Supplemental Materials for

"How Exploitation Launched Human Cooperation"

in *Behavioral Ecology and Sociobiology*

Rahul Bhui, Maciej Chudek, Joseph Henrich*

* Corresponding Author: Department of Human Evolutionary Biology, Harvard University, Divinity Avenue, Cambridge, MA; email: henrich@fas.harvard.edu

1. Further Approximations to Stability Conditions

The method used in the main text to yield interpretable analytical approximations to the stability conditions can be applied to yield many different expressions. Some expressions are easier to interpret but less accurate, while others are harder to interpret but more accurate. Here, we document and compare these alternative approximations to demonstrate the robustness of insights from the approximations focused on in the main text, and to explore this method further.

1a. Stability of Reputational Cooperator population against Defector/Stingy invasion

As shown in the main text, *R* is stable against *D* when

$$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right). \tag{S1}$$

(The same inequality, except without *t*, describes stability against *S*.)

Recall that the probability of an *R* type having a good reputation is given by

$$G_R = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}, \tag{S2}$$

and hence

$$\frac{1}{G_R} = \frac{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}{\rho(1-\varepsilon)} = \left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{1-\rho}{\rho}\frac{\eta}{1-\varepsilon}G_R. \tag{S3}$$

The approximations are derived by iteratively substituting the expanded expression for G_R into the stability condition, starting with substituting (S3)(S2) into (S1), and then applying the fact that G_R must be between 0 and 1. From this method we obtain four approximations: two upper bounds and two lower bounds.

The approximation in the main text results from one level of substitution:

$$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}G_R. \quad \text{(S4)}$$

Since $G_R \leq 1$, as in the main text, the RHS must be bounded above by

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}. \quad \text{(S5)}$$

Furthermore, since $G_R \geq 0$, the RHS must be bounded below by

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]. \quad \text{(S6)}$$

We refer to (S5) as Upper Bound 1 and (S6) as Lower Bound 1.

More refined approximations can be obtained by a second level of substitution, of (S2) into (S4):

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}\left(\frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}\right)$$
$$=\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)G_R\eta+\big(1-\varepsilon(1-\zeta)\big)}. \quad \text{(S7)}$$

The upper bound of this expression is

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon(1-\zeta)}. \quad \text{(S8)}$$

and the lower bound is

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\big(1-\varepsilon(1-\zeta)\big)}, \quad \text{(S9)}$$

We refer to (S9) as Upper Bound 2 and (S8) as Lower Bound 2. While any number of even more accurate bounds can be gained by iterating this process again and again, further expressions decline in ease of interpretation, and simulations indicate that the ones above are sufficiently faithful to the exact solution for our purposes. In fact, they are exact when $\eta = 0$.

All four bounds are collected in Table S1 and compared visually in Figure S1. As can be seen, more complex approximations tend to be more accurate; in this case, Lower Bound 2 is the most accurate. However, all approximations are generally quite good, with the partial exception of the simplest Lower Bound 1, which neglects to capture the noise parameter η at all.

1b. Stability of Reputational Cooperator population against Mafioso invasion

As shown in the main text, *R* is stable against *M* when

$$\frac{d}{t} > \frac{G_R}{G_R - G_M}, \tag{S10}$$

and

$$G_M = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R}, \tag{S11}$$

leading to

$$\frac{G_R}{G_R - G_M} = \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\left(\frac{1-\varepsilon(1-\zeta)}{G_R}\right)\right]. \tag{S12}$$

Using the same method as above, we start by substituting (S3)(S2) into (S12), and then applying the fact that G_R must be between 0 and 1.

This first level of substitution yields

$$\begin{aligned}\frac{G_R}{G_R - G_M} &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\big(1-\varepsilon(1-\zeta)\big)\left(\frac{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}{\rho(1-\varepsilon)}\right)\right] \\ &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}G_R\eta\right].\end{aligned} \tag{S13}$$

Since $G_R \leq 1$, this must be bounded above by

$$\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\right], \tag{S14}$$

and since $G_R \geq 0$, it must be bounded below by

$$\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon}\right]. \tag{S15}$$

We refer to (S5) as Upper Bound 1 and (S6) as Lower Bound 1.

Other approximations can be obtained by a second level of substitution, of (S2) into (S13)(S4):

$$\begin{aligned}&\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\left(\frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}\right)\right] \\ &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \frac{\eta\big(1-\varepsilon(1-\zeta)\big)}{\big(1-\varepsilon(1-\zeta)\big) + \left(\frac{1-\rho}{\rho}\right)G_R\eta}\right].\end{aligned} \tag{S16}$$

The upper bound of this expression (occurring when $G_R \to 0$) is presented in the main text:

$$\frac{d}{t} > \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon} + \eta\right], \tag{S17}$$

and the lower bound is

$$\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon} + \frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta + \left(1-\varepsilon(1-\zeta)\right)}\right]. \tag{S18}$$

We refer to (S17) as Upper Bound 2 and (S8) as Lower Bound 2. Again, while this process can be continually iterated, further expressions are less interpretable, and the ones above appear close to the exact solution. (Indeed, they are exact when $\eta = 0$.)

