12-2008

The Evolution of Reciprocity in Sizable Human Groups

Casey G. Rothschild
Middlebury College

Citation

Casey Rothschild (2009). The evolution of reciprocity in sizable human groups. Journal of Theoretical Biology 257, 609-617.

The Evolution of Reciprocity in Sizable Human Groups

Casey G. Rothschild[1]

Department of Economics, Middlebury College VT, 05753

Abstract

The scale and complexity of human cooperation is an important and unresolved evolutionary puzzle. This article uses the finitely repeated n person Prisoners' Dilemma game to illustrate how sapience can greatly enhance group-selection effects and lead to the evolutionary stability of cooperation in large groups. This affords a simple and direct explanation of the human "exception."

Key words: Human Cooperation, Group Selection, Reciprocal Altruism, Sapience

[1] Present Address: MIT Department of Economics, 50 Memorial Drive, E52-262f, Cambridge, MA 02142. Tel: 617.324.7502. Fax: 617.253.1330. Email Address: caseyr@mit.edu. Thanks: to Muhamet Yildiz and Drew Fudenberg for comments on an early draft; to Jeffrey Carpenter and Peter Matthews for re-inspiring me to pick this idea up again; to two anonymous referees for useful advice; and to Yanmin Choo and Phuong Le for many helpful comments and assistance.

Preprint submitted to Elsevier Science *29 December 2008*

The evolution of human cooperation is a long-standing puzzle that has received much recent attention. Research has focused on three nested questions. First: how can altruistic behavior survive evolutionary pressures at all (Hamilton, 1964a,b; Fletcher and Zwick, 2004; Nowak, 2006)? Second: how can cooperation evolve in large groups of unrelated individuals (Boyd and Richerson, 1988; Boyd et al. 2003)? Third: why does large-scale cooperation among unrelated individuals seem to be a distinctly human phenomenon (Bernhard et al., 2006; Fehr and Fischbacher, 2003; Fehr and Gächter, 2002; Gintis 2000; Bowles, 2006)?

A number of compelling arguments have been forwarded for resolving the first two questions. These include: kin selection and inclusive fitness (Hamilton, 1964a,b); reciprocal altruism (Trivers, 1971; Axelrod, 1984); altruistic punishment (Fehr and Gächter, 2002; Henrich and Boyd, 2001) and reputation (Rockenback and Milinski (2006)); group and multi-level selection (Maynard-Smith, 1964; Wilson and Sober, 1994); assortative matching (Wright, 1921; Bergstrom, 2002); and spatial effects and imitation (Nowak and May, 1992; Grim, 1995; Nowak and Sigmund, 2004; Boyd and Richerson, 2002; Langer et al. 2008).

The puzzle of human exceptionalism is a particularly active topic of research (Gintis, 2000; Bowles, 2006; Boyd, 2006; Wilson and Wilson, 2007; Johnson et al., 2008). As articulated by Fehr and Fischbacher (2003), a fully successful resolution requires showing that large-scale cooperation is possible precisely because of some characteristic "quantitatively, or probably even qualitatively, unique" (page 785) to humans. The core of this paper is a very simple model with precisely this feature.

The paper is structured as follows. Section 2 considers a completely standard finitely repeated public goods game (n player Prisoners' Dilemma) model of evolutionary dynamics. In this game, reciprocal altruists compete with unconditional defectors, as in Cohen and Eshel (1976). Evolutionary dynamics can support cooperation in small groups in this model, but, as in Boyd and Richerson (1988), cooperation becomes unsustainable as group size grows.

Section 3 then tweaks the model by replacing the unconditional defectors with non-cooperative agents who are more recognizably human: they are intrinsically non-cooperative in the sense that they are purely self-interested, but they play strategically. So, even though they are not intrinsically cooperative, they can *behave* cooperatively when it is in their own selfish best interest. This simple modification completely reverses the standard result: identical evolutionary dynamics not only *can* support cooperative behavior in large groups, but cooperative behavior is actually *ensured* in sufficiently large groups.

This result is similar to Johnson et al.'s (2008) discussion of the consequences of replacing "Tit for Tat" reciprocators with "continuous" reciprocators in Boyd and Richerson's (1988) model, in that it highlights the sensitivity of Boyd and Richerson's conclusions to the specification of strategies of the players. It differs in several important respects, however.

First, the model presented here provides a simple and clean example to illustrate how sapience (specifically: strategic foresight) can greatly enhance the survival of cooperation in large groups. In this respect, it directly addresses why large-group cooperation might be particular to humans.

Second, Johnson et al. (2008) interpret the sensitivity of Boyd and Richerson's results as evidence in favor of individual rather than group selection as the underlying mechanism for human cooperation. The present results are based on a model with similar dynamics, but here they can be interpreted as supporting group selection mechanisms. Specifically, a particularly intriguing feature that appears in this model is a stark dichotomy between phenotypically cooperative *behavior* of individuals in groups and the genotypically *uncooperative* nature of the majority of those individuals. This is an example of a point emphasized by Wilson (2004): in the context of complicated phenotype-genotype relationships, absolute fitness advantages can fail to translate into relative fitness advantages, even with randomly formed groups. In Section 3's model, strategically rational types have an absolute fitness advantage, but, ironically, their strategic foresight helps ensure the continued survival of genotypically cooperative non-strategic individuals by facilitating broad cooperation in groups within which the later represent only a distinct minority. This allows the cooperative genotype to maintain a relative fitness advantage despite representing only a small proportion of the population.

Group or "multi-level selection" explanations for human cooperation espoused, for example, by Bowles (2006) and Wilson and Wilson (2007) emphasize the importance of social control and culturally transmitted norms in facilitating group or multi-level selection in human populations. Sapience clearly plays an important, if indirect, role in these sorts of stories. The present results indicate how sapience can *directly* enhance the evolution of cooperation in large groups, even in the absence of complex systems of social control or cultural and moral norms. Presumably, these direct effects reinforce the cultural norm effects.

