Supplementary Materials:
Corruption Drives the Emergence of Civil Society

1 Calculation of Stationary Distribution

Our model of cooperation follows the common formulation of evolutionary dynamics simulations [1]. Specifically we consider a set of M agents each subscribing to one of d strategies. At each time step a random sample of N agents are chosen to play a public goods game. The payoffs received by each agent are determined by the number of each type of strategy. At each time step 2 agents are randomly chosen and their payoffs are compared. The probability of one agent imitating the other is determined by a logistic function of the difference in payoffs and an imitation strength s. There is also a small probability μ that a randomly chosen agent will undergo a mutation to a different strategy.

In order to calculate the stationary distribution of strategies in our evolutionary dynamics we consider, in common with previous work on life-death processes [2], the rates of transitions between homogeneous states in which all agents subscribe to a single strategy. Under deterministic dynamics these homogeneous states may be absorbing i.e. Once cooperation has collapsed and defectors have taken over, the system cannot return to a homogeneous state of cooperators. However random mutation allows mixing between homogeneous states via *mutation* and subsequent *fixation*.

Consider a population of agents each subscribing to strategy X. The probability that the system makes the transition to the state of all agents subscribing to a different strategy Y depends on the product of two quantities;

1. The probability that a random mutation introduces an agent with strategy Y ($\mu_{X,Y}$)
2. The probability that this single mutant can invade the population and lead all agents to switch to strategy Y; this is known as the fixation probability ($\rho_{X,Y}$).

In this formulation we assume that the mutation rate is low so that each mutation event leads either to fixation of a new homogeneous state or reversion to the same homogeneous state before the next mutation event occurs. Therefore, at any given time, at most two strategies are present.

Addressing (1), mutations occur in the population at a rate μ. The resultant strategy is chosen from the $d-1$ other strategies at random, giving a mutation probability

$$\mu_{X,Y} = \frac{\mu}{(d-1)} \tag{1}$$

Addressing (2), the fixation probability can be expressed explicitly from the product of the probability of each agent, after the first mutant agent, successively imitating the invading strategy. This requires a detailed description of the payoffs and imitation probabilities (section 1.2). Alternatively, (2) can be inferred simply in the limit of strong imitation (section 1.1).

Once we have an expression for the transition matrix between the homogeneous states, we can find the stationary distribution of the system of agents as the dominant eigenvector. This is a vector of values of size d which represents the long run probabilities of finding the system in a given state. We require that the transition matrix T be row normalised i.e. If the system is found in state X it must either remain in state X or transition to state $k \neq X$. Because the stationary distribution tells us the *relative proportions* of each state and the fact that the mutation probability does not depend on the source or target states, the *actual numerical value* of μ is not important and it is convenient to omit it from T.

For a simple system of $d = 3$ states X,Y and Z representing cooperators, defectors and non-participants respectively, we can construct T

$$T = \begin{pmatrix} 1 - \frac{1}{2}\rho_{X,Y} - \frac{1}{2}\rho_{X,Z} & \frac{1}{2}\rho_{X,Y} & \frac{1}{2}\rho_{X,Z} \\ \frac{1}{2}\rho_{Y,X} & 1 - \frac{1}{2}\rho_{Y,X} - \frac{1}{2}\rho_{Y,Z} & \frac{1}{2}\rho_{Y,Z} \\ \frac{1}{2}\rho_{Z,X} & \frac{1}{2}\rho_{Z,Y} & 1 - \frac{1}{2}\rho_{Z,X} - \frac{1}{2}\rho_{Z,Y} \end{pmatrix} \tag{2}$$

The factor of $\frac{1}{2}$ corresponds to $\frac{1}{d-1}$.

1.1 Strong Imitation Limit

The individual entries of T can be populated by simple arguments under the limit $s \to \infty$ (and under suitable conditions for other parameters such as punishment strength or cost) so that a strategy with a superior payoff will always be imitated and an inferior payoff will not. There are in fact only 3 possible values for the fixation probabilities $\rho_{i,j}$

$\rho_{i,j} = 0$: If $P_j < P_i$ for a single mutant with strategy j, then the mutation cannot invade and the fixation probability is 0.

$\rho_{i,j} = 1$: If $P_j > P_i$ for a single mutant with strategy j, then the mutation is beneficial and induces transition to a homegenous state j

$\rho_{i,j} = \frac{1}{2}$: This is peculiar to a single cooperator attempting to invade non-participants. The non-participants receive a fixed payoff of σ but a single cooperator will also receive a payoff σ since she has no partner with which to participate in a PGG. At the next imitation event involving the mutant cooperator, the cooperator will have the opportunity to imitate a non-participant. Since the payoffs are identical, the cooperator will revert to a non-participant with probability $\frac{1}{2}$, but is equally likely to convert a non-participant to cooperation under a neutral drift. Once two or more cooperators are present, this strategy is dominant and they invade with probability 1.

Our intuitive understanding of PGGs tells us that in the absence of punishment, free-riding always pays ($\rho_{X,Y} = 1$) and that unilateral cooperation in the face of defection does not ($\rho_{Y,X} = 0$). When cooperation is underway, it pays to participate ($\rho_{X,Z} = 0$) and due to the argument above, cooperators are slow to take over non-participants ($\rho_{Z,X} = \frac{1}{2}$). Finally, if no-one is playing the PGG then something is better than nothing ($\rho_{Y,Z} = 1$ and $\rho_{Z,Y} = 0$). Therefore T reduces to

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{3}{4} \end{pmatrix} \tag{3}$$

Leading to a stationary probability $\left[\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right]$; the systems spends half of its time in a state of non-participation and an equal one quarter both as all cooperators or defectors. Intuitively there is a single cycle from full cooperation, which may only be invaded by defectors (under the assumption that $\sigma < \frac{Ncr}{N-1}$). Defectors in turn may only be invaded by non-participants. Once in a state of full non-participation, the population may only *slowly* be invaded by cooperators due to the argument above leading to a fixation probability of $\frac{1}{2}$. Therefore non-participation predominates over long time averages as seen in simulation.

