Supporting Information for

The co-evolution of social institutions, demography, and large-scale human cooperation

Simon T.Powers[*]

Laurent Lehmann[†]

This file includes:

Appendices S1-S5

Figures S1–S7

[*] Department of Ecology & Evolution, University of Lausanne, Switzerland, Simon.Powers@Unil.ch

[†] Department of Ecology & Evolution, University of Lausanne, Switzerland, Laurent.Lehmann@Unil.ch

SI Appendicies

We investigate here the effect of relaxing several assumptions of our model, which were briefly explained in the *Sensitivity to other model assumptions* section of the *Main text.*

Appendix S1: Alternative mechanisms of institution formation

We investigated two alternative aggregation functions for forming an institution from individual preferences: 1. taking the majority preference of social individuals; 2. having a leader create the institutional rules from its own preference.

To set the institutional h-value from the majority preference of social individuals, we divided the h-preferences of social individuals on each patch into ten equally spaced bins. The institutional h-value was then taken as the midpoint of the bin containing the greatest frequency of individual preferences. To implement a leader, we randomly chose a social individual and set the institutional h-value to be the h-preference of this individual. The h-preferences of the other group members were then set to be equal to the sum of the leader's h-preference and a normally distributed random variable with mean zero and variance 0.05, where the addition of the random variable represents a copying error.

We found that forming the institution in either of these ways did not qualitatively affect the results compared to taking the mean h-preference of social individuals.

Appendix S2: Failure to form an institution if a consensus is not reached

Our model assumed that social individuals were always able to agree upon an institutional h-value. However, in reality, such an agreement may not be possible if the variance in individual preferences is too large. For example, Kosfeld *et al.* (2009) considered a unanimity rule, in which all individuals must agree on the form of an institution before one can be created. To implement the possibility of failure in institution formation if the preferences of group members are too divergent, we added a threshold parameter, U, such that institution formation is unsuccessful on patch j if

$$\frac{1}{n_{\mathrm{c}j}(t) + n_{\mathrm{d}j}(t)} \sum_{i=1}^{n_j(t)} s_{ij}(t) \left(h_{ij}(t) - h_j(t)\right)^2 > U. \tag{S1}$$

Here, $s_{ij}(t) = 1$ if individual i on patch j is a cooperator or defector, $s_{ij}(t) = 0$ otherwise, and $h_j(t)$ is given by Eq. 4 of the main text, so that the above left member is the variance of h-preferences among socials in patch j.

This condition says that institution formation fails due to disagreement if the variance in h-preferences of social individuals on patch j exceeds U. If institution formation fails, then the growth rates and "carrying capacity" of socials on patch j are set as follows:

$$\begin{aligned} r_{\mathrm{c}j}(t) &= r_{\mathrm{d}j}(t) = r_{\mathrm{a}} \\ K_{\mathrm{s}j}(t) &= K_{\mathrm{a}}. \end{aligned} \tag{S2}$$

The effect of varying U in our model is shown in Fig. S2. Unless U is very small (less than 0.002), then cooperative institutions are still able to invade and then be maintained in large groups. However, if U becomes very small then institutions cannot be reliably maintained.

Appendix S3: Option to not pay the cost of institution formation

We investigated a version of the model in which individuals need not take part in negotiating the form of the institution, and hence not pay the cost I. In that case, their h-preferences are not counted when setting the institutional h-value. This represents the fact that the dilemma facing the individuals that provision an institution is different from the dilemma faced by individuals about whether to cooperate or not once an institution is in place (Ostrom, 1990). To implement this we added a third locus to the model, with two variants. The first variant means that the individual participates in institution formation: their h-preference is counted when setting the institutional h-value, and they pay the cost I. These are the "administrators", while individuals with the second variant do not pay the cost I, and their h-preferences are not counted when forming the institution. Note that asocials also carry this locus, but do not express either variant since they do not join an institution. The h-value on patch j is then set according to:

$$h_j(t) = \frac{1}{\sum_{i=1}^{n_j(t)} s_{ij}(t)\iota_{ij}(t)} \sum_{i=1}^{n_j(t)} s_{ij}(t)\iota_{ij}(t)h_{ij}(t), \tag{S3}$$