The stark reversal of the large-scale survival of cooperative behavior with sapient actors in place of unconditional defectors is striking. There are a number of reasons to interpret this result with some caution, however. For example, the formal results in Section 3 rely on a number of simplifications, including: (i) the observability of type; (ii) a continuum of individuals; (iii) a zero mutation rate; and (iv) a significant (though not exorbitant) level of reasoning capacity for the strategic agents. Section 4 the extent to which these are reasonable abstractions.

Section 5 offers some brief conclusions. Proofs of key results appear in the Appendix.

1 The Evolutionary Model

The evolutionary model is a version of the Haystack model (Maynard-Smith, 1964). Each generation t consists of a unit measure of individuals. Individuals are sorted randomly (non-assortatively) into groups (the haystacks) with n players. The n individuals within each group play an M times repeated public goods game. At the end of the M periods, generation t individuals reproduce asexually. The fraction of generation $t+1$ players who are the offspring of a given generation t individual is proportional to the generation t individual's M-period payoff. The generation $t+1$ individuals are then assigned randomly to new groups (haystacks) of size n, play the repeated public goods game, reproduce... and so on.

In each of the M periods of the public goods, each of the n individuals chooses whether to cooperate (C) or to defect (D). If j individuals play C in a given period, those who cooperate and those who defect receive period payoffs of βj and $\beta j + 1$, respectively. The interpretation is that each player has a unit endowment which she can consume (playing D) or contribute to a public good (playing C). The latter strategy provides a benefit β to each individual in the group, including herself. Contributions are assumed to be socially productive ($\beta n > 1$) but individually harmful ($\beta < 1$). For technical reasons, β is further assumed to satisfy the slightly stronger condition: $\beta n \geq (2 - \beta) > 1$.

Individuals' strategies are determined by their genetic "type," which is transmitted without mutation from parent to child. (The no-mutation assumption is inessential but expositionally convenient.) Specifying the set of "types" determines the play of the game and hence the evolutionary dynamics. The following two sections analyze these dynamics with two distinct sets of types.

2 The Baseline Model

This section considers a model with two types. Uncooperative types (U-types) are unconditional defectors: they play D each period. Reciprocal altruists, or "Tit-for-Tats" (T-types), play C as long as all members of their group cooperated in the preceding period.[2]

2.1 Payoffs

Let j index the number of T-types in a given group. All individuals will cooperate in all n periods in groups with $j = n$. In groups with $j < n$, the j T-types will cooperate in period 1, the $(n - j)$ U-types will defect in period 1, and all individuals will defect in periods $2, \cdots, M$. Payoffs $u_T(j)$ and $u_U(j)$ to T and U types are therefore given as follows:

$$\begin{array}{c|ccc} & j = 0 & 0 < j < n & j = n \\ \hline u_T(j) & -- & \beta j + M - 1 & \beta M n \\ u_U(j) & M & \beta j + M & -- \\ \hline \end{array} \tag{1}$$

For notational ease, define $u_U(n) = 0$ and $u_T(0) = M - 1$. Then $u_U(j) =$

[2] Boyd and Richerson (1988) consider generalized T-types who cooperate so long as fewer than a members defected in the preceding period. Allowing for these types is expositionally more cumbersome, but it does not materially affect the results.

$u_T(j) + 1$ for all j.

2.2 Dynamics

Let p_t denote the fraction of T types in generation t. Random assignment implies that the fraction of generation t individuals who are in groups with j T-types is given by the binomial density,

$$f(n, j, p_t) = \binom{n}{j} p_t^j (1 - p_t)^{n-j}. \tag{2}$$

It will also be useful to use the cumulative distribution function (the proportion of the population in groups with no more than j types):

$$F(n, j, p) = \sum_{i=0}^{j} f(n, j, p). \tag{3}$$

Evolutionary dynamics are given by the map $D_n : [0, 1] \to [0, 1]$ from the proportion p_t of T-types in generation t to the proportion p_{t+1} in the subsequent generation:

$$D_n : p_t \to p_{t+1} = \frac{\sum_{j=0}^{n} f(n, j, p_t) j u_T(j)}{\sum_{j=0}^{n} f(n, j, p_t) \left[j u_T(j) + (n - j) u_U(j) \right]}. \tag{4}$$

(These are simply the "replicator" dynamics described in, e.g., Taylor and Jonker (1978).) Theorem 1 summarizes the properties of these dynamics.

Theorem 1 (Baseline Dynamics) *The mapping D_n described in Equation (4) has exactly three fixed points: $\underline{p} = 0$, $\overline{p} = 1$, and $p^* = \left(\frac{1-\beta}{(\beta n-1)(M-1)}\right)^{\frac{1}{n-1}}$. Points $\underline{p}$ and $\overline{p}$ are stable. Point p^* is unstable.*

[Insert figure 1(a) and (b) about here]

The Appendix contains a formal proof of this well known result.

Theorem 1 implies that, for any n, evolutionary dynamics lead (almost surely) to a homogenous population. The unstable interior fixed point p^* marks the cutoff between initial population fractions that will lead to a population with all defectors ($p < p^*$) or all Tit-for-Tats ($p > p^*$). Figure 1 provides an illustration for $n = 25$ and $n = 200$. Note that the basin of attraction for the cooperative steady state is smaller for $n = 200$. This is an illustration of evisceration of cooperation in large groups discovered by Boyd and Richerson (1988) and formalized in the following corollary. The corollary, which uses D_n^t to denote the t-times iterate dynamics, follows directly from a the observation that $\lim_{n\to\infty} p^* = 1$.

Corollary 2 (Evisceration of Cooperation in Large Groups) *For all $p < 1$ there exists n^* such that*

$$\lim_{t\to\infty} D_n^t(p) = 0$$

for all $n \geq n^*$.

Boyd and Richerson view their analog of Corollary 2 as a puzzle in need of explanation, saying:

> This result satisfies the natural historian's conventional wisdom: large, cooperative, groups composed of distantly related individuals are unusual in nature. But it leaves human cooperation unexplained.