1.2 Explicit Calculation of Transition Probabilities (Intermediate Imitation Strength)

The dynamics of the evolution of cooperation amongst a finite-sized population of agents diverges from the behaviour of mean-field treatments such as replicator dynamics. Now the stochastic effects of mutation become significant [3]. The fixation probability of an l mutant in an otherwise homogeneous population of k agents, (2), can be calculated explicitly from the theory of birth-death processes [1] as

$$\rho_{k,l} = \frac{1}{1 + \sum_{q=1}^{M-1} \Pi_{N_l=1}^{q} \frac{\tau_{l \to k(N_l)}}{\tau_{k \to l(N_l)}}} \tag{4}$$

Where M is the size of the population and the number of agents with strategy k or l respectively is given by N_k and N_l with $M = N_k + N_l$. Here $\tau_{l \to k(N_l)}$ represents the probability that one of the N_k players will convert to strategy l via imitation. This transition probability for a single agent can be written explicitly for a Moran process obeying a logistic imitation probability.

$$\tau_{l \to k}(N_l) = \frac{N_l}{M} \frac{M - N_l}{M} \frac{1}{1 + \exp\left[-s(P_k - P_l)\right]} \tag{5}$$

Where s is the imitation strength and P_k and P_l are the payoffs of strategies k and l which depend on the number of k and l players. Thankfully the fixation probability simplifies to

$$\rho_{k,l} = \frac{1}{1 + \sum_{q=1}^{M-1} \exp[-s \sum_{N_l=1}^{q} (P_k - P_l)]} \tag{6}$$

Although there is no analytical expression for this at intermediate values of s, the sums can be readily evaluated and the entries of T calculated. In turn the stationary distribution can be calculated.

Henceforth, unless otherwise specified, we use the following parameter values.

PGG contribution	c	1.0
PGG multiplier	r	3.0
Population size	M	100
Sample size	N	5
Imitation strength	s	1000
Non-participation payoff	σ	1.0
Pool punishment effect	B	0.7
Pool punishment cost	G	0.7
Peer punishment effect	β	0.7
Peer punishment cost	γ	0.7
Bribe as proportion of tax	K	0.5

2 Replicating Results of Sigmund *et al*

Sigmund *et al* [4] calculate the stationary distributions of their simulations in an analagous way. However the introduction of new punishing strategies introduces a fourth possible value for the fixation probability. When a peer-punishing mutant arises in a homogeneous population of cooperators, there is neutral drift since peer-punishers have no-one to punish so enjoy the same benefits as cooperators with no additional costs. This leads to a fixation probability of $\frac{1}{M}$ [1]. In this scheme, the possible strategies are:

Cooperators (X): Participate and contribute c to the PGG

Defectors (Y): Participate but do not contribute to the PGG

Loners (Z): Neither participate nor contribute to PGG

Peer Punishers (W): Participate and contribute to the PGG (cooperate) and pay a fixed cost per defector γ to punish defectors if encountered (the more the defectors, the more the cost).

Pool Punishers (V): Participate and contribute to the PGG (cooperate) and pay a fixed a prior cost G toward a punishment pool (central authority), which will punish defectors if defectors appear.

The payoff is determined by choosing a sample population of size N to play the public good game. Below is the payoff calculations for the different strategies. It is important to note here that we assume here weak altruism (self-returning) not strong altruism (others-only) [5], since it is more common in models of public goods games.

The second order punishment terms inflicted by pool-punishers and peer-punishers are also worth noting. Peer-punishers inflict the fine $\beta.\frac{(N-1).W}{M-1}.(1-P_{second})$ on cooperators, which is proportional to the number of defectors (term P_{second}). Pool-punishers inflict the fine $B \times V \times \frac{N-1}{M-1}$ on cooperators and peer-punishers regardless of defectors existence. This is consistent with the original work of Sigmund and et al [4].

$$\begin{aligned}
P_\sigma &= \frac{\binom{Z}{N-1}}{\binom{M-1}{N-1}} \\
P_{second} &= \frac{\binom{M-Y-2}{N-2}}{\binom{M-2}{N-2}} \\
Y \text{ payoff} &= (P_\sigma.\sigma) + (1-P_\sigma).r.c.\frac{M-Z-Y-C}{M-Z} - B(N-1)\frac{V+H}{M-1} - \beta.\frac{(N-1).W+H}{M-1} \\
X \text{ payoff} &= (P_\sigma\sigma) + (1-P_\sigma).c.\left(r.\frac{M-Z-Y-C}{M-Z} - 1\right) - B(N-1)\frac{V+H}{M-1} \\
&\quad -\beta.\frac{(N-1).W}{M-1}.(1-P_{second}) \\
Z \text{ payoff} &= \sigma \\
W \text{ payoff} &= (P_\sigma\sigma) + (1-P_\sigma).c.\left(r.\frac{M-Z-Y-C}{M-Z} - 1\right) - (N-1).\frac{Y+C}{M-1}.\gamma \\
&\quad -\frac{(N-1)X}{M-1}.\gamma.(1-P_{second}) - B(N-1)\frac{V+H}{M-1} \\
V \text{ payoff} &= (P_\sigma\sigma) + (1-P_\sigma).\left(c.\left[r.\frac{M-Z-Y-C}{M-Z} - 1\right] - G\right)
\end{aligned}$$