All four bounds are collected in Table S1 and compared visually in Figures S2 (high ζ) and S3 (low ζ). Again, the more complex approximations are generally more accurate; here, Lower Bound 2 has the lowest error. However, all approximations are generally good.

Table S1: Approximate stability conditions for population of reputational cooperators

	Rare invader in population of *Reputational Cooperators*	
	Defector / Stingy	*Mafioso*
Upper Bound 1	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\right]$
Upper Bound 2	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon(1-\zeta)}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\eta\right]$
Lower Bound 2	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\left(1-\varepsilon(1-\zeta)\right)}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\left(1-\varepsilon(1-\zeta)\right)}\right]$
Lower Bound 1	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}\right]$

Note: Stability conditions assume $d > t$ and $c > rb$. Expressions for *Stingy* are same as for *Defector* except with $t = 0$. In the main text, Upper Bound 1 is presented for the *Defector* invasion and Upper Bound 2 for the *Mafioso* invasion.

Figure S1. Minimum threshold values of *d-t/c-rb* required for reputational cooperation to be stable against rare defectors. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

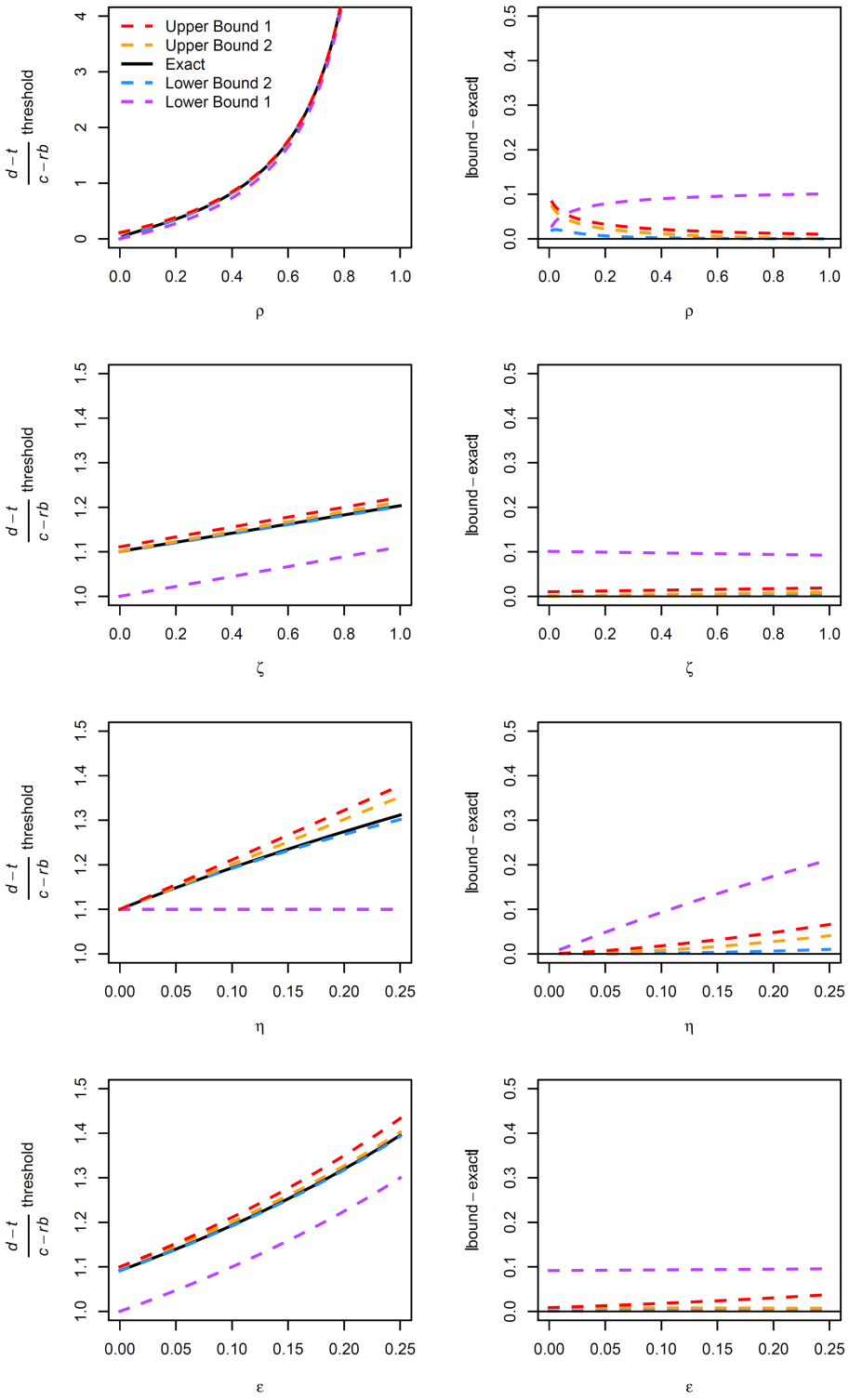


Figure S2. Minimum threshold values of *d/t* required for reputational cooperation to be stable against rare Mafiosos. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