The next section modifies this simple model in order to illustrate a potential resolution of this puzzle.

3 A Modified Model with *Homo Sapiens*

The baseline model has two types: the intrinsically cooperative Tit-for-Tats, and unconditional defectors. The defectors are actively *un*-cooperative: they are hard-wired to make anti-social choices *even when it runs counter to their own narrow self-interest.* A distinguishing feature of humans is their ability to reason strategically and avoid such mistakes. This section considers how the dynamics change when the irrational unconditional defectors from the baseline model are replaced with selfishly motivated but reasoning players.

3.1 Types

Tit-for-Tat types are exactly as in the baseline model. The new genotypically non-cooperative types are now strategic (S)-types: they are the perfectly rational, forward looking players of standard non-cooperative game theory. Type is common knowledge within a given group. [3]

[3] Section 4 discusses the robustness of the conclusions to these assumptions. In particular, it notes that for low numbers of repetitions M, "perfect rationality" is a much stronger assumption than is necessary for the results to hold.

3.2 Payoffs

Payoffs are determined by a subgame perfect equilibria (SPE; Selten, 1965). Define $j^* = min\{j | j \geq (1-\beta)/\beta\}$ and consider the following strategies:

- If $j < j^*$: defect in every period.
- If $j \geq j^*$: cooperate in periods $1, \cdots, M-1$; defect in period M.

These strategies describe the highest payoff SPE of the game. To wit: defecting is clearly a strictly dominant strategy in period M. In period $M-1$, S types who deviate and play D gain $1-\beta$. This induces the T-types to defect in period M, reducing the deviator's utility by βj in groups with $j < n$ T-types. Cooperation at period $M-1$ is thus consistent with a SPE if and only if $\beta j \geq 1-\beta$, i.e., if and only if $j \geq j^*$. If $j \geq j^*$, a similar argument shows that cooperating in $M-2$, $M-3$, $\cdots$, 1 is an SPE strategy if $j \geq j^*$. If $\beta j < 1-\beta$, then backwards induction reveals that "defect in every period" is the unique SPE strategy for strategic types. Hence, the proposed strategies constitute a SPE, and no other SPE can involve more cooperation.

Taking this "best" SPE to be outcome of the game,[4] the M-period payoffs as a function of the number j of T-types in a group are given by:

[4] Focusing on the best equilibrium is common, and it has been shown to have firm theoretical justifications in some contexts (Fudenberg and Maskin, 1990; Kim and Sobel, 1995). The justification for making it in this particular setting, however, is purely pragmatic: it makes the argument as clean and simple as possible.

$$\begin{array}{c|cc} & 0 \leq j < j^* & j^* \leq j \leq n \\ \hline u_T(j) & \beta j + M - 1 & (M-1)\beta n + \beta j \\ u_S(j) & \beta j + M & (M-1)\beta n + \beta j + 1 \\ \hline \end{array} \tag{5}$$

For notational convenience, (5) implicitly uses $u_T(0) = M - 1$ and $u_S(n) = M\beta n + 1$, so that $u_S(j) = u_T(j) + 1$ for all j.

3.3 Dynamics

Other than the different payoffs to the two types, evolutionary dynamics are the same as in the baseline model. Let p_t denote the fraction of T-types in generation t, let $f(n, j, p_t)$ be defined as in Equation (2), and let $\tilde{D}_n$ denote the mapping $p_t \rightarrow p_{t+1}$, i.e.,

$$\tilde{D}_n : p_t \rightarrow p_{t+1} = \frac{\sum_{j=0}^{n} f(n, j, p_t) j u_T(j)}{\sum_{j=0}^{n} f(n, j, p_t) \left[j u_T(j) + (n - j) u_S(j) \right]}. \tag{6}$$

The following theorem summarizes the key properties of this mapping. A formal proof appears in the Appendix.

Theorem 3 (Modified Dynamics) *The points $\underline{p} = 0$ and $\overline{p} = 1$ are fixed points of the mapping $\tilde{D}_n$. Point $\overline{p}$ is unstable; point $\underline{p}$ is stable. Furthermore:*

(1) For all p_0, $\lim_{t \to \infty} \tilde{D}_n^t(p_0)$ exists.

(2) $\{\tilde{D}^t(p)\}_{t=0}^{\infty}$ *is monotone in* t*: the sequence is either non-increasing for all* t *or else is non-decreasing for all* t*.*

(3) $\exists\ N$ *such that for* $n \geq N$ *and for all* $p_0 \geq \frac{j^*}{n}$*,* $\lim_{t\to\infty} \tilde{D}_n^t(p_0) > \frac{j^*}{n}$*.*

[Insert figure 2 (a) and (b) around here]

Properties (1) and (2) of Theorem 3 state that the evolutionary dynamics are "nice" in the sense that the fraction of T-types will converge monotonically to some stable level (which may depend on the initial fraction of T-types). Panel Figure 2(a) illustrates for $n = 25$, $M = 2$, and $\beta = .2$. It shows four fixed points: the stable points 0 and p^* and the unstable points p' and 1. The basin of attraction of the interior fixed point p^*—at which T-types "survive" evolutionary pressures—is the range $(p', 1)$. When group size increases to $n = 500$, as in Figure 2(b), the dynamics are qualitatively similar, but both interior fixed points have moved left; the basin of attraction for the "T-types survive" fixed point p^* has correspondingly increased; and the fraction of T-types at p^* has decreased.

Simulations suggest that the dynamics for sufficiently large n *always* have the same qualitative two-basin-of-attraction structure. They also indicate that the interior "peak" visible in both panels of Figure 2 occurs near $\frac{j^*}{n}$—the "critical fraction" of T-types required for cooperation—and that the peak gets narrower and narrower as n grows.[5] It thus appears to be true that there is a unique

[5] If $-\frac{dF(n,j,p)}{dp}$ is a single peaked function of p—which seems intuitively correct and has been borne out by all simulations, but for which there is no obvious proof—then

stable interior fixed point p^* with the following properties: (a) for sufficiently large n it is always to the right of $\frac{j^*}{n}$; (b) it converges to zero as group size grows; and (c) its basin of attraction converges to (0,1).