where $\iota_{ij}(t) = 1$ if individual i on patch j is an administrator, $\iota_{ij}(t) = 0$ otherwise. If there are no social individuals that are administrators on a patch, then no institution is formed and the growth rates and carrying capacity of social individuals are given by Eq. S2. The effects of introducing the option to not pay I are shown in Fig. S3. Social individuals are still able to invade asocials under small initial patch sizes, and maintain institutions as the carrying capacity increases through co-evolution (Fig. S3a). The proportion of administrators in such cases is between 3 and 5% (Fig. S3b). This can be viewed as a division of labor, in which only a few individuals take on the administrator role.

Why are the administrators (individuals paying I) not driven extinct? In patches with no administrators, the institution collapses and socials individuals receive the same carrying capacity as asocials. Such patches are thus less productive than other patches with administrators. Consequently, the institution quickly becomes re-established by immigrant administrators from other patches that do have institutions. The result is that institutions are globally stable, even with the option to not pay the cost of their formation while still receiving their benefits.

As in the base model, social individuals remain at close to zero frequency for larger initial patch sizes. Recall that asocials carry the variants at the locus of the administrator trait, but do not express either phenotype, since they do not join an institution and hence do not pay the cost regardless of their trait value at this locus. Hence, the two variants among asocials are neutral. This explains why the frequency of the administrator variant (across all individuals, social and asocial) is around 50% for larger initial patch sizes, where asocials are at very high frequency (Fig. S3).

We have assumed here that individuals have the option of paying a zero cost for I. This is in fact a worst case assumption, since in reality there is always likely to be some cost of social living relative to an asocial lifestyle (e.g. increased parasite load). Thus, individuals that do not take part in institutional negotiations would still pay some non-zero I, but one lower than individuals that do negotiate. In this case, selection against paying I within a single patch would be even weaker, and so the institution on a single patch would collapse even less frequently.

Appendix S4: Varying the efficiency of punishment

To implement a varying efficiency of punishment, we introduced an efficiency constant P (range $[0, 1]$), such that the growth rate of defectors is given by

$$r_{\mathrm{d}j}(t) = r_{\mathrm{a}} - I - P\frac{(1 - h_j(t))\, n_{\mathrm{c}j}(t)B}{n_{\mathrm{d}j}\,(t)}, \tag{S4}$$

while the growth rate of a cooperator is still given by Eq. 2 of the main text.

Figure S4 shows the effect of varying P on the range of patch sizes over which the cooperative equilibrium is reached. This range is unaffected from the base model, unless P is less than 0.25 (meaning that the reduction in the growth rate of defectors is less than one quarter of the investment in sanctioning, corresponding to a very inefficient sanctioning technology) . For P less than 0.2, cooperative institutions are not able to reliably become established even under patch sizes of 10. However, such a low value of P is unlikely to be plausible. Indeed, many models (Boyd *et al.*, 2003, 2010; Sigmund *et al.*, 2010) and behavioral economics experiments (Fehr & Gächter, 2002; Herrmann *et al.*, 2008) assume that P is greater than 1 (e.g. ratios of 3 units of punishment per unit of investment are common in public goods experiments in behavioral economics).

Appendix S5: Introducing punishment free-riders that pay for cooperation but not sanctioning

Finally, we considered a version of the model similar to “pool punishment” models (Sigmund *et al.*, 2010; Perc, 2012), where we have two type of individuals contributing to the public good: those who contribute B at a cost C to themselves (our initial "cooperators"), and a novel type we call "punishment free-riders" which invest the agreed amount into the benefits of cooperation, but not into sanctioning. Punishment free-riders thus produce a reduced public good benefit of $h_j(t)B$ (all of which is allocated to carrying capacity enhancement), at a reduced cost of $h_j(t)C$. We first consider the case where punishment free-riders are sanctioned in the same way as defectors (as is also common in pool punishment models), such that investment in punishment is shared equally between punishment free-riders and ordinary defectors. The growth rate of punishment free-riders is then given by

