_{I,J}g_{J,I}\Big] \\ &= (1-u_x)\Big[b(f^X + f^Z g^Z) - c\hat{g}\Big]. \end{aligned}$$

By identical reasoning we will have

$$\begin{aligned}\Pi^X &= (1-u_x)\Big[b(f^X + f^Z g^X) - c\Big],\\ \Pi^Y &= (1-u_x)\Big[b(f^X + f^Z g^Y)\Big].\end{aligned}$$

5.3. Equally sized groups. Suppose now that every group has the same size $1/K$ and the insularities are given by $\omega_{I,J} = \delta_{I,J} + (1-\delta_{I,J})\omega$, i.e., individuals always interact with fellow in-group members but only interact with out-group individuals with probability ω. By symmetry, reputational views will only differ depending on whether the observer is in the donor's in-group or out-group. Define $g_{\text{in}} = g_{I,I}$ and $g_{\text{out}} = g_{I,J}|_{I\neq J}$, i.e., g_{in} is an individual's view of their in-group and g_{out} their out-group. The average reputation g (Eq. [14]) can be expanded out thus:

$$\begin{aligned} g = \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} g_{I,J} &= \left(\frac{1}{K}\right)^2 \frac{1}{1/K + \omega(K-1)/K} \sum_I \sum_J \omega_{I,J} g_{I,J} \\ &= \left(\frac{1}{K}\right)^2 \frac{K}{1+\omega(K-1)} (K g_{\text{in}} + \omega K(K-1) g_{\text{out}}) \\ &= \frac{g_{\text{in}} + \omega(K-1) g_{\text{out}}}{1+\omega(K-1)}. \end{aligned} \tag{15}$$

In a population of discriminators, g_{in} and g_{out} can be expanded out:

$$\begin{aligned} g_{\text{in}} &= \gamma_{I,I}(P^{GC} - P^{BD}) + P^{BD} \\ &= \frac{1}{\mathcal{M}_I} \sum_L \nu_L \omega_{L,I} g_{L,I} (P^{GC} - P^{BD}) + P^{BD} \\ &= \frac{1}{1+\omega(K-1)} \sum_L \omega_{L,I} g_{L,I} (P^{GC} - P^{BD}) + P^{BD} \\ &= \frac{1}{1+\omega(K-1)} \big[g_{\text{in}} + (K-1)\omega g_{\text{out}}\big] (P^{GC} - P^{BD}) + P^{BD} \end{aligned} \tag{16}$$

and

$$\begin{aligned} g_{\text{out}} &= \gamma_{I,J}(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + \gamma_{I,J}(P^{GD} - P^{BD}) + \gamma_{I,I}(P^{BC} - P^{BD}) + P^{BD} \\ &= \frac{1}{\mathcal{M}_I} \sum_L \nu_L \omega_{L,I} g_{L,I} g_{L,J} (P^{GC} - P^{GD} - P^{BC} + P^{BD}) \\ &\quad + \frac{1}{\mathcal{M}_I} \sum_L \nu_L \omega_{L,I} g_{L,J} (P^{GD} - P^{BD}) \\ &\quad + \frac{1}{\mathcal{M}_I} \sum_L \nu_L \omega_{L,I} g_{L,I} (P^{BC} - P^{BD}) \\ &\quad + P^{BD} \\ &= \frac{1}{1+\omega(K-1)} \big[(1+\omega) g_{\text{in}} g_{\text{out}} + (K-2)\omega (g_{\text{out}})^2\big] (P^{GC} - P^{GD} - P^{BC} + P^{BD}) \\ &\quad + \frac{1}{1+\omega(K-1)} \Big[\omega g_{\text{in}} + \big(1 + (K-2)\omega\big) g_{\text{out}}\Big] (P^{GD} - P^{BD}) \\ &\quad + \frac{1}{1+\omega(K-1)} \big[g_{\text{in}} + (K-1)\omega g_{\text{out}}\big] (P^{BC} - P^{BD}) \\ &\quad + P^{BD}. \end{aligned} \tag{17}$$

These equations can be solved, but their general solution is not very informative. Note that sending $u_x \to 0$ and defining $\theta = 1/[1+\omega(K-1)]$ yields the corresponding expressions from (2); see SI table S3. A useful simplification is to send

$\omega \to 0$:

$$\begin{aligned}
g_{\text{in}}|_{\omega\to 0} &= \frac{P^{BD}}{1-P^{GC}+P^{BD}} \\
&= \begin{cases} \dfrac{u_a}{1-\epsilon+u_a} = \dfrac{u_a}{2u_a+u_x-2u_xu_a} & \textit{Shunning, Scoring}, \\ \dfrac{1-u_a}{2-\epsilon+u_a} = \dfrac{1-u_a}{1+u_x-2u_xu_a} & \textit{Stern Judging, Simple Standing}, \end{cases} \\
g_{\text{out}}|_{\omega\to 0} &= \frac{P^{BD}(1+P^{BC}-P^{GC})}{P^{BD}(2+P^{BC}-2P^{GC})+(1-P^{GC})(1-P^{GD})} \\
&= \begin{cases} \dfrac{u_a(1+u_a-\epsilon)}{u_a(2+u_a-2\epsilon)+(1-\epsilon)(1-u_a)} = \dfrac{u_a[2u_a(1-u_x)+u_x]}{u_a(1+2u_a)(1-u_x)+u_x} & \textit{Shunning}, \\ \dfrac{u_a}{1-\epsilon+u_a} = \dfrac{u_a}{2u_a(1-u_x)+u_x} & \textit{Scoring}, \\ \dfrac{(1-u_a)(2-2\epsilon)}{(1-u_a)(4-4\epsilon)} = \dfrac{1}{2} & \textit{Stern Judging}, \\ \dfrac{(1-u_a)(2-u_a-\epsilon)}{(1-u_a)(3-u_a-2\epsilon)+(1-\epsilon)(1-u_a)} = \dfrac{1+u_x-2u_au_x}{1+2u_a+3u_x-6u_au_x} & \textit{Simple Standing}. \end{cases}
\end{aligned} \tag{18}$$

The expressions for $g_{\text{in}}|_{\omega\to 0}$ are the same as the main text expressions for g with $K = 1$ and $f^Z = 1$.

5.4. Invasibility of equally sized groups by defectors. Given Eqs. [16] and [17], we can determine when discriminators resist invasion by defectors. We require (when $f_Z = 1$)

$$\begin{aligned}
\Pi^Z &> \Pi^Y \\
bg^Z - c\hat{g} &> bg^Y \\
(b-c)g &> bg^Y \\
\therefore \frac{b}{c} &> \frac{g}{g-g^Y}, \text{with} \\
g^Y &= \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} \big[\gamma_{I,J}(P^{GD}-P^{BD})+P^{BD}\big].
\end{aligned} \tag{19}$$

Solving for g^Y requires that we compute the sum at the end of Eq. [19]. Let ω_{out} be the out-group interaction parameter and ω_{in} the probability of in-group interactions (these are ω and 1 respectively; these terms are used solely for bookkeeping).

norm	g_{in}	g_{out}
Stern Judging $(K \geq 3)$	$\frac{2(1-u_a)(1+[K-1]\omega) - u_x(2+[K-1]\omega)}{2(1+[K-1]\omega)}$	$\frac{K-2-u_x}{2(K-2)}$
Stern Judging $(K = 2, u_x > 0)$	$1 - \frac{2u_x + (2+\omega)u_a}{2\sqrt{1+\omega}}$	$\frac{1}{4\omega\sqrt{1+\omega}}$ $\times \Big\{ 4(1+\omega-\sqrt{1+\omega}) + 2u_x(2+\omega-2\sqrt{1+\omega})$ $+u_a[8-8\sqrt{1+\omega}+\omega(8+\omega-4\sqrt{1+\omega})] \Big\}$
Stern Judging $(K = 2, u_x = 0)$	$1-u_a$	$1/2$
Simple Standing	$1-u_a-u_x$	$1-(u_a+u_x)[2+(K-1)\omega]$
Shunning	$\frac{u_a(1+2[K-1]\omega)}{(K-1)\omega}$	u_a

Table S3. In-group and out-group reputations for all-discriminator populations consisting of K equally sized groups, to *first order* in u_a and u_x. These agree with the expressions provided in (2) given $\theta = 1/[1+\omega(K-1)]$ and $u_x \to 0$. The expression for *Shunning* is novel to our analysis: in their notation, the *Shunning* in-group reputation is $(2-\theta)u_a/(1-\theta)$, and both the in-group and out-group reputations include $\mathcal{O}(u_a^2)$ terms that we ignore. The *Shunning* in-group approximation breaks down as $\omega \to 0$; it appears to be valid only for $\omega^r \gg u_x$. Otherwise, the exact $\omega = 0$ expression (Eq. [18]) is a useful approximations for g_{in}. Likewise, the *Stern Judging* ($K = 2, u_x > 0$) out-group term fails when $\omega \gg u_x$. The $u_x = 0$ out-group expression is a better approximation. Finally, *Scoring* is omitted from this table because the exact expression $g = u_a/(1-\epsilon+u_a)$ is always valid irrespective of group structure and insularity.

The sum is

$$\begin{aligned}
\sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} \gamma_{I,J} &= \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_L \nu_L \omega_{L,I} g_{L,J} \\
&= \sum_I \sum_I \sum_J \sum_L (\nu_I)^2 \Big(\frac{1}{\mathcal{M}_I}\Big)^2 \nu_J \nu_L \omega_{I,J} \omega_{L,I} g_{L,J} \\
&= \Big(\frac{1}{K}\Big)^4 \Big(\frac{1}{1/K + \omega(K-1)/K}\Big)^2 \sum_I \sum_I \sum_J \sum_L \omega_{I,J} \omega_{L,I} g_{L,J} \\
&= \Big(\frac{1}{K}\Big)^4 \Big(\frac{K}{1+\omega(K-1)}\Big)^2 \\
&\quad \times \sum_I \Big[K(K-1)(K-2)\omega_{\text{out}}\omega_{\text{out}} g_{\text{out}} + K(K-1)\omega_{\text{in}}\omega_{\text{out}} g_{\text{out}} \\
&\quad + K(K-1)\omega_{\text{out}}\omega_{\text{in}} g_{\text{out}} + K(K-1)\omega_{\text{out}}\omega_{\text{out}} g_{\text{in}} + K\omega_{\text{in}}\omega_{\text{in}} g_{\text{in}} \\
&= \Big(\frac{1}{K}\Big)^4 \Big(\frac{K}{1+\omega(K-1)}\Big)^2 \\
&\quad \times \sum_I \Big[K(K-1)(K-2)\omega^2 g_{\text{out}} + 2K(K-1)\omega g_{\text{out}} \\
&\quad + K(K-1)\omega^2 g_{\text{in}} + K g_{\text{in}} \Big] \\
&= \Big(\frac{1}{K}\Big)^3 \Big(\frac{K}{1+\omega(K-1)}\Big)^2 \\
&\quad \times \Big[K(K-1)(K-2)\omega^2 g_{\text{out}} + 2K(K-1)\omega g_{\text{out}} + K(K-1)\omega^2 g_{\text{in}} + K g_{\text{in}} \Big] \\
&= \Big(\frac{1}{1+\omega(K-1)}\Big)^2 \\
&\quad \times \Big[(K-1)(K-2)\omega^2 g_{\text{out}} + 2(K-1)\omega g_{\text{out}} + (K-1)\omega^2 g_{\text{in}} + g_{\text{in}} \Big] \\
&= \frac{g_{\text{out}}\omega(K-1)[(K-2)\omega+2] + g_{\text{in}}[(K-1)\omega^2+1]}{[1+\omega(K-1)]^2}.
\end{aligned} \qquad [20]$$

Thus,

$$g^Y = \frac{g_{\text{out}}\omega(K-1)[(K-2)\omega+2] + g_{\text{in}}[(K-1)\omega^2+1]}{[1+\omega(K-1)]^2}(P^{GD} - P^{BD}) + P^{BD}. \qquad [21]$$

Substituting Eqs. [21], [15], [16], and [17] into Eq. [19] yields the condition that b/c must be greater than a fraction whose numerator is given by

$$[1 + (K-1)\omega][g_{\text{in}} + g_{\text{out}}(K-1)\omega]$$

and denominator by

$$\begin{aligned}
&g_{\text{in}}\big[\big((K-1)\omega^2+1\big)(P^{BD} - P^{GD}) + (K-1)\omega + 1\big] \\
&\quad + g_{\text{out}}(K-1)\omega\big[\big((K-2)\omega+2\big)(P^{BD} - P^{GD}) + (K-1)\omega + 1\big] \\
&\quad - P^{BD}\big[(K-1)\omega+1\big]^2.
\end{aligned}$$

These are consistent with the expressions from (2). The b/c condition can also be expressed in terms of the weighted average reputation g. We solve for g_{in} and g_{out} self-consistently via

$$\begin{aligned} g_{\text{in}}|_{f^Z=1} &= \gamma_{I,I}(P^{GC} - P^{BD}) + P^{BD} \\ &= \Big(\frac{1}{\mathcal{M}_I}\sum_L \nu_L \omega_{L,I} g_{L,I}\Big)(P^{GC} - P^{BD}) + P^{BD} \\ &= \frac{K}{1+\omega(K-1)}\Big(\frac{1}{K}g_{\text{in}} + \omega\frac{K-1}{K}g_{\text{out}}\Big)(P^{GC} - P^{BD}) + P^{BD} \\ &= \frac{g_{\text{in}} + \omega(K-1)g_{\text{out}}}{1+\omega(K-1)}(P^{GC} - P^{BD}) + P^{BD} \\ \therefore g_{\text{in}}\Big(1 - \frac{1}{1+\omega(K-1)}\Big) &= \frac{\omega(K-1)g_{\text{out}}}{1+\omega(K-1)}(P^{GC} - P^{BD}) + P^{BD} \\ \therefore g_{\text{in}}\frac{\omega(K-1)}{1+\omega(K-1)} &= \frac{\omega(K-1)g_{\text{out}}}{1+\omega(K-1)}(P^{GC} - P^{BD}) + P^{BD} \\ \therefore g_{\text{in}} &= g_{\text{out}}(P^{GC} - P^{BD}) + \frac{1+\omega(K-1)}{\omega(K-1)}P^{BD} \end{aligned} \tag{22}$$

or, equivalently,

$$g_{\text{in}} = g(P^{GC} - P^{BD}) + P^{BD}.$$

Likewise

$$\begin{aligned} \frac{g_{\text{in}} + \omega(K-1)g_{\text{out}}}{1+\omega(K-1)} &= g \\ \therefore \frac{\omega(K-1)g_{\text{out}}}{1+\omega(K-1)} &= g - \frac{g_{\text{in}}}{1+\omega(K-1)} \\ \therefore g_{\text{out}} &= \frac{g[1+\omega(K-1)] - g_{\text{in}}}{\omega(K-1)} \\ &= \frac{g[1+\omega(K-1) - P^{GC} + P^{BD}] - P^{BD}}{\omega(K-1)} \end{aligned}$$

Combining Eqs. [22], [20], and [13] yields

$$\begin{aligned} &(b-c)g > b\Big[\frac{g_{\text{out}}\omega(K-1)[(K-2)\omega+2] + g_{\text{in}}[(K-1)\omega^2+1]}{[1+\omega(K-1)]^2}(P^{GD} - P^{BD}) + P^{BD}\Big] \\ &\therefore \frac{b-c}{b}g > \frac{\{g[1+\omega(K-1) - P^{GC} + P^{BD}] - P^{BD}\}[(K-2)\omega+2] + [g(P^{GC} - P^{BD}) + P^{BD}][(K-1)\omega^2+1]}{[1+\omega(K-1)]^2} \\ &\qquad\quad \times \left(P^{GD} - P^{BD}\right) + P^{BD} \\ &\therefore \frac{b-c}{b}g > g\frac{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{GD} - P^{BD})}{1+\omega(K-1)} + \frac{(1-\omega)P^{BD}(P^{GD} - P^{BD})}{1+\omega(K-1)} + P^{BD}. \end{aligned}$$

This can be rearranged to yield

$$g\Big(1 - \frac{c}{b} + \frac{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{BD} - P^{GD})}{1+\omega(K-1)}\Big) > \frac{(1-\omega)P^{BD}(P^{GD} - P^{BD})}{1+\omega(K-1)} + P^{BD}.$$

Thus, discriminators resist invasion by defectors provided

$$\begin{aligned} g &> \frac{(1-\omega)P^{BD}(P^{BD} - P^{GD}) + [1+\omega(K-1)]P^{BD}}{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{BD} - P^{GD}) + (1-c/b)[1+\omega(K-1)]}, \text{ or} \\ g &> \frac{P^{BD}[K\omega + (1-\omega)(1 - P^{GD} + P^{BD})]}{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{BD} - P^{GD}) + (1-c/b)[1+\omega(K-1)]}. \end{aligned}$$

Setting $\omega = 1$ yields Eq. [5].

5.5. Invasibility of equally sized groups by cooperators. Bolstered by our preceding analysis, we also consider when discriminators resist invasion by cooperators; such invasion can occur, e.g., under *Simple Standing*. We require

$$\begin{aligned}
\Pi^Z &> \Pi^X \\
(b-c)g &> bg^X - c \\
\frac{b}{c} &> \frac{g-1}{g-g^X}, \text{with} \\
g^X &= \sum_I \nu_I \frac{1}{\mathcal{M}_I} \sum_J \nu_J \omega_{I,J} \big[\gamma_{I,J}(P^{GC} - P^{BC}) + P^{BC}\big] \\
&= \frac{g_{\text{out}}\omega(K-1)[(K-2)\omega+2] + g_{\text{in}}[(K-1)\omega^2+1]}{[1+\omega(K-1)]^2}(P^{GC} - P^{BC}) + P^{BC}.
\end{aligned}$$

The critical b/c value thus simplifies to a fraction whose numerator is given by

$$[1+(K-1)\omega][g_{\text{in}} + (g_{\text{out}}-1)(K-1)\omega - 1]$$

and denominator by

$$\begin{aligned}
&g_{\text{in}}\Big[\big((K-1)\omega^2+1\big)(P^{BC} - P^{GC}) + (K-1)\omega + 1\Big] \\
&\quad + g_{\text{out}}(K-1)\omega\Big[\big((K-2)\omega+2\big)(P^{BC} - P^{GC}) + (K-1)\omega + 1\Big] \\
&\quad - P^{BC}\big[(K-1)\omega+1\big]^2,
\end{aligned}$$

which again is consistent with (2). We can likewise express this condition in terms of g:

$$\begin{aligned}
(b-c)g &> b\Big[\frac{g_{\text{out}}\omega(K-1)[(K-2)\omega+2] + g_{\text{in}}[(K-1)\omega^2+1]}{[1+\omega(K-1)]^2}(P^{GC} - P^{BC}) + P^{BC}\Big] - c \\
\therefore \frac{b-c}{b}g &> \frac{\{g[1+\omega(K-1) - P^{GC} + P^{BD}] - P^{BD}\}[(K-2)\omega+2] + [g(P^{GC} - P^{BD}) + P^{BD}][(K-1)\omega^2+1]}{[1+\omega(K-1)]^2} \\
&\qquad \times \big(P^{GC} - P^{BC}\big) + P^{BC} - \frac{c}{b} \\
\therefore \frac{b-c}{b}g &> g\frac{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{GC} - P^{BC})}{1+\omega(K-1)} + \frac{(1-\omega)P^{BD}(P^{GC} - P^{BC})}{1+\omega(K-1)} + P^{BC} - \frac{c}{b}.
\end{aligned}$$

This can be rearranged to yield

$$g\Big(1 - \frac{c}{b} + \frac{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{BC} - P^{GC})}{1+\omega(K-1)}\Big) > \frac{(1-\omega)P^{BD}(P^{GC} - P^{BC})}{1+\omega(K-1)} + P^{BC} - \frac{c}{b}.$$

Thus, discriminators resist invasion by cooperators provided

$$g > \frac{(1-\omega)P^{BD}(P^{GC} - P^{BC}) + [1+\omega(K-1)](P^{BC} - c/b)}{[2 - P^{GC} + P^{BD} + \omega(K-2+P^{GC} - P^{BD})](P^{BC} - P^{GC}) + (1-c/b)[1+\omega(K-1)]}$$

for *Stern Judging* and *Shunning*. For *Scoring* and *Simple Standing*, the sign of the inequality is reversed, as the denominator is negative.

