JOURNAL OF THE ROYAL SOCIETY Interface

Corruption drives the emergence of civil society

Sherief Abdallah, Rasha Sayed, Iyad Rahwan, Brad L. LeVeck, Manuel Cebrian, Alex Rutherford and James H. Fowler

J. R. Soc. Interface 2014 **11**, 20131044, published 29 January 2014

JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Report

Cite this article: Abdallah S, Sayed R, Rahwan I, LeVeck BL, Cebrian M, Rutherford A, Fowler JH. 2014 Corruption drives the emergence of civil society. *J. R. Soc. Interface* **11**: 20131044.
http://dx.doi.org/10.1098/rsif.2013.1044

Received: 12 November 2013
Accepted: 6 January 2014

Subject Areas:
computational biology

Keywords:
social learning, evolutionary dynamics, politics

Author for correspondence:
Sherief Abdallah
e-mail: shario@ieee.org

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2013.1044 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing

Corruption drives the emergence of civil society

Sherief Abdallah[1,2,3], Rasha Sayed[1], Iyad Rahwan[4,2], Brad L. LeVeck[5], Manuel Cebrian[6,7], Alex Rutherford[4,8] and James H. Fowler[5]

[1]Informatics Department, The British University in Dubai, Dubai, United Arab Emirates
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
[3]Faculty of Computers and Information, Cairo University, Cairo, Egypt
[4]Department of Electrical Engineering and Computer Science, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[5]Political Science Department, University of California, San Diego, CA, USA
[6]National Information and Communications Technology Australia, Melbourne, Victoria 3010, Australia
[7]Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia
[8]United Nations Global Pulse

Centralized sanctioning institutions have been shown to emerge naturally through social learning, displace all other forms of punishment and lead to stable cooperation. However, this result provokes a number of questions. If centralized sanctioning is so successful, then why do many highly authoritarian states suffer from low levels of cooperation? Why do states with high levels of public good provision tend to rely more on citizen-driven peer punishment? Here, we consider how corruption influences the evolution of cooperation and punishment. Our model shows that the effectiveness of centralized punishment in promoting cooperation breaks down when some actors in the model are allowed to bribe centralized authorities. Counterintuitively, a weaker centralized authority is actually more effective because it allows peer punishment to restore cooperation in the presence of corruption. Our results provide an evolutionary rationale for why public goods provision rarely flourishes in polities that rely only on strong centralized institutions. Instead, cooperation requires both decentralized and centralized enforcement. These results help to explain why citizen participation is a fundamental necessity for policing the commons.

A centuries-old debate exists on how to best govern society and promote cooperation: is cooperation best maintained by a central authority [1,2] or is it better handled by more decentralized forms of governance [3,4]? The debate is still unresolved, and identifying mechanisms that promote cooperation remains one of the most difficult challenges facing society and policymakers today [4].

Decentralized, individual sanctioning of non-cooperators (also known as free-riders or defectors) is one of the main tools used by societies to promote and maintain cooperation [5]. Individuals can sanction free-riders implicitly via behavioural reciprocity (as in the case of the highly successful tit-for-tat strategy [6]) or explicitly via costly punishment [7]. Both of these forms of peer punishment have been widely studied using evolutionary models and behavioural experiments [8–10,6,11].

Recently, however, Sigmund *et al.* [12] showed that centralized institutions can have an evolutionary advantage over peer punishment because, unlike peer-punishers, these institutions may eliminate 'second-order' free-riding. Second-order free-riders cooperate with other players but they do not pay the cost of punishing defectors and this can allow defectors to re-emerge [13–15]. To address this problem, Sigmund *et al.* present a model of 'pool' punishment, where agents commit resources to a centralized authority that sanctions free-riders [12,16]. Pool punishment avoids the second-order free-rider problem because the centralized authority punishes *any* individual who does not

contribute to the punishment pool (including cooperators and peer-punishers). This allows pool-punishers to quickly take over a population, displacing both free-riders and peer-punishers [12]. These advantages help to explain why human societies frequently delegate punishment to centralized institutions [12,17,16]. They also help to explain why centralized institutions acquire an increasing monopoly over legitimate punishment over time by stigmatizing [18] and criminalizing [19, p. 371,372] various forms of peer punishment.

However, the dominance of pool punishment in the Sigmund *et al.* model [12] also creates three puzzles. First, the results imply that increasing the severity of centralized pool punishment always increases cooperation. Yet, many authoritarian states, which have the ability to severely punish citizens, suffer from low levels of participation and public goods provision [20]. Meanwhile, states with high levels of public goods, such as western democracies [20–22], typically limit the government's ability to punish individuals and tolerate more forms of peer punishment.

Second, centralized pool punishment quickly takes over a population and completely displaces peer punishment [12] in the Sigmund *et al.* model [12], but many (if not most) societies exhibit a mix of centralized and decentralized punishment strategies. Even in societies with centralized punishment, citizens engage in costly acts of protest against agents who harm the public good. As recent events—from the Occupy protests to the Arab Spring—illustrate, this occurs even when the government punishes protestors [23–25]. What unmodelled factors might allow peer punishment to evolve alongside centralized enforcement institutions—even when these institutions are actively hostile towards various forms of peer punishment?

Third, the Sigmund *et al.* model [12] assumes that the centralized authority punishes all forms of peer punishment. This is because peer-punishers in their model, by definition, do not contribute to the centralized authority. However, many societies with centralized enforcement also recognize certain forms of peer punishment as legitimate. For instance, civil litigation, jury duty, anti-incumbent voting and other forms of political participation are also instances of altruistic peer punishment [26–28]. In these and other cases, citizens engage in a *hybrid* peer-pool punishment strategy. These individuals pay taxes to a central authority but also engage in selective acts of peer punishment that are individually costly, but not punished by a central power. Given all the costs they bear, it is unclear how such hybrid strategies may evolve.