The transition matrix is given by:

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \end{array} \begin{array}{c} \begin{array}{ccccc} X & Y & Z & V & W \end{array} \\ \begin{pmatrix} T_{XX} & T_{XY} & T_{XZ} & T_{XV} & T_{XW} \\ T_{YX} & T_{YY} & T_{YZ} & T_{YV} & T_{YW} \\ T_{ZX} & T_{ZY} & T_{ZZ} & T_{ZV} & T_{ZW} \\ T_{VX} & T_{VY} & T_{VZ} & T_{VV} & T_{VW} \\ T_{WX} & T_{WY} & T_{WZ} & T_{WV} & T_{WW} \end{pmatrix} \end{array} \tag{7}$$

Where

$$T_{ij} \begin{cases} \frac{1}{4}\mu\rho_{ij} & \text{if } i \neq j \\ 1 - \frac{1}{4}\mu\sum_{k\neq i}\rho_{ik} & \text{if } i = j \end{cases} \tag{8}$$

This reduces to

$$\begin{array}{c|ccccc} & X & Y & Z & V & W \\ \hline X & \frac{3}{4} - \frac{1}{4M} & \frac{1}{4} & 0 & 0 & \frac{1}{4M} \\ Y & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ Z & \frac{1}{8} & 0 & \frac{5}{8} & \frac{1}{8} & \frac{1}{8} \\ V & \frac{1}{4} & 0 & 0 & \frac{1}{2} & \frac{1}{4} \\ W & \frac{1}{4M} & 0 & 0 & 0 & 1 - \frac{1}{4M} \end{array} \tag{9}$$

With the stationary distribution $\frac{1}{3M+23}\,[6, 6, 4, 1, 3M + 6]$ i.e. Peer-punishers predominate. See Fig(2).

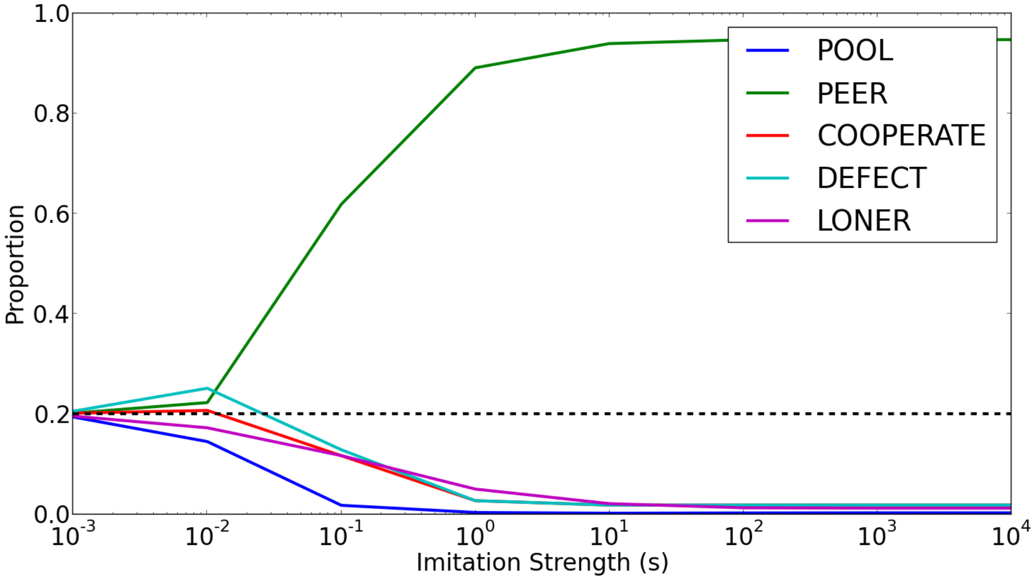

Figure 1: Stationary distributions of states as a function of imitation strength. The dashed line represents equal distribution between the d states.

Including second order punishment leads to pool punishers dominating. Pool punishers now punish defectors, cooperators and peer punishers for not contributing to the pool. Peer-punishers continue to punish defectors and cooperators.

The main differences introduced is that there is no longer a neutral drift between cooperators and peer punishers ($\rho_{X,W} \to 0$), cooperators no longer invade pool-punishers ($\rho_{V,X} \to 0$) or peer-punishers ($\rho_{V,W} \to 0$).

The transition matrix becomes

$$\begin{array}{cccccc} & X & Y & Z & V & W \\ X & \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ Y & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ Z & \frac{1}{8} & 0 & \frac{5}{8} & \frac{1}{8} & \frac{1}{8} \\ V & 0 & 0 & 0 & 1 & 0 \\ W & \frac{1}{4M} & 0 & 0 & 0 & 1-\frac{1}{4M} \end{array} \tag{10}$$

Since there is no flow out of a state of full pool-punishers, but flow into it; the stationary distribution becomes $[0, 0, 0, 1, 0]$. (See Fig(2)). Thus the presence of second-order punishment of second-order free-riders (cooperators and peer-punishers) determines whether pool-punishers or peer-punishers will prevail. The latter outcome is preferable since pool-punishers have clear dominance, whereas without second order punishment cooperation is susceptible to breaking down (See [4] Fig 3a, main paper)