Lacking a formal proof of this conjecture, Property (3) of Theorem 3 makes a weaker (but still sufficiently strong) claim: it asserts instead that the interval $[\frac{j^*}{n}, 1)$ is contained within the union of the basins of attraction of all of the steady states strictly to the right of $\frac{j^*}{n}$. Hence, as n grows, the basin of attraction of all fixed points with surviving T-types converges to $(0, 1)$, and the survival of *some* T-types is ensured.

It follows that the T-types' survival is ensured in large groups—though apparently only as small fractions of large groups. As the following corollary establishes, however, the genotypically cooperative T-types survive in sufficient numbers to ensure that cooperative behavior is the norm in large groups. It asserts formally that, starting from *any* initial fraction of T-types, most groups in every generation will be cooperative—in the sense that all individuals in these groups will cooperate in periods $1, ..., M-1$—so long as the group size is sufficiently large. The formal proof is provided in the Appendix.

Corollary 4 (Survival of Cooperation in Large Groups) *For any* $p_0 > 0$ *there exists an* N *such that*

$$n > N \Rightarrow 1 - F(n, j^* - 1, \tilde{D}_n^t(p_0)) > 0.5 \quad \forall t \geq 0.$$

this topological structure on the dynamics would follow easily.

Figure 3 illustrates this corollary. It fixes $p_0 = .95$ and computes the $\lim_{t\to\infty} p_t$ for both the baseline dynamics (panel (a)) and the modified dynamics (panel (b)), and plots this limit as a function of group size. In the baseline dynamics, T-types go extinct when groups are larger than $n = 46$. In the modified model, the surviving fraction of T-types also decreases towards zero as group size grows, but it remains strictly positive. The second curve in panel (b) plots the limiting fraction of types in cooperative groups—i.e. those with at least j^* T-types and which therefore cooperate for at least $M - 1$ periods. It suggests an even stronger result than Corollary 4 establishes: as group size grows, the fraction of cooperative groups appears to approach one—so that almost *everybody* behaves cooperatively. All simulations have borne this stronger result out, though a formal proof has remained elusive.

[Insert Figure 3 (a) and (b) around here]

4 Discussion and Caveats

The difference between Corollaries 2 and 4 is striking: for a fixed p_0, the former states that cooperation is completely eviscerated in sufficiently large groups; the latter states that cooperative behavior becomes the norm in large groups. The intuition behind this reversal is straightforward. As formalized by Price (1970), the long-run evolutionary stability of cooperation is determined by a horse-race between "within group" effects, which favor the genotypically non-cooperative types and "between group" effects, which favor groups with more genotypically cooperative T-types. The evisceration of cooperation formalized in baseline model obtains because the number of T-types needed to sustain cooperative behavior in any given group grows with the size of the group. Groups with sufficiently many T-types become increasingly rare as group size grows, so the between-group effects become negligible.

In the modified model, by contrast, the fraction of T-types required to sustain cooperative behavior in a given group decreases with group size. This is because only a small number (j^*) of T-types in a group is needed to induce strategic cooperation by the self-interested S-types. Combined with the growing gross public benefit of cooperation (i.e., $\beta n - 1$), this ensures that between-group evolutionary forces come to dominate as group size grows.

This intuition makes it clear that the reversal is robust to several modifications, such as allowing Tit-for-Tat types to have some (fixed or slowly growing) tolerance for the number of defectors, or having the public benefit β decrease with group size (so long as the gross public benefit of pro-social behavior grows sufficiently quickly). Similarly, the qualitative results are robust to the introduction of a small exogenous probability of "mutation" to the opposite type during reproduction.

A number of other modeling assumptions raise potentially more significant concerns about the practical interpretation the results. First, one might worry that invoking subgame perfection requires endowing strategic types with an implausible amount of rationality. Second, Section 3 takes type to be observable.

Since observability plays a critical role in generating cooperation in groups with strategic players, it is important explore the extent to which the result hinges on this assumption. Third, the joint assumptions of a continuum of individuals and a deterministic environment rule out the possibility of "accidental" extinction of the genotypically cooperative T-types *via* drift. This is a particular concern since the fraction of T-types in the interior equilibrium identified in Theorem 3 shrinks to zero as the group size n grows.

We consider each of these concerns in turn.

4.1 Rationality Assumptions

The baseline model of Section 2 departs from Boyd and Richerson's (1988) model by assuming that the number of repetitions of the game M is finite rather than infinite. With non-rational actors, this distinction is unimportant. With strategic actors it is.

Although formal results that are qualitatively similar to those in Section 3 could be derived in an infinitely repeated version of the model, the finitely repeated version relies on substantially less stringent cognitive capability assumptions for equilibrium play. When $M = 2$, for example, the proposed equilibrium requires only that strategic types are able to calculate one period ahead (and second order mutual knowledge of this fact and of rationality). This is far less restrictive than the "common knowledge of rationality" assumption that would be required to ensure equilibrium in the infinitely repeated version. Qualitatively, with finite repetitions, strategic types only have to be *homo sapiens*, not *homo economicus*.

4.2 Observability of Type

The assumption that type is perfectly observable is analytically important: it is what allows strategic types to condition their play on the number of genotypically cooperative T-types in their group. After discussing the extent to which it is conceptually important, this section describes how and when allowing self-reporting can replace the observability assumption.

4.2.1 Consequences of Relaxing Observability

To highlight the importance of the observability assumption, consider a polar opposite case: type is completely unobservable and strategic types know only the population fraction p of T-types.

On the one hand, individual *behavior* here is quite similar to behavior in the "perfect observability" case: as in Kreps et al. (1982), there is a (Bayesian) equilibrium with the property that, for sufficiently large p, strategic types cooperate with high probability in most rounds of the game. Since cooperation is probabilistic in this equilibrium, cooperation will, on average, be higher in groups with more cooperative T-types, just as in the perfect observability case.