Here, we show that allowing for *corruption* in the model can help to explain both why societies want to limit the severity of centralized punishment, and why peer punishment frequently evolves alongside centralized punishment institutions. We investigate the effect of corrupt players who can bribe a central authority to avoid punishment. The results show that when pool-punishers dominate a system, the central authority becomes a single point of failure, which is highly vulnerable to corruption. This gives an opportunity for individuals playing a hybrid peer-pool strategy to evolve because peer punishment becomes relatively more effective under these circumstances by helping to increase the overall level of cooperation.

In summary, given the possibility of corruption, Leviathans can promote cooperation, but only if they also allow individuals to take action against actors who harm the public good. Our model therefore provides an evolutionary rationale for why public goods provision and cooperation rarely flourish in polities with strong centralized punishment alone. Instead, cooperation rests on an authority that protects a fundamental aspect of civil society, citizen participation in policing the commons [29,30].

Our baseline model is a public good game (PGG) with both peer and pool punishments [12]. The PGG is a simple model for studying contributions to a project with non-excludable positive externalities, which may include everything from the provision of social insurance to the protection of the environment. Let M denote the population size and let $N \leq M$ denote the number of individuals who are randomly chosen in a given round to play a PGG. In the game, each individual is faced with a choice: whether or not to contribute a fixed amount, $c > 0$, to the common pool. Once each individual chooses her action, each individual will obtain $rc(N_c/N)$, where r is a factor greater than 1, N_c is the number of contributors to the common pool and N is the total number of participants (whether they contributed or not). If all individuals contribute, $N_c = N$, then the social welfare is maximized and each individual obtains rc. However, each actor gains an equal share of $rc(N_c/N)$, whether or not they contribute, making it a dominant strategy for each individual to free-ride by contributing 0 (the pay-offs are written explicitly in the electronic supplementary material).

The population includes X *cooperators*, who contribute c and Y *defectors* (*free-riders*), who do not. Consistent with previous work, we also assume that the game is not compulsory and some players may choose not to participate in the PGG [12,31–34]. These *loners* earn a fixed small pay-off, σ. In addition, W *peer-punishers* cooperate by contributing c to the PGG but also impose a fine, β, on each free-rider at a cost γ [5]. In other words, each free-rider pays a total fine βN_w, where N_w is the number of peer-punishers in the group, and every peer-punisher incurs an extra cost γN_f, where N_f is the number of free-riders in the group. Furthermore, peer-punishers inflict a penalty on cooperators proportional to the number of defectors (second-order punishment). We also have V *pool-punishers* who, instead of directly punishing free-riders, contribute a fixed amount, G, to a punishment pool before participating in the game and then contribute c to the PGG. Those who do not contribute to the pool (including free-riders, cooperators and peer-punishers) are then fined BN_v each, where N_v is the number of pool-punishers in the group. We also introduce C *corruptors* to the model. A corruptor pays the central authority a fixed fee KG to avoid being punished for not contributing to the PGG (this only makes sense if the fee is less than the total contributions paid by pool-punishers, $KG < G + c$).

We study the equilibria of fully mixed populations of fixed size M and variable composition by computing the pay-offs obtained by players using these strategies, assuming that agents play in randomly sampled groups of size N. The difference in pay-offs, together with the parameter $s \geq 0$, determines the rate at which individuals with lower pay-offs are replaced by types with higher pay-offs. As in other evolutionary models, this process can be interpreted either as evolution or social learning. We also allow for random switching of strategies with a mutation rate $\mu \geq 0$. We derive equilibria as the long-run distribution of different strategies both analytically (in the limit of strong imitation) and using numerical simulation. For details, see the electronic supplementary material.

Figure 1 shows sample runs of a numerical simulation of the model. Without corruption, pool punishment eventually