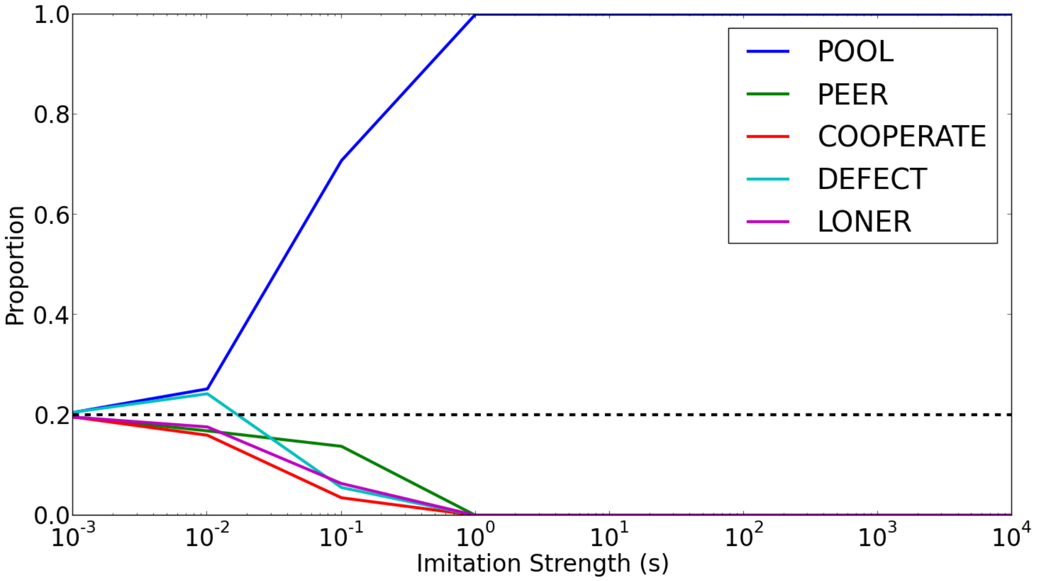

Figure 2: Stationary distributions of states as a function of imitation strength. The dashed line represents equal distribution between the d states.

3 Corruptors

We now introduce a fifth strategy into the model of Sigmund *et al*:

Corruptors (C): A corruptor pays the central authority a fixed fee $KG < G + c$ to avoid punishment for defecting from the PGG. Parameter $K \in [0, 1]$ here is a new parameter that controls bribe as percentage of G, the fee paid by pool punishers (well-behaving citizens).

The payoff of corruptors is:

$$C \text{ payoff} \quad = \quad (P_\sigma \sigma) + (1 - P_\sigma).\left(c.r.\frac{M - Z - Y - C}{M - Z} - KG\right) - \beta.(N-1)\frac{W + H}{M - 1}$$

This leads to the larger transition matrix:

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \end{array}
\begin{array}{c} \begin{array}{cccccc} X & Y & Z & V & W & C \end{array} \\
\begin{pmatrix}
T_{XX} & T_{XY} & T_{XZ} & T_{XV} & T_{XW} & T_{XC} \\
T_{YX} & T_{YY} & T_{YZ} & T_{YV} & T_{YW} & T_{YC} \\
T_{ZX} & T_{ZY} & T_{ZZ} & T_{ZV} & T_{ZW} & T_{ZC} \\
T_{VX} & T_{VY} & T_{VZ} & T_{VV} & T_{VW} & T_{VC} \\
T_{WX} & T_{WY} & T_{WZ} & T_{WV} & T_{WW} & T_{WC} \\
T_{CX} & T_{CY} & T_{CZ} & T_{CV} & T_{CW} & T_{CC}
\end{pmatrix} \end{array} \qquad (11)$$

3.1 Weak Pool Punishment (Low B)

When second-order punishment is weak (low values of B), peer punishers are stable with respect to pool-punishers. Substitution for the fixation probabilities leads to

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \end{array}
\begin{array}{c} \begin{array}{cccccc} X & Y & Z & V & W & C \end{array} \\
\begin{pmatrix}
\frac{3}{5} & \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} \\
0 & \frac{4}{5} & \frac{1}{5} & 0 & 0 & 0 \\
\frac{1}{10} & 0 & \frac{7}{10} & \frac{1}{10} & \frac{1}{10} & 0 \\
0 & 0 & 0 & \frac{4}{5} & 0 & \frac{1}{5} \\
\frac{1}{5M} & 0 & 0 & 0 & 1 - \frac{1}{5M} & \\
0 & \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{3}{5}
\end{pmatrix} \end{array} \qquad (12)$$

The stationary distribution is now $\frac{1}{M+7}[1, 2, 2, 1, M, 1]$ (using a population size $M = 100$ this is approximately $[0.01, 0.02, 0.02, 0.01, 0.93, 0.01]$) confirming clear dominance of peer-punishers.

3.2 Strong Pool Punishment (High B)

However, under extremely high second-order punishment cooperation breaks down with pool punishers dominating followed by loners and corrupt. Modifying (12) yields

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \end{array}
\begin{array}{c} \begin{array}{cccccc} X & Y & Z & V & W & C \end{array} \\
\begin{pmatrix}
\frac{2}{5} - \frac{1}{5M} & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5M} & \frac{1}{5} \\
0 & \frac{3}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 \\
\frac{1}{10} & 0 & \frac{7}{10} & \frac{1}{10} & \frac{1}{10} & 0 \\
0 & 0 & 0 & \frac{4}{5} & 0 & \frac{1}{5} \\
\frac{1}{5M} & 0 & 0 & \frac{1}{5} & \frac{4}{5} - \frac{1}{5M} & 0 \\
0 & \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{3}{5}
\end{pmatrix} \end{array} \qquad (13)$$

This leads to a stationary distribution of

$$\frac{1}{\frac{77}{16}+\frac{33(2+3M)}{16(22+17M)}}\Bigg[\frac{3}{8}-\frac{9(2+3M)}{8(22+17M)},\frac{11}{16}-\frac{9(2+3M)}{16(22+17M)},$$
$$\frac{9}{8}-\frac{3(2+3M)}{8(22+17M)},\frac{13}{8}+\frac{9(2+3M)}{8(22+17M)},\frac{3(2+3M)}{(22+17M)},1\Bigg] \quad (14)$$

This can be evaulated with $M = 100$ as $[0.034, 0.114, 0.204, 0.352, 0.102, 0.193]$ i.e. pool-punishers predominate, followed by loners and corruptors.