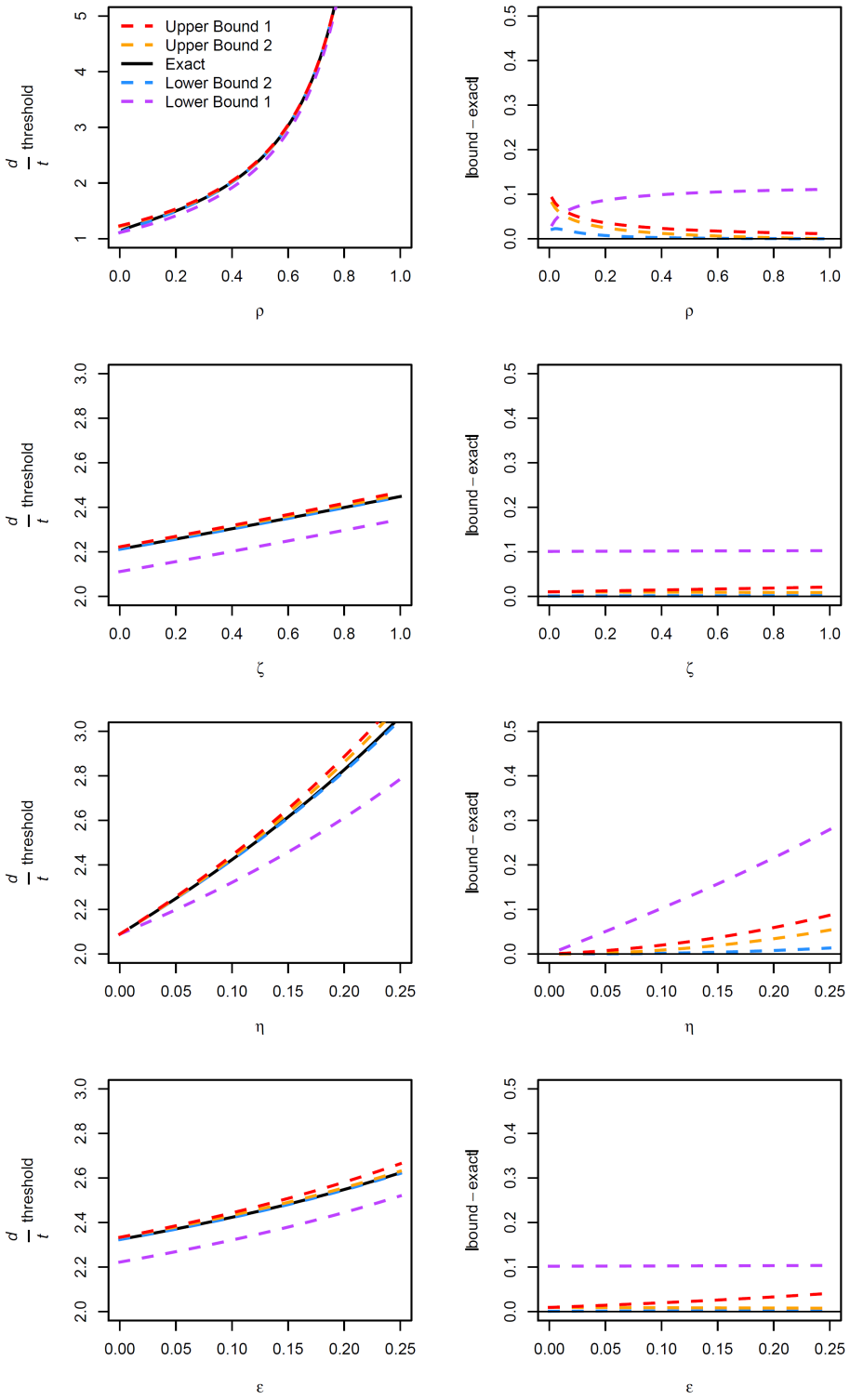