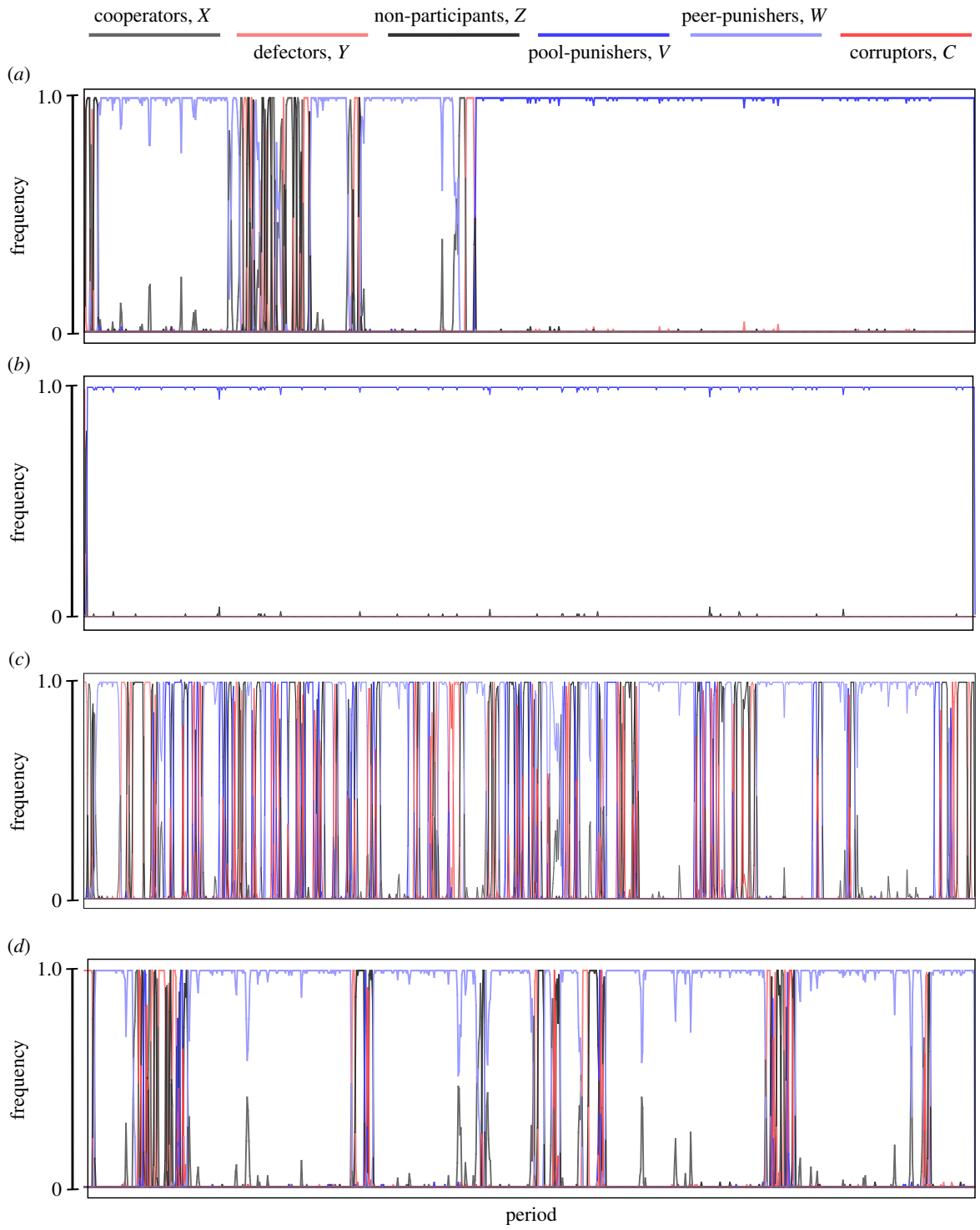

Figure 1. Sample simulation runs showing the effect of the corruptor strategy. For all the runs, the following parameter values were used (please refer to the electronic supplementary material for more details): $S = 100\,000$, $M = 100$, $N = 5$, $\mu = 0.001$, $\sigma = 1.0$, $c = 1.0$, $r = 3.0$, $\beta = 0.7$, $\gamma = 0.7$, $k = 0.5$ and $G = 0.7$. The severity of institutional punishment is controlled via parameter B, which is set to either 0.7 or 7. In (*a*), without the corruptor strategy, the results are consistent with the results reported in the previous work [12], where pool-punishers predominate. In (*b*), the predominance of the pool-punishers becomes decisive as the severity of the institutional punishment escalates. In (*c*), with the corruptor strategy added to the mix of available strategies, and with the severity of institutional punishment set to ($B = 7$), the pool-punishers are no longer stable and cooperation deteriorates in general. Finally, in (*d*), as institutional punishment becomes more lenient, peer-punishers emerge and largely maintain cooperation ($B = 0.7$), even in the presence of corruptors.

takes over the population [12], and does so even earlier when B, the severity of second-order pool punishment, is higher.

The situation changes dramatically when we introduce corruptors. Pool punishment is no longer a dominant strategy, as shown in a sample simulation run (figure 1*c*). Interestingly, figure 1*d* shows that weakening pool punishment (lowering the fine B) allows peer-punishers to re-emerge as a relatively stable strategy that restores cooperation in the presence of corruption.

We investigate this further in figure 2*a*, which shows the proportion of different strategies as a function of second-order punishment severity. For low values of B, peer-punishers dominate and prevent the corruptor strategies from gaining ground. As B increases, peer-punishers disappear and pool-punishers