$$r_{\mathrm{p}j}(t) = r_{\mathrm{a}} - I - h_j(t)C - P\frac{(1 - h_j(t))\, n_{\mathrm{c}j}(t)B}{n_{\mathrm{d}j}\,(t) + n_{\mathrm{p}j}(t)}, \tag{S5}$$

where $n_{\mathrm{p}j}(t)$ is the number of punishment free-riders on patch j at time t. Likewise, the growth rate of defectors is given by

$$r_{\mathrm{d}j}(t) = r_{\mathrm{a}} - I - P\frac{(1 - h_j(t))\, n_{\mathrm{c}j}(t)B}{n_{\mathrm{d}j}\,(t) + n_{\mathrm{p}j}(t)}, \tag{S6}$$

which depends on the efficiency of punishment P, while the growth rate of a cooperator is still given by Eq. 2 of the main text. The "carrying capacity" of socials is then given by

$$K_{\mathrm{s}j}(t) = K_{\mathrm{a}} + \beta\left[1 - \exp\left(-\gamma h_j(t)\left(n_{\mathrm{c}j}(t) + n_{\mathrm{p}j}(t)\right)B\right)\right]. \tag{S7}$$

Figure S5 demonstrates that for $P = 1$, the introduction of punishment free-riders that are themselves punished does not reduce the range of patch sizes over which institutions can be created and maintained. Indeed, we found that they were driven extinct, bar the effects of recurrent mutation (fig. S7c). Thus where they are themselves punished, the introduction of this type of "second-order free-rider" does not qualitatively affect our results.

We also investigated the case where punishment free-riders are *not* themselves punished. In that case, the growth rate of punishment free-riders is given by

$$r_{\mathrm{p}j}(t) = r_{\mathrm{a}} - I - h_j(t)C, \tag{S8}$$

while that of defectors is assumed to be

$$r_{\mathrm{d}j}(t) = r_{\mathrm{a}} - I - P\frac{(1 - h_j(t))\, n_{\mathrm{c}j}(t)B}{n_{\mathrm{d}j}\,(t)}, \tag{S9}$$

which depends on the efficiency of punishment P. The growth rate of a cooperator is still as given by Eq. 2 of the main text, while the carrying capacity of socials is given by Eq. S7.

The outcome of evolution then depends upon the efficiency of punishment, P. If the efficiency of punishment is high, then only a small number of individuals need to invest into punishment in order to maintain cooperation as an equilibrium. Thus for $P \geq 0.7$ cooperation remains stable (fig. S6): the interaction of demography, population structure and migration provides selection pressure to maintain sufficient investment into punishment. However, smaller values of P require greater investment into punishment in order to maintain cooperation. In that case, we find that cooperation becomes unstable (fig. S7b) as punishment free-riders increase in frequency (fig. S7d). Figure S6 shows the long-run mean frequency of cooperators and defectors for varying P (this is the frequency averaged over time during a single run of 3×10^6 generations). A long-run frequency of cooperators greater than 0.95 (i.e. stable cooperation) requires much larger P if punishment free-riders are not themselves punished.

References

Boyd, R., Gintis, H. & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328, 617–620.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P.J. (2003). The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 3531–3535.

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.

Herrmann, B., Thöni, C. & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319, 1362–1367.

Kosfeld, M., Okada, A. & Riedl, A. (2009). Institution formation in public goods games. *Am. Econ. Rev.*, 99, 1335–1355.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions).* Cambridge University Press, Cambridge, UK.

Perc, M. (2012). Sustainable institutionalized punishment requires elimination of second-order free-riders. *Sci. Rep.*, 2.

Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466, 861–863.