On the other hand, the evolutionary dynamics are much different: they inevitably lead to complete evisceration of cooperation. Observability confers a distinct disadvantage on S-types, since it can lead to the unraveling of cooperative behavior as other S-types anticipate the last-round defections (and then penultimate round defections, etc...). Removing observability ensures that S-types will achieve at least weakly, and sometimes strictly higher payoffs than T-types (since they can always imitate T-types).

Whether or not evolutionary forces with intermediate levels of observability will generally eviscerate cooperation is an open—and analytically challenging—question. Suppose, for example, type is observable, but there is some positive probability of "recognition" errors. Then groups with more T-types will be more likely to be cooperative than groups with fewer—just as in the "perfect observability" case. This induces the qualitative correlation between group composition and group payoff that is necessary for between group forces to potentially overwhelm within group forces. At the same time, imperfect observability blurs the sharp cutoff at j^* between cooperation and the lack thereof, reducing the *quantitative* magnitude of this correlation. Monte Carlo simulations indicate that that the "perfect observability" dynamics are robust to the introduction of modest recognition errors for a fixed group size. Whether or not a fixed error rate undermines cooperation in sufficiently large groups remains an open question.

4.2.2 Self-Revelation of Type

One might hope to rely on self-reporting of type rather than assuming observability. This can be modeled by looking for a truth-telling equilibrium in a pre-game round wherein each individual in a group simultaneously states his type, and where play in the M-round repeated game then follows the equilibrium strategies from Section 3 (with the *reported* number j of T-types).[6]

If deception is costless and impossible to observe, then truth telling is *not* an equilibrium, since if (and only if) a strategic type happens to be in a group with exactly j^*-1 T-types, unilateral mis-reporting by an S-types will improve his payoff by inducing cooperation in rounds $1, \cdots, M-1$ by his opponents.

Now suppose deception involves some small cost $\varepsilon(n)$ and that there is some probability $P(n,j)$ of a group successfully "sniffing out" a deception. The gross benefit of unilaterally misreporting is:

$$(1-P(n,j^*-1))\,[f(n,j^*-1,p)]\,[(M-1)(\beta n-1)+(2-\beta j^*).] \qquad (7)$$

Expression (7) assumes that an unsuccessful deception leads the group to play according to the true rather than the reported j. The first term is the probability of not being sniffed out. The second term is the probability of being in a group with exactly j^*-1 individuals (the only time successful unilateral deception matters). The third term is the benefit conditional on being in such

[6] Grégoire and Robson (2003) consider a qualitatively similar "pre-play" signalling game in a model with different dynamics.

a group and successfully deceiving; it is computed under the assumption that an S-type who gets away with deception subsequently cooperates in periods $1, ..., M-2$ and defects in $M-1$ and M (which is optimal).

Since $f(n, j^*-1, p)$ is maximized at $p = \frac{j^*-1}{n}$ and

$$\lim_{n\to\infty} f\left(n, j^*-1, \frac{j^*-1}{n}\right) = \frac{(j^*-1)^{j^*-1}}{(j^*-1)!} e^{-j^*+1} \tag{8}$$

is finite, the cost of deception will outweigh the expected benefit in large groups whenever $\frac{\varepsilon(n)}{1-P(n,j)}$ grows slightly faster than n. In this case, engaging in deception will be undesirable for sufficiently large n, and the central results of Section 3 will continue to hold. This condition is plausible. It will hold, for example, if ε is independent of n and each strategic type has an independent and arbitrarily small probability $\eta > 0$ of sniffing out a defection. Alternatively, it will hold if deception has a per group member cost and if the likelihood of detection increases, even arbitrarily slowly, with n.

4.3 Drift

Focusing on a model with a continuum of individuals is analytically convenient. One might reasonably have concerns about the appropriateness of this abstraction, however, especially since the population fraction of T-types in the interior fixed point identified in Corollary 4 shrinks zero as group size grows. In particular, random fluctuations could reduce the realized fraction of T-types in a given generation into the basin of attraction of the steady state at $p = 0$

(*viz* Figure 2) and lead to their eventual extinction.

A back of the envelope calculation is useful for assessing the quantitative importance of this concern. A generous time-frame for human evolution is on the order of 100,000 generations. To have a better than 50% chance of surviving this long, the probability of extinction in any given generation should be less than approximately 10^{-5}. So if fluctuations large enough to eliminate T-types in a given generation are "5-sigma" events, then they are not quantitatively problematic.

Consider mean-zero shocks which cause the realized fraction of p of T-types in a given generation to be approximately normally distributed around the expected value p^* (i.e., the interior equilibrium). If the stochasticity is individual-specific, the variance of this distribution will scale as $\eta/(p^* N)$, where N is the total population size, and η is a measure of the individual-level fluctuations, which we conservatively take to be 1.[7] (For a concrete example, suppose that each T-type "birth" has a probability $\eta/2$ each of producing twins or of being stillborn.)

Extinction of T-types will result if random fluctuations lead their population fraction to fall below $(1-\alpha)p^*$ for some α. (In Figure 2, $\alpha = \frac{p'}{p^*}$.) This will be

[7] Focusing on individual-specific shocks is reasonable here since the aggregate population shocks that uniformly both T- and S-types will not affect p.

a 5-sigma event when:

$$\frac{\alpha p^*}{\sqrt{1/(p^*N)}} = \frac{1}{2}\left(\frac{1-\beta}{\beta n}\right)^{\frac{3}{2}}\sqrt{N} \geq 5. \tag{9}$$

Or, for small β, approximately when $(5/\alpha)^2(\beta n)^3 \leq N$. Boyd et al. (2003) use $\beta n = 2, 4$ and 8; these imply robustness to drift if $N > 800/\alpha^2, 6400/\alpha^2$, and $51,200/\alpha^2$ for $\beta n = 2, 4$ and 8, respectively, so long as the interior steady state exists.[8] For $b = 8$, simulations indicate that $\alpha \approx .5$ when $n = 100$ (which Boyd et al. (2003) suggest is reasonable for representing evolution in small scale societies). Then cooperation is robust to drift as long as the total population N is on the order of $500,000$. For $b = 4$ and $n = 100$ and 1000, drift is unproblematic even for $N = 20,000$ and $N = 100,000$, respectively. So the drift-free model appears to be a reasonable abstraction.