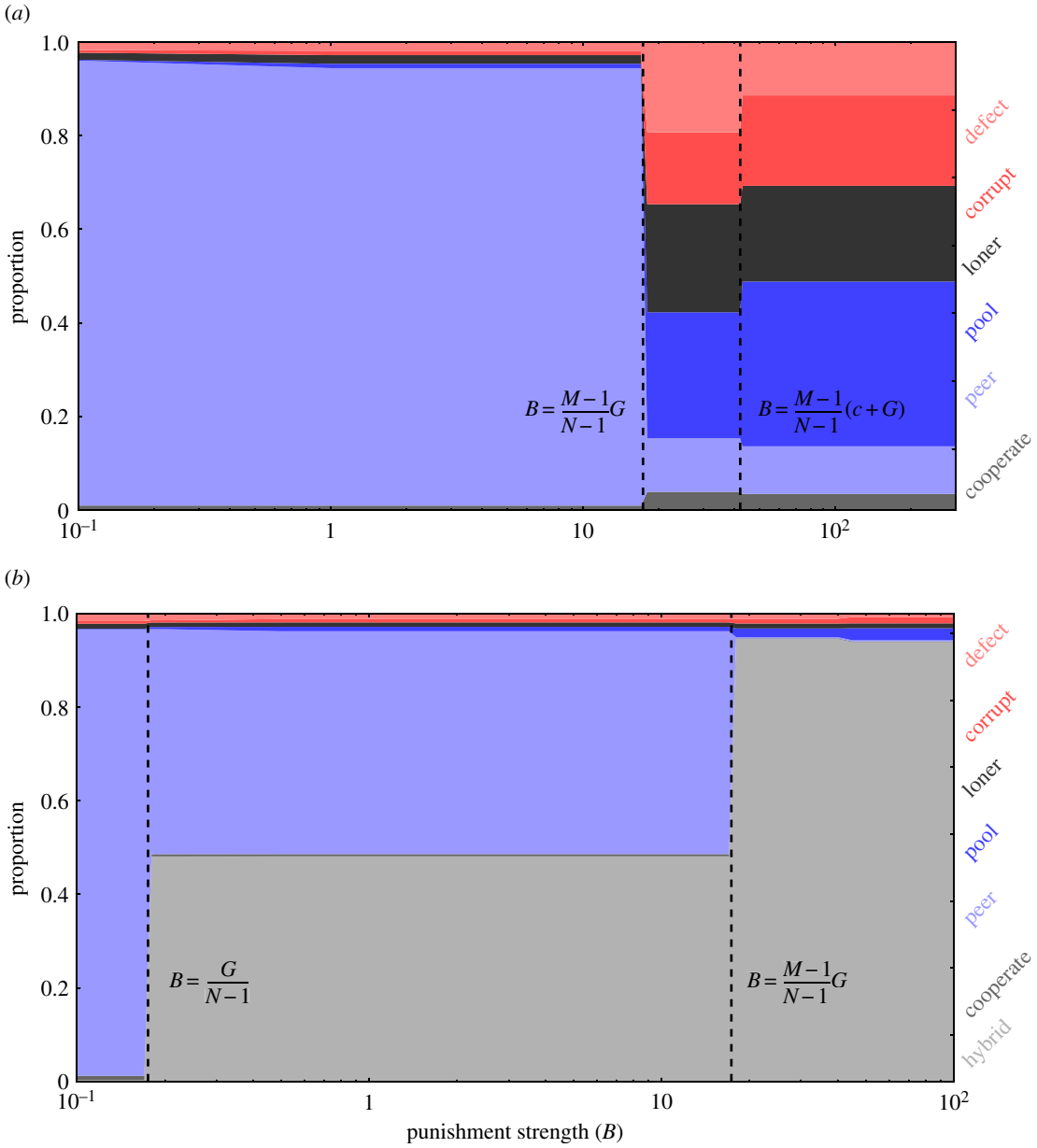

Figure 2. Stationary distributions of strategies as a function of institutional punishment severity (parameter B). In (*a*), the corruptor strategy is included in the set of available strategies and we observe the adverse effect of institutional punishment. The greater B, the greater the percentage of corruption. A clear phase transition happens when $B > (M - 1/N - 1)G$, when the expected punishment exerted by a single pool-punisher (in a sample of N) exceeds the punishment cost for the pool-punisher, G. This allows a pool-punisher to severely suppress peer-punishers, which in turn allows corruptors, defectors and loners to grow in the population. In (*b*), both the corruptor and the hybrid strategies are included. As a result, increasing B no longer backfires, and the same level of cooperation is maintained. The hybrid strategy becomes dominant for $B > (M - 1/N - 1)G$, when the expected punishment exerted by a single peer-punisher (in a sample of N) exceeds the punishment cost for the pool-punisher, G.

become more prevalent. However, with even higher values of B, the prevalence of corruptors also increases. This causes the total number of cooperative individuals to decline. We confirm these results by analytical computation of the long-run frequencies of strategies in the (X, Y, Z, V, W, C) subpopulation (for methods, see the electronic supplementary material). With low B, the frequencies, respectively, are $1/(M + 7)$ [1,2,2,1,M,1], confirming the clear dominance of peer-punishers (with a population of $M = 100$, this is approx. [0.01, 0.02, 0.02, 0.01, 0.93, 0.01]). Strong central punishment, however, yields the distribution [0.034, 0.114, 0.204, 0.352, 0.102, 0.193], i.e. ineffective (corrupted) pool-punishers dominate, followed by loners and corruptors.

Strong centralized punishment allows corruptors to exploit pool-punishers in two ways: pool-punishers contribute to a public good, at the same time funding a flawed institution that corruptors use to their advantage. Weak centralized punishment, on the other hand, provides an opportunity for peer-punishers to counteract both corruptors and defectors.

Lastly, we introduce H *hybrid punishers*. In addition to contributing c to the public good, individuals using this strategy pay both γ to punish defectors directly and G to the punishment pool, and as such they are not punished by the central authority. Hybrid individuals can be thought of as upstanding citizens that pay their taxes but also engage in forms of 'legitimate' peer sanctioning.

Figure 2*b* shows that, unlike peer punishment alone, this hybrid strategy dominates the population when centralized punishment is severe. This occurs even though the hybrid strategy pays a higher average cost compared with pool-punishers. Setting $M = 100$, the long-run distribution of strategies in the (X, Y, Z, V, W, C, H) subpopulation is [0.001, 0.004, 0.008, 0.013, 0.004, 0.007, 0.96]

As a consequence, a high level of cooperation is maintained across all levels of centralized punishment. The dominance of the hybrid strategy is robust against different parameter values (including β, γ and K, as we show in the electronic supplementary information).

Of course, one might wonder why individuals would create a second-order punishment institution in the first place, as figure 2*b* also shows that second-order punishment does not increase the overall level of cooperation; nor does it make cooperation significantly more stable than peer punishment alone. Our relatively simple model is unlikely to fully answer this very general question, as we have left out many features that could cause centralized institutions to remain advantageous. For example, these institutions might aggregate views on who should be punished; and this aggregation could cause perceptual errors (which are not in our model) to cancel out [35].

It is also possible that institutions may further evolve to deal with this remaining instability. Analytical results in the electronic supplementary information show that when second-order punishment is strong, hybrid punishers are only destabilized by neutral-drift towards pool-punishers (who then allow corruptors and defectors to emerge). Institutions may therefore want to screen and punish pure pool-punishers; and it is interesting that many justice systems have evolved rules that fine people who merely pay their taxes but do not register for various forms of hybrid punishment, for example jury duty.