5.6. Norm competition with insularity and variable costs and benefits. If benefits and costs vary, so that b^{in} and c^{in} are the benefit and cost associated with an intra-group interaction and b^{out} and c^{out} are the cost associated with an inter-group interaction, then ν_1 grows if

$$\begin{aligned}
\dot{\nu}_1 &> 0 \\
\therefore \Pi_1^Z &> \Pi_2^Z \\
\therefore \frac{1}{\mathcal{M}_1}\big[\nu_1(b^{\text{in}}g_{1,1} - c^{\text{in}}g_{1,1}) + \nu_2\omega(b^{\text{out}}g_{1,2} - c^{\text{out}}g_{2,1})\big] &> \frac{1}{\mathcal{M}_2}\big[\nu_1\omega(b^{\text{out}}g_{2,1} - c^{\text{out}}g_{1,2}) + \nu_2(b^{\text{in}}g_{2,2} - c^{\text{in}}g_{2,2})\big] \\
\therefore \nu_1\Big[\frac{1}{\mathcal{M}_1}(b^{\text{in}}g_{1,1} - c^{\text{in}}g_{1,1}) - \frac{1}{\mathcal{M}_2}\omega(b^{\text{out}}g_{2,1} - c^{\text{out}}g_{1,2})\Big] &> \nu_2\Big[\frac{1}{\mathcal{M}_2}(b^{\text{in}}g_{2,2} - c^{\text{in}}g_{2,2}) - \frac{1}{\mathcal{M}_1}\omega(b^{\text{out}}g_{1,2} - c^{\text{out}}g_{2,1})\Big] \\
\therefore \nu_1\big[\mathcal{M}_2(b^{\text{in}}g_{1,1} - c^{\text{in}}g_{1,1}) - \mathcal{M}_1\omega(b^{\text{out}}g_{2,1} - c^{\text{out}}g_{1,2})\big] &> \nu_2\big[\mathcal{M}_1(b^{\text{in}}g_{2,2} - c^{\text{in}}g_{2,2}) - \mathcal{M}_2\omega(b^{\text{out}}g_{1,2} - c^{\text{out}}g_{2,1})\big] \\
\therefore \frac{\nu_1}{\nu_2} &> \frac{\mathcal{M}_1(b^{\text{in}}g_{2,2} - c^{\text{in}}g_{2,2}) - \mathcal{M}_2\omega(b^{\text{out}}g_{1,2} - c^{\text{out}}g_{2,1})}{\mathcal{M}_2(b^{\text{in}}g_{1,1} - c^{\text{in}}g_{1,1}) - \mathcal{M}_1\omega(b^{\text{out}}g_{2,1} - c^{\text{out}}g_{1,2})}
\end{aligned}$$

When we hold c^{in} and c^{out} constant but set $b^{\text{out}} > b^{\text{in}}$, it becomes *harder* for *Stern Judging* (group 1) to beat *Shunning* (group 2). When $\nu_1 = 1/2$, we have (for $u_a = u_x = .02$) $g_{1,1} = .96, g_{2,1} = .42, g_{1,2} = .09, g_{2,2} = .13$. What this means is that group 1 cooperates with group 2 more than the reverse, so increases in b^{out} are not reciprocated. Thus, raising b^{out} rather than b^{in} increases the rate of unreciprocated fitness gain by group 2, allowing it to outcompete group 1.

5.7. Dependence of group fitness on insularity. In this section, we consider how the fitness of a group of discriminators depends on the insularity parameter ω, under the assumptions of well-mixed copying, K equally sized groups, and $\omega_{I,J} = \delta_{I,J} + (1 - \delta_{I,J})\omega$. This can shine light on how insularity might be expected to evolve in a group-level selection scenario, in which an entire group can be replaced by a group with a different level of insularity. (The problem of how insularity might evolve at the individual level is deferred to Section 5.8.) We have

$$\begin{aligned}
\mathcal{M}_I &= \sum_L \nu_L \omega_{I,L} \\
&= \frac{1 + \omega(K-1)}{K} = \mathcal{M} \text{ (no dependence on } I\text{)} \\
\therefore \Pi^Z &= \sum_I \nu_I \Pi_I^Z \\
&= \sum_I \nu_I \frac{1}{\mathcal{M}_I} \Big\{ \sum_J \nu_J \omega_{I,J} \big[b g_{I,J} - c g_{J,I} \big] \Big\} \text{ (dropping the } (1 - u_x) \text{ prefactor)} \\
&= \frac{1}{\mathcal{M}} \frac{1}{K} \big[b g_{\text{in}} - c g_{\text{in}} + (K-1)\omega(b g_{\text{out}} - c g_{\text{out}}) \big] \\
&= \frac{1}{1 + \omega(K-1)} (b - c) \big[g_{\text{in}} + \omega(K-1) g_{\text{out}} \big] \\
&= (b - c) g_{\text{in}} \frac{1 + \omega(K-1) g_{\text{out}}/g_{\text{in}}}{1 + \omega(K-1)}.
\end{aligned}$$

We have written this in a suggestive form. Because the $\omega(K-1)$ term in the numerator has a factor $g_{\text{out}}/g_{\text{in}}$ attached to it, and because this factor is (for every social norm besides *Scoring*) less than 1, the numerator will generally shrink relative to the denominator as ω increases and grow as ω decreases. The fitness of a group thus generally increases with decreasing ω. This sensitively depends on our decision to normalize by dividing by $\mathcal{M}$, which is necessary to ensure that interactions that do not happen make no contribution to fitness (instead of, e.g., contributing zero fitness, which would be indistinguishable from mutual defection). If we did not divide by $\mathcal{M}$, we would instead have

$$\begin{aligned}
\Pi^Z &= \frac{b - c}{K} \big[g_{\text{in}} + \omega(K-1) g_{\text{out}} \big] \\
&= \left(\frac{(b - c) g_{\text{in}}}{K} \right) \big[1 + \omega(K-1) g_{\text{out}}/g_{\text{in}} \big],
\end{aligned}$$

which is monotonically increasing in ω. Finally, if we allow out-group and in-group interactions to have different payoffs, we have

$$\begin{aligned}
\Pi^Z &= \frac{1}{\mathcal{M}} \frac{1}{K} \big[b_{\text{in}} g_{\text{in}} - c_{\text{in}} g_{\text{in}} + \omega(K-1)(b_{\text{out}} g_{\text{out}} - c_{\text{out}} g_{\text{out}}) \big] \\
&= \frac{1}{1 + \omega(K-1)} \big[(b_{\text{in}} - c_{\text{in}}) g_{\text{in}} + \omega(K-1)(b_{\text{out}} - c_{\text{out}}) g_{\text{out}} \big] \\
&= \frac{(b_{\text{in}} - c_{\text{in}}) g_{\text{in}} + \omega(K-1)(b_{\text{out}} - c_{\text{out}}) g_{\text{out}}}{1 + \omega(K-1)} \\
&= (b_{\text{in}} - c_{\text{in}}) g_{\text{in}} \frac{1 + \omega(K-1)(b_{\text{out}} - c_{\text{out}}) g_{\text{out}}/[(b_{\text{in}} - c_{\text{in}}) g_{\text{in}}]}{1 + \omega(K-1)}
\end{aligned}$$

The relevant ratio is now $(b_{\text{out}} - c_{\text{out}}) g_{\text{out}}/[(b_{\text{in}} - c_{\text{in}}) g_{\text{in}}]$. If this ratio is greater than 1 (for example, because out-group interactions are more rewarding than in-group interactions), it is possible for insularity to be selected against at the group level, as higher ω results in increased fitness.

5.8. Evolution of insularity. When groups are fixed and only strategies evolve, we saw that insularity can defray the destabilizing effects of group structure on cooperation. This raises the question of how insularity itself will evolve, in this setting. To study this, we first analyze the effects of a fixed level of insularity on fitness in a group-structured population, finding that more insular populations which prefer in-group interactions generally enjoy greater mean fitness in each group. This happens because high insularity increases the rate of interactions with in-group members who are most likely to share reputational views. We also performed an invasibility analysis to determine whether a mutant with a higher level of insularity can spread in a resident population that is less insular. For all norms in which fitness has any dependence on group identity, we find that a more insular mutant can always invade a less insular resident, so that a population will always evolve towards greater insularity. However, if out-group social interactions are potentially more rewarding than in-group interactions (e.g., $b^{\text{out}} > b^{\text{in}}$), then we will see that a population may resist invasion by insular types, or it may evolve to stable intermediate levels of insularity.

To study the evolution of insularity, we consider a population fixed for discriminators, but the resident population is facing potential invasion by a mutant discriminator with a different level of insularity. Residents have out-group interaction parameter ω^r, and mutants have out-group interaction parameter ω^m; both mutants always interact with their in-groups. We set $\omega^m < \omega^r$, i.e., the mutant is more insular (less likely to engage in out-group interactions) than the resident, and we posit that potential interactions between out-group mutants and residents occur only with probability $\psi(\omega^r, \omega^m)$. Two natural forms for $\psi(\omega^r, \omega^m)$ are

$$\begin{aligned} \psi(\omega^r, \omega^m) &= \frac{\omega^r + \omega^m}{2} \text{ (the arithmetic mean)} \\ \psi(\omega^r, \omega^m) &= \sqrt{\omega^r \omega^m} \text{ (the geometric mean)}. \end{aligned}$$

The geometric mean formulation has the advantage that $\sqrt{\omega^s}$ can be thought of as the probability that an individual of type s proposes an out-group interaction and $\sqrt{\omega_{s'}}$ the probability that their out-group partner accepts; however, it is more difficult to work with than the arithmetic mean. Thus, we use the arithmetic mean for the remainder of this analysis; using the geometric mean instead effects minor quantitative but not qualitative changes.

norm	g^m_{in}
Stern Judging $(K \geq 3)$	$\frac{2(1-u_x-u_a)+(K-1)(2-2u_a-u_x)\psi(\omega^r,\omega^m)}{2(1+[K-1]\psi(\omega^r,\omega^m))}$
Stern Judging $(K = 2, u_x > 0)$	$\frac{2(1-u_a-u_x)\omega^r+[2(u_x[1-\sqrt{1+\omega^r}]+\omega^r)+u_a(2-(2+\omega^r)\sqrt{1+\omega^r})]\psi(\omega^r,\omega^m)}{2\omega^r(1+\psi(\omega^r,\omega^m))}$
Stern Judging $(K = 2, u_x = 0)$	$1 - u_a$
Simple Standing	$1 - u_a - u_x$
Shunning	$\frac{u_a(1+3[K-1]\omega^r+2\omega^r\psi(\omega^r,\omega^m)[K-1]^2)}{(K-1)\omega^r(1+[K-1]\psi(\omega^r,\omega^m))}$

Table S4. Approximate in-group reputations for a *mutant* with different out-group interaction parameter ω^m from the resident; in this scenario, the mutant is invading a population consisting of K equally sized groups under well-mixed strategic imitation. Expressions are to first order in u_a and u_x. See the caveats in the caption of table S3.

We assume well-mixed copying and K groups of equal size $1/K$. Let f_r be the frequency of the resident and f_m the frequency of the mutant; because we are concerned with the invasibility of the mutant, we will set $f_r = 1$. The total number of interactions the two types engage in is

$$\begin{aligned} \mathcal{M}^r_I &= \sum_J \nu_J \left[f^r \omega^r_{I,J} + f^m \psi(\omega^r_{I,J}, \omega^m_{I,J}) \right] \\ &= \frac{1 + \omega^r (K-1)}{K}, \\ \mathcal{M}^m_I &= \sum_J \nu_J \big[f^r \Big[\psi(\omega^r_{I,J}, \omega^m_{I,J}) + f^m \omega^m_{I,J} \big] \\ &= \frac{1 + \psi(\omega^r, \omega^m)(K-1)}{K}. \end{aligned}$$

We will drop the I subscript moving forward, as there is no dependence on I. The fitnesses of the resident and mutant are

norm	g^m_{out}
Stern Judging $(K \geq 3)$	$\dfrac{K-2-u_x}{2(K-2)}$
Stern Judging $(K=2, u_x > 0)$	$\dfrac{1}{4\omega\sqrt{1+\omega^r}(1+\psi(\omega^r,\omega^m))}$ $\times\Big\{4(1+\omega^r-\sqrt{1+\omega^r})+u_a(8(1-\sqrt{1+\omega^r})+\omega^r(4-\omega^r))$ $+u_x(4(1-\sqrt{1+\omega^r}+\omega^r(4\sqrt{1+\omega^r}-2)$ $+\Big[4(1+\omega^r-\sqrt{1+\omega^r})+2u_x(4(1-\sqrt{1+\omega^r})+\omega^r)$ $+u_a\omega^r(10+\omega^r-4\sqrt{1+\omega^r})+12u_a(1-\sqrt{\omega^r})\Big]\psi(\omega^r,\omega^m)\Big\}$
Stern Judging $(K=2, u_x = 0)$	$1/2$
Simple Standing	$\dfrac{1-2(u_a+u_x)+(K-1)[1-3u_x-3u_a-(K-1)(u_x+u_a)\omega^r]\psi(\omega^r,\omega^m)}{1+(K-1)\psi(\omega^r,\omega^m)}$
Shunning	u_a

Table S5. Approximate out-group reputations for a *mutant* with different out-group interaction parameter ω^m from the resident; in this scenario, the mutant is invading a population consisting of K equally sized groups under well-mixed strategic imitation. Expressions are to first order in u_a and u_x. See the caveats in the caption of table S3.

given respectively by

$$\begin{aligned}
\Pi^r|_{f^r=1} &= (1-u_x)\sum_I \frac{1}{\mathcal{M}^r}\Big\{\sum_J \nu_J \sum_s f^s_J \psi(\omega^r_{I,J}\omega^s_{I,J})\big[b_{I,J}g^r_{I,J} - cg^s_{J,I}\big]\Big\}\Big|_{f_r=1} \\
&= (1-u_x)\frac{1}{1+\omega^r(K-1)}\sum_I\Big\{\sum_J \omega^r_{I,J}\big[b_{I,J}g^r_{I,J} - cg^r_{J,I}\big]\Big\} \\
&= (1-u_x)\frac{b_{\text{in}}g^r_{\text{in}} - c_{\text{in}}g^r_{\text{in}} + (K-1)\omega^r(b_{\text{out}}g^r_{\text{out}} - c_{\text{out}}g^r_{\text{out}})}{1+\omega^r(K-1)} \\
\Pi_m|_{f_r=1} &= (1-u_x)\sum_I \frac{1}{\mathcal{M}_m}\Big\{\sum_J \nu_J \sum_s f^s_J \psi(\omega^m_{I,J},\omega^s_{I,J})\big[b_{I,J}g^m_{I,J} - c_{I,J}g^s_{J,I}\big]\Big\}\Big|_{f_r=1} \\
&= (1-u_x)\frac{1}{1+\psi(\omega^m,\omega^r)(K-1)}\sum_I\Big\{\sum_J \psi(\omega^m_{I,J},\omega^s_{I,J})\big[b_{I,J}g^m_{I,J} - c_{I,J}g^r_{J,I}\big]\Big\} \\
&= (1-u_x)\frac{b_{\text{in}}g^m_{\text{in}} - c_{\text{in}}g^r_{\text{in}} + (K-1)\psi(\omega^m,\omega^r)(b_{\text{out}}g^m_{\text{out}} - c_{\text{out}}g^r_{\text{out}})}{1+\psi(\omega^m,\omega^r)(K-1)}.
\end{aligned}$$

When $f^r = 1$, reputations are given by

$$\begin{aligned}
g^r_{\text{in}} &= \frac{1}{\mathcal{M}^r}\frac{1}{K}(g^r_{\text{in}} + \omega^r(K-1)g^r_{\text{out}})(P^{GC} - P^{BD}) + P^{BD}, \\
g^r_{\text{out}} &= \frac{1}{\mathcal{M}^r}\frac{1}{K}\Big\{\big[(1+\omega^r)g^r_{\text{in}}g^r_{\text{out}} + (K-2)\omega^r(g^r_{\text{out}})^2\big](P^{GC} - P^{GD} - P^{BC} + P^{BD}) \\
&\quad + \big[\omega^r g^r_{\text{in}} + (1+(K-2)\omega^r)g^r_{\text{out}}\big](P^{GD} - P^{BD}) \\
&\quad + \big[g^r_{\text{in}} + (K-1)\omega^r g^r_{\text{out}}\big](P^{BC} - P^{BD})\Big\} + P^{BD}, \\
g^m_{\text{in}} &= \frac{1}{\mathcal{M}^m}\frac{1}{K}(g^r_{\text{in}} + \psi(\omega^m,\omega^r)(K-1)g^r_{\text{out}})(P^{GC} - P^{BD}) + P^{BD}, \\
g^m_{\text{out}} &= \frac{1}{\mathcal{M}^m}\frac{1}{K}\Big\{\big[(1+\psi(\omega^m,\omega^r)(K-1))g^r_{\text{in}}g^r_{\text{out}} \\
&\quad + (K-2)\psi(\omega^m,\omega^r)(g^r_{\text{out}})^2\big](P^{GC} - P^{GD} - P^{BC} + P^{BD}) \\
&\quad + \big[\psi(\omega^m,\omega^r)g^r_{\text{in}} + (1+(K-2)\psi(\omega^m,\omega^r))g^r_{\text{out}}\big](P^{GD} - P^{BD}) \\
&\quad + \big[g^r_{\text{in}} + (K-1)\psi(\omega^m,\omega^r)g^r_{\text{out}}\big](P^{BC} - P^{BD})\Big\} + P^{BD}.
\end{aligned}$$

Mutant invasibility requires that $\Pi^m|_{f^r=1} - \Pi^r|_{f^r=1} > 0$. Dropping the $(1-u_x)$ prefactor allows us to rewrite this invasibility condition as

$$\begin{aligned} \Pi^m|_{f^r=1} - \Pi^r|_{f^r=1} > 0 \\ b_{\text{out}}(g^m_{\text{out}} - g^r_{\text{out}}) + \frac{(b_{\text{out}} - c_{\text{out}})g^r_{\text{out}} - (b_{\text{in}} - c_{\text{in}})g^r_{\text{in}}}{1+\omega^r(K-1)} + \frac{b_{\text{in}}g^m_{\text{in}} - b_{\text{out}}g^m_{\text{out}} - c_{\text{in}}g^r_{\text{in}} + c_{\text{out}}g^r_{\text{out}}}{1+\psi(\omega^r,\omega^m)(K-1)} > 0. \end{aligned} \quad [23]$$

The first term in Eq. [23] is the difference in mutant and resident fitness due to being on the receiving end of out-group cooperation events. The second term is the difference in the fitness the *resident* accrues as a result of *out-group versus in-group* interactions. The third is the difference in the fitness the *mutant* accrues as a result of *in-group versus out-group* interactions. And so in summary, for the mutant to invade the resident, some combination of the following must be true:

1. The mutant must be targeted by more (and potentially more rewarding) out-group cooperation events than the resident.

2. The resident must, on net, suffer as a result of out-group cooperation events (which may be more rewarding but are less likely to be reciprocated: consider that, in general, $g^r_{\text{out}} < g^r_{\text{in}}$).