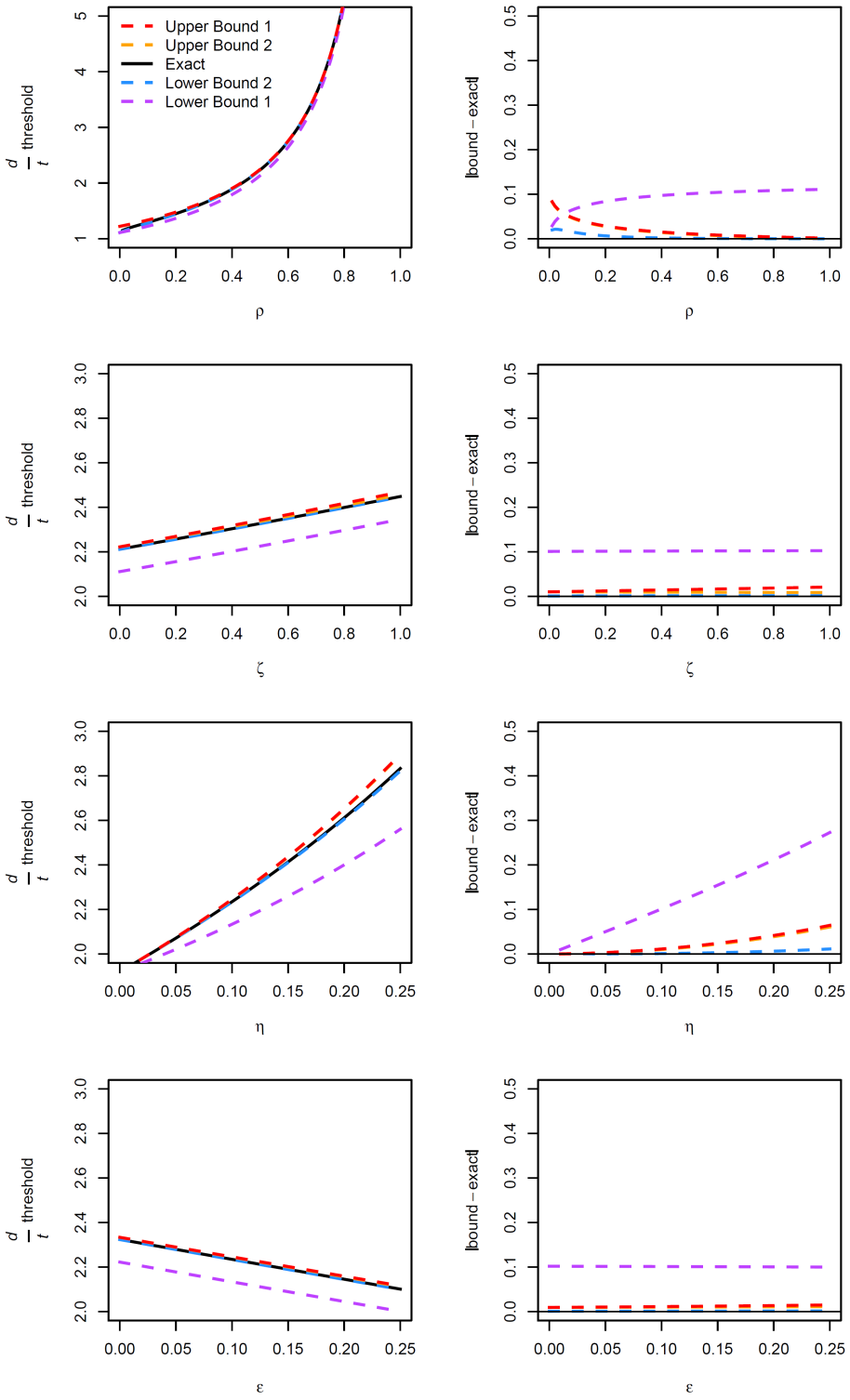