Importantly, however, we have shown that simply adding the risk of corruption can help to explain why centralized and decentralized forms of punishment frequently coexist. No additional appeal to civic norms or civic culture is needed. Which is not to say that these things do not exist or that they do not further promote citizen participation in policing the commons. Rather, our model shows that independent of other virtues, peer-punishment strategies can have a fitness advantage over pool punishment alone. In the face of corruption, peer and hybrid punishment strategies better promote cooperation because they are competitive. If one punisher fails to punish a corrupt individual, another might step in; and this result may help to explain why polities who want to control corruption and promote cooperation often become more tolerant to various forms of decentralized sanctioning [36].

Funding statement. S.A. is funded by the British University in Dubai. M.C. is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence programme.

References

1. Hobbes T. 1960 *Leviathan: or the matter, forme and power of a commonwealth ecclesiasticall and civil*. New Haven, CT: Yale University Press.
2. Hardin G. 1968 The tragedy of the commons. *Science* **162**, 1244–1248. (doi:10.1126/science.162.3859.1243)
3. Kropotkin PA. 1907 *Mutual aid: a factor of evolution*. London, UK: W. Heinemann.
4. Dietz T, Ostrom E, Stern PC. 2003 The struggle to govern the commons. *Science* **302**, 1907–1912. (doi:10.1126/science.1091015)
5. Nowak MA. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
6. Axelrod R. 2006 *The evolution of cooperation: revised edition*. New York, NY: Basic books.
7. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
8. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)
9. Fehr E, Gächter S. 2000 Cooperation and punishment in public goods experiments. *Am. Econom. Rev.* **90**, 980–994.
10. Egas M, Riedl A. 2008 The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. B* **275**, 871–878. (doi:10.1098/rspb.2007.1558)
11. Ohtsuki H, Hauert C, Lieberman E, Nowak MA. 2006 A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441**, 502–505. (doi:10.1038/nature04605)
12. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
13. Panchanathan K, Boyd R. 2004 Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502. (doi:10.1038/nature02978)
14. Fowler JH. 2005 Human cooperation: second-order free-riding problem solved? *Nature* **437**, E8. (doi:10.1038/nature04201)
15. Dreber A, Rand D, Fudenberg D, Nowak M. 2008 Winners don't punish. *Nature* **452**, 348–351. (doi:10.1038/nature06723)
16. Traulsen A, Röhl T, Milinski M. 2012 An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* **279**, 3716–3721. (doi:10.1098/rspb.2012.0937)
17. Baldassarri D, Grossman G. 2011 Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl Acad. Sci. USA* **108**, 11 023–11 027. (doi:10.1073/pnas.1105456108)
18. Rosenbaum T. 2011 Justice? Vengeance? You need both. *The New York Times*. See http://www.nytimes.com/2011/07/28/opinion/28rosenbaum.html.
19. Hallam H. 1853 *View of the state of Europe during the middle ages*. New York, NY: Harper & Brothers.
20. Acemoglu D, Robinson J. 2012 *Why nations fail: the origins of power, prosperity, and poverty*. New York, NY: Crown Publishing Group.
21. Deacon RT. 2009 Public good provision under dictatorship and democracy. *Public Choice* **139**, 241–262. (doi:10.1007/s11127-008-9391-x)
22. Lake DA, Baum MA. 2001 The invisible hand of democracy political control and the provision of public services. *Comp. Political Stud.* **34**, 587–621. (doi:10.1177/0010414001034006001)
23. Harcourt BE. 2011 Occupy Wall Street's 'political disobedience'. *New York Times*, 11 October 2011, p. 13.
24. Morsi M. 2013 Egypt president issues stern warnings to opposition. *Ahram Online*. See http://english.ahram.org.eg/NewsContent/1/64/67627/Egypt/Politics-/Egypt-president-warns-opposition-against-promoting.aspx.
25. Moghadam VM. 2012 *Globalization and social movements: Islamism, feminism, and the global justice movement*. Lanham, MD: Rowman and Littlefield Publishers.
26. Fowler JH, Kam CD. 2007 Beyond the self: social identity, altruism, and political participation. *J. Polit.* **69**, 813–827. (doi:10.1111/j.1468-2508.2007.00577.x)
27. Grechenig K, Nicklisch A, Thöni C. 2010 Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. *J. Empir. Legal Stud.* **7**, 847–867. (doi:10.1111/j.1740-1461.2010.01197.x)
28. Smirnov O, Dawes CT, Fowler JH, Johnson T, McElreath R. 2010 The behavioral logic of collective action: partisans cooperate and punish more than nonpartisans. *Polit. Psychol.* **31**, 595–616. (doi:10.1111/j.1467-9221.2010.00768.x)

29. Putnam RD, Leonardi R, Nanetti RY. 1994 *Making democracy work: civic traditions in modern Italy*. Princeton, NJ: Princeton university press.
30. Verba S, Schlozman KL, Brady HE. 2005 *Voice and equality: civic voluntarism in American politics*. Cambridge, MA: Harvard University Press.
31. Fowler JH. 2005 Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102**, 7047–7049. (doi:10.1073/pnas.0500938102)
32. Hauert C, De Monte S, Hofbauer J, Sigmund K. 2002 Volunteering as Red Queen mechanism for cooperation in public goods games. *Science* **296**, 1129–1132. (doi:10.1126/science.1070582)
33. Hauertwz C, De Montewy S, Hofbauerw J, Sigmund K. 2002 Replicator dynamics for optional public good games. *J. Theor. Biol.* **218**, 187–194. (doi:10.1006/jtbi.2002.3067)
34. Semmann D, Krambeck HJ, Milinski M. 2003 Volunteering leads to rock–paper–scissors dynamics in a public goods game. *Nature* **425**, 390–393. (doi:10.1038/nature01986)
35. McLennan A. 1998 Consequences of the Condorcet jury theorem for beneficial information aggregation by rational agents. *Am. Polit. Sci. Rev.* **92**, 413–418. (doi:10.2307/2585673)
36. Egorov G, Guriev S, Sonin K. 2009 Why resource-poor dictators allow freer media: a theory and evidence from panel data. *Am. Polit. Sci. Rev.* **103**, 645. (doi:10.1017/S0003055409990219)