3. The mutant must, on net, benefit as a result of forgoing out-group cooperation events.

These three conditions establish that, in general, more rewarding out-group interactions favor less insularity (i.e., higher ω) and thus make it harder for more insular mutants to invade. How much harder will depend on the social norm, in particular the values of g^r_{in}, g^r_{out}, g^m_{in}, and g^m_{out}. The first two can be read off of SI table S3 (with $\omega = \omega^r$); the provided expressions are valid for g^r_{in} and g^r_{out}. The latter two can be found in SI tables S4 and S5.

Simplifying Eq. [23] in a useful way is difficult, but we can show that, in general, it will be possible for insular mutants to invade *unless* out-group cooperation is much more rewarding than in-group cooperation. This is intuitive, as insular mutants forgo out-group interactions; for insularity to be selected against, the interactions they forgo must be especially rewarding. Moreover, because it is generally easier to be seen as good by one's in-group than by one's out-group, we expect that if $b_{\text{in}} = b_{\text{out}}$ and $c_{\text{in}} = c_{\text{out}}$, it will not be possible for populations to resist invasion by more insular mutants.

We demonstrate that, when interactions' costs and benefits do not depend on group identity, insular mutants can essentially always invade. To do so, we expand [23] in u_a and u_x; we drop terms that are $\mathcal{O}(u_a^2)$, $\mathcal{O}(u_x^2)$, and $\mathcal{O}(u_a u_x)$. (We do not address *Scoring* here, as insularity does not affect reputations, and therefore it does not affect fitnesses.) For Shunning, we obtain the following condition for a mutant with out-group interaction parameter ω^m to invade a resident with parameter ω^r:

$$\begin{aligned} 0 < &\frac{u_a(\omega^r - \omega^m)}{\omega^r[1+(K-1)\omega^r][2+(K-1)(\omega^r+\omega^m)]^2} \\ &\times \Big((\omega^r - \omega^m)u_a\big[(b_{\text{in}} - c_{\text{in}})[1+2(K-1)\omega^r][2+(K-1)(\omega^r+\omega^m)] \\ &+ 2b_{\text{in}}[1+(K-1)\omega^r] - (b_{\text{out}} - c_{\text{out}})(K-1)\omega^r[2+(K-1)(\omega^r+\omega^m)]\big]\Big). \end{aligned}$$

When in-group and out-group interactions are indistinguishable, we have

$$\frac{u_a(\omega^r - \omega^m)\big[(b-c)[2+(K-1)(\omega^r+\omega^m)] + 2b\big]}{\omega^r[2+(K-1)(\omega^r+\omega^m)]^2},$$

which is always positive for $\omega^r > \omega^m$.

For Simple Standing,

$$\begin{aligned} 0 < &\frac{(K-1)(\omega^r - \omega^m)}{[1+(K-1)\omega^r][2+(K-1)(\omega^r+\omega^m)]^2} \\ &\times \Big((b_{\text{in}} - c_{\text{in}})(1-u_a-u_x)[2+(K-1)(\omega^r+\omega^m)] \\ &- (b_{\text{out}} - c_{\text{out}})[2+(K-1)(\omega^r+\omega^m)][1-2u_x-2u_a-(u_x+u_a)(K-1)\omega^r] \\ &+ b_{\text{out}}(K-1)[1+(K-1)\omega^r](u_a+u_x)(\omega^m+\omega^r)\Big). \end{aligned}$$

For identical in- and out-group interactions, this becomes

$$\frac{(\omega^r - \omega^m)(K-1)(u_a+u_x)\big[(b-c)[2+(K-1)(\omega^r+\omega^m)] + b(K-1)(\omega^m+\omega^r)\big]}{[2+(K-1)(\omega^r+\omega^m)]^2}$$

This, too, is always positive provided $\omega^r > \omega^m$.

For Stern Judging, a general expression can be obtained for $K > 2$:

$$\begin{aligned}
0 < {} & \frac{(K-1)(\omega^r - \omega^m)}{2[1+(K-1)\omega^r]^2[2+(K-1)(\omega^r+\omega^m)]^2(K-2)} \times \\
& \Big((b_{\text{in}} - c_{\text{in}})(K-2)[2+(K-1)(\omega^r+\omega^m)][(K-1)\omega^r(2-2u_a-u_x)+2(1-u_a-u_x)] \\
& - 2b_{\text{in}}(K-2)u_x[1+(K-1)\omega^r] \\
& - (b_{\text{out}} - c_{\text{out}})(K-u_x-2)[2+(K-1)(\omega^r+\omega^m)] \Big).
\end{aligned} \tag{24}$$

We can gain a better understanding of how in-group and out-group interactions affect the invasion of insular mutants by taking the limit $u_x \to 0$:

$$\frac{[2(1-u_a)(b_{\text{in}}-c_{\text{in}}) - (b_{\text{out}}-c_{\text{out}})](K-1)(\omega^r-\omega^m)}{2[1+(K-1)\omega^r][2+(K-1)(\omega^r+\omega^m)]} > 0, \tag{25}$$

which is always satisfied for $\omega^r > \omega^m$ *unless*

$$b_{\text{out}} - c_{\text{out}} > 2(1-u_a)(b_{\text{in}} - c_{\text{in}}).$$

That is, residents resist invasion by higher-insularity mutants only when out-group interactions are much more rewarding than in-group interactions. It is intriguing to note that this agrees perfectly with the condition in SI section 5.7 for group fitness, i.e., $(b_{\text{out}} - c_{\text{out}})g_{\text{out}}/[(b_{\text{in}} - c_{\text{in}})g_{\text{in}}] > 1$, since (under *Stern Judging* and with $u_x = 0$) we have $g_{\text{in}} = 1 - u_a$ and $g_{\text{out}} = 1/2$. (Evaluating the left hand side of Eq. [23] numerically reveals that this approximation slightly underestimates the value of the ratio $(b_{\text{out}} - c_{\text{out}})/(b_{\text{in}} - c_{\text{in}})$ required for a population to resist invasion.)

Setting in-group and out-group costs and benefits equal to each other in Eq. [24] yields

$$\begin{aligned}
0 < {} & \frac{(K-1)(\omega^r - \omega^m)}{2[1+(K-1)\omega^r]^2[2+(K-1)(\omega^r+\omega^m)]^2(K-2)} \times \\
& \Big((b-c)[2+(K-1)(\omega^r+\omega^m)]\big[(K-1)\omega^r\{(K-2)(1-2u_a)-(K-3)u_x\} \\
& \quad + K(1-2u_x-2u_a)+4u_a+5u_x-2\big] - 2bu_x[1+(K-1)\omega^r] \Big).
\end{aligned}$$

Sending $u_x \to 0$ reduces this to a specific form of Eq. [25]:

$$\frac{(b-c)(K-1)(1-2u_a)(\omega^r-\omega^m)}{2[1+(K-1)\omega^r][2+(K-1)(\omega^r+\omega^m)]} > 0.$$

This condition is always satisfied provided $\omega^r > \omega^m$ and $u_a < 1/2$, meaning that, when interactions' costs and benefits do not depend on group membership, a higher-insularity mutant (i.e., one that interacts less with its out-group) can always invade the resident population under *Stern Judging*, just as we found with *Shunning* and *Simple Standing*.

Finally, for *Stern Judging*, the condition for an insular mutant to invade when $K = 2$ is different from the general case:

$$\begin{aligned}
0 < {} & \frac{\omega^r - \omega^m}{4\omega^r(1+\omega^r)^{3/2}(2+\omega^r+\omega^m)^2} \\
& \times \Big(2\omega^r(b_{\text{in}} - c_{\text{in}})(2+\omega^r+\omega^m)(2\sqrt{1+\omega^r} - u_a(2+\omega^r) - 2u_x) \\
& + 4b_{\text{in}}(1+\omega^r)[u_a\omega^r - 2(u_a+u_x)(\sqrt{1+\omega^r}-1)] \\
& - (b_{\text{out}} - c_{\text{out}})(2+\omega^r+\omega^m)\big[u_a(\omega_2^r - 4\omega^r\sqrt{1+\omega^r} + 8\{1+\omega^r-\sqrt{1+\omega^r}\}) \\
& \quad + 2u_x(2+\omega^r-2\sqrt{1+\omega^r}) + 4(1+\omega^r-\sqrt{1+\omega^r})\big] \\
& - 2b_{\text{out}}(1+\omega^r)(\omega^r+\omega^m)(u_a\omega^r - 2(u_x+u_a)(\sqrt{1+\omega^r}-1)) \Big).
\end{aligned} \tag{26}$$

Sending both error rates to zero is informative:

$$\frac{(\omega^r-\omega^m)\big[(b_{\text{in}}-c_{\text{in}})\omega^r - (b_{\text{out}}-c_{\text{out}})(\sqrt{1+\omega^r}-1)\big]}{\omega^r(1+\omega^r)(2+\omega^m+\omega^r)} > 0.$$

This condition is satisfied for $\omega^r > \omega^m$ provided

$$\begin{aligned} &(b_{\text{in}} - c_{\text{in}})\omega^r > (b_{\text{out}} - c_{\text{out}})(\sqrt{1+\omega^r} - 1) \\ \therefore\ &\frac{b_{\text{in}} - c_{\text{in}}}{b_{\text{out}} - c_{\text{out}}} > \frac{\sqrt{1+\omega^r} - 1}{\omega^r}. \end{aligned}$$

The ratio on the right is between $1/2$ (for $\omega^r \to 0$) and $\sqrt{2} - 1 \approx .41$ (for $\omega^r \to 1$), meaning that insular mutants can invade unless out-group cooperation is a little more than twice as beneficial as in-group cooperation, a condition similar to the $K > 2$ case.

For equal in-group and out-group costs and benefits, Eq. [26] becomes

$$\begin{aligned} 0 < &\frac{\omega^r - \omega^m}{4\omega^r(1+\omega^r)^{3/2}(2+\omega^r+\omega^m)^2} \times \\ &\Big((b-c)(2+\omega^r+\omega^m)\big[4(1+\omega^r)(\sqrt{1+\omega^r}-1) - u_a(3[\omega^r]^2 - 4\omega^r\sqrt{1+\omega^r} + 8\omega^r - 8\sqrt{1+\omega^r} + 8) \\ &\quad - u_x(6\omega^r - 4\sqrt{1+\omega^r} + 4)\big] \\ &+ 2b(1+\omega^r)(2-\omega^r-\omega^m)(u_a\omega^r - 2(u_a+u_x)(\sqrt{1+\omega^r}-1)\Big) \end{aligned}$$

Sending the error rates to zero allows us to show that this is, in general, positive for $\omega^r > \omega^m$:

$$\frac{(\omega^r - \omega^m)(b-c)(\sqrt{1+\omega^r}-1)}{\omega^r\sqrt{1+\omega^r}(2+\omega^r+\omega^m)} > 0.$$

In general, out-group reputations compared to in-group reputations are very low, middling, and fairly high under *Shunning*, *Stern Judging*, and *Simple Standing* respectively. We thus expect that, if we hold b_{in}, c_{in}, and c_{out} constant, we will need to raise b_{out} more under *Shunning* than under *Stern Judging*, and more under *Stern Judging* than under *Simple Standing*, for a population to resist invasion by insular mutants. We verify this numerically by checking the sign of Eq. [23]. Results are shown in SI figure S5. In white areas, the difference between mutant and resident fitness is positive, meaning that the mutant can invade; in black areas, the difference is negative. We find that when $b_{\text{out}} = b_{\text{in}}$, it is impossible to resist invasion by successively more insular mutants, irrespective of the social norm. For *Simple Standing*, this situation reverses quickly with increased b_{out}; it reverses more slowly for *Stern Judging* and *Shunning*. Some norm and parameter combinations can support stable intermediate values of the out-group interaction parameter ω, but in most cases the population will either evolve to be fully mixed or fully insular ($\omega = 1$ or 0, respectively).

6. Third-order norms and the remaining "leading eight" norms

All of our analysis hitherto has focused on second-order norms. We now turn to the interesting question of third-order norms, in which the reputation of a donor may be updated according to not only the donor's action and the recipient's reputation, but also the donor's current reputation. In this space there are not 4 but 16 possible behavioral strategies, which can be represented as four bits, corresponding to the donor's action when both their reputation and the recipient's is good, when their own reputation is good but the donor's is bad, and so on. We first consider the simplest case of publicly shared reputations, then we generalize to group-wise reputations.

6.1. Properties of the leading eight. The "leading eight" social norms consist of reputation dynamics (assessment rules, i.e., rules for assigning "good" and "bad" reputations) and behavioral strategies (action rules, i.e., rules for deciding whether to defect or cooperate based on one's own reputation and that of the recipient) that satisfy the following conditions (7):

1. Either *good* or *bad* cooperating with *good* is *good*.

2. Either *good* or *bad* defecting with *good* is *bad*.

3. *Good* defecting with *bad* is *good*.

4. Either *good* or *bad* individuals will cooperate with *good*.

5. *Good* individuals will defect with *bad*.

6. Iff *bad* cooperating with *bad* is *good* and *bad* defecting with *bad* is *bad*, then *bad* will cooperate with *bad*: otherwise, *bad* will defect with *bad*.

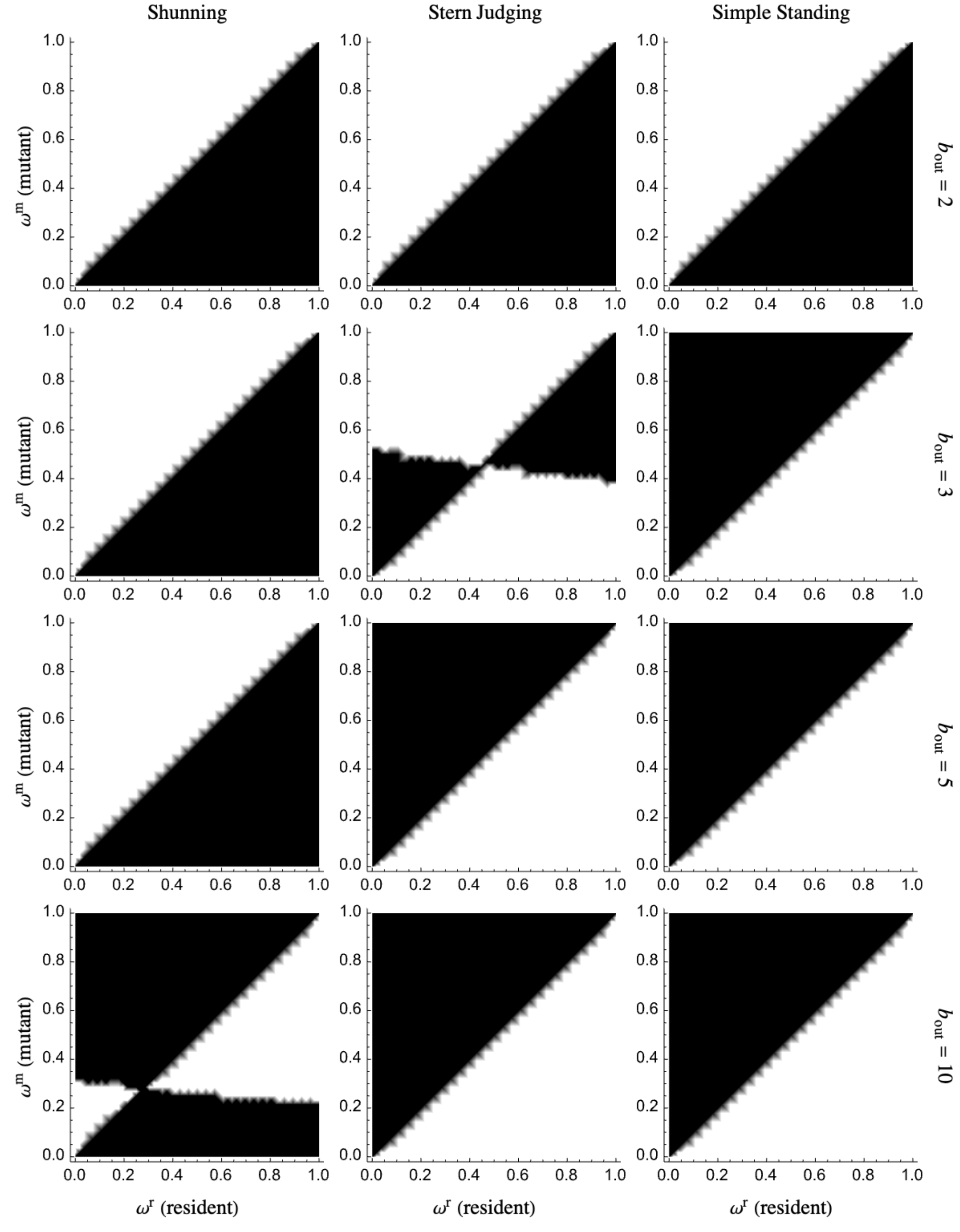

Fig. S5. The invasibility of a resident population with out-group interaction parameter ω^r by a mutant with parameter ω^m under different norms and out-group benefits. The value of b_{out} is along the right side of each row: other parameter values are $K = 2$, $u_a = u_x = 0.02$, $b_{\text{in}} = 2$, and $c_{\text{in}} = c_{\text{out}} = 1$. White corresponds to the mutant being able to invade: black corresponds to the resident resisting invasion. For *Stern Judging* at $b_{\text{out}} = 3$, an intermediate value of ω can be achieved by successive invasion of mutants. For *Shunning* at $b_{\text{out}} = 10$, there is bistability. For all other norms and parameter combinations shown, the population evolves either toward full mixing ($\omega = 1$) or full insularity ($\omega = 0$).