Figure S3. Minimum threshold values of *d/t* required for reputational cooperation to be stable against rare Mafiosos, with low ζ. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{1}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.



2. Basins of Attraction for Reputational Cooperation

Here we suppose that *R* types are not necessarily predominant, but constitute proportion p_R of the population. Then we determine the minimum initial value of p_R such that the population will converge to entirely *R* (i.e., the basin of attraction). In this section, we assume $\eta = 0$ for simplicity.

2a. Basin of Attraction against Defectors

In a population with fraction p_R *R*s and $p_D = 1 - p_R$ *D*s, and recognizing that $G_D = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\{t(p_R(1-G_R) + p_D) - d(p_R(1-G_R) + p_D)\}$$

$$= \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)(t-d)(1 - p_R G_R),$$

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t - d(p_R(1-G_D) + p_D)\}$$

$$= \rho\{b(1-r)p_R\} + (1-\rho)(t-d).$$

Then $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)(d-t)p_R G_R > 0$$

$$p_R > \frac{c - rb}{d - t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right).$$

Here, $G_R = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, so $\frac{1}{G_R} = 1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)$ and the condition becomes

$$p_R > \frac{c - rb}{d - t}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right).$$

Accordingly, p_R must meet some minimum threshold.

2b. Basin of Attraction against Stingy

In a population with fraction p_R *R*s and $p_S = 1 - p_R$ *S*s, and recognizing that $G_S = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\big\{t\big(p_R(1-G_R) + p_S(1-G_S)\big) - d(1-G_R)\big\}$$

$$= \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\big(t(1 - p_R G_R) - d(1-G_R)\big),$$

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)\big\{t\big(p_R(1-G_R) + p_S(1-G_S)\big) - d(1-G_S)\big\}$$

$$= \rho\{b(1-r)p_R\} + (1-\rho)(t(1 - p_R G_R) - d).$$

Then $w_R > w_S$ when

$$\rho(rb - c) + (1 - \rho)dG_R > 0$$

$$\frac{d}{c - rb} > \frac{\rho}{1 - \rho}\left(\frac{1}{G_R}\right).$$

Since $\frac{1}{G_R} = 1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)$, we need

$$1 > \frac{c - rb}{d}\left(\frac{\rho}{1 - \rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1 - \varepsilon}\right)\right).$$

which does not depend on p_R.

2c. Basin of Attraction against Mafiosos

In a population with fraction p_R *R*s and $p_M = 1 - p_R$ *M*s, we have

$$w_R = \rho\{b - c\} + (1 - \rho)\{t\big(p_R(1 - G_R) + p_M(1 - G_M)\big) - d(p_R(1 - G_R) + p_M)\}$$

$$= \rho\{b - c\} + (1 - \rho)\{t\big((1 - G_M) - p_R(G_R - G_M)\big) - d(1 - p_R G_R)\},$$

$$w_M = \rho\{b - c\} + (1 - \rho)\{t - d(p_R(1 - G_M) + p_M)\}$$

$$= \rho\{b - c\} + (1 - \rho)\big(t - d(1 - p_R G_M)\big).$$

Then $w_R > w_M$ when

$$dp_R(G_R - G_M) > t\big(G_M + p_R(G_R - G_M)\big)$$

$$p_R > \frac{t}{d - t}\left(\frac{G_M}{G_R - G_M}\right).$$

Here,

$$G_M = \frac{\rho(1 - \varepsilon)}{\rho\big(1 - \varepsilon(1 - \zeta)\big) + (1 - \rho)(p_R G_R + p_M G_M)},$$

and observing that $\frac{G_M}{G_R - G_M} = \frac{G_R}{G_R - G_M} - 1 = \frac{\rho}{1-\rho}\left(\frac{1-\varepsilon(1-\zeta)}{p_R G_R + p_M G_M}\right)$, we have

$$p_R > \frac{t}{d - t}\left(\frac{\rho}{1 - \rho}\right)\left(\frac{1 - \varepsilon(1 - \zeta)}{p_R G_R + p_M G_M}\right).$$

However, since G_M is the solution to $(1 - \rho)p_M G_M^2 + \big[\rho\big(1 - \varepsilon(1 - \zeta)\big) + (1 - \rho)p_R G_R\big]G_M - \rho(1 - \varepsilon) = 0$, this does not permit an easily interpretable analytical solution and must be numerically computed.

2d. Basin of Attraction against Defectors and Stingy

In a population with fraction p_R *R*s, p_S *S*s, and $p_D = 1 - p_R - p_S$ *D*s, and recognizing that $G_S = G_D = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\big(t(1 - p_R G_R) - d(1 - (1-p_D)G_R)\big)$$

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)(t(1 - p_R G_R) - d)$$

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)(t-d).$$

Then $w_R > w_S$ when

$$\rho(rb - c) + (1-\rho)(1-p_D)dG_R > 0$$

$$1 - p_D > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right)$$

$$1 - p_D > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right),$$

and $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)(d(1-p_D)G_R - tp_R G_R) > 0$$

$$(1-p_D)d - tp_R > (c - rb)\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right)$$

$$1 - p_D - \left(\frac{t}{d}\right)p_R > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right),$$

so p_D cannot be too large. Figure S4 shows the barycentric plot considering all three of these strategies.

2e. Basin of Attraction against Defectors and Mafiosos

In a population with fraction p_R *R*s, p_M *M*s, and $p_D = 1 - p_R - p_M$ *D*s, and recognizing that $G_D = 0$, we have

$$w_R = \rho\big\{b\big(r + (1-r)(1-p_D)\big) - c\big\} + (1-\rho)\big(t(1 - p_R G_R - p_M G_M) - d(1 - p_R G_R)\big)$$

$$w_M = \rho\big\{b\big(r + (1-r)(1-p_D)\big) - c\big\} + (1-\rho)\big(t - d(1 - p_R G_M)\big)$$

$$w_D = \rho\{b(1-r)(1-p_D)\} + (1-\rho)(t-d).$$

Then $w_R > w_M$ when

$$dp_R(G_R - G_M) - t(p_R G_R + p_M G_M) > 0$$

$$p_R[G_R(d-t) - G_M d] > p_M G_M t$$

$$p_R\left(\frac{G_R(d-t) - G_M d}{G_M t}\right) > p_M$$

$$p_R\left(\left(\frac{d-t}{t}\right)\left(\frac{G_R - G_M}{G_M}\right) - 1\right) > p_M.$$