Supplementary figures

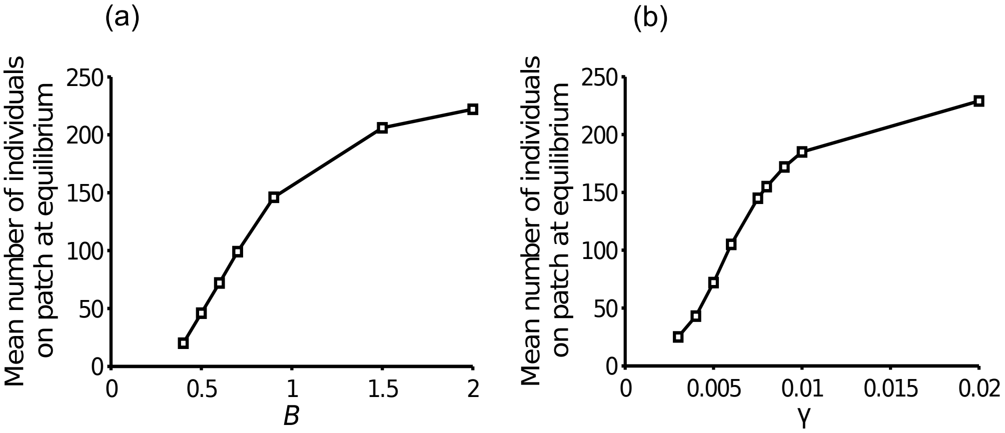

Figure S1: The equilibrium patch size after cooperation invades is affected by (a) the per capita benefit of cooperation, B; (b) the gradient of the benefit from cooperation function,γ. Parameters: $K_a = 15$, $m = 0.1$, $\gamma = 0.0075$, $I = 0.1$ (a), $B = 0.9$ (b).

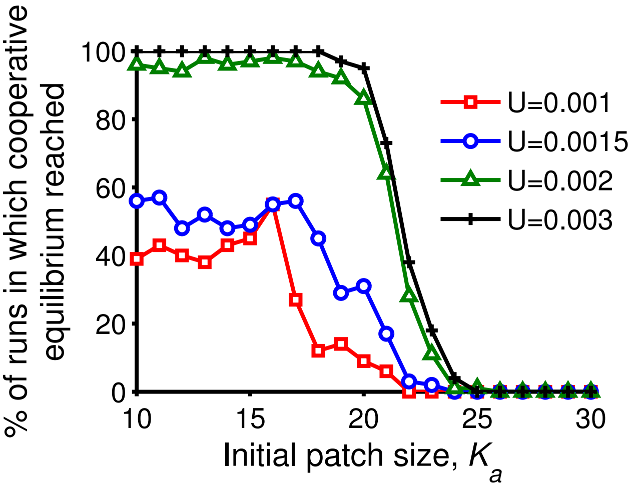

Figure S2: Effect of varying the consensus threshold parameter, U, on the number of trials (out of 100) in which the cooperative equilibrium was reached from a population initially fixed for asocials. Parameters: $B = 0.9$, $m = 0.1$, $I = 0.1$, $\gamma = 0.0075$.