There are eight such combinations of reputation dynamics and action rule (behavioral strategy), summarized in table S6. The action rule is represented by c_{UV}, which are the probabilities that an individual with reputation U cooperates against an individual with reputation V. These values are always either $1 - u_x$ (because individuals who intend to cooperate can accidentally defect with probability u_x) or 0 (because individuals who intend to defect can never accidentally cooperate). The reputation dynamics are specified by the values n_{UAV}, which are the probability of earning a good reputation by having reputation U and performing action A against an individual with reputation V. These values are always either $1 - u_a$ or u_a, allowing for assessment error. (This notation is slightly different from our treatment of second-order social norms, where P_{AV} is the probability of earning a good reputation by *intending* to perform action A against a recipient with reputation V. The difference is that P_{CV} includes the possibility of both successful cooperation *and* accidental defection against an individual with reputation V, whereas n_{UCV} does not: successful cooperation and accidental defection are treated separately.)

6.2. Public reputations. When the whole population follows the same reputation dynamics and is fixed for the same action rule, the mean proportion of individuals with good reputations is given by

$$\begin{aligned} g = {} & g^2(c^{GG}n^{GCG} + [1 - c^{GG}]n^{GDG}) + g(1-g)(c^{GB}n^{GCB} + [1 - c^{GB}]n^{GDB} + c^{BG}n^{BCG} + [1 - c^{BG}]n^{BDG}) \\ & + (1-g)^2(c^{BB}n^{BCB} + [1 - c^{BB}]n^{BDB}). \end{aligned} \quad [27]$$

Under *Stern Judging*, we will have

$$\begin{aligned} c^{GG} &= c^{BG} = 1 - u_x, \\ c^{GB} &= c^{BB} = 0, \\ n^{GCG} &= n^{BCG} = 1 - u_a, \\ n^{GCB} &= n^{BCB} = u_a, \\ n^{GDG} &= n^{BDG} = u_a, \\ n^{GDB} &= n^{BDB} = 1 - u_a, \end{aligned}$$

so Eq. [27] simplifies to

$$\begin{aligned} g &= g^2([1 - u_x][1 - u_a] + u_x u_a) + g(1-g)(1 - u_a + [1 - u_x][1 - u_a] + u_x u_a) + (1-g)^2(1 - u_a) \\ &= g([1 - u_x][1 - u_a] + u_x u_a) + (1 - g)(1 - u_a), \end{aligned}$$

as expected; this equation is also the same under *Simple Standing* (where $n^{GCB} = n^{BCB} = 1 - u_a$ rather than u_a, but this term does not appear in the reputation dynamics, because $c^{GB} = c^{BB} = 0$). Note that *Stern Judging* and *Simple Standing* are norms s_6 and s_3 (respectively) in table S6.

Under *Scoring*, we have

$$\begin{aligned} c^{GG} &= c^{BG} = 1 - u_x, \\ c^{GB} &= c^{BB} = 0, \\ n^{GCG} &= n^{BCG} = 1 - u_a, \\ n^{GCB} &= n^{BCB} = 1 - u_a, \\ n^{GDG} &= n^{BDG} = u_a, \\ n^{GDB} &= n^{BDB} = u_a, \end{aligned}$$

so Eq. [27] is

$$\begin{aligned} g &= g^2([1 - u_x][1 - u_a] + u_x u_a) + g(1-g)(u_a + [1 - u_x][1 - u_a] + u_x u_a) + (1-g)^2 u_a \\ &= g([1 - u_x][1 - u_a] + u_x u_a) + (1 - g)u_a, \end{aligned}$$

again as expected: it would be the same under *Shunning*, for which the only difference is $n^{GCB} = n^{BCB} = u_a$. (Scoring and *Shunning* are not part of the "leading eight": we present those equations only for completeness.)

norm	n^{GCG}	n^{GDG}	n^{GCB}	n^{GDB}	n^{BCG}	n^{BDG}	n^{BCB}	n^{BDB}	c^{GG}	c^{GB}	c^{BG}	c^{BB}
s_1	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_a$	u_a	$1-u_a$	u_a	$1-u_x$	0	$1-u_x$	$1-u_x$
s_2	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_a$	u_a	$1-u_a$	u_a	$1-u_x$	0	$1-u_x$	$1-u_x$
s_3	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_x$	0	$1-u_x$	0
s_4	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_x$	0	$1-u_x$	0
s_5	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_x$	0	$1-u_x$	0
s_6	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_x$	0	$1-u_x$	0
s_7	$1-u_a$	u_a	$1-u_a$	$1-u_a$	$1-u_a$	u_a	u_a	u_a	$1-u_x$	0	$1-u_x$	0
s_8	$1-u_a$	u_a	u_a	$1-u_a$	$1-u_a$	u_a	u_a	u_a	$1-u_x$	0	$1-u_x$	0

Table S6. The "leading eight" social norms and action rules, modeled after table 2 of (8). Here, n_{UAV} is the probability that a donor with reputation U, who performs action A against a recipient with reputation V, earns a good reputation, and c_{UV} is the probability that a donor with reputation U will cooperate with a recipient with reputation V. As elsewhere in our model, we allow for assessment error with probability u_a and asymmetric execution error with probability u_x (individuals can accidentally defect but not accidentally cooperate). Norm s_3 is *Simple Standing*, and norm s_6 is *Stern Judging*; both are symmetric with respect to the reputation of the donor and, thus, are second-order norms.

6.3. Group-wise reputations. We now consider the possibility that the population is divided into groups, which each follow their own third-order social norm, i.e., each group has its own rule for assigning reputations and is fixed for its own particular behavioral strategy. We assume that an individual in group I acts according to their own view of themselves and the recipient, as well as their own action rule, but that J judges them according to J's reputation dynamics and J's view of the recipient. Thus, when $I = J$, we will have

$$\begin{aligned} g_{I,J} = {} & g_{I,I} g_{\bullet,J}(c_I^{GG} n_J^{GCG} + [1 - c_I^{GG}] n_J^{GDG}) + g_{I,I}(1 - g_{\bullet,J})(c_I^{GB} n_J^{GCB} + [1 - c_I^{GB}] n_J^{GDB}) \\ & + (1 - g_{I,I}) g_{\bullet,J}(c_I^{BG} n_J^{BCG} + [1 - c_I^{BG}] n_J^{BDG}) + (1 - g_{I,I})(1 - g_{\bullet,J})(c_I^{BB} n_J^{BCB} + [1 - c_I^{BB}] n_J^{BDB}). \end{aligned}$$

When $I \neq J$, we account for the fact that I and J may have different views of both the donor I and recipient L. We enumerate these possibilities. When group J observes an interaction by an individual from group I, they can form a good reputation of I in the following ways. With probability ν_L, an interaction between I and L is observed, and:

1. with probability $g_{I,I} g_{L,I}$, the donor sees themselves and the recipient as good.
    (a) with probability $g_{I,J} g_{L,J}$, the observer thinks the donor and recipient are both good. If the donor cooperates (probability c_I^{GG}), the observer considers that good with probability n_J^{GCG}. If the donor defects (probability $1 - c_I^{GG}$), the observer considers that good with probability n_J^{GDG}.
    (b) with probability $g_{I,J}(1 - g_{L,J})$, the observer thinks the donor is good but the recipient is bad. If the donor cooperates (probability c_I^{GG}), the observer considers that good with probability n_J^{GCB}. If the donor defects (probability $1 - c_I^{GG}$), the observer considers that good with probability n_J^{GDB}.
    (c) with probability $(1 - g_{I,J}) g_{L,J}$, the observer thinks the donor is bad but the recipient is good. If the donor cooperates (probability c_I^{GG}), the observer considers that good with probability n_J^{BCG}. If the donor defects (probability $1 - c_I^{GG}$), the observer considers that good with probability n_J^{BDG}.
    (d) with probability $(1 - g_{I,J})(1 - g_{L,J})$, the observer thinks the donor and recipient are both bad. If the donor cooperates (probability c_I^{GG}), the observer considers that good with probability n_J^{BCB}. If the donor defects (probability $1 - c_I^{GG}$), the observer considers that good with probability n_J^{BDB}.
2. with probability $g_{I,I}(1 - g_{L,I})$, the donor sees themselves as good and the recipient as bad.
    (a) with probability $g_{I,J} g_{L,J}$, the observer thinks the donor and recipient are both good. If the donor cooperates (probability c_I^{GB}), the observer considers that good with probability n_J^{GCG}. If the donor defects (probability $1 - c_I^{GB}$), the observer considers that good with probability n_J^{GDG}.
    (b) with probability $g_{I,J}(1 - g_{L,J})$, the observer thinks the donor is good but the recipient is bad. If the donor cooperates (probability c_I^{GB}), the observer considers that good with probability n_J^{GCB}. If the donor defects (probability $1 - c_I^{GB}$), the observer considers that good with probability n_J^{GDB}.
    (c) with probability $(1 - g_{I,J}) g_{L,J}$, the observer thinks the donor is bad but the recipient is good. If the donor cooperates (probability c_I^{GB}), the observer considers that good with probability n_J^{BCG}. If the donor defects (probability $1 - c_I^{GB}$), the observer considers that good with probability n_J^{BDG}.
    (d) with probability $(1 - g_{I,J})(1 - g_{L,J})$, the observer thinks the donor and recipient are both bad. If the donor cooperates (probability c_I^{GB}), the observer considers that good with probability n_J^{BCB}. If the donor defects (probability $1 - c_I^{GB}$), the observer considers that good with probability n_J^{BDB}.
3. with probability $(1 - g_{I,I}) g_{L,I}$, the donor sees themselves as bad and the recipient as good.
    (a) with probability $g_{I,J} g_{L,J}$, the observer thinks the donor and recipient are both good. If the donor cooperates (probability c_I^{BG}), the observer considers that good with probability n_J^{GCG}. If the donor defects (probability $1 - c_I^{BG}$), the observer considers that good with probability n_J^{GDG}.
    (b) with probability $g_{I,J}(1 - g_{L,J})$, the observer thinks the donor is good but the recipient is bad. If the donor cooperates (probability c_I^{BG}), the observer considers that good with probability n_J^{GCB}. If the donor defects (probability $1 - c_I^{BG}$), the observer considers that good with probability n_J^{GDB}.
    (c) with probability $(1 - g_{I,J}) g_{L,J}$, the observer thinks the donor is bad but the recipient is good. If the donor cooperates (probability c_I^{BG}), the observer considers that good with probability n_J^{BCG}. If the donor defects (probability $1 - c_I^{BG}$), the observer considers that good with probability n_J^{BDG}.
    (d) with probability $(1 - g_{I,J})(1 - g_{L,J})$, the observer thinks the donor and recipient are both bad. If the donor cooperates (probability c_I^{BG}), the observer considers that good with probability n_J^{BCB}. If the donor defects (probability $1 - c_I^{BG}$), the observer considers that good with probability n_J^{BDB}.

4. with probability $(1-g_{I,I})(1-g_{L,I})$, the donor sees themselves and the recipient as bad.
   (a) with probability $g_{I,J}g_{L,J}$, the observer thinks the donor and recipient are both good. If the donor cooperates (probability c_I^{BB}), the observer considers that good with probability n_J^{GCG}. If the donor defects (probability $1-c_I^{BB}$), the observer considers that good with probability n_J^{GDG}.
   (b) with probability $g_{I,J}(1-g_{L,J})$, the observer thinks the donor is good but the recipient is bad. If the donor cooperates (probability c_I^{BB}), the observer considers that good with probability n_J^{GCB}. If the donor defects (probability $1-c_I^{BB}$), the observer considers that good with probability n_J^{GDB}.
   (c) with probability $(1-g_{I,J})g_{L,J}$, the observer thinks the donor is bad but the recipient is good. If the donor cooperates (probability c_I^{BB}), the observer considers that good with probability n_J^{BCG}. If the donor defects (probability $1-c_I^{BB}$), the observer considers that good with probability n_J^{BDG}.
   (d) with probability $(1-g_{I,J})(1-g_{L,J})$, the observer thinks the donor and recipient are both bad. If the donor cooperates (probability c_I^{BB}), the observer considers that good with probability n_J^{BCB}. If the donor defects (probability $1-c_I^{BB}$), the observer considers that good with probability n_J^{BDB}.

Summing over all possible groups L yields

$$
\begin{aligned}
g_{I,J} = \delta_{I,J}\Bigg\{ & g_{I,I}g_{\bullet,J}(c_I^{GG}n_J^{GCG}+[1-c_I^{GG}]n_J^{GDG}) + g_{I,I}(1-g_{\bullet,J})(c_I^{GB}n_J^{GCB}+[1-c_I^{GB}]n_J^{GDB}) \\
& + (1-g_{I,I})g_{\bullet,J}(c_I^{BG}n_J^{BCG}+[1-c_I^{BG}]n_J^{BDG}) + (1-g_{I,I})(1-g_{\bullet,J})(c_I^{BB}n_J^{BCB}+[1-c_I^{BB}]n_J^{BDB})\Bigg\} \\
& + (1-\delta_{I,J})\Bigg\{\sum_L \nu_L\Bigg(g_{I,I}g_{L,I}\Bigg[g_{I,J}g_{L,J}(c_I^{GG}n_J^{GCG}+[1-c_I^{GG}]n_J^{GDG}) + g_{I,J}(1-g_{L,J})(c_I^{GG}n_J^{GCB}+[1-c_I^{GG}]n_J^{GDB}) \\
& + (1-g_{I,J})g_{L,J}(c_I^{GG}n_J^{BCG}+[1-c_I^{GG}]n_J^{BDG}) + (1-g_{I,J})(1-g_{L,J})(c_I^{GG}n_J^{BCB}+[1-c_I^{GG}]n_J^{BDB})\Bigg] \\
& + g_{I,I}(1-g_{L,I})\Bigg[g_{I,J}g_{L,J}(c_I^{GB}n_J^{GCG}+[1-c_I^{GB}]n_J^{GDG}) + g_{I,J}(1-g_{L,J})(c_I^{GB}n_J^{GCB}+[1-c_I^{GB}]n_J^{GDB}) \\
& + (1-g_{I,J})g_{L,J}(c_I^{GB}n_J^{BCG}+[1-c_I^{GB}]n_J^{BDG}) + (1-g_{I,J})(1-g_{L,J})(c_I^{GB}n_J^{BCB}+[1-c_I^{GB}]n_J^{BDB})\Bigg] \\
& + (1-g_{I,I})g_{L,I}\Bigg[g_{I,J}g_{L,J}(c_I^{BG}n_J^{GCG}+[1-c_I^{BG}]n_J^{GDG}) + g_{I,J}(1-g_{L,J})(c_I^{BG}n_J^{GCB}+[1-c_I^{BG}]n_J^{GDB}) \\
& + (1-g_{I,J})g_{L,J}(c_I^{BG}n_J^{BCG}+[1-c_I^{BG}]n_J^{BDG}) + (1-g_{I,J})(1-g_{L,J})(c_I^{BG}n_J^{BCB}+[1-c_I^{BG}]n_J^{BDB})\Bigg] \\
& + (1-g_{I,I})(1-g_{L,I})\Bigg[g_{I,J}g_{L,J}(c_I^{BB}n_J^{GCG}+[1-c_I^{BB}]n_J^{GDG}) + g_{I,J}(1-g_{L,J})(c_I^{BB}n_J^{GCB}+[1-c_I^{BB}]n_J^{GDB}) \\
& + (1-g_{I,J})g_{L,J}(c_I^{BB}n_J^{BCG}+[1-c_I^{BB}]n_J^{BDG}) + (1-g_{I,J})(1-g_{L,J})(c_I^{BB}n_J^{BCB}+[1-c_I^{BB}]n_J^{BDB})\Bigg]\Bigg)\Bigg\}.
\end{aligned}
\tag{28}
$$

We now allow individuals to switch between gossip groups. We assume that, when an individual switches groups, it adopts both the reputation dynamics *and* action rule (behavioral strategy) of that group. The fitness of group I can be expressed as

$$
\begin{aligned}
\Pi_I = b\sum_J \nu_J\big[& g_{J,J}g_{I,J}c_J^{GG} + g_{J,J}(1-g_{I,J})c_J^{GB} + (1-g_{J,J})g_{I,J}c_J^{BG} + (1-g_{J,J})(1-g_{I,J})c_J^{BB}\big] \\
- c\sum_J \nu_J\big[& g_{I,I}g_{J,I}c_I^{GG} + g_{I,I}(1-g_{J,I})c_I^{GB} + (1-g_{I,I})g_{J,I}c_I^{BG} + (1-g_{I,I})(1-g_{J,I})c_I^{BB}\big].
\end{aligned}
$$

When there are two equally sized groups following different norms, fitnesses are given by (dropping the $1/2$ prefactor)

$$\begin{aligned}
\Pi_1 = b\big[&(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB} \\
&+ g_{2,2}g_{1,2}c_2^{GG} + g_{2,2}(1-g_{1,2})c_2^{GB} + (1-g_{2,2})g_{1,2}c_2^{BG} + (1-g_{2,2})(1-g_{1,2})c_2^{BB}\big] \\
&- c\big[(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB} \\
&+ g_{1,1}g_{2,1}c_1^{GG} + g_{1,1}(1-g_{2,1})c_1^{GB} + (1-g_{1,1})g_{2,1}c_1^{BG} + (1-g_{1,1})(1-g_{2,1})c_1^{BB}\big] \\
\Pi_2 = b\big[&g_{1,1}g_{2,1}c_1^{GG} + g_{1,1}(1-g_{2,1})c_1^{GB} + (1-g_{1,1})g_{2,1}c_1^{BG} + (1-g_{1,1})(1-g_{2,1}c_1^{BB} \\
&+ (g_{2,2})^2c_2^{GG} + g_{2,2}(1-g_{2,2})(c_2^{GB}+c_2^{BG}) + (1-g_{2,2})^2c_2^{BB}\big] \\
&- c\big[g_{2,2}g_{1,2}c_2^{GG} + g_{2,2}(1-g_{1,2})c_1^{GB} + (1-g_{2,2})g_{1,2}c_1^{BG} + (1-g_{2,2})(1-g_{1,2})c_1^{BB} \\
&+ (g_{2,2})^2c_2^{GG} + g_{2,2}(1-g_{2,2})(c_2^{GB}+c_2^{BG} + (1-g_{2,2})^2c_2^{BB}\big].
\end{aligned}$$