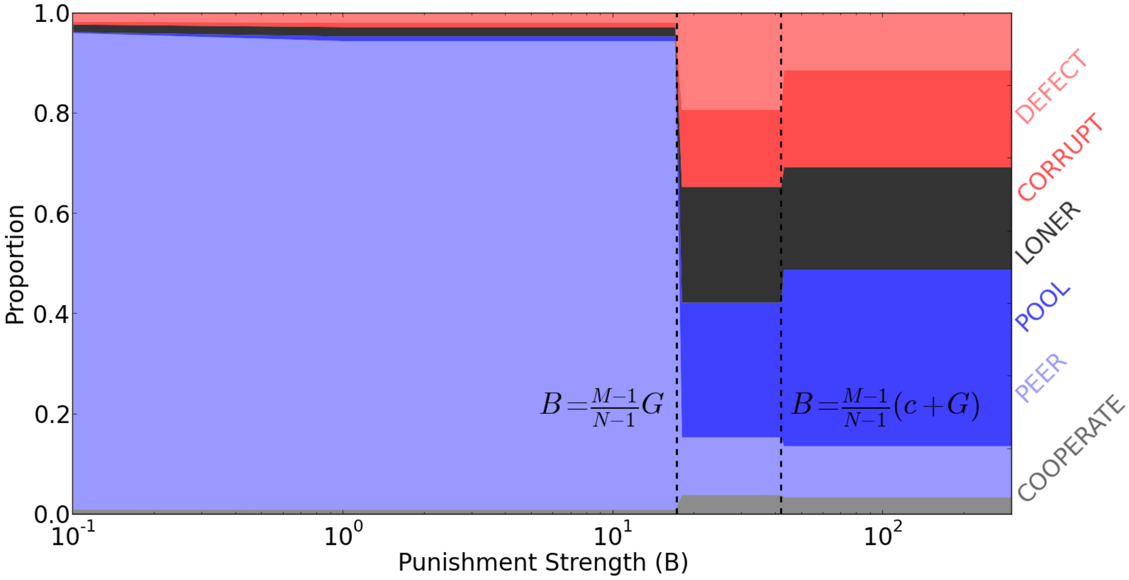

Figure 3: Stationary distributions of states as a function of pool punishment strength. As B increases cooperation breaks down.

We see 2 very clear discontinuities at $B \approx 17$ and $B \approx 40$ when the proportion of peer punishers drops to be replaced by pool-punishers. Above the first threshold, pool-punishing agents will invade peer-punishers ($\rho_{WV} \to 1$). Above the second threshold, pool-punishing agents will also take over defectors. These points are explained below.

Firstly, at intermediate values of B the expected pool-punishment (calculated from the probability of being selected with the single pool-punisher in the sample of N) is low. Therefore, the low probability of being matched with a pool-punisher doesn't incentivise the payment of G. However, once B is sufficiently high, the threat of pool-punishment *even from a single pool-punishing hybrid player* is too high of a risk and all non-pool-punishing strategies can be invaded by pool-punishing strategies ($\rho_{WH}, \rho_{WV} \to 1$). The condition for this is given by the expected cost of receiving pool-punishment when a single pool-punishing hybrid agent is present in a population

$$(\frac{N-1}{M-1})B \quad (15)$$

When this is equal to G it is cheaper to pay tax than to risk pool-punishment

$$G < \frac{(N-1)}{M-1}B \tag{16}$$

$$B^{*} = (\frac{M-1}{N-1})G \tag{17}$$

Substituting $N = 5$,$M = 100$ and $G = 0.7$ gives a critical value when $B^{*} = 17.325$.

Adressing the second threshold; as the pool-punishment term becomes very large, the expected value of pool-punishment for a homogeneous population of defectors being punished by a single pool-punisher becomes so large that pool-punishers may invade defectors, despite the pool-punisher making a heavy loss in the PGG.

$$c + G < \frac{N-1}{M-1}B \tag{18}$$

$$B^{*} = \frac{M-1}{N-1}(c + G) \tag{19}$$

Substituting for M, N, c and G gives $B^{*} = 42.075$.

4 Corruptors and Hybrid Punishers

Finally, we add Hybrid-Punishers (H) to the set of possible strategies.

Hybrid-Punishers (H): These players participate and contribute to the PGG (cooperate), pay a fixed cost per defector γ, and pay a fixed a prior cost G toward a punishment pool.

The payoff of hybrid punishers is then defined by the following equation:

$$H \text{ payoff} \quad = \quad (P_\sigma \sigma) + (1 - P_\sigma).\left(c.\left[r.\frac{M - Z - Y - C}{M - Z} - 1\right] - G\right) - (N-1).\frac{Y + C}{M - 1}.\gamma$$

We now have the following transition matrix.