This must be numerically computed; observe that

$$G_R = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$$

$$G_M = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)(p_R G_R + p_M G_M)}$$

$$\frac{G_R}{G_M} = \frac{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)(p_R G_R + p_M G_M)}{\rho\big(1-\varepsilon(1-\zeta)\big)} = 1 + \left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right),$$

so we have

$$p_R\left(\left(\frac{d-t}{t}\right)\left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right) - 1\right) > p_M.$$

Also $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)\big(dp_R G_R - t(p_R G_R + p_M G_M)\big) > 0$$

$$p_R G_R(d-t) - p_M G_M t > (c - rb)\left(\frac{\rho}{1-\rho}\right)$$

$$p_R\left(\frac{G_R}{G_M}\right) > \frac{c-rb}{d-t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_M}\right) + p_M\left(\frac{t}{d-t}\right).$$

$$p_R\left[1 + \left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right)\right]$$

$$> \frac{c-rb}{d-t}\left[\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right) + \frac{p_R G_R + p_M G_M}{1-\varepsilon}\right] + p_M\left(\frac{t}{d-t}\right).$$

Again, numerical solutions must be obtained. Figure S5 shows the barycentric plot considering all three of these strategies.

Figure S4. Barycentric plot reflecting evolution of R, D, and S strategies. Parameters are set at $\rho = \frac{1}{3}$, $b = 2$, $c = 1$, $r = \frac{1}{10}$, $d = 2$, $t = 1$, $\zeta = \frac{1}{2}$, and $\varepsilon = \frac{1}{10}$. The entire S-D axis consists of equilibria.

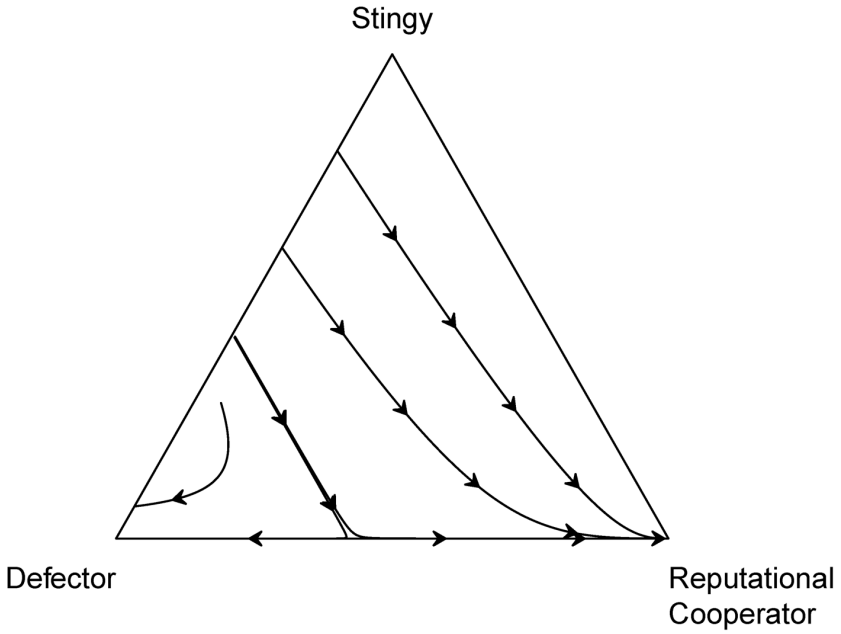
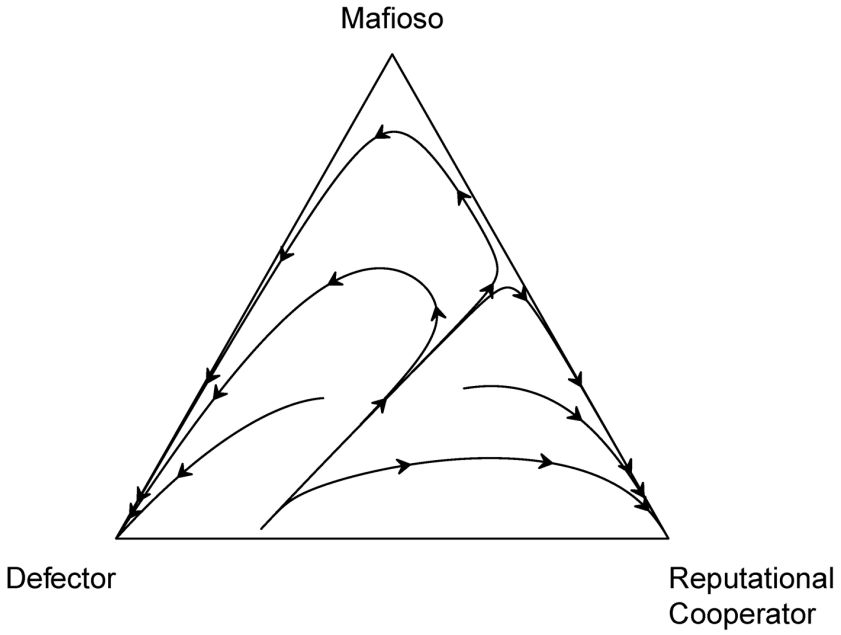