The fitness gain of group 1 due to self-interactions is given by

$$\begin{aligned}
\pi_{1,1} &= b\big[(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB}\big] \\
&\quad - c\big[(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB}\big] \\
&= (b-c)\big[(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB}\big].
\end{aligned}$$

The fitness gain of group 1 due to interactions with group 2 is

$$\begin{aligned}
\pi_{1,2} = b\big[&(g_{2,2}g_{1,2}c_2^{GG} + g_{2,2}(1-g_{1,2})c_2^{GB} + (1-g_{2,2})g_{1,2}c_2^{BG} + (1-g_{2,2})(1-g_{1,2})c_2^{BB}\big] \\
&- c\big[g_{1,1}g_{2,1}c_1^{GG} + g_{1,1}(1-g_{2,1})c_1^{GB} + (1-g_{1,1})g_{2,1}c_1^{BG} + (1-g_{1,1})(1-g_{2,1})c_1^{BB}\big].
\end{aligned}$$

Likewise,

$$\begin{aligned}
\pi_{2,2} &= (b-c)\big[(g_{2,2})^2c_2^{GG} + g_{2,2}(1-g_{2,2})(c_2^{GB}+c_2^{BG}) + (1-g_{2,2})^2c_2^{BB}\big], \\
\pi_{2,1} &= b\big[g_{1,1}g_{2,1}c_1^{GG} + g_{1,1}(1-g_{2,1})c_1^{GB} + (1-g_{1,1})g_{2,1}c_1^{BG} + (1-g_{1,1})(1-g_{2,1}c_1^{BB}\big] \\
&\quad - c\big[g_{2,2}g_{1,2}c_2^{GG} + g_{2,2}(1-g_{1,2})c_1^{GB} + (1-g_{2,2})g_{1,2}c_1^{BG} + (1-g_{2,2})(1-g_{1,2})c_1^{BB}\big]
\end{aligned}$$

We thus have

$$\begin{aligned}
\Pi_1 &> \Pi_2 \\
\Pi_1 - \Pi_2 &> 0 \\
\pi_{1,1} + \pi_{1,2} - \pi_{2,1} - \pi_{2,2} &> 0 \\
\therefore (b-c)\Big(\big[(g_{1,1})^2c_1^{GG} + g_{1,1}(1-g_{1,1})(c_1^{GB}+c_1^{BG}) + (1-g_{1,1})^2c_1^{BB}\big] & \\
-\big[(g_{2,2})^2c_2^{GG} + g_{2,2}(1-g_{2,2})(c_2^{GB}+c_2^{BG}) + (1-g_{2,2})^2c_2^{BB}\big]\Big) & \\
+ (b+c)\Big(\big[(g_{2,2}g_{1,2}c_2^{GG} + g_{2,2}(1-g_{1,2})c_2^{GB} + (1-g_{2,2})g_{1,2}c_2^{BG} + (1-g_{2,2})(1-g_{1,2})c_2^{BB}\big] & \\
-\big[g_{1,1}g_{2,1}c_1^{GG} + g_{1,1}(1-g_{2,1})c_1^{GB} + (1-g_{1,1})g_{2,1}c_1^{BG} + (1-g_{1,1})(1-g_{2,1}c_1^{BB}\big]\Big) &> 0.
\end{aligned}$$

This is a form very similar to Eq. (3) from the main text: the $(b-c)$ term is the difference in the fitnesses of groups 1 and 2 due to within-group interactions, and the $(b+c)$ term is the difference in their fitnesses due to between-group interactions, i.e., fitness differences due to unreciprocated between-group cooperation.

Next, we consider competition between groups following different norms in a manner similar to the main text. We focus on the two "leading eight" norms that happen to be second-order, namely *Stern Judging* and *Simple Standing*. The results can be seen in figure S6. While *Simple Standing* is readily out-competed by most of the other leading eight, *Stern Judging* is not: for *Stern Judging*, we have $\nu_1^* < 1/2$ in almost every scenario, with the exception of competition against s_8, for which it is greater than $1/2$ for small values of b and never moves far below $1/2$. s_8 is very similar to *Stern Judging*, with the exception that it regards bad individuals who defect against other bad individuals as bad, not good.

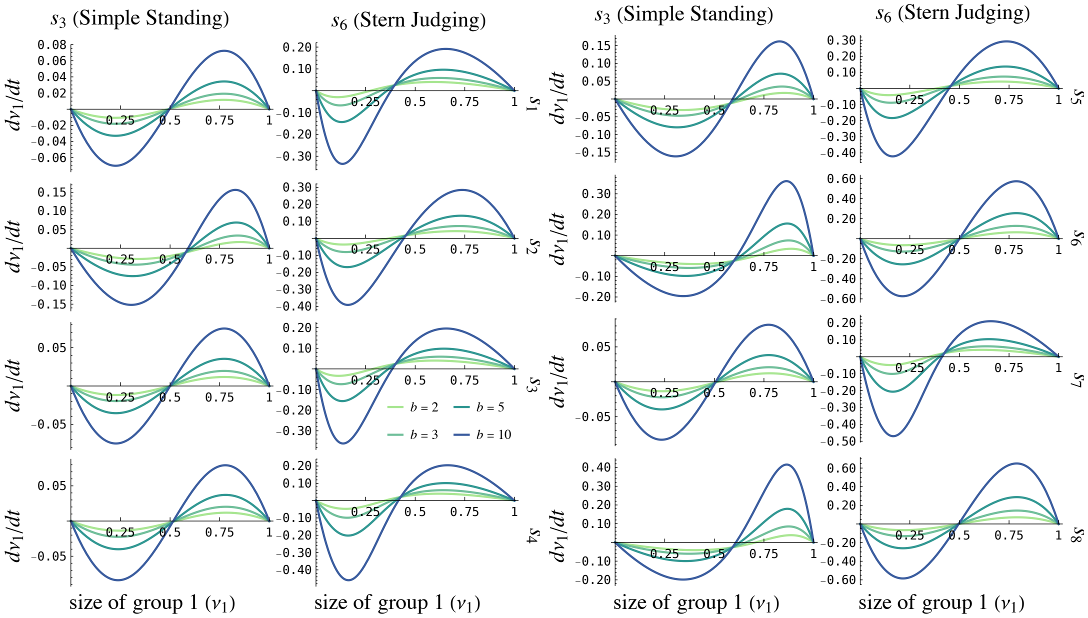

Fig. S6. Group size dynamics for $K = 2$ groups and varying values of the benefit b, when one group follows a second-order norm (either *Stern Judging* or *Simple Standing*) and the other follows a different “leading eight” norm, which may be third-order. In each pair of columns, the norm used in group 1 is along the top: the norm used in group 2, along the right. Values of b are as inset in the $s_6 - s_3$ figure. In all plots, $c = 1$, $u_a = u_x = 0.02$.

6.4. The many-group limit: private reputations. If the number of groups gets large ($K \to \infty$), all groups follow the same norm, and all groups are of the same size $1/K$, then we can reason similarly to subsection 3.1: self-interactions almost never occur, so the $I = J$ term of Eq. [28] drops out, and the remaining $g_{I,J}$ converge to a common value g. This yields

$$\begin{aligned}
g = g^2 \big(& g^2[c^{GG} n^{GCG} + (1 - c^{GG}) n^{GDG}] + g(1 - g)[c^{GG}(n^{GCB} + n^{BCG}) \\
& + (1 - c^{GG})(n^{GDB} + n^{BDG})] + (1 - g)^2[c^{GG} n^{BCB} + (1 - c^{GG}) n^{BDB}]\big) \\
& + g(1 - g)\big(g^2[(c^{GB} + c^{BG}) n^{GCG} + (2 - c^{GB} - c^{BG}) n^{GDG}] + g(1 - g)[(c^{GB} + c^{BG})(n^{GCB} + n^{BCG}) \\
& + (2 - c^{GB} - c^{BG})(n^{GDB} + n^{BDG})] + (1 - g)^2[(c^{GB} + c^{BG}) n^{BCB} + (2 - c^{GB} - c^{BG}) n^{BDB}]\big) \\
& + (1 - g)^2\big(g^2[c^{BB} n^{GCG} + (1 - c^{BB}) n^{GDG}] + g(1 - g)[c^{BB}(n^{GCB} + n^{BCG}) \\
& + (1 - c^{BB})(n^{GDB} + n^{BDG})] + (1 - g)^2[c^{BB} n^{BCB} + (1 - c^{BB}) n^{BDB}]\big).
\end{aligned}$$

If we specify that the action rule does not depend on the reputation of the actor, we have $c^{GG} = c^{BG} = c^G$ and $c^{GB} = c^{BB} = c^B$. We obtain a familiar expression:

$$\begin{aligned}
g = g^2 \big(& g^2[c^G n^{GCG} + (1 - c^G) n^{GDG}] + g(1 - g)[c^G(n^{GCB} + n^{BCG}) \\
& + (1 - c^G)(n^{GDB} + n^{BDG})] + (1 - g)^2[c^G n^{BCB} + (1 - c^G) n^{BDB}]\big) \\
& + g(1 - g)\big(g^2[(c^B + c^G) n^{GCG} + (2 - c^B - c^G) n^{GDG}] + g(1 - g)[(c^B + c^G)(n^{GCB} + n^{BCG}) \\
& + (2 - c^B - c^G)(n^{GDB} + n^{BDG})] + (1 - g)^2[(c^B + c^G) n^{BCB} + (2 - c^B - c^G) n^{BDB}]\big) \\
& + (1 - g)^2\big(g^2[c^B n^{GCG} + (1 - c^B) n^{GDG}] + g(1 - g)[c^B(n^{GCB} + n^{BCG}) \\
& + (1 - c^B)(n^{GDB} + n^{BDG})] + (1 - g)^2[c^B n^{BCB} + (1 - c^B) n^{BDB}]\big) \\
= g^3 \big(& c^G[n^{GCG} - n^{GCB} - n^{BCG} + n^{BCB} - n^{GDG} + n^{GDB} + n^{BDG} - n^{BDB}] \\
& - c^B[n^{GCG} - n^{GCB} - n^{BCG} + n^{BCB} - n^{GDG} + n^{GDB} + n^{BDG} - n^{BDB}]\big) \\
& + g^2\big(c^G[n^{GCB} + n^{BCG} - 2n^{BCB} - n^{GDB} - n^{BDG} + 2n^{BDB}] \\
& + c^B[n^{GCG} - 2n^{GCB} - 2n^{BCG} + 3n^{BCB} - n^{GDG} + 2n^{GDB} + 2n^{BDG} - 3n^{BDB}] \\
& + n^{GDG} - n^{GDB} - n^{BDG} + n^{BDB}\big) \\
& + g\big(c^G(n^{BCB} - n^{BDB}) + c^B(n^{GCB} + n^{BCG} - 3n^{BCB} - n^{GDB} - n^{BDG} + 3n^{BDB}) \\
& + n^{GDB} + n^{BDG} - 2n^{BDB}\big) \\
& + c^B(n^{BCB} - n^{BDB}) + n^{BDB}.
\end{aligned}$$

This expression is equivalent to equation 20 from the supplement of (9), which considered third-order norms with private information but only one action rule (the classic “discriminator” strategy: cooperate with good and defect with bad, irrespective of one’s own reputation).

7. Multiple groups with disjoint strategic imitation

In the preceding analyses, we have assumed that individuals freely copy strategies across groups (well-mixed copying), so that the impact of population structure was to partition reputation information and game play 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.

7.1. Strategic type dynamics. When strategies cannot be copied between groups, we must independently track the frequencies and fitnesses of types within each group. We zero in on the case of two groups. Fitnesses are given by

$$\begin{aligned}
\Pi_1^X &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{1,1}^X] + \nu_2[f_2^X + f_2^Z g_{1,2}^X]\big) - c\Big] \\
\Pi_1^Y &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{1,1}^Y] + \nu_2[f_2^X + f_2^Z g_{1,2}^Y]\big)\Big] \\
\Pi_1^Z &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{1,1}^Z] + \nu_2[f_2^X + f_2^Z g_{1,2}^Z]\big) - cg_{\bullet,1}\Big] \\
\Pi_2^X &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{2,1}^X] + \nu_2[f_2^X + f_2^Z g_{2,2}^X]\big) - c\Big] \\
\Pi_2^Y &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{2,1}^Y] + \nu_2[f_2^X + f_2^Z g_{2,2}^Y]\big)\Big] \\
\Pi_2^Z &= (1-u_x)\Big[b\big(\nu_1[f_1^X + f_1^Z g_{2,1}^Z] + \nu_2[f_2^X + f_2^Z g_{2,2}^Z]\big) - cg_{\bullet,2}\Big].
\end{aligned}$$

Armed with these fitness expressions, we can study how the strategic composition of one group affects the other. We first consider the behavior of strategies in group 1, when group 2 is exogenously fixed for either DISC or ALLD. Figure S7 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 $u_a = u_x = .02$. When strategic types are copied in a well-mixed manner (top row), *Shunning* cannot sustain cooperation, whereas *Stern Judging* and *Simple Standing* both maintain sizeable basins of attraction toward cooperation. When group 2 is exogenously fixed for DISC, this remains the case, as good behavior in group 1 from the perspective of group 2 is rewarded. When group 2 is exogenously fixed for ALLD, none of the three norms we consider can sustain cooperation, as discriminators waste fitness on cooperative acts with group 2 that will never be repaid in kind.

In the subsequent section (7.2), we 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

$$\frac{b}{c} > \frac{1}{\nu_1(P^{GC} - P^{GD})} = \frac{1}{\nu_1(\epsilon - u_a)}.$$

(This reduces to the one-group case in the limit $\nu_1 \to 1$.) In the example shown in Figure S7, 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 S7). 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 S7. 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 strategic compositions to resemble each other.

7.2. Behavior of equilibria. We now turn to an analytical treatment of the equilibria seen in Figure S7, specifically those along the DISC-ALLD edge; this will allow us to glean some insight into under what circumstances group 1 can be invaded when group 2's strategic composition is fixed. When DISC and ALLD are the only two strategic types present, their

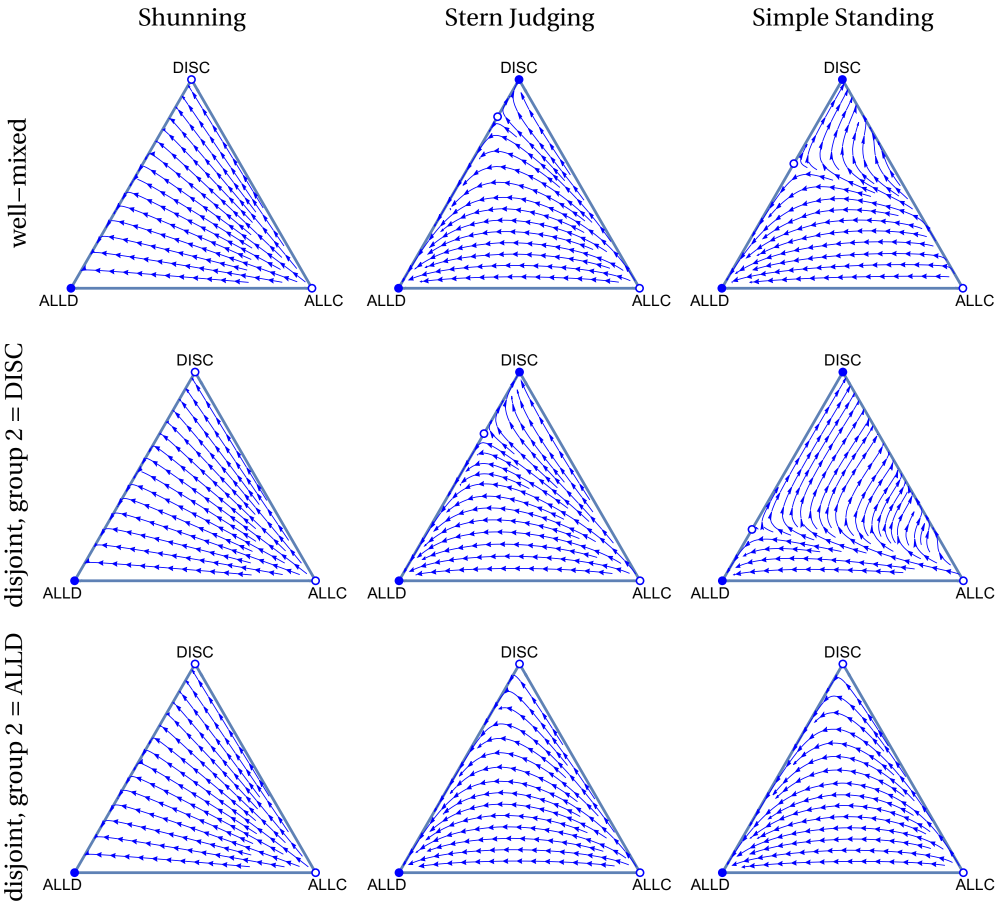

Fig. S7. Strategy frequency dynamics for one of two equally-sized groups ($\nu_1 = \nu_2 = 1/2$) under *Shunning*, *Stern Judging*, and *Simple Standing*. In all panels, the dynamics of strategy frequencies in group 1 are shown. The top row corresponds to well-mixed copying; the bottom two rows correspond 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, $b = 2$, $c = 1$, $u_a = u_x = 0.02$.

fitnesses are given by

$$\begin{aligned}
\Pi_1^Y &= (1-u_x)\Big[b\big(\nu_1 f_1^Z g_{1,1}^Y + \nu_2 f_2^Z g_{1,2}^Y\big)\Big]\\
\Pi_1^Z &= (1-u_x)\Big[b\big(\nu_1 f_1^Z g_{1,1}^Z + \nu_2 f_2^Z g_{1,2}^Z\big) - c g_{\bullet,1}\Big]\\
\Pi_2^Y &= (1-u_x)\Big[b\big(\nu_1 f_1^Z g_{2,1}^Y + \nu_2 f_2^Z g_{2,2}^Y\big)\Big]\\
\Pi_2^Z &= (1-u_x)\Big[b\big(\nu_1 f_1^Z g_{2,1}^Z + \nu_2 f_2^Z g_{2,2}^Z\big) - c g_{\bullet,2}\Big].
\end{aligned}$$

7.2.1. Both groups fixed for ALLD. Let $f = f_1^Z$ be the frequency of Z in group 1. For DISC to invade ALLD in group 1, we would require

$$\begin{gathered}
(\partial_f \dot{f})|_{f=0} > 0\\
\Pi_1^Z|_{f_1^Y=1, f_2^Y=1} > \Pi_1^Y|_{f_1^Y=1, f_2^Y=1}\\
\therefore -c g_{\bullet,1} > 0.
\end{gathered}$$

Since c and $g_{\bullet,2}$ are both positive numbers, Z cannot invade.