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \\ H \end{array} \begin{array}{c} \begin{array}{ccccccc} X & Y & Z & V & W & C & H \end{array} \\ \begin{pmatrix} T_{XX} & T_{XY} & T_{XZ} & T_{XV} & T_{XW} & T_{XC} & T_{XH} \\ T_{YX} & T_{YY} & T_{YZ} & T_{YV} & T_{YW} & T_{YC} & T_{YH} \\ T_{ZX} & T_{ZY} & T_{ZZ} & T_{ZV} & T_{ZW} & T_{ZC} & T_{ZH} \\ T_{VX} & T_{VY} & T_{VZ} & T_{VV} & T_{VW} & T_{VC} & T_{VH} \\ T_{WX} & T_{WY} & T_{WZ} & T_{WV} & T_{WW} & T_{WC} & T_{WH} \\ T_{CX} & T_{CY} & T_{CZ} & T_{CV} & T_{CW} & T_{CC} & T_{CH} \\ T_{HX} & T_{HY} & T_{HZ} & T_{HV} & T_{HW} & T_{HC} & T_{HH} \end{pmatrix} \end{array} \tag{20}$$

4.1 Weak Pool Punishment (Low B)

Assuming a low value of B, results in the transition matrix below.

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \\ H \end{array} \begin{array}{c} \begin{array}{ccccccc} X & Y & Z & V & W & C & H \end{array} \\ \begin{pmatrix} \frac{3}{6} - \frac{1}{6M} & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6M} & \frac{1}{6} & 0 \\ 0 & \frac{4}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\ \frac{1}{12} & 0 & \frac{9}{12} & \frac{1}{12} & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{6} - \frac{1}{6M} & 0 & \frac{1}{6} & \frac{1}{6M} \\ \frac{1}{6M} & 0 & 0 & 0 & 1 - \frac{1}{6M} & 0 & 0 \\ 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{2}{3} & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6M} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} - \frac{1}{6M} \end{pmatrix} \end{array} \tag{21}$$

The stationary distribution in this case is given as

$$\frac{1}{\Gamma}\left[\frac{24}{13} + \frac{15M}{13}, \frac{31}{15} + \frac{55M}{26}, \frac{46}{13} + \frac{45M}{13}, 1 + 5M, 3(16 + 34M + 15M^2, 5(5 + 8M)), 1\right] \tag{22}$$

Where the normalisation factor is given as

$$\Gamma = \frac{127}{13} + \frac{305M}{26} + \frac{5}{13}(5 + 8M) + \frac{3}{26}(16 + 34M + 15M^2) \tag{23}$$

This evaluates to $[0.01, 0.017, 0.016, 0.008, 0.94, 0.006, 0.001]$. Peer punishers overwhelmingly predominate, followed by defectors, loners and cooperators (agrees with low B limit of Fig 4 of corruption paper).

4.2 Strong Pool Punishment (High B)

When B is very large the transition matrix becomes

$$\begin{array}{c} \\ X \\ Y \\ Z \\ V \\ W \\ C \\ H \end{array} \begin{array}{c} \begin{array}{ccccccc} X & Y & Z & V & W & C & H \end{array} \\ \begin{pmatrix} \frac{2}{6}-\frac{1}{6M} & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6M} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{4}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\ \frac{1}{12} & 0 & \frac{2}{3} & \frac{1}{12} & \frac{1}{12} & 0 & \frac{1}{12} \\ 0 & 0 & 0 & \frac{5}{6}-\frac{1}{6M} & 0 & \frac{1}{6} & \frac{1}{6M} \\ \frac{1}{6M} & 0 & 0 & 0 & \frac{5}{6}-\frac{1}{6M} & 0 & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{6M} & 0 & 0 & 1-\frac{1}{6M} \end{pmatrix} \end{array} \tag{24}$$

The stationary distribution can be expressed as

$$\frac{1}{\Gamma}\left[\frac{3(2+M)}{(70+86M+27M^2)}, \frac{-22-17M}{(70+86M+27M^2)}, \frac{6(5+4M)}{(70+86M+27M^2)}, \right. \\ \left. \frac{-70-59M}{(70+86M+27M^2)}, \frac{6(1+2M)}{(70+86M+27M^2)}, \frac{-38-31M}{(70+86M+27M^2)}, 1\right] \tag{25}$$

Where the normalisation factor is given as

$$\Gamma = \frac{1}{1-\frac{-70-59M}{70+86M+27M^2}-\frac{-38-31M}{70+86M+27M^2}-\frac{-22-17M}{70+86M+27M^2}-\frac{3(2+M)}{70+86M+27M^2}-\frac{6(1+2M)}{70+86M+27M^2}-\frac{6(5+4M)}{70+86M+27M^2}} \tag{26}$$

With the stationary distribution as follows $[0.001, 0.0059, 0.008, 0.020, 0.004, 0.011, 0.950]$; hybrid punishers predominate (in agreement with the high B limit of Fig 4 of corruption paper). The proportions are plotted as a function of B below in Fig (4.2).

We see 2 very clear discontinuities at $B \approx 0.2$ and $B \approx 17$ when the proportion of peer punishers drops to be replaced by hybrid strategies. Above the first threshold; hybrid strategies may no longer be invaded by peer-punishers ($\rho_{HW} \to 0$). Above the second threshold, hybrid agents will also invade peer-punishers ($\rho_{WH} \to 1$). The explanation for the second threshold is the same as section 3 and the first threshold is explained below.

For a single peer-punishing mutant to invade hybrid players, the saving from paying the tax G must outweigh any possible second order pool-punishment. Since, apart from the mutant herself, only pool-punishers are present this has an expected value of $B(N-1)$ i.e. punishment from all the other players in the sample.

$$G < B(N-1) \tag{27}$$

Leading to a threshold value for $B^* = 0.175$.