Figure S5. Barycentric plot reflecting evolution of R, D, and M strategies. Parameters are set at $\rho = \frac{1}{2}$, $b = 3$, $c = 1$, $r = \frac{1}{10}$, $d = 4$, $t = 1$, $\zeta = \frac{1}{2}$, and $\varepsilon = \frac{1}{10}$.



3. Exact Solution to Frequency of *R* in Good Standing

As laid out previously, in a population of *R*, $G_R = \frac{\rho(1-\varepsilon)}{\rho(1-\varepsilon(1-\zeta))+(1-\rho)G_R\eta}$ is the solution to the quadratic equation $(1-\rho)\eta G_R^2 + \rho\big(1-\varepsilon(1-\zeta)\big)G_R - \rho(1-\varepsilon) = 0$, which is

$$G_R = \frac{-\rho\big(1-\varepsilon(1-\zeta)\big) + \sqrt{\rho^2\big(1-\varepsilon(1-\zeta)\big)^2 + 4\rho(1-\rho)(1-\varepsilon)\eta}}{2(1-\rho)\eta}.$$

Because the solution must be non-negative, only the positive result is accepted.

4. Obligate (Non-Reputational) Cooperation is Dominated by Reputational Cooperation

In a population of type *j*, the payoff of an individual who always cooperates (never exploits regardless of partner's standing, and always contribute) is

$$w_{OC} = \rho\big\{b\big(r + (1-r)p_{OC} + (1-r)p_jY_j\big) - c\big\} + (1-\rho)\{-d(1-G_{OC})\}.$$

The payoff of an individual who cooperates based on reputation (exploits only when partner is in bad standing, and always contribute) is

$$w_R = \rho\big\{b\big(r + (1-r)p_R + (1-r)p_jY_j\big) - c\big\} + (1-\rho)\{t(1-G_R) - d(1-G_R)\}.$$

Intuitively, the former behaves exactly like the latter except they relinquish the free takings from exploiting those in bad standing. Since this exploitation has no reputational consequences, both types have good standing identical fractions of the time (i.e., $G_{OC} = G_R$), and thus in any given population, reputational cooperators always perform at least as well as obligate cooperators.

5. Stability Conditions for Populations of Other Strategies

In the main text, stability conditions were presented for populations of reputational cooperators R resisting invasion by rare defectors D, stingy types S, and Mafiosos M. Here, we present stability conditions for populations comprising each one of the other strategies to flesh out the full space of combinations.

5a.i. Stability of Defector population against Reputational Cooperator invasion

D will earn the public benefit *b* only when meeting a rare *R* in the contribution game, which happens with probability $(1-r)p_R$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game due to virtually always meeting defectors. Thus,

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t-d\}.$$

R will earn *b* when meeting another *R* in the contribution game, which happens with probability $r+(1-r)p_R$, and they will always pay the cost *c*. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1-G_D$, and will always suffer damage *d*. So,

$$w_R = \rho\{b(r+(1-r)p_R) - c\} + (1-\rho)\{t(1-G_D) - d\}.$$

Here, $G_i = \frac{\rho Y_i(1-\varepsilon)}{\rho[Y_i(1-\varepsilon)(1-\zeta)+\zeta]+(1-\rho)G_D[X_i+(1-X_i)\eta]}$. This entails $G_D = 0$, so *D* is stable when

$$w_D > w_R \Leftrightarrow rb < c.$$

5a.ii. Stability of Defector population against Mafioso invasion

D will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t-d\}.$$

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r+(1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r+(1-r)p_R) - c\} + (1-\rho)\{t-d\}.$$

Since $G_D = 0$, *D* is stable when

$$w_D > w_M \Leftrightarrow rb < c.$$

5a.iii. Stability of Defector population against Stingy invasion

D will never earn the public benefit *b* in the contribution game, nor will they ever pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = (1-\rho)\{t-d\}.$$

S will also never earn *b* or pay the cost *c* in the contribution game. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1-G_D$, and will always suffer damage *d*.

$$w_S = (1-\rho)\{t(1-G_D)-d\}$$

Since $G_D = 0$, *S* has equal fitness to *D* here.