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 $\rho_{i,j}$

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

$\boldsymbol{\rho_{i,j} = 1}$: 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

$\boldsymbol{\rho_{i,j} = \frac{1}{2}}$: 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

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

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

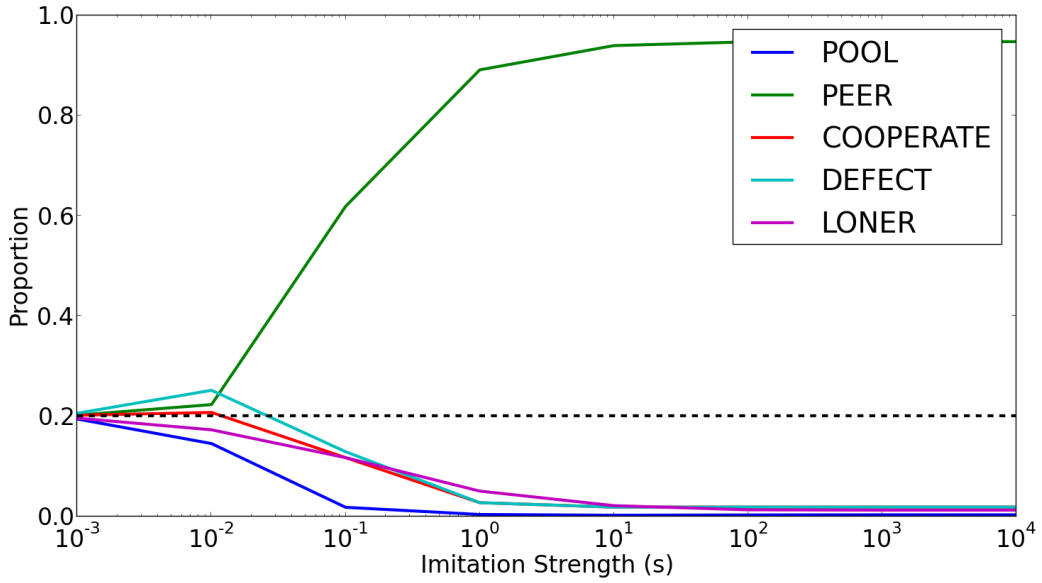

Figure 1: Stationary distributions of states as a function of imitation strength. The dashed line represents equal distribution between the d states.

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

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

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)

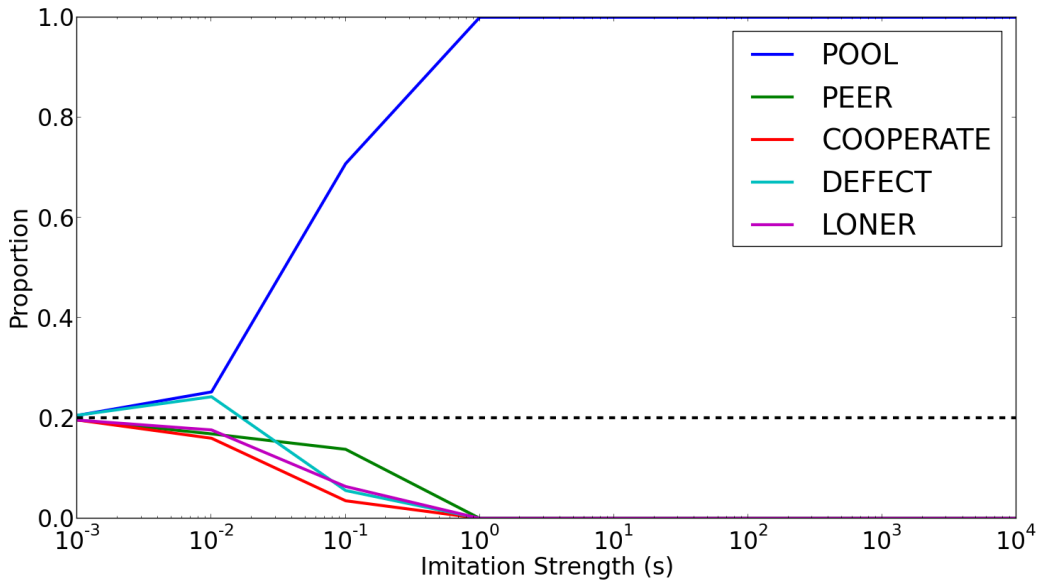

Figure 2: Stationary distributions of states as a function of imitation strength. The dashed line represents equal distribution between the d states.

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:

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

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

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

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

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

This leads to a stationary distribution of

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

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.