The transition matrix in (24) also shows that when second-order punishment is strong, hybrid punishers are only destablized by neutral drift towards pool-punishers, who can then be exploited by corruptors. One interpretation is that this form of instability represents a risk that exists in the real world. When

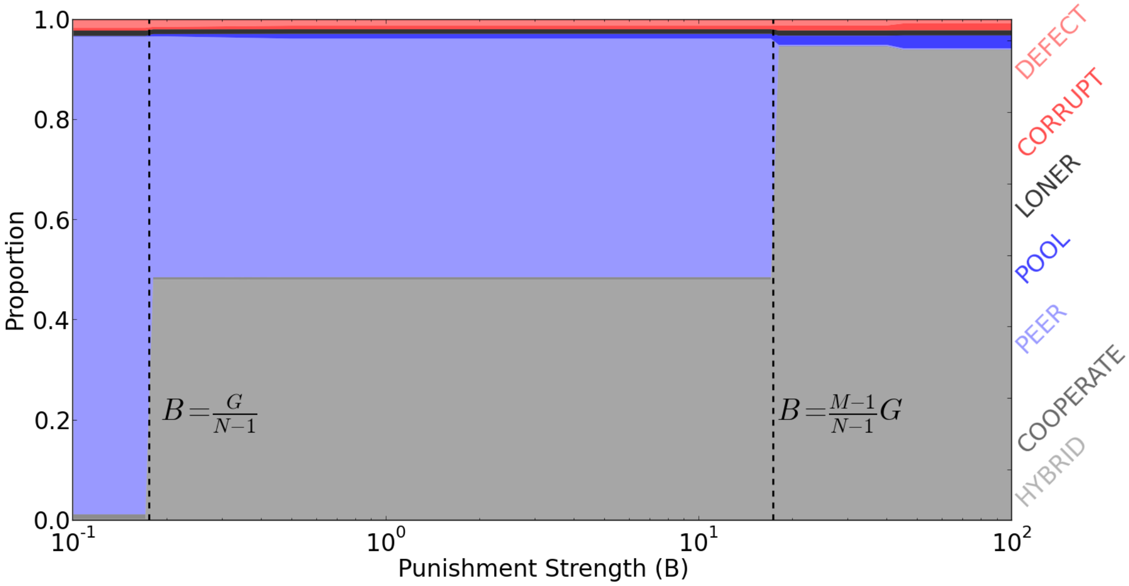

Figure 4: Stationary distributions of states as a function of pool punishment strength. As B increases hybrid punishers become dominant.

there is high cooperation, individuals might become lax in their propensity to altruistically punish defection, and this can destablize cooperation. As mentioned in the main text, this risk may motivate goverments to sometimes mandate that citizens to sign up for certain peer punishment duties, like jury duty, and punish those who merely pay their taxes. If pool-punishers were also punished by second-order punishment, then there would be no neutral drift towards this strategy, and the stationary distribution would be $[0, 0, 0, 0, 0, 0, 1]$, as there would be no flows away from the hybrid punisher state.

It is worth noting that dominance of the hybrid strategy is robust against change in different premeters for high values of B (effect of pool punishement). Figure 4.2 shows that unless the cost of peer punishment (γ) is too steep, the hybrid strategy dominates. Simlarly, Figure 4.2 shows that the hybrid strategy dominates the population unless the severity of peer punishment (β) is too small. In both cases, when the hybrid strategy can not dominate, corruption, defection, and non-participation increase significantly. Finally, with respect to the cost of corruption (K), hybrid strategy dominates unless the cost of corruption is too high.

References

[1] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, 2006.

[2] D. Fudenberg and L. A. Imhof, "Imitation processes with small mutations," *Journal of Economic Theory*, vol. 131, no. 1, pp. 251 – 262, 2006.

[3] A. Traulsen, M. A. Nowak, and J. M. Pacheco, "Stochastic dynamics of invasion and fixation," *Phys. Rev. E*, vol. 74, p. 011909, Jul 2006.

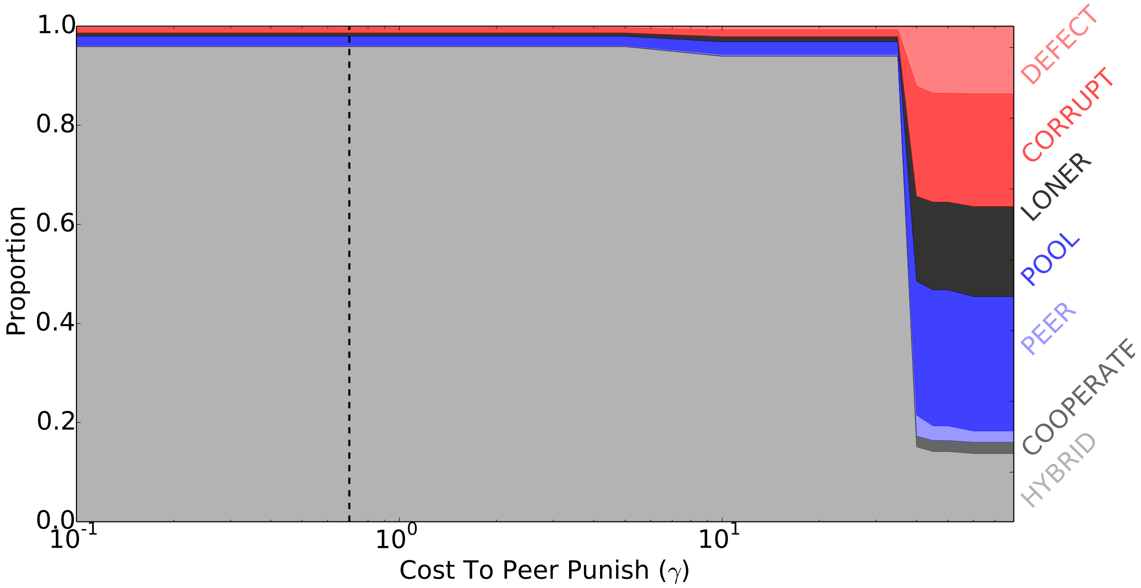

Figure 5: Stationary distributions of states as a function of γ with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \beta = 0.7$, and $K = 0.5$.