5 Conclusions

Theorem 3 and its corollary show how the received wisdom that cooperative behavior is evolutionarily unstable in large groups is highly sensitive to modeling assumptions: simply replacing the biological automata of standard models with forward-looking strategic players, cooperation in large groups becomes the *norm*, completely overturning standard "impossibility of cooperation large

[8] Boyd et al. (2003) fix $b \equiv \beta n$ instead of β. Theorem 3 assumes a fixed β, so it does not guarantee an interior steady state for large n; such an interior steady state *will* always exist for sufficiently large M.

groups" results. Though suggestive, the results herein can hardly be regarded as dispositive regarding the evolutionary causes of human cooperation: the model on which they are based is simply too stylized.[9]

Instead, the central results of this paper are best viewed as illustrative of two important ideas. First, the *potential* importance of human facultative reasoning skills—arguably the defining characteristic of *homo sapiens*—should receive greater emphasis in explanations of why our species is essentially unique in exhibiting large scale cooperation among unrelated individuals. But this explanation is probably best viewed as a complement to rather than a substitute for other recently proposed resolutions of the puzzle of human exceptionalism.[10] Gintis (2000), for example, argues that the combination of strong reciprocity (punishment of non-cooperative behavior) and uniquely human abilities which make punishment "cheaper" (such as tool-making, hunting ability,

[9] For example, enriching the set of types and interactions would almost certainly yield a more complex set of evolutionary dynamics and complicate the clean cut results of this streamlined model. As pointed out by Doebeli and Hauert (2005), alternative games can yield significantly different results; and actual evolutionary processes almost certainly involve a multiplicity of different types of "games." Furthermore, Lindgren's (1991) tournaments and theoretical studies of complex dynamical systems indicate that highly complex and difficult to analyze dynamics are likely to be the norm in real-world dynamical processes.

[10] Even if it were *the* explanation for large scale human cooperation, sapience would't fully resolve the "puzzle" of human exceptionalism; it merely reduces the question to why our *sapience* is exceptional.

and stone throwing) can help to resolve it.[11] It seems likely that strategic punishment by sapient types—the analog of strategic cooperation—would further enhance the evolutionary case for strong reciprocity.

Bowles (2006) emphasizes group selection in the context of cultural transmission of reproductive leveling institutions such as monogamy and points out that this transmission is only possible in light of humans' cognitive and linguistic abilities. Closely related is Wilson and Wilson's (2007) and Wilson et al.'s (2008) strong advocacy for multi-level selection theory. They argue that that humans have undergone a "major transition" to a fundamentally "groupish" nature so that human *groups* effectively behave as evolutionary units. This major transition was theoretically facilitated by social control mechanisms associated with moral systems which "suppress[ed] fitness differences within groups and made it possible for between-group selection to become an important evolutionary force" (Wilson and Wilson, 2007, page 343).

There are several ways in which strategically rational players may have been instrumental in helping humans to undergo such a major transition. For example, sapient types' imitation of genotypic cooperators represents a form of the reproductive leveling within groups that is central to allowing between-group forces to become operative. And the non-linear phenotype-genotype relationship which results as strategic types imitate genotypically cooperative

[11] There is some debate over the evolutionary stability. See, e.g., Dreber et al. (2008) and Gächter et al. (2008).

types undermines much of the force of the argument against group selection (Williams, 1966) and can strongly enhance group-selection pressures (Wilson, 2004). Furthermore, strategic rationality may have been helpful in developing moral systems or institutions for providing "top-down" rewards (as in Cuesta et al., 2008).

The second key idea in this paper is the identification and distillation of an heretofore under-appreciated *symbiosis* between self-interest and reciprocal altruism. The literal "prediction" of Section 3's model—that human societies will consist of many purely self-interested individuals induced to cooperate by the presence of a small proportion of intrinsically cooperative individuals—is not correct; most humans clearly have both cooperative and strategically self-interested proclivities. But, the model provides a clear illustration of how these two proclivities can be mutually reinforcing.

The synergy between strategic self-interest and innate cooperativeness indicates the potential utility of reconciling the different modeling conventions used by economists and evolutionary social scientists. Economists typically assume the individuals in their models to be rational and ruthlessly self-interested cognitive supermen (with occasional, if misguided, appeals to evolutionary metaphors to "justify" this assumption). In contrast, evolutionary social scientists have typically employed evolutionary game theory to show how evolutionary forces affect biological automata *sans* reasoning skills. The observation that large groups of unrelated humans frequently cooperate in real-world public goods settings poses a puzzle for economists and evolutionary theorists alike: in economists' models, rational self-interest is predicted to undermine cooperation. Evolutionary social scientists have found that evolutionary models with automata have been unsupportive of cooperative behavior in large unrelated groups. The results herein suggest that synthesizing these two approaches with agents with intermediate cognitive abilities can potentially help to resolve both sides of the puzzle.

References

[1] José. A. Adell and Pedro Jodrá. The median of the poisson distribution. *Metrika*, 61:337–346, 2005.

[2] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[3] Theodore Bergstrom. The algebra of assortative encounters and the evolution of cooperation. *International Game Theory Review*, 5:221–228, 2003.

[4] Helen Bernhard, Urs Fischbacher, and Ernst Fehr. Parochial altruism in humans. *Nature*, 442:912–915, 2006.

[5] Samuel Bowles. Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314:1569–1572, 2006.

[6] Robert Boyd. The puzzle of human sociality. *Science*, 314:1555–1556, 2006.

[7] Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100:3531–3535, 2003.

[8] Robert Boyd and Joseph Henrich. Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208:79–89, 2001.

[9] Robert Boyd and Peter J. Richerson. The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, 132:337–356, 1988.