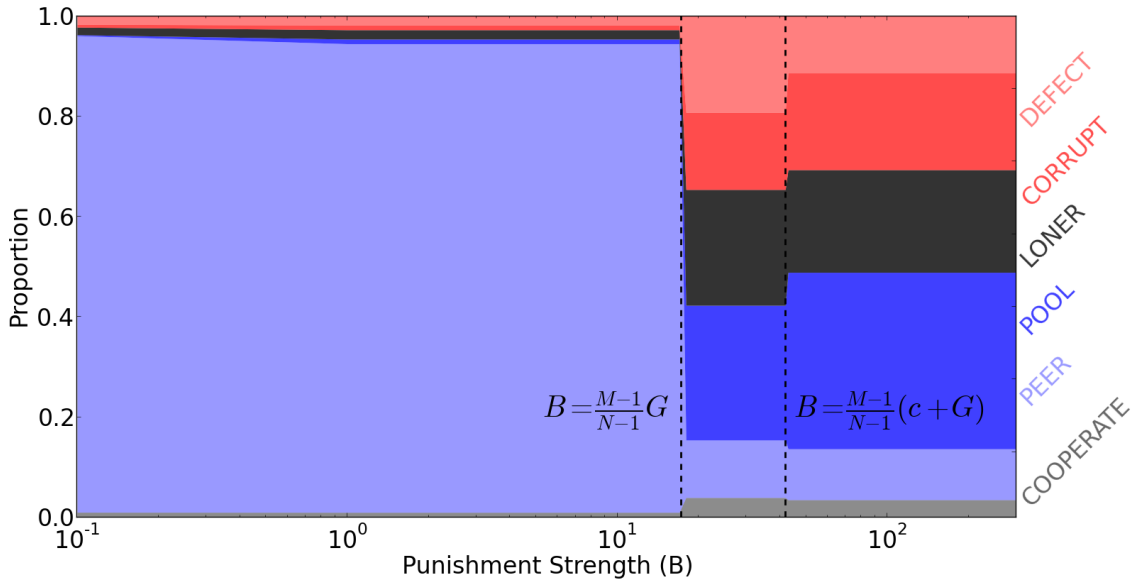

Figure 3: Stationary distributions of states as a function of pool punishment strength. As B increases cooperation breaks down.

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

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

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

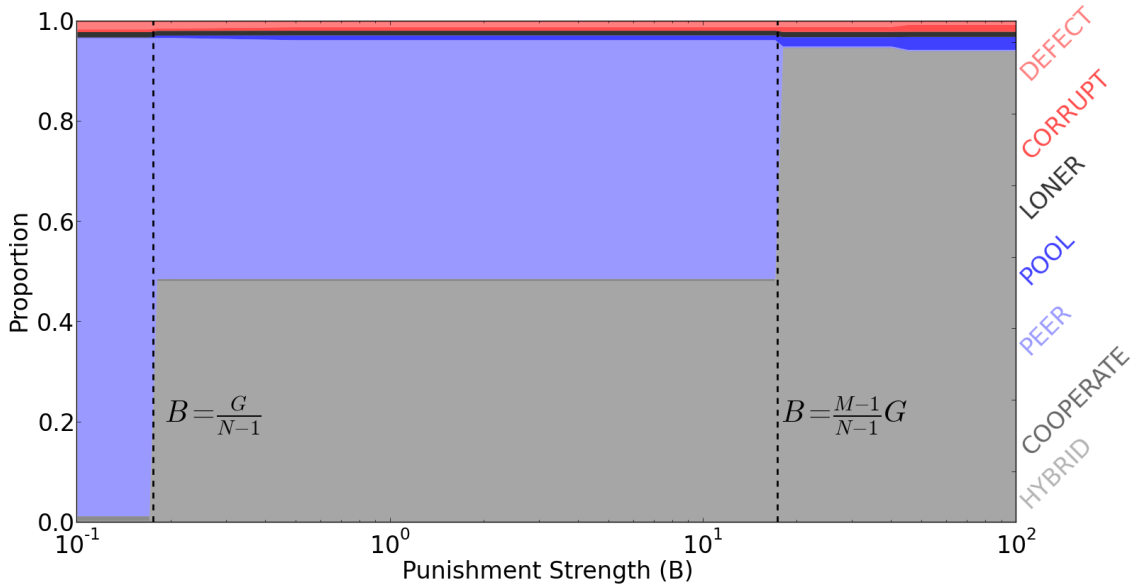

Figure 4: Stationary distributions of states as a function of pool punishment strength. As B increases hybrid punishers become dominant.

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.