7.2.2. Both groups fixed for DISC. Assume now that both groups are fixed for DISC. For ALLD to invade DISC in group 1, we would require

$$\begin{aligned}
\Pi_1^Y|_{f_1^Z=1,f_2^Z=1} &> \Pi_1^Z|_{f_1^Z=1,f_2^Z=1} \\
b\big(\nu_1 g_{1,1}^Y + \nu_2 g_{1,2}^Y\big)|_{f_1^Z=1,f_2^Z=1} &> b\big(\nu_1 g_{1,1}^Z + \nu_2 g_{1,2}^Z\big)|_{f_1^Z=1,f_2^Z=1} - cg_{\bullet,1}|_{f_1^Z=1,f_2^Z=1} \\
\frac{b}{c} &< \frac{g_{\bullet,1}}{\nu_1\big(g_{1,1}^Z - g_{1,1}^Y\big) + \nu_2\big(g_{1,2}^Z - g_{1,2}^Y\big)} \\
\therefore \frac{b}{c} &< \frac{g_{\bullet,1}}{G_{1,2}\nu_2(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g_{\bullet,1}\big[\nu_1(P^{GC} - P^{GD}) + \nu_2(P^{BC} - P^{BD})\big]}.
\end{aligned}$$

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.

7.2.3. 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^Y). ALLD can invade 1 provided

$$\begin{aligned}
(\partial_f \dot{f})|_{f=0} &> 0 \\
\therefore (\partial_f[f(\Pi_1^Y - \bar{\Pi}_1)])|_{f=0} &> 0 \\
\therefore (\partial_f[f(\Pi_1^Y - f\Pi_1^Y - (1-f)\Pi_1^Z)])|_{f=0} &> 0 \\
\therefore (\partial_f[(f-f^2)\Pi_1^Y - (f-f^2)\Pi_1^Z)])|_{f=0} &> 0 \\
\therefore (\partial_f[f-f^2][\Pi_1^Y - \Pi_1^Z)])|_{f=0} &> 0 \\
\therefore ([1-2f][\Pi_1^Y - \Pi_1^Z)])|_{f=0} & \\
\therefore \Pi_1^Y|_{f=0} &> \Pi_1^Z|_{f=0} \\
\therefore b\big(\nu_1 f_1^Z g_{1,1}^Y + \nu_2 f_2^Z g_{1,2}^Y\big)\Big|_{f_1^Y=0,f_2^Y=1} &> \Big[b\big(\nu_1 f_1^Z g_{1,1}^Z + \nu_2 f_2^Z g_{1,2}^Z\big) - cg_{\bullet,1}\Big]\Big|_{f_1^Y=0,f_2^Y=1} \\
\therefore b\nu_1 g_{1,1}^Y &> b\nu_1 g_{1,1}^Z - cg_{\bullet,1}\Big|_{f_1^Y=0,f_2^Y=1} \\
\therefore b\nu_1(g_{1,1}^Z - g_{1,1}^Y) &< cg_{\bullet,1} \\
\therefore \frac{b}{c} &< \frac{g_{\bullet,1}}{\nu_1(g_{1,1}^Z - g_{1,1}^Y)} \\
\therefore \frac{b}{c} &< \frac{g_{\bullet,1}}{\nu_1[g_{\bullet,1}P^{GC} + (1-g_{\bullet,1})P^{BD} - g_{\bullet,1}P^{GD} - (1-g_{\bullet,1})P^{BD}]} \\
\therefore \frac{b}{c} &< \frac{g_{\bullet,1}}{\nu_1 g_{\bullet,1}(P^{GC} - P^{GD})} \\
\therefore \frac{b}{c} &< \frac{1}{\nu_1(P^{GC} - P^{BD})} = \frac{1}{\nu_1(\epsilon - u_a)}.
\end{aligned}$$

Letting $\nu_1 \to 1$ allows us to recover the one-group condition, $b/c < 1/(P^{GC} - P^{GD}) = 1/(\epsilon - u_a)$. 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_2^Z). DISC being able to invade requires

$$\begin{aligned}
(\partial_f \dot{f})|_{f=0} &> 0 \\
\therefore \Pi_2^Z|_{f=0} &> \Pi_2^Y|_{f=0} \\
\therefore b\Big[\big(\nu_1 f_1^Z g_{2,1}^Z + \nu_2 f_2^Z g_{2,2}^Z - c g_{\bullet,2}\big)\Big]\Big|_{f_1^Z=1, f_2^Z=0} &> b\big(\nu_1 f_1^Z g_{2,1}^Y + \nu_2 f_2^Z g_{2,2}^Z\big)\Big|_{f_1^Z=1, f_2^Z=0} \\
\therefore b\nu_1(g_{2,1}^Z - g_{2,1}^Y) &> c g_{\bullet,1}|_{f_1^Z=1, f_2^Z=0} \\
\therefore \frac{b}{c} &> \frac{g_{\bullet,1}}{\nu_1(g_{2,1}^Z - g_{2,1}^Y)} \\
\therefore \frac{b}{c} &> \frac{g_{\bullet,1}}{\nu_1\big[G_{2,1}(P^{GC} - P^{GD} - P^{BC} + P^{GD}) + g_{\bullet,2}(P^{BC} - P^{BD})\big]}.
\end{aligned}$$

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.

8. Derivation of replicator equation under different copying models

Here we explicitly derive the replicator dynamics for various group-wise strategy copying scenarios. In this section we use i and j to denote strategic types, as opposed to s and s'.

8.1. 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 $\phi(\Pi^j, \Pi^i) = 1/(1 + \exp[\beta(\Pi^j - \Pi^i)]$. 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,

$$\begin{aligned}
\mathbb{E}[\Delta f^i] &= \frac{1}{N}\mathbb{P}\Big(\Delta f^i = \frac{1}{N}\Big) - \frac{1}{N}\mathbb{P}\Big(\Delta f^i = -\frac{1}{N}\Big) \\
&= \frac{1}{N}\left(\sum_j f^j f^i \phi(\Pi^j, \Pi^i) - f^i \sum_j f^j \phi(\Pi^i, \Pi^j)\right) \\
&= \frac{1}{N} f^i \left(\sum_j f^j \Big[\phi(\Pi^j, \Pi^i) - \phi(\Pi^i, \Pi^j)\Big]\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\phi(\Pi^j, \Pi^i) = \frac{1}{1 + \exp[\beta(\Pi^j - \Pi^i)]} &\approx \frac{1}{1 + \exp[\beta(\Pi^j - \Pi^i)]}\Big|_{\beta=0} + \beta\Big(\frac{d}{d\beta}\Big[\frac{1}{1 + \exp[\beta(\Pi^j - \Pi^i)]}\Big]\Big)\Big|_{\beta=0} + \mathcal{O}(\beta^2) \\
&= \frac{1}{2} + \beta\Big[\frac{(\Pi^i - \Pi^j)\exp[\beta(\Pi^j - \Pi^i)]}{(1 + \exp[\beta(\Pi^j - \Pi^i)])^2}\Big]\Big|_{\beta=0} + \mathcal{O}(\beta^2) \\
&= \frac{1}{2} + \beta\frac{\Pi^i - \Pi^j}{4} + \mathcal{O}(\beta^2).
\end{aligned}$$

Hence

$$\phi(\Pi^j, \Pi^i) - \phi(\Pi^i, \Pi^j) \approx \beta\frac{\Pi^i - \Pi^j}{2} + \mathcal{O}(\beta^2).$$

We therefore have

$$\begin{aligned}\mathbb{E}[\Delta f^i] &= \frac{1}{N} f^i \sum_j \left[\frac{\beta}{2} f^j (\Pi^i - \Pi^j) + \mathcal{O}(\beta^2)\right] \\ &\approx \frac{\beta}{2N} f^i \sum_j f^j (\Pi^i - \Pi^j) \\ &= \frac{\beta}{2N} f^i (\Pi^i \sum_j f^j - \sum_j f^j \Pi^j) \\ &= \frac{\beta}{2N} f^i (\Pi^i - \bar{\Pi}).\end{aligned}$$

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

$$\dot{f^i} = f^i (\Pi^i - \bar{\Pi})$$

8.2. Multiple groups, copying only group identity. We now consider that there is more than one group ($K > 1$) and the entire population is fixed for the same strategy, but individuals can copy the group identity of others. This turns out to be almost identical to the one-group case outlined in section 8.1, except that the relevant transition probability is instead

$$\phi(\Pi_J, \Pi_I) = \frac{1}{1 + \exp[\beta(\Pi_J - \Pi_I)]}.$$

The remainder of the argument proceeds identically, except with f^i and f^j replaced with ν_I and ν_J, and we obtain

$$\begin{aligned}\dot{\nu}_I &= \nu_I (\Pi_I - \bar{\Pi}), \text{with} \\ \Pi_I &= \sum_i f_I^i \Pi_I^i.\end{aligned}$$

8.3. Multiple groups, disjoint strategic imitation. When there is more than one group ($K > 1$), the analysis of SI Section 8.1 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

$$\dot{f}_I^i = f_I^i (\Pi_I^i - \bar{\Pi}_I). \quad [29]$$

8.4. Multiple groups, well-mixed strategic imitation. 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 ν_I be the frequency of group I, and let $n_I^i = N \nu_I f_I^i$ 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 $\nu_I f_I^j$. With probability $\nu_J f_J^i$, 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 $\phi(\Pi_I^j, \Pi_J^i)$. n_I^i increases by 1.

2. **Decrease**. A type i individual in group I is chosen to update with probability $\nu_I f_I^i$. With probability $\nu_J f_J^j$, 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 $\phi(\Pi_I^i, \Pi_J^j)$. n_I^i decreases by 1.

Thus

$$\begin{aligned}
\mathbb{E}\Big[\Delta n_I^i\Big] &= \mathbb{P}\Big(\Delta n_I^i = 1\Big) - \mathbb{P}\Big(\Delta n_I^i = -1\Big) \\
&= \nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \phi(\Pi_I^j, \Pi_J^i) - \nu_I f_I^i \sum_j \sum_J \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j) \\
&\approx \nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \Big(\frac{1}{2} + \beta \frac{\Pi_J^i - \Pi_I^j}{4}\Big) - \nu_I f_I^i \sum_j \sum_J \nu_J f_J^j \Big(\frac{1}{2} + \beta \frac{\Pi_J^j - \Pi_I^i}{4}\Big)\Big] \\
&= \nu_I \frac{1}{2}\Big[\sum_j f_I^j \sum_J \nu_J f_J^i - f_I^i \sum_j \sum_J \nu_J f_J^j\Big] \\
&\quad + \nu_I \frac{\beta}{4} \sum_j f_I^j \Big[\sum_J \nu_J f_J^i (\Pi_J^i - \Pi_I^j) - f_I^i \sum_J \nu_J f_J^j (\Pi_J^j - \Pi_I^i)\Big] \\
&= \nu_I \frac{1}{2}\Big[\sum_J \nu_J \big(f_J^i \sum_j f_I^j - f_I^i \sum_j f_J^j\big)\Big] \\
&\quad + \nu_I \frac{\beta}{4}\Big[\sum_J \nu_J \sum_j \Big(f_I^j f_J^i (\Pi_J^i - \Pi_I^j) - f_I^i f_J^j (\Pi_J^j - \Pi_I^i)\Big)\Big]. \\
&= \nu_I \frac{1}{2}\Big[\sum_J \nu_J \big(f_J^i - f_I^i\big)\Big] \\
&\quad + \nu_I \frac{\beta}{4}\Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \sum_j f_I^j \Pi_I^j) + f_I^i (\Pi_I^i - \sum_j f_J^j \Pi_J^j)\big)\Big].
\end{aligned}$$

Rescaling time allows us to recast this as an equation for $\dot{n}_I^i$. Recalling that $n_I^i = N\nu_I f_I^i$, and dropping the $1/2$ prefactor, we have

$$\dot{f}_I^i \propto \sum_J \nu_J \Big[f_J^i - f_I^i\Big] + \frac{\beta}{2}\Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \sum_j f_I^j \Pi_I^j) + f_I^i (\Pi_I^i - \sum_j f_J^j \Pi_J^j)\big)\Big].$$

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

$$\begin{aligned}
\dot{f}^i &= \sum_I \nu_I \dot{f}_I^i \\
&\propto \sum_I \nu_I \Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \sum_j f_I^j \Pi_I^j) + f_I^i (\Pi_I^i - \sum_j f_J^j \Pi_J^j)\big)\Big] \\
&= \sum_J \nu_J f^i \Pi_J^i - f^i \sum_I \nu_I \sum_j f^j \Pi_I^j + \sum_I \nu_I f^i \Pi_I^i - f^i \sum_J \nu_J \sum_j f^j \Pi_J^j \\
&= f^i \Big(\sum_J \nu_J \Pi_J^i + \sum_I \nu_I \Pi_I^i - \sum_j f^j \sum_I \nu_I \Pi_I^j - \sum_j f^j \sum_J \nu_J \Pi_J^j\Big) \\
&\propto f^i \sum_J \nu_J \big(\Pi_J^i - \Pi_J\big).
\end{aligned}$$

Rescaling time allows us to write this as an equality:

$$\begin{aligned}\dot{f}^i &= f^i \sum_J \nu_J \big(\Pi^i_J - \Pi_J\big)\\ &= f^i\Big(\sum_J \nu_J \Pi^i_J - \bar{\Pi}\Big) = f^i\big(\Pi^i - \bar{\Pi}\big), \text{with}\\ \Pi^i &= \sum_L \nu_L \Pi^i_L,\\ \bar{\Pi} &= \sum_L \nu_L \sum_i f^i \Pi^i_L = \sum_L \nu_L \Pi_L = \sum_i f^i \Pi^i.\end{aligned} \tag{30}$$

8.5. 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 (disjoint imitation) 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_I, the following events are possible.

1. **Increase**. A type j individual in group I is chosen to update with probability $\nu_I f^j_I$. With probability $(1-m)f^i_I$, the compared individual is type i, I, and with probability $\nu_J m f^i_J$, the compared individual is type i, J (J can be I). The update happens with probability $\phi(\Pi^j_I, \Pi^i_I)$ (for i, I) or $\phi(\Pi^j_I, \Pi^i_J)$ (for type i, J). In either case, n^i_I increases by 1.

2. **Decrease**. A type i individual in group I is chosen to update with probability $\nu_I f^i_I$. With probability $(1-m)f^j_I$, the compared individual is type $j(\neq i), I$, and with probability $\nu_J m f^j_J$, the compared individual is type $j(\neq i), J$ (J can be I). The update happens with probability $\phi(\Pi^i_I, \Pi^j_I)$ (for j, I) or $\phi(\Pi^i_I, \Pi^j_J)$ (for j, J). In either case, n^i_I decreases by 1.

We thus have

$$\begin{aligned}
\mathbb{E}\Big[\Delta n_I^i\Big] &= \mathbb{P}\Big(\Delta n_I^i = 1\Big) - \mathbb{P}\Big(\Delta n_I^i = -1\Big) \\
&= (1-m)\Big[\nu_I \sum_j f_I^j f_I^i \phi(\Pi_I^j, \Pi_I^i) - \nu_I f_I^i \sum_j f_I^j \phi(\Pi_I^i, \Pi_I^j)\Big] \\
&\quad + m\Big[\nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \phi(\Pi_I^j, \Pi_J^i) - \nu_I f_I^i \sum_j \sum_J \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j)\Big] \\
&\approx (1-m)\Big[\nu_I \sum_j f_I^j f_I^i \big(\frac{1}{2} + w\frac{\Pi_I^i - \Pi_I^j}{4}\big) - \nu_I f_I^i \sum_j f_I^j \big(\frac{1}{2} + w\frac{\Pi_I^j - \Pi_I^i}{4}\big)\Big] \\
&\quad + m\Big[\nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \big(\frac{1}{2} + w\frac{\Pi_J^i - \Pi_I^j}{4}\big) - \nu_I f_I^i \sum_j \sum_J \nu_J f_J^j \big(\frac{1}{2} + w\frac{\Pi_J^j - \Pi_I^i}{4}\big)\Big] \\
&= m\nu_I \frac{1}{2}\Big[\sum_j f_I^j \sum_J \nu_J f_J^i - f_I^i \sum_j \sum_J \nu_J f_J^j\Big] \\
&\quad + (1-m)\nu_I \frac{w}{2} \sum_j f_I^j f_I^i (\Pi_I^i - \Pi_I^j) \\
&\quad + m\nu_I \frac{w}{4} \sum_j f_I^j \Big[\sum_J \nu_J f_J^i (\Pi_J^i - \Pi_I^j) - f_I^i \sum_J \nu_J f_J^j (\Pi_J^j - \Pi_I^i)\Big] \\
&= m\nu_I \frac{1}{2}\Big[\sum_J \nu_J \big(f_J^i \sum_j f_I^j - f_I^i \sum_j f_J^j\big)\Big] \\
&\quad + (1-m)\nu_I \frac{w}{2} f_I^i \big(\Pi_I^i - \sum_j f_I^j \Pi_I^j\big) \\
&\quad + m\nu_I \frac{w}{4}\Big[\sum_J \nu_J \sum_j \big(f_I^j f_J^i (\Pi_J^i - \Pi_I^j) - f_I^i f_J^j (\Pi_J^j - \Pi_I^i)\big)\Big]. \\
&= m\nu_I \frac{1}{2}\Big[\sum_J \nu_J \big(f_J^i - f_I^i\big)\Big] \\
&\quad + (1-m)\nu_I \frac{w}{2} f_I^i \big(\Pi_I^i - \Pi_I\big) \\
&\quad + m\nu_I \frac{w}{4}\Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \Pi_I) + f_I^i (\Pi_I^i - \Pi_J)\big)\Big].
\end{aligned}$$

Recalling that $n_I^i = N\nu_I f_I^i$, the replicator dynamics are given by

$$\begin{aligned}
\dot{f}_I^i &\propto m\frac{1}{2}\Big[\sum_J \nu_J \big(f_J^i - f_I^i\big)\Big] \\
&\quad + (1-m)\frac{w}{2} f_I^i \big(\Pi_I^i - \Pi_I\big) \\
&\quad + m\frac{w}{4}\Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \Pi_I) + f_I^i (\Pi_I^i - \Pi_J)\big)\Big] \\
&\propto \frac{m}{w}\Big[\sum_J \nu_J \big(f_J^i - f_I^i\big)\Big] \\
&\quad + (1-m) f_I^i \big(\Pi_I^i - \Pi_I\big) \\
&\quad + \frac{m}{2}\Big[\sum_J \nu_J \big(f_J^i (\Pi_J^i - \Pi_I) + f_I^i (\Pi_I^i - \Pi_J)\big)\Big]
\end{aligned}$$

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 Eq. [29] and setting $m = 1$ yields Eq. [30], subject to rescaling.