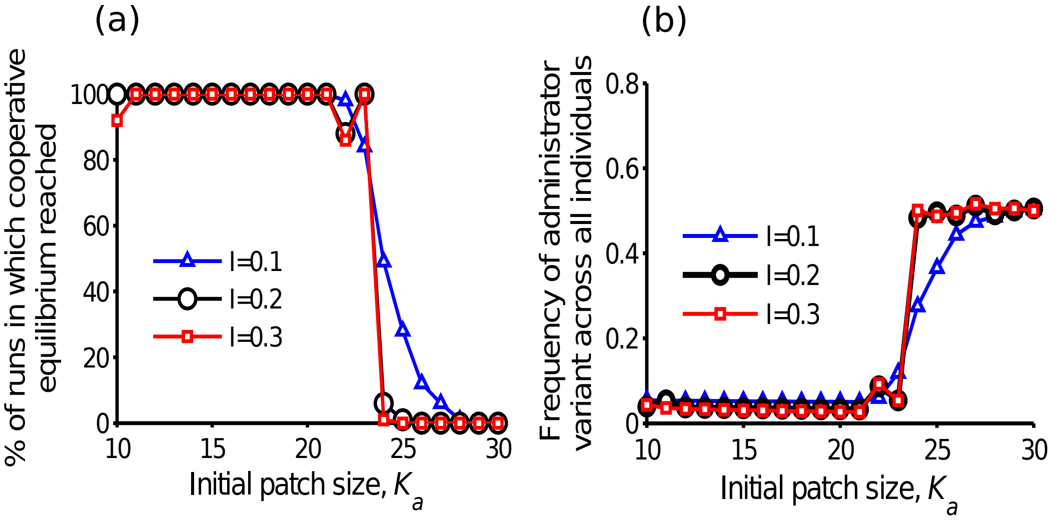

Figure S3: Effect of introducing the option for an individual to not pay the cost I of institution formation, and not have their h-preference affect the institutional h-value. Social individuals with the "administrator" variant pay the cost I, and their h-preference is counted when setting the institutional h-value. Asocials carry the locus for the administrator trait, but do not express either phenotype since they do not join an institution. (a) The number of trials (out of 100) in which the cooperative equilibrium was reached from a population initially fixed for asocials. (b) The global frequency of the administrator variant during the last 1000 generations (across all individuals, social and asocial, averaged over 100 trials). Parameters: $B = 0.9$, $m = 0.1$, $\gamma = 0.0075$.

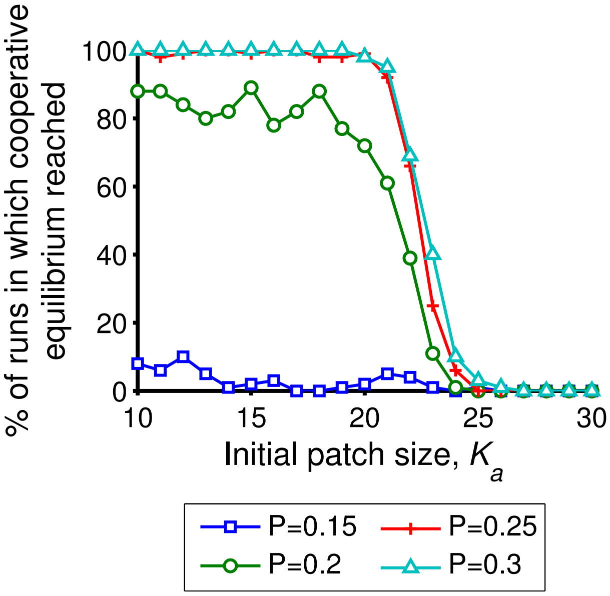

Figure S4: Effect of varying the efficiency of punishment, P, on the range of patch sizes over which the cooperative equilibrium is reached from a population initially fixed for asocials (over 100 trials). Parameters: $B = 0.9$, $m = 0.1$, $\gamma = 0.0075$, $I = 0.1$.

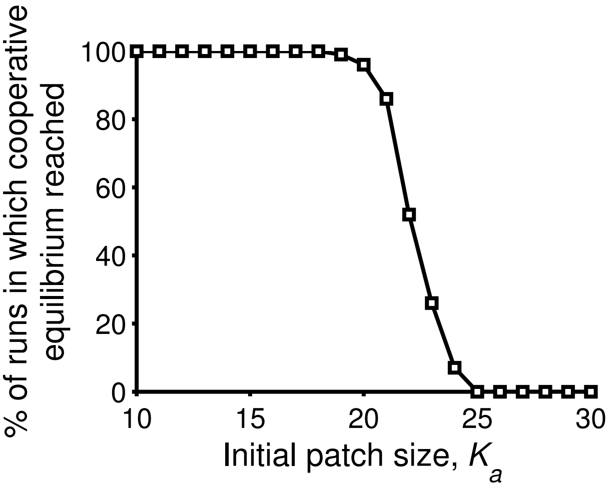

Figure S5: Effect of introducing punishment free-riders that invest the agreed amount into cooperation but not into sanctioning (but are punished for doing so), on the range of patch sizes over which the cooperative equilibrium is reached from a population initially fixed for asocials (over 100 trials). Parameters: $B = 0.9$, $m = 0.1$, $\gamma = 0.0075$, $I = 0.1$, $P = 1$.