[4] K. Sigmund, H. De Silva, A. Traulsen, and C. Hauert, "Social learning promotes institutions for governing the commons," *Nature*, vol. 466, pp. 861–863, Aug. 2010.

[5] J. A. Fletcher and M. Zwick, "Strong altruism can evolve in randomly formed groups," *Journal of Theoretical Biology*, vol. 228, no. 3, pp. 303–313, 2004.

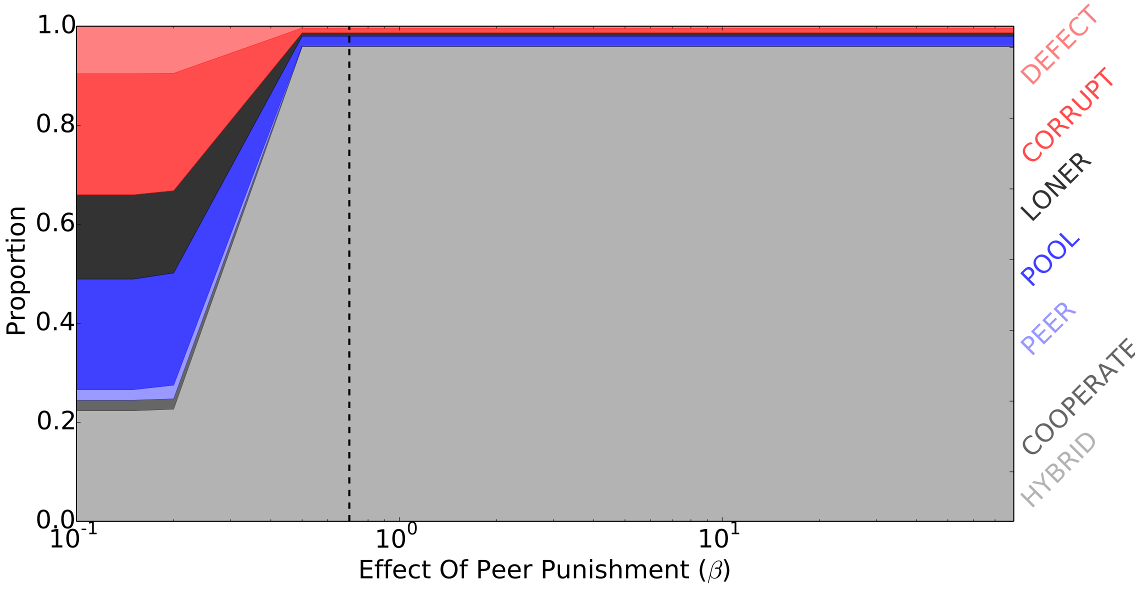

Figure 6: Stationary distributions of states as a function of β with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \gamma = 0.7$, and $K = 0.5$.

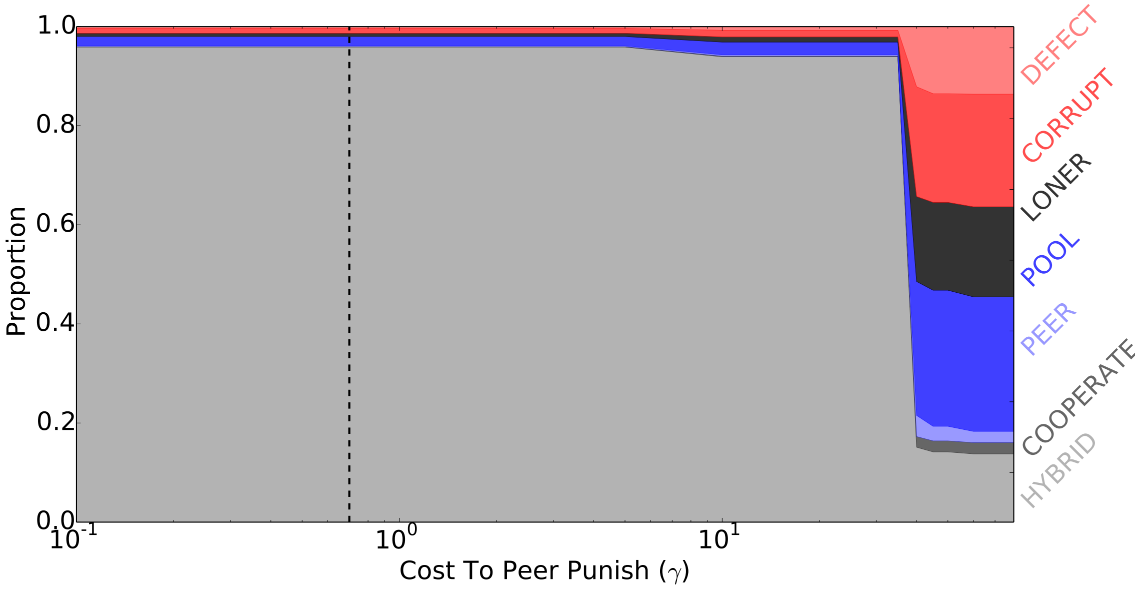

Figure 7: Stationary distributions of states as a function of K with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \gamma = 0.7$, and $\beta = 0.7$.