8.6. Multiple groups, copying both strategy and group identity. We now assume that individuals engage in both well-mixed strategic copying *and* copying of group identity. With probability τ, an individual resolves to update their group identity; with probability $1-\tau$, they resolve to update their behavioral strategy. We consider here the possible change in n_I^i.

1. **Increase...**

   (a) **...by changing group identity**. A type i individual in group $J \in \{1 \dots K\}$ is chosen to update with probability $\nu_J f_J^i$. With probability τ, they choose to update their group identity. With probability $\nu_I f_I^j$, the comparison partner is type j (any strategy) and in group I. The update happens with probability $\phi(\Pi_J^i, \Pi_I^j)$.

   (b) **...by changing behavioral strategy**. A type j (any strategy) individual in group I is chosen to update with probability $\nu_I f_I^j$. With probability $1-\tau$, they choose to update their behavioral strategy. With probability $\nu_J f_J^i$, the comparison partner is type i and in group $J \in \{1 \dots K\}$. The update happens with probability $\phi(\Pi_I^j, \Pi_J^i)$.

2. **Decrease...**

   (a) **...by changing group identity**. A type i individual in group I is chosen to update with probability $\nu_I f_I^i$. With probability τ, they choose to update their group identity. With probability $\nu_J f_J^j$, the comparison partner is type j (any strategy) and in group $J \in \{1 \dots K\}$. The update happens with probability $\phi(\Pi_I^i, \Pi_J^j)$.

   (b) **...by changing behavioral strategy**. A type i individual in group I is chosen to update with probability $\nu_I f_I^i$. With probability $1-\tau$, they choose to update their behavioral strategy. With probability $\nu_J f_J^j$, the comparison partner is type j (any strategy) and in group $J \in \{1 \dots K\}$. The update happens with probability $\phi(\Pi_I^i, \Pi_J^j)$.

We have

$$\begin{aligned}
\mathbb{E}\Big[\Delta n_I^i\Big] &= \mathbb{P}\Big(\Delta n_I^i = 1\Big) - \mathbb{P}\Big(\Delta n_I^i = -1\Big) \\
&= \tau\Big(\sum_J \nu_J f_J^i \nu_I \sum_j f_I^j \phi(\Pi_J^i, \Pi_I^j) - \nu_I f_I^i \sum_J \sum_j \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j)\Big) \\
&\quad + (1-\tau)\Big(\nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \phi(\Pi_I^j, \Pi_J^i) - \nu_I f_I^i \sum_J \sum_j \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j)\big) \\
&= \tau\Big(\nu_I \sum_J \nu_J f_J^i \sum_j f_I^j \phi(\Pi_J^i, \Pi_I^j) - \nu_I f_I^i \sum_J \sum_j \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j)\Big) \\
&\quad + (1-\tau)\Big(\nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \phi(\Pi_I^j, \Pi_J^i) - \nu_I f_I^i \sum_J \sum_j \nu_J f_J^j \phi(\Pi_I^i, \Pi_J^j)\big) \\
&\approx \tau \nu_I \sum_J \nu_J f_J^i \sum_j f_I^j \Big(\frac{1}{2} + \beta \frac{\Pi_I^j - \Pi_J^i}{4}\Big) + (1-\tau)\nu_I \sum_j f_I^j \sum_J \nu_J f_J^i \Big(\frac{1}{2} + \beta \frac{\Pi_J^i - \Pi_I^j}{4}\Big) \\
&\quad - \nu_I f_I^i \sum_J \sum_j \nu_J f_J^j \Big(\frac{1}{2} + \beta \frac{\Pi_J^j - \Pi_I^i}{4}\Big) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I \frac{\beta}{4} \sum_J \nu_J f_J^i \sum_j f_I^j \big[\tau(\Pi_I^j - \Pi_J^i) + (1-\tau)(\Pi_J^i - \Pi_I^j)\big] \\
&\quad - \nu_I f_I^i \frac{\beta}{4} \sum_J \sum_j \nu_J f_J^j (\Pi_J^j - \Pi_I^i) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I \frac{\beta}{4} \sum_J \nu_J f_J^i \sum_j f_I^j (1-2\tau)(\Pi_J^i - \Pi_I^j) \\
&\quad - \nu_I f_I^i \frac{\beta}{4} \sum_J \sum_j \nu_J f_J^j (\Pi_J^j - \Pi_I^i) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I (1-2\tau) \frac{\beta}{4} \Big(\sum_J \nu_J f_J^i \sum_j f_I^j \Pi_J^i - \sum_J \nu_J f_J^i \sum_j f_I^j \Pi_I^j\Big) \\
&\quad - \nu_I f_I^i \frac{\beta}{4} \Big(\sum_J \sum_j \nu_J f_J^j \Pi_J^j - \sum_J \sum_j \nu_J f_J^j \Pi_I^i\Big) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I (1-2\tau) \frac{\beta}{4} \Big(\sum_J \nu_J f_J^i \Pi_J^i \sum_j f_I^j - \sum_j f_I^j \Pi_I^j \sum_J \nu_J f_J^i\Big) \\
&\quad - \nu_I f_I^i \frac{\beta}{4} \Big(\sum_J \sum_j \nu_J f_J^j \Pi_J^j - \sum_J \sum_j \nu_J f_J^j \Pi_I^i\Big) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I (1-2\tau) \frac{\beta}{4} \Big(\sum_J \nu_J f_J^i \Pi_J^i - f^i \sum_j f_I^j \Pi_I^j\Big) \\
&\quad - \nu_I f_I^i \frac{\beta}{4} \Big(\sum_J \sum_j \nu_J f_J^j \Pi_J^j - \Pi_I^i\Big) \\
&= \nu_I \frac{1}{2}(f^i - f_I^i) + \nu_I (1-2\tau) \frac{\beta}{4} \Big(\sum_J \nu_J f_J^i \Pi_J^i - f^i \Pi_I\Big) - \nu_I f_I^i \frac{\beta}{4} \Big(\bar{\Pi} - \Pi_I^i\Big) \\
&\propto \nu_I \big(f^i - f_I^i + \frac{\beta}{2} \Big[(1-2\tau)\Big(\sum_J \nu_J f_J^i \Pi_J^i - f^i \Pi_I\Big) + f_I^i \Big(\Pi_I^i - \bar{\Pi}\Big)\Big]\big).
\end{aligned}$$

Observe that the $f^i - f_I^i$ term lacks a prefactor and therefore will dominate, so we expect that all f_I^i will rapidly equilibrate to a common value f^i. The result is actually a system of equations in both f_I^i and ν_I, since $\nu_I = \sum_j n_I^j / N$

and $f^i = n^i/(N\nu_I)$. Thus

$$\begin{aligned}
\nu_I &= \frac{1}{N}\sum_j n_I^j \\
\therefore \dot{\nu}_I &= \frac{1}{N}\sum_j \dot{n}_I^j \\
&\propto \frac{1}{N}\sum_j \nu_I \big(f^j - f_I^j + \frac{\beta}{2}\Big[(1-2\tau)\Big(\sum_J \nu_J f_I^j \Pi_I^j - f^j \Pi_I\Big) + f_I^j\Big(\Pi_I^j - \bar{\Pi}\Big)\Big]) \\
&= \frac{1}{N}\nu_I \frac{\beta}{2}\Big[(1-2\tau)\sum_j \Big(\sum_J \nu_J f_I^j \Pi_I^j - f^j \Pi_I\Big) + \sum_j f_I^j (\Pi_I^j - \bar{\Pi})\Big] \\
&= \frac{1}{N}\nu_I \frac{\beta}{2}\Big[(1-2\tau)\big(\bar{\Pi} - \Pi_I\big) + \Pi_I - \bar{\Pi}\Big] \\
&= \frac{1}{N}\nu_I \beta\tau(\Pi_I - \bar{\Pi}).
\end{aligned}$$

As expected, when $\tau \to 0$ (i.e., individuals never update their group identity), this vanishes. For positive τ, ν_I changes at a rate that depends on the difference between ν_I's fitness and the population average. We can take advantage of the fact that the f_I^i equilibrate rapidly to a common value f^i and average out the fact that fitnesses Π_I^i may differ by group, by considering only

$$
\begin{aligned}
f^i &= \sum_I \nu_I f_I^i \\
\therefore \dot{f}^i &= \frac{d}{dt}\Big(\sum_I \nu_I f_I^i) \\
&= \sum_I (\dot{\nu}_I f_I^i + \nu_I \dot{f}_I^i) \\
&= \sum_I \nu_I \dot{f}_I^i \\
&= \sum_I \nu_I \frac{1}{N} \frac{d}{dt} \frac{n_I^i}{\nu_I} \\
&= \frac{1}{N} \sum_I \nu_I \frac{\dot{n}_I^i \nu_I - n_I^i \dot{\nu}_I}{(\nu_I)^2} \\
&= \frac{1}{N} \sum_I \nu_I \frac{\dot{n}_I^i \nu_I - N \nu_I f_I^i \dot{\nu}_I}{(\nu_I)^2} \\
&= \frac{1}{N} \sum_I \Big(\dot{n}_I^i - N f_I^i \dot{\nu}_I\Big) \\
&= \frac{1}{N} \sum_I \nu_I \big(f^i - f_I^i + \frac{\beta}{2}\Big[(1-2\tau)\Big(\sum_J \nu_J f_J^i \Pi_J^i - f_I^i \Pi_I\Big) + f_I^i \Big(\Pi_I^i - \bar{\Pi}\Big)\Big] - f_I^i \beta\tau(\Pi_I - \bar{\Pi})\big) \\
&= \frac{1}{N} \sum_I \nu_I \big(\frac{\beta}{2}\Big[(1-2\tau)\Big(\sum_J \nu_J f_J^i \Pi_J^i - f_I^i \Pi_I\Big) + f_I^i \Big(\Pi_I^i - \bar{\Pi}\Big)\Big] - f_I^i \beta\tau(\Pi_I - \bar{\Pi})\big) \\
&= \frac{1}{N} \sum_I \nu_I \big(\frac{\beta}{2}\Big[(1-2\tau)\Big(\sum_J \nu_J f^i \Pi_J^i - f^i \Pi_I\Big) + f^i \Big(\Pi_I^i - \bar{\Pi}\Big)\Big] - f^i \beta\tau(\Pi_I - \bar{\Pi})\big) \\
&= \frac{1}{N} \big(\frac{\beta}{2}\Big[(1-2\tau)\Big(f^i \sum_I \nu_I \sum_J \nu_J \Pi_J^i - f^i \sum_I \nu_I \Pi_I\Big) + f^i \Big(\sum_I \nu_I \Pi_I^i - \sum_I \nu_I \bar{\Pi}\Big)\Big] \\
&\quad - f^i \beta\tau(\sum_I \nu_I \Pi_I - \sum_I \nu_I \bar{\Pi})\big) \\
&= \frac{1}{N} f^i \big(\frac{\beta}{2}\Big[(1-2\tau)\Big(\sum_I \nu_I \Pi^i - \bar{\Pi}\Big) + \Big(\Pi^i - \bar{\Pi}\Big)\Big] - \beta\tau(\bar{\Pi} - \bar{\Pi})\big) \\
&= \frac{1}{N} f^i \beta \Big[(1-\tau)(\Pi^i - \bar{\Pi})\Big].
\end{aligned}
$$

Sending $\tau \to 1$ (i.e., individuals only ever update their group identity, not their behavioral strategy) yields $\dot{f}^i = 0$ due to selection. (The leading term in the expression for $\dot{n}_I^i$ still has no β prefactor and, thus, corresponds to the f_I^i equilibrating as a result of random group identity switching, even in the absence of behavioral strategy updating, so the necessary assumption that the f_I^i equilibrate rapidly is not violated.) Rescaling time again so as to drop the β/N prefactor yields the system of equations

$$
\begin{aligned}
\dot{\nu}_I &= \nu_I \tau(\Pi_I - \bar{\Pi}), \\
\dot{f}^i &= f^i(1-\tau)(\Pi^i - \bar{\Pi}).
\end{aligned}
$$

8.7. Multiple groups, with an immediate switching cost. We now impose a fitness cost to switching groups, so that individuals are less likely to switch groups. Call the cost α. Imposing this fitness cost is tantamount to amending the form of $\phi(\Pi_I, \Pi_J)$, so that

$$\phi(\Pi_J, \Pi_I) = \frac{1}{1 + \exp[\beta(\Pi_J - \Pi_I - \alpha)]}.$$

Thus,

$$\begin{aligned}
\mathbb{E}[\Delta\nu_I] &= \frac{1}{N}\mathbb{P}\Big(\Delta\nu_I = \frac{1}{N}\Big) - \frac{1}{N}\mathbb{P}\Big(\Delta\nu_I = -\frac{1}{N}\Big) \\
&= \frac{1}{N}\left(\sum_J \nu_J\nu_I\phi(\Pi_J,\Pi_I) - \nu_I\sum_J \nu_J\phi(\Pi_I,\Pi_J)\right) \\
&= \frac{1}{N}\nu_I\left(\sum_J \nu_J\Big[\phi(\Pi_J,\Pi_I) - \phi(\Pi_I,\Pi_J)\Big]\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\phi(\Pi_J,\Pi_I) = \frac{1}{1+\exp[\beta(\Pi_J-\Pi_I-\alpha)]} &\approx \frac{1}{1+\exp[\beta(\Pi_J-\Pi_I-\alpha)]}\Big|_{\beta=0} + \beta\Big(\frac{d}{d\beta}\Big[\frac{1}{1+\exp[\beta(\Pi_J-\Pi_I-\alpha)]}\Big]\Big)\Big|_{\beta=0} + \mathcal{O}(\beta^2) \\
&= \frac{1}{2} + \beta\Big[\frac{(\Pi_I-\Pi_J-\alpha)\exp[\beta(\Pi_J-\Pi_I-\alpha)]}{(1+\exp[\beta(\Pi_J-\Pi_I-\alpha)])^2}\Big]\Big|_{\beta=0} + \mathcal{O}(\beta^2) \\
&= \frac{1}{2} + \beta\frac{\Pi_I-\Pi_J-\alpha}{4} + \mathcal{O}(\beta^2).
\end{aligned}$$

Hence

$$\begin{aligned}
\phi(\Pi_J,\Pi_I) - \phi(\Pi_I,\Pi_J) &\approx \beta\frac{\Pi_I-\Pi_J-\alpha}{4} - \beta\frac{\Pi_J-\Pi_I-\alpha}{4} + \mathcal{O}(\beta^2) \\
&\approx \beta\frac{\Pi_I-\Pi_J}{2} + \mathcal{O}(\beta^2).
\end{aligned}$$

We therefore have

$$\begin{aligned}
\mathbb{E}[\Delta\nu_I] &= \frac{1}{N}\nu_I\sum_J\Big[\frac{\beta}{2}\nu_J(\Pi_I-\Pi_J) + \mathcal{O}(\beta^2)\Big] \\
&\approx \frac{\beta}{2N}\nu_I\sum_J \nu_J(\Pi_I-\Pi_J) \\
&= \frac{\beta}{2N}\nu_I(\Pi_I-\bar{\Pi}).
\end{aligned}$$

This model with an instantaneous switching cost is identical to equation 8.2, i.e., the case of no switching cost.

8.8. Multiple groups, with a transient switching cost. We now posit that there is a “cost” to switching group membership, but that the cost continues to impacts an individual’s fitness for some duration after the switch. The cost is α, and the rate at which individuals transition from the “new” state to the “established” state is σ. In other words, an individual who switches to a new group pays a fitness cost α for a typical duration of time $1/\sigma$ – after which the individual becomes an established member of the group and no longer pays the cost of having switched.

We also posit that, when an individual switches back to their old group, they no longer suffer the cost. Individuals can end up in the “new” category in their opposing group by being “old”‘ in their current group and comparing their fitness to that of either “old” or “new”‘ individuals in the opposing group. Individuals can end up in the “old” category in their opposing group either by being “new” in that group and transitioning (at rate σ) or by being “new” in their current group and (via selection) switching groups.

Under this model, consider the dynamics of ν_1^{new}. It can change by:

1. **Increase**. A group $2,$ old individual (probability ν_2^{old}) compares themselves to any group 1 individual, either old or new (probability ν_1^{new} or ν_1^{old}, respectively). The update then happens with probability $\phi(\Pi_2^{\text{old}},\Pi_1^{\text{new}})$ or $\phi(\Pi_2^{\text{old}},\Pi_1^{\text{old}})$ respectively.

2. **Decrease**. A group $1,$ new individual (probability ν_1^{new}) compares themselves to any group 2 individual, either old or new (probability ν_2^{new} or ν_2^{old}, respectively). The update then happens with probability $\phi(\Pi_1^{\text{new}},\Pi_2^{\text{new}})$ or $\phi(\Pi_1^{\text{new}},\Pi_2^{\text{old}})$ respectively. Otherwise, ν_1^{new} also constantly decreases at rate σ.

Recalling that

$$\phi(\Pi_2^{\text{old}}, \Pi_1^{\text{new}}) \approx \frac{1}{2} + \beta\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{old}}}{4} + \mathcal{O}(\beta^2),$$

we thus have

$$\begin{aligned}
\mathbb{E}[\Delta\nu_1^{\text{new}}] &= \frac{1}{N}\mathbb{P}\Big(\Delta\nu_1^{\text{new}} = \frac{1}{N}\Big) - \frac{1}{N}\mathbb{P}\Big(\Delta\nu_1^{\text{new}} = -\frac{1}{N}\Big) \\
&= \frac{1}{N}\Big(\nu_2^{\text{old}}[\nu_1^{\text{old}}\phi(\Pi_2^{\text{old}}, \Pi_1^{\text{old}}) + \nu_1^{\text{new}}\phi(\Pi_2^{\text{old}}, \Pi_1^{\text{new}})] \\
&\quad - \nu_1^{\text{new}}[\nu_2^{\text{old}}\phi(\Pi_1^{\text{new}}, \Pi_2^{\text{old}}) + \nu_2^{\text{new}}\phi(\Pi_1^{\text{new}}, \Pi_2^{\text{new}}) + \sigma]\Big) \\
&\approx \frac{1}{N}\Big(\nu_2^{\text{old}}[\nu_1^{\text{old}}(\frac{1}{2} + \beta\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{old}}}{4}) + \nu_1^{\text{new}}(\frac{1}{2} + \beta\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{old}}}{4})] \\
&\quad - \nu_1^{\text{new}}[\nu_2^{\text{old}}(\frac{1}{2} + \beta\frac{\Pi_2^{\text{old}} - \Pi_1^{\text{new}}}{4}) + \nu_2^{\text{new}}(\frac{1}{2} + \beta\frac{\Pi_2^{\text{new}} - \Pi_1^{\text{new}}}{4}) + \sigma]\Big) \\
\therefore \dot{\nu}_1^{\text{new}} &= \nu_2^{\text{old}}\nu_1^{\text{new}}\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{old}}}{2} + \nu_2^{\text{old}}\nu_1^{\text{old}}\big(\frac{1}{2\beta} + \frac{\Pi_1^{\text{old}} - \Pi_2^{\text{old}}}{4}\big) \\
&\quad - \nu_1^{\text{new}}\nu_2^{\text{new}}\big(\frac{1}{2\beta} + \frac{\Pi_2^{\text{new}} - \Pi_1^{\text{new}}}{4}\big) - \frac{\sigma}{\beta}\nu_1^{\text{new}}.
\end{aligned}$$

In the low-β limit, this is

$$\dot{\nu}_1^{\text{new}} \approx \nu_2^{\text{old}}\nu_1^{\text{old}} - \nu_1^{\text{new}}\nu_2^{\text{new}} - \sigma\nu_1^{\text{new}}.$$

after rescaling σ.