5b.i. Stability of Stingy population against Defector invasion

S will never earn the public benefit *b* in the contribution game, nor will they ever pay the cost *c*. They will take t when meeting someone in bad standing, which happens with probability $1-G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1-G_S$. Thus,

$$w_S = (1-\rho)\{t(1-G_S)-d(1-G_S)\}.$$

D will also never earn *b* or pay the cost *c* in the contribution game. In the theft game, they will always take *t*, and suffer damage *d* when they are in bad standing.

$$w_S = (1-\rho)\{t-d(1-G_D)\}$$

Here, $G_i = \frac{\rho Y_i(1-\varepsilon)}{\rho[Y_i(1-\varepsilon)(1-\zeta)+\zeta]+(1-\rho)G_S[X_i+(1-X_i)\eta]}$, This entails $G_S = G_D = 0$, and so *D* has equal fitness to *S* here.

5b.ii. Stability of Stingy population against Reputational Cooperator invasion

S will earn the public benefit *b* only when meeting a rare *R* in the contribution game, which happens with probability $(1-r)p_R$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens with probability $1-G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1-G_S$. Thus,

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)\{t(1-G_S)-d(1-G_S)\}.$$

R will earn *b* when meeting another *R* in the contribution game, which happens with probability $r+(1-r)p_R$, and they will always pay the cost *c*. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1-G_S$, and will suffer damage *d* when they are themselves in bad standing, occurring with probability $1-G_R$. So,

$$w_R = \rho\{b(r+(1-r)p_R)-c\} + (1-\rho)\{t(1-G_S)-d(1-G_R)\}.$$

Note that $G_S = 0$ and $G_R = \frac{\rho(1-\varepsilon)}{\rho(1-\varepsilon(1-\zeta))+(1-\rho)G_S\eta} = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, hence *S* is stable when

$$w_S > w_R \Leftrightarrow \rho(c - rb) > (1-\rho)d\left[\frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}\right]$$

$$\Leftrightarrow \frac{d}{c-rb} < \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right].$$

5b.iii. Stability of Stingy population against Mafioso invasion

S will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens with probability $1 - G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1 - G_S$. Thus,

$$w_S = \rho\{b(1-r)p_M\} + (1-\rho)\{t(1-G_S) - d(1-G_S)\}.$$

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r + (1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t*, and will suffer damage *d* when they are themselves in bad standing, occurring with probability $1 - G_M$. So,

$$w_M = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\{t - d(1-G_M)\}.$$

Note that $G_S = 0$ and $G_M = \frac{\rho(1-\varepsilon)}{\rho(1-\varepsilon(1-\zeta))+(1-\rho)G_S} = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, hence *S* is stable when

$$w_S > w_M \Leftrightarrow \rho(c - rb) > (1-\rho)d\left[\frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}\right]$$

$$\Leftrightarrow \frac{d}{c-rb} < \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right].$$

5c.i. Stability of Mafioso population against Defector invasion

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r + (1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r + (1-r)p_M) - c\} + (1-\rho)\{t - d\}.$$

D will earn the public benefit *b* when meeting an *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = \rho\{b(1-r)p_M\} + (1-\rho)\{t-d\}.$$

Thus *M* is stable when

$$w_M > w_D \Leftrightarrow rb > c.$$

5c.ii. Stability of Mafioso population against Reputational Cooperator invasion

M will always earn *b* and pay *c* in the contribution game. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b-c\} + (1-\rho)\{t-d\}.$$

R will also always earn *b* and pay *c* in the contribution game. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1 - G_M$, and will always suffer damage *d*. So,

$$w_R = \rho\{b-c\} + (1-\rho)\{t(1-G_M) - d\}.$$

Since $G_M > 0$, a population full of *M* is always stable against *R*.

5c.iii. Stability of Mafioso population against Stingy invasion

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r + (1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r + (1-r)p_M) - c\} + (1-\rho)\{t-d\}.$$

S will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens here with probability $1 - G_M$, and will always suffer damage d in the theft game. Thus,

$$w_S = \rho\{b(1-r)p_M\} + (1-\rho)\{t(1-G_M) - d\}.$$

Recognizing that $\frac{1}{G_M} = \frac{\rho(1-\varepsilon(1-\zeta))+(1-\rho)G_M}{\rho(1-\varepsilon)} = \left[1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{1-\rho}{\rho}\frac{G_M}{1-\varepsilon}$, *M* is stable when

$$w_M > w_S \Leftrightarrow \frac{t}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_M}\right)$$

$$\Leftrightarrow \frac{t}{c-rb} > \frac{\rho}{1-\rho}\left[1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{G_M}{1-\varepsilon}.$$