[10] Robert Boyd and Peter J. Richerson. Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology*, 215:287–296, 2002.

[11] Daniel Cohen and Ilan Eshel. On the founder effect and the evolution of altruistic traits. *Theoretical Population Biology*, 10:276–302, 1976.

[12] José A. Cuesta, Raúl Jiménez, Haydeé Lugo, and Angel Sánchez. The shared reward dilemma. *Journal of Theoretical Biology*, 251:253–263, 2008.

[13] Micahel Doebeli and Christoph Hauert. Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. *Ecology Letters*, 8:748–766, 2005.

[14] Anna Dreber, David Rand, Drew Fudenberg, and Martin A. Nowak. Winers don't punish. *Nature*, 452:348–351, 2008.

[15] Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, 425:785–791, 2003.

[16] Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415:137–140, 2002.

[17] Jeffrey Fletcher and Martin Zwick. Strong altruism can evolve in randomly formed groups. *Journal of Theoretical Biology*, 228:303–313, 2004.

[18] Simon Gächter, Elke Renner, and Martin Sefton. The long-run benefits of punishment. *Science*, 322:1510, 2008.

[19] Herbert Gintis. Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206:169–179, 2000.

[20] Phillipe Grégoire and Arthur Robson. Imitation, group selection and cooperation. *International Game Theory Review*, 5:229–247, 2003.

[21] Patrick Grim. The greater generosity of the spatialized Prisoner's Dilemma. *Journal of Theoretical Biology*, 173:353–359, 1995.

[22] William Donald Hamilton. The genetical evolution of social behavior I. *Journal of Theoretical Biology*, 8:1–16, 1964.

[23] William Donald Hamilton. The genetical evolution of social behavior II. *Journal of Theoretical Biology*, 8:17–52, 1964.

[24] Dominic Johnson, Michael Price, and Masanori Takezawa. Renaissance of the Individual: Reciprocity, Positive Assortmentment, and the Puzzle of Human Cooperation, in In C. Crawford and D. Krebs, Eds., *Foundations of Evolutionary Psychology.* Lawrence Erlbaum, New York, 2008.

[25] Yong-Gwan Kim and Joel Sobel. An evolutionary approach to pre-play communication. *Econometrica*, 63:1181–1193, 1995.

[26] David Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[27] Philipp Langer, Martin Nowak, and Christoph Hauert. Spatial invasion of cooperation. *Journal of Theoretical Biology*, 250:634–641, 2008.

[28] Kristian Lindgren. Evolutionary Phenomenon in Simple Dynamics, in C.G. Langton et al. (eds.) *Artificial Life II.* Addison-Wesley, Reading, MA, 1991.

[29] Eric Maskin and Drew Fudenberg. Evolution and cooperation in noisy repeated games. *American Economic Review*, 80:274–279, 1990.

[30] John Maynard Smith. Group selection and kin selection. *Nature*, 201:1145–1147, 1964.

[31] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314:1560–1563, 2006.

[32] Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, 1992.

[33] Martin A. Nowak and Karl Sigmund. Evolutionary dynamics of biological games. *Science*, 303:793–799, 2004.

[34] George R. Price. Selection and covariance. *Nature*, 227:520–521, 1970.

[35] Bettina Rockenbach and Manfred Milinski. The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444:718–723, 2006.

[36] Reinhard Selten. Spieltheoretische behandlung eines oligopolmodells mit nachfragetragheit. *Zeitschrift für Gesamte Staatswissenschaft*, 121:301–324, 1965.

[37] Peter Taylor and Leo Jonker. Selection and covariance. *Mathematical Biosciences*, 40:145–156, 1978.

[38] Robert L. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, 1971.

[39] George Christopher Williams. *Adaption and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton University Press, Princeton, NJ, 1966.

[40] David Sloan Wilson. What's wrong with absolute individual fitness? *Trends in ecology and evolution*, 19:245–248, 2004.

[41] David Sloan Wilson and Elliot Sober. Re-introducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences*, 17:585–654, 1994.

[42] David Sloan Wilson, Mark Van Vugt, and Rick O'Gorman. Multilevel selection theory and major evolutionary transitions: implications for psychological science. *Current directions in psychological science*, In press, 2008.

[43] David Sloan Wilson and Edward Osborne Wilson. Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology*, 82:327–348, 2007.

[44] Sewall Green Wright. Systems of mating iii: Assortative mating based on somatic resemblance. *Genetics*, 6:144–161, 1921.

6 Appendix

PROOF. [Proof of Theorem 1] First note that payoffs are strictly positive. Define global average payoffs

$$\bar{u}_t \equiv (1/n)\sum_{j=0}^{n}\left(ju_T(j) + (n-j)u_U(j)\right) f(n,j,p_t)$$

and note that $(n-j)u_U(j) = (n-j)(u_T(j)+1)$. Then

$$\begin{aligned}(p_{t+1}-p_t)\,n\bar{u}_t =& \textstyle\sum_{j=0}^{n} j u_T(j) f(n,j,p_t) \\ & -p_t \textstyle\sum_{j=0}^{n} \left(j u_T(j) + (n-j)(u_T(j)+1)\right) f(n,j,p_t) \\ =& \textstyle\sum_{j=0}^{n}(j-p_t n) u_T(j) f(n,j,p_t) - p_t \textstyle\sum_{j=0}^{n}(n-j) f(n,j,p_t) \\ =& \textstyle\sum_{j=0}^{n} \left((j-p_t n)\left(\beta j + M - 1\right) f(n,j,p_t)\right) + \\ & (n-p_t n)\left(\beta M n - (\beta n + M - 1)\right) f(n,n,p_t) - n p_t(1-p_t)\end{aligned} \tag{10}$$