Likewise, consider ν_1^{old}. It can change by:

1. **Increase**. A group 2, new individual (probability ν_2^{new}) compares themselves to any group 1 individual, either old or new (probability ν_1^{new} or ν_1^{old}, respectively). The update then happens with probability $\phi(\Pi_2^{\text{new}}, \Pi_1^{\text{new}})$ or $\phi(\Pi_2^{\text{new}}, \Pi_1^{\text{old}})$ respectively. The increase also happens secularly at rate $\sigma\nu_1^{\text{new}}$.

2. **Decrease**. A group 1, old individual (probability ν_1^{old}) compares themselves to any group 2 individual, either old or new (probability ν_2^{new} or ν_2^{old}, respectively). The update then happens with probability $\phi(\Pi_1^{\text{old}}, \Pi_2^{\text{new}})$ or $\phi(\Pi_1^{\text{old}}, \Pi_2^{\text{old}})$ respectively.

Thus,

$$\begin{aligned}
\mathbb{E}[\Delta\nu_1^{\text{old}}] &= \frac{1}{N}\mathbb{P}\Big(\Delta\nu_1^{\text{old}} = \frac{1}{N}\Big) - \frac{1}{N}\mathbb{P}\Big(\Delta\nu_1^{\text{old}} = -\frac{1}{N}\Big) \\
&= \frac{1}{N}\Big(\nu_2^{\text{new}}[\nu_1^{\text{old}}\phi(\Pi_2^{\text{new}}, \Pi_1^{\text{old}}) + \nu_1^{\text{new}}\phi(\Pi_2^{\text{new}}, \Pi_1^{\text{new}})] \\
&\quad - \nu_1^{\text{old}}[\nu_2^{\text{old}}\phi(\Pi_1^{\text{old}}, \Pi_2^{\text{old}}) + \nu_2^{\text{new}}\phi(\Pi_1^{\text{old}}, \Pi_2^{\text{new}})] + \nu_1^{\text{new}}\sigma\Big) \\
&\approx \frac{1}{N}\Big(\nu_2^{\text{new}}[\nu_1^{\text{old}}(\frac{1}{2} + \beta\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{new}}}{4}) + \nu_1^{\text{new}}(\frac{1}{2} + \beta\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{new}}}{4})] \\
&\quad - \nu_1^{\text{old}}[\nu_2^{\text{old}}(\frac{1}{2} + \beta\frac{\Pi_2^{\text{old}} - \Pi_1^{\text{old}}}{4}) + \nu_2^{\text{new}}(\frac{1}{2} + \beta\frac{\Pi_2^{\text{new}} - \Pi_1^{\text{old}}}{4})] + \nu_1^{\text{new}}\sigma\Big) \\
\therefore \dot{\nu}_1^{\text{old}} &= \nu_2^{\text{new}}\nu_1^{\text{old}}\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{new}}}{2} + \nu_2^{\text{new}}\nu_1^{\text{new}}\big(\frac{1}{2\beta} + \frac{\Pi_1^{\text{new}} - \Pi_2^{\text{new}}}{4}\big) \\
&\quad - \nu_1^{\text{old}}\nu_2^{\text{old}}\big(\frac{1}{2\beta} + \frac{\Pi_2^{\text{old}} - \Pi_1^{\text{old}}}{4}\big) + \frac{\sigma}{\beta}\nu_1^{\text{new}}.
\end{aligned}$$

In the low-β limit, this is

$$\dot{\nu}_1^{\text{old}} \approx \nu_2^{\text{new}}\nu_1^{\text{new}} - \nu_1^{\text{old}}\nu_2^{\text{old}} + \sigma\nu_1^{\text{new}}.$$

The total $\dot{\nu}_1$ is

$$\begin{aligned}
\dot{\nu}_1 &= \dot{\nu}_1^{\text{new}} + \dot{\nu}_1^{\text{old}} \\
&= \nu_2^{\text{old}}\nu_1^{\text{new}}\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{old}}}{2} + \nu_2^{\text{old}}\nu_1^{\text{old}}\Big(\frac{1}{2\beta} + \frac{\Pi_1^{\text{old}} - \Pi_2^{\text{old}}}{4}\Big) \\
&\quad - \nu_1^{\text{new}}\nu_2^{\text{new}}\Big(\frac{1}{2\beta} + \frac{\Pi_2^{\text{new}} - \Pi_1^{\text{new}}}{4}\Big) - \frac{\sigma}{\beta}\nu_1^{\text{new}} \\
&\quad + \nu_2^{\text{new}}\nu_1^{\text{old}}\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{new}}}{2} + \nu_2^{\text{new}}\nu_1^{\text{new}}\Big(\frac{1}{2\beta} + \frac{\Pi_1^{\text{new}} - \Pi_2^{\text{new}}}{4}\Big) \\
&\quad - \nu_1^{\text{old}}\nu_2^{\text{old}}\Big(\frac{1}{2\beta} + \frac{\Pi_2^{\text{old}} - \Pi_1^{\text{old}}}{4}\Big) + \frac{\sigma}{\beta}\nu_1^{\text{new}} \\
&= \nu_2^{\text{old}}\nu_1^{\text{new}}\frac{\Pi_1^{\text{new}} - \Pi_2^{\text{old}}}{2} + \nu_2^{\text{new}}\nu_1^{\text{old}}\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{new}}}{2} + \nu_2^{\text{old}}\nu_1^{\text{old}}\frac{\Pi_1^{\text{old}} - \Pi_2^{\text{old}}}{2} + \nu_1^{\text{new}}\nu_2^{\text{new}}\frac{\Pi_2^{\text{new}} - \Pi_1^{\text{new}}}{2} \\
&= \nu_1\nu_2\bar{\Pi}_1 - \nu_1\nu_2\bar{\Pi}_2 \\
&= \nu_1(\bar{\Pi}_1 - \bar{\Pi}),
\end{aligned}$$

after rescaling time to eliminate the $1/2$ constant and defining

$$\begin{aligned}
\bar{\Pi}_1 &= (\nu_1^{\text{new}}\Pi_1^{\text{new}} + \nu_1^{\text{old}}\Pi_1^{\text{old}})/\nu_1, \\
\bar{\Pi}_2 &= (\nu_2^{\text{new}}\Pi_2^{\text{new}} + \nu_2^{\text{old}}\Pi_2^{\text{old}})/\nu_2.
\end{aligned}$$

Note that the expression for $\dot{\nu}_1$ is independent of β, but the individual expressions for $\dot{\nu}_1^{\text{new}}$ and $\dot{\nu}_1^{\text{old}}$ contain inverse powers of β; the approximate ("low-β") expressions are independent of the fitnesses. This means we may make the *Ansatz* that, while ν_1 changes only slowly (over timescales associated with selection), ν_1^{new} and ν_1^{old} equilibrate quickly in response to changes in ν_1; the separation of timescales is set by $1/\beta$. Define ρ_1 and ρ_2, the fraction of groups 1 and 2 that are in the "new" state, so that $\nu_1^{\text{new}} = \rho_1\nu_1$. Then

$$\begin{aligned}
\dot{\nu}_1^{\text{new}} &= \dot{\rho}_1\nu_1 + \rho_1\dot{\nu}_1 \\
&\approx \dot{\rho}_1\nu_1 \\
\therefore \dot{\rho}_1 &= \dot{\nu}_1^{\text{new}}/\nu_1 \\
\therefore \dot{\rho}_1 &= \big(\nu_2^{\text{old}}\nu_1^{\text{old}} - \nu_1^{\text{new}}\nu_2^{\text{new}} - \sigma\nu_1^{\text{new}}\big)/\nu_1 \\
&= (1-\rho_2)(1-\rho_1)\nu_2 - \rho_1\rho_2\nu_2 - \sigma\rho_1 \\
&= \nu_2(1-\rho_1-\rho_2) - \sigma\rho_1 \\
&= (1-\nu_1)(1-\rho_1-\rho_2) - \sigma\rho_1, \\
\dot{\rho}_2 &= \nu_1(1-\rho_1-\rho_2) - \sigma\rho_2.
\end{aligned}$$

These equations have an equilibrium at

$$\rho_1 = \frac{1-\nu_1}{1+\sigma}, \rho_2 = \frac{\nu_1}{1+\sigma},$$

which is stable.

9. The "main character" effect

A common feature of social networks, especially in social media, is the existence of an individual who is assigned a bad reputation by the entire population, for example due to high-visibility acts of bad behavior. On social networks like Twitter, such individuals are referred to as "main characters". In this section, we consider how the existence of such a "main character" affects the rate of cooperation in a group-structured population of discriminators (DISC). And we also study whether unconditional defectors (ALLD) can invade a population of discriminators, when a "main character" is present.

We model the main character effect as follows. We assume that, with probability $1-m$, each individual's reputation is assessed based on how they behaved toward a random other individual in the population, as in prior models we have studied. However, with probability m, the individual is instead assessed based on how they behaved towards the singular main character. (For simplicity, we do not consider insular social interactions in this version of the model.) Since the main character is considered bad by the entire population, a discriminator will assuredly defect with them. The reputational profile of discriminators is thus given by

$$\begin{aligned} g_{\text{in}} &= (1-m)\big[gP^{GC} + (1-g)P^{BD}\big] + mP^{BD} \\ &= (1-m)g(P^{GC} - P^{BD}) + P^{BD}, \\ g_{\text{out}} &= (1-m)\big[GP^{GC} + (g-G)(P^{GD} + P^{BC}) + (1-2g+G)P^{BD}\big] + mP^{BD} \\ &= (1-m)\big[G(P^{GC} - P^{GD} - P^{BC} + P^{BD}) + g(P^{GD} + P^{BC} - 2P^{BD})\big] + P^{BD}, \end{aligned}$$

in which

$$\begin{aligned} g &= \frac{g_{\text{in}} + (K-1)g_{\text{out}}}{K}, \\ G &= \frac{2g_{\text{in}}g_{\text{out}} + (K-2)(g_{\text{out}})^2}{K}. \end{aligned}$$

Here, g is the average reputation (which is also the average cooperation rate in the remainder of the population, i.e., excluding the main character), and G is a particular form of $G_{I,J}$, the probability that an observer regards an out-group donor's action as good.

Next, we consider the temptation to become an unconditional defector. Discriminators resist this temptation provided

$$\begin{aligned} \Pi^Z|_{f^Z=1} &> \Pi^Y|_{f^Z=1} \\ (b-c)g &> bg^Y \\ \therefore \frac{b}{c} &> \frac{g}{g-g^Y}, \end{aligned}$$

in which

$$\begin{aligned} g^Y &= (1-m)\big[gP^{GD} + (1-g)P^{BD}\big] + mP^{BD} \\ &= (1-m)g(P^{GD} - P^{BD}) + P^{BD}. \end{aligned}$$

The effect of a "main character" is thus twofold. By increasing the relative importance of the P^{BD} term in g_{in} and g_{out}, it can improve the overall rate of cooperation in a population of discriminators, provided that defecting against a bad individual is considered good (i.e., P^{BD} is large, which is true of *Stern Judging* and *Simple Standing* but not *Shunning*). However, it has a similar effect on the reputational profile of defectors. Thus, as m becomes large, g^Y and g both approach P^{BD}, and the critical b/c ratio needed to sustain a population of discriminators, $(b/c)^*$, goes to infinity.

We zero in on the unique case of *Stern Judging*, in which there is a large gap between g_{in} and g_{out}. For simplicity, we set $u_x = 0$. Some results can be seen in SI Figure S8. In this case, when $m = 0$, we have

$$\begin{aligned} g_{\text{in}} &= 1 - u_a, \\ g_{\text{out}} &= \frac{1}{2}, \\ g &= \frac{1-2u_a}{K} + \frac{1}{2}, \\ g^Y &= \frac{1}{2} - \frac{(1-2u_a)^2}{K}, \\ \left(\frac{b}{c}\right)^* &= \frac{1-2u_a+K}{2(1-u_a)(1-2u_a)}. \end{aligned}$$

As expected, the cooperation rate g becomes dominated by g_{out} as K grows. When $K \to \infty$, we have $g = g^Y = 1/2$, and thus it is impossible to sustain cooperation, as $(b/c)^* \to \infty$.

On the other hand, when $m \to 1$, all g terms approach $1 - u_a$; thus, while a high rate of cooperation in a population of discriminators is assured for any K, cooperation cannot be sustained, as defectors are guaranteed to invade. In phenomenological terms, the population cannot distinguish between individuals who defect against the main character because they are discriminators punishing someone they see as bad and individuals who defect against the main character because they are unconditional defectors; the defectors accrue slightly higher fitness because they never cooperate with the rest of the population.

For intermediate values of m, an intriguing picture emerges. While the cooperation rate g always grows with m, the critical ratio $(b/c)^*$ is non-monotonic in m; in general there is a unique value of m, which we call m^*, that minimizes

$(b/c)^*$ and thus makes it easiest for populations to resist invasion by defectors. For example, when $K = 2$ and $m = 0$, we have

$$\left(\frac{b}{c}\right)^* = \frac{3 - 2u_a}{2 - 6u_a + 4u_a^2}. \qquad [31]$$

The critical value of m is

$$m^* = \frac{\sqrt{u_a}(1 - 2\sqrt{u_a} - 2u_a)}{(1 + \sqrt{u_a})(1 - 2u_a)},$$

at which

$$\left(\frac{b}{c}\right)^* = \frac{1 - u_a}{(1 - \sqrt{u_a})^2}. \qquad [32]$$

For $u_a = 0.02$, Eq. [31] is approximately 1.57, whereas Eq. [32] is approximately 1.33; the critical value m^* is approximately 0.09.

As $K \to \infty$, cooperation can never be sustained when $m = 0$. The critical value of m is given by

$$m^* = \frac{1}{2},$$

at which

$$\left(\frac{b}{c}\right)^* = \frac{2}{1 - 2\sqrt{u_a(1 - u_a)}}.$$

For $u_a = 0.02$, this is approximately 2.78.

Thus, somewhat remarkably, a “main character” can stabilize cooperation even in the limit of private assessment ($K \to \infty$), by providing a common enemy and fomenting consensus among otherwise fragmented gossip groups. However, when main character interactions become too important in determining one’s reputation, this advantage dissipates and it becomes more difficult to sustain cooperation. Second-order effects of a main character, such as the possibility of a “bounty” for identifying a main character (and the corresponding incentive to seek one out), are exciting potential topics for future research. Notably absent from our analysis is any consideration of the welfare of the main character themself, who is always targeted for defection irrespective of their behavior.

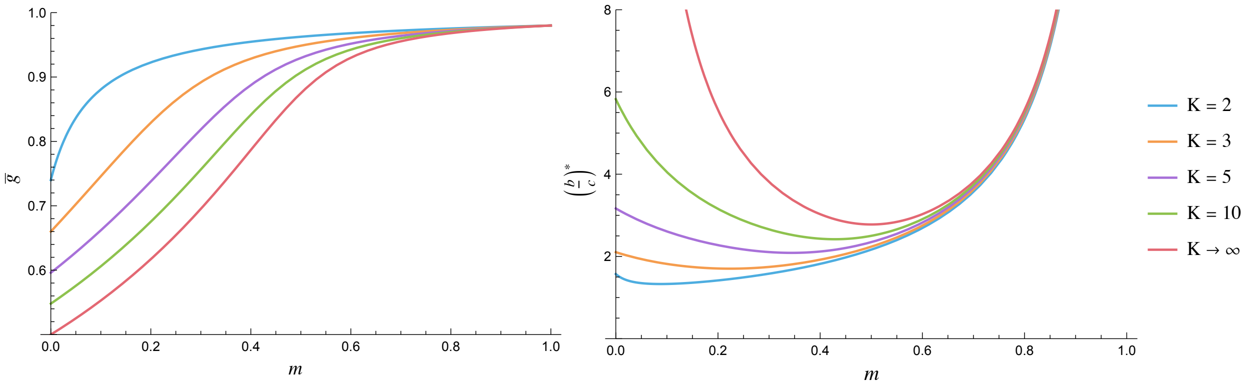

Fig. S8. The effect of a "main character", a high-visibility individual who is considered bad by the entire population. In both plots, the norm is *Stern Judging*, and error rates are $u_a = 0.02$, $u_x = 0$. Left plot: the average cooperation rate in the remainder of the population g, when the population consists entirely of discriminators. As individuals are more likely to be judged on their interactions with the main character (i.e., as m increases), g increases. Right: the critical value $(b/c)^*$ needed for discriminators to resist the temptation to become unconditional defectors. A moderate value of m can decrease this value, but a yet higher value of m pushes $(b/c)^*$ towards infinity, rendering it impossible to sustain cooperation.

References

1. T Sasaki, I Okada, Y Nakai, The evolution of conditional moral assessment in indirect reciprocity. *Sci. Reports* **7**, 41870 (2017).
2. M Nakamura, N Masuda, Groupwise information sharing promotes ingroup favoritism in indirect reciprocity. *BMC Evol. Biol*. **12**, 213 (2012).
3. AL Radzvilavicius, AJ Stewart, JB Plotkin, Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
4. S Uchida, T Sasaki, Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos, Solitons & Fractals* **56**, 175–180 (2013) Collective Behavior and Evolutionary Games.
5. AL Radzvilavicius, TA Kessinger, JB Plotkin, Adherence to public institutions that foster cooperation. *Nat. Commun*. **12**, 3567 (2021).
6. N Masuda, Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *J. Theor. Biol*. **311**, 8–18 (2012).
7. H Ohtsuki, Y Iwasa, The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol*. **239**, 435–444 (2006).
8. S Podder, S Righi, K Takács, Local reputation, local selection, and the leading eight norms. *Sci. Reports* **11**, 16560 (2021).
9. C Perret, M Krellner, TA Han, The evolution of moral rules in a model of indirect reciprocity with private assessment. *Sci. Reports* **11**, 23581 (2021).