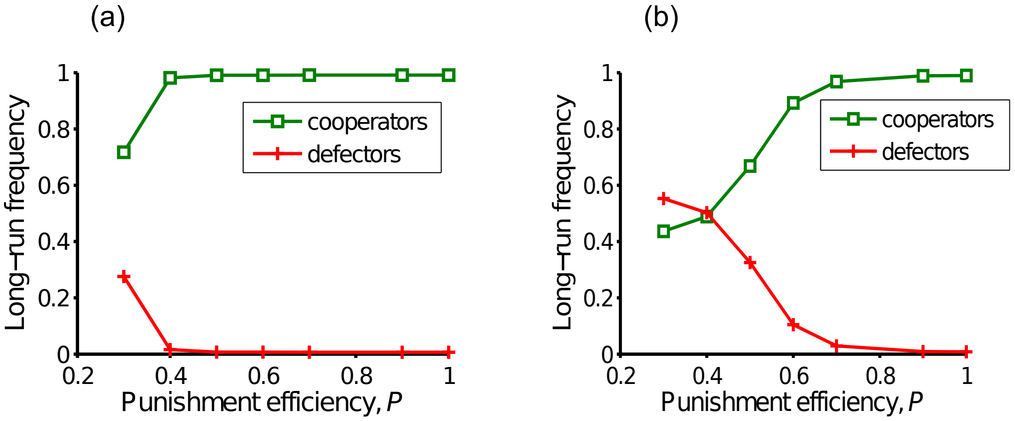

Figure S6: Long-run frequencies (mean over 3×10^6 generations) of cooperators and defectors, in the presence of punishment free-riders. (a) Punishment free-riders are punished the same as defectors. (b) Punishment free-riders are not punished. Parameters: $K_{\mathrm{a}} = 15$, $B = 0.9$, $m = 0.1$, $\gamma = 0.0075$, $I = 0.1$.

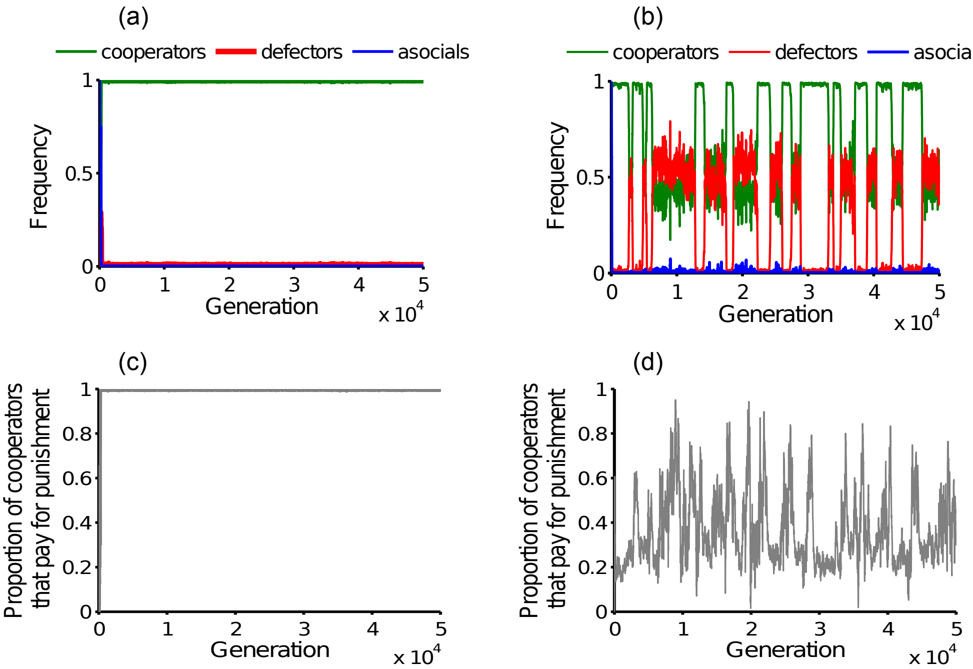

Figure S7: Dynamics during a run with inclusion of punishment free-riders, with punishment efficiency $P = 0.5$. Upper plots show type frequencies. (a) Punishment free-riders are punished in the same way as defectors. (b) Punishment free-riders are not punished. Lower plots show the proportion of cooperators that invest into punishment. (c) Punishment free-riders are punished in the same way as defectors. (d) Punishment free-riders are not punished. Parameters: $K_\mathrm{a} = 15$, $B = 0.9$, $m = 0.1$, $\gamma = 0.0075$, $I = 0.1$.