The variance and expected value of j are $\sum_{j=0}^{n}\left((j-p_t n) j f(n,j,p_t)\right) = np_t(1-p_t)$ and $\sum_{j=0}^{n}\left(j f(n,j,p_t)\right) = np_t$, respectively, and $f(n,n,p_t) = p_t^n$. We can therefore re-write Equation (10) as

$$(p_{t+1}-p_t)\,n\bar{u}_t = np_t(1-p_t)\left((\beta n - 1)(M-1)p_t^{n-1} - (1-\beta)\right). \tag{11}$$

Denote the right-hand-side of Equation (11) by $\Delta(p_t)$. Note that $\Delta(p_t) \geq 0 \Leftrightarrow p_{t+1} > p_t$. $\Delta(p_t)$ is a continuous function of p_t with three real zeros at $p = 0$, $p = 1$ and $p = p^*$. Our assumption that $n > (2-\beta)/\beta$ ensures $p^* \in (0,1)$. Also, $\Delta'(0) = -n(1-\beta) < 0$; similarly, $\Delta'(1) < 0$ and $\Delta'(p^*) > 0$. Hence, $p_{t+1} < p_t$ $\forall p \in (0,p^*)$ and $p_{t+1} > p_t$ $\forall p \in (p^*,1)$, completing the proof.

PROOF. [Proof of Theorem 3] As in the proof of Theorem 1, define global average payoffs $\bar{u}_t$ and note that:

$$\begin{aligned}(p_{t+1}-p_t)\, n\bar{u}_t =& \textstyle\sum_{j=0}^{n}(j-p_t n)u_T(j)f(n,j,p_t) - p_t \sum_{j=0}^{n}(n-j)f(n,j,p_t) \\ =& \textstyle\sum_{j=0}^{n}\left((j-p_t n)\left(\beta j + M - 1\right) f(n,j,p_t)\right) + \\ & \textstyle\sum_{j=j^*}^{n}(j-p_t n)(M-1)(\beta n - 1)f(n,j,p_t) - np_t(1-p_t) \\ =& (M-1)(\beta n-1)p_t(1-p_t)\left(-\tfrac{dF(n,j^*-1,p)}{dp}\right) - (1-\beta)np_t(1-p_t),\end{aligned} \tag{12}$$

where the last step uses the following two observations:

(1) $\sum_{j=0}^{n} f(n,j,p_t)j(j-p_t n)$ is the variance, $np_t(1-p_t)$, of the binomial distribution, .

(2) $\frac{df(n,j,p)}{dp} = f(n,j,p)\frac{(j-np)}{p(1-p)}$ (as is easily verified by direct computation).

This directly confirms the (obvious) fact that $\underline{p}=0$ and $\overline{p}=1$ are fixed points.

When $p_t \in (0,1)$, the sign of $p_{t+1}-p_t$ is equal to the sign of

$$-\frac{1-\beta}{(M-1)(\beta-\frac{1}{n})} + \left[-\frac{dF(N,j^*-1,p_t)}{dp}\right]. \tag{13}$$

The first term is strictly negative, and independent of p_t. The second term is strictly positive. The stability properties will follow by establishing that $-\frac{dF(n,j^*-1,p)}{dp} \to 0$ as $p \to 0$ or $p \to 1$.

Stability properties of $\underline{p}$ and $\overline{p}$: For $p_t \approx 1$:

$$F(n,j^*-1,p) \approx \frac{n!}{(n-j^*+1)!,(j^*-1)!}(1-p)^{n-j^*+1}$$

(plus higher order terms in $(1-p)$; intuitively, the tails of the binomial distribution fall off fast, so the bulk of the mass to the left of j^* is concentrated at j^*-1.) Since $n - j^* + 1 \geq 2$ (which follows from $\beta n > 2 - \beta > 1$), $\frac{dF(n,j^*-1,p_t)}{dp} \to 0$ as $p \to 1$. Similarly, $\frac{dF(n,j^*-1,p_t)}{dp} \to 0$ as $p_t \to 0$. We conclude that $p_{t+1} - p_t$ is negative as $p_t \to 0$ or $p_t \to 1$, so that $\underline{p}$ and $\overline{p}$ are stable and unstable, respectively.

Existence of (monotone) limits (Properties (1) and (2)): It is straightforward to establish from Equations (5) and (6) (e.g., with some tedious algebra) that

$$p \geq q \Rightarrow \tilde{D}_n p \geq \tilde{D}_n q. \tag{14}$$

Taking any p_0 with $\tilde{D}_n(p_0) = p_1 \geq p_0$, this ensures that $p_t \equiv \tilde{D}_n^t p_0$ is a non-decreasing sequence, whereby $p^\infty \equiv \lim_{t\to\infty} \tilde{D}_n^t(p)$ exists. A similar argument applies if $\tilde{D}_n(p_0) < p_0$.

Convergence to limits greater than $\frac{j^*}{n}$ (Property (3)). For sufficiently large N, we will show that $\tilde{D}_n(j^*/n) > \frac{j^*}{n}$ for all $n \geq N$; Property (3) will then follow directly from the monotonicity of the dynamics.

Explicitly computing:

$$\begin{aligned} -\frac{1}{N}\frac{dF(N, j^*-1, \frac{j^*}{N})}{dp} &= \frac{N}{j^*(N-j^*)}\sum_{j=0}^{j^*-1}\binom{N}{j}\left(\frac{j^*}{N}\right)^j\left(1-\frac{j^*}{N}\right)^{N-j}(j^*-j) \\ &\geq \frac{N}{j^*(N-j^*)}\left(1-\frac{j^*}{N}\right)^N j^*. \end{aligned}$$

Hence, $\lim_{N\to\infty} -\frac{1}{N}\frac{dF(N,j^*-1,\frac{j^*}{N})}{dp} \geq e^{-j^*}$, and $\lim_{N\to\infty} -\frac{dF(N,j^*-1,\frac{j^*}{N})}{dp} = \infty$. Using Equation (13), we can therefore find a sufficiently large N so that $\tilde{D}_n\left(\frac{j^*}{n}\right) > \frac{j^*}{n}$

whenever $n \geq N$, completing the proof.

PROOF. [Proof of Corollary 4] From Theorem 3, there exists $\bar{N}$