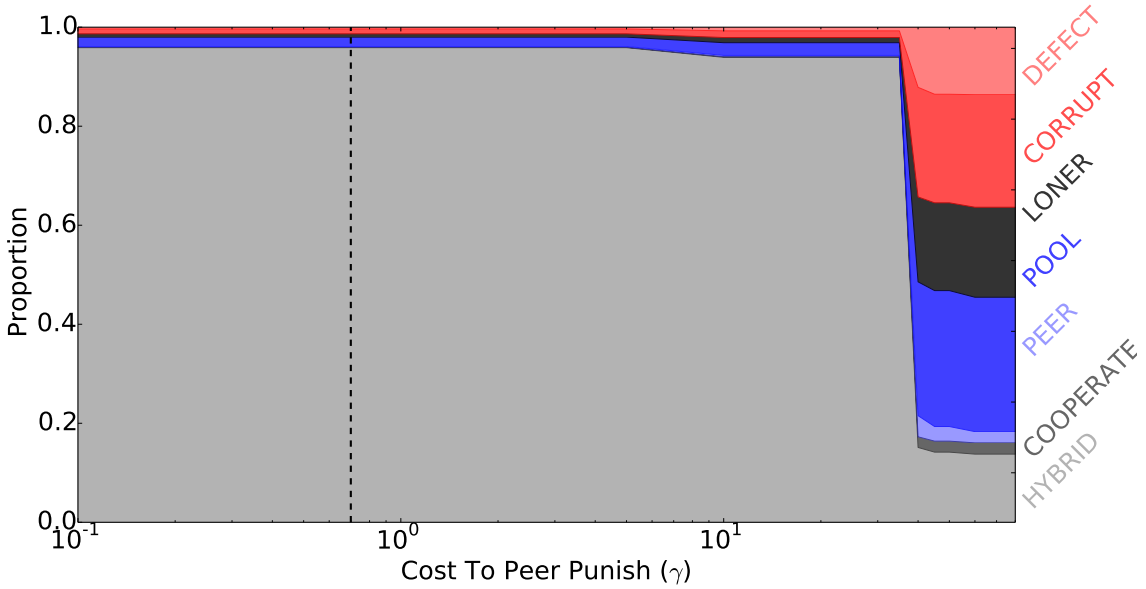

Figure 5: Stationary distributions of states as a function of γ with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \beta = 0.7$, and $K = 0.5$.

[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.

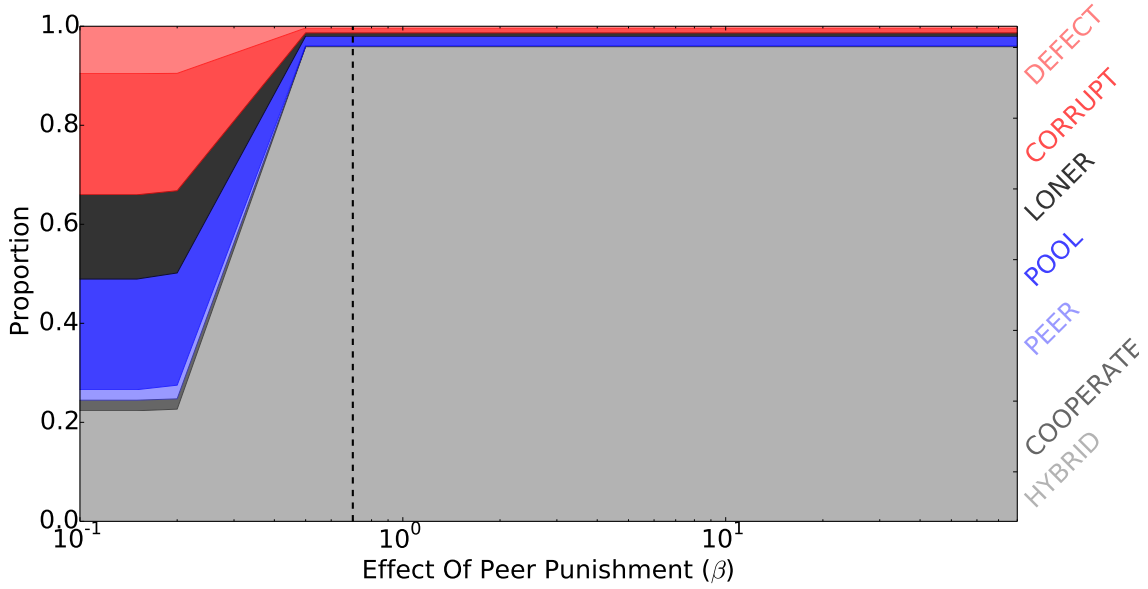

Figure 6: Stationary distributions of states as a function of β with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \gamma = 0.7$, and $K = 0.5$.

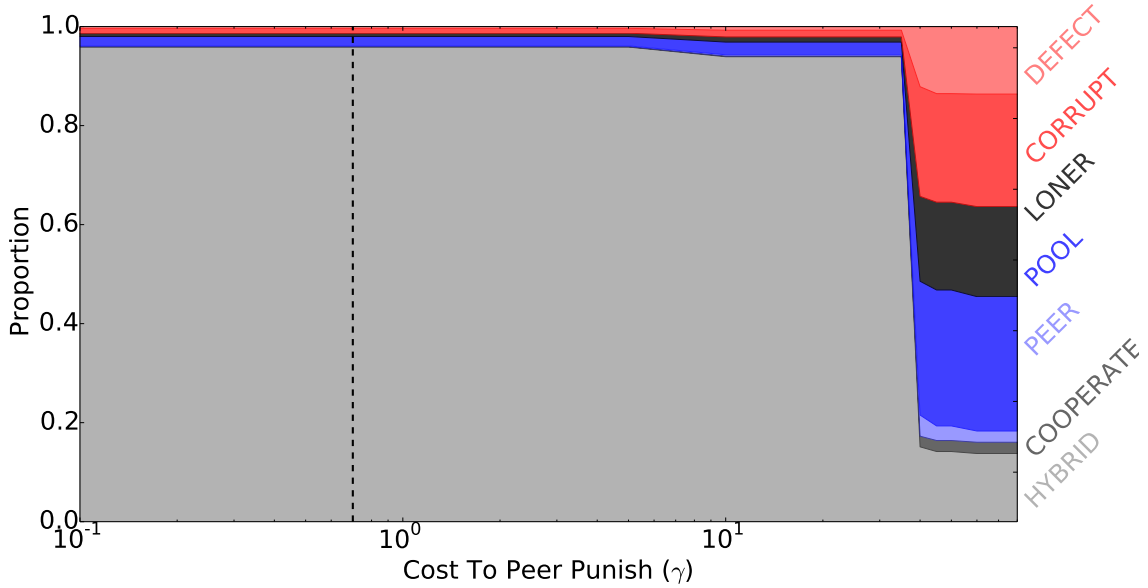

Figure 7: Stationary distributions of states as a function of K with the following settings: $M = 100, N = 5, r = 3, c = 1, \sigma = 1, G = 0.7, B = 1000, \gamma = 0.7$, and $\beta = 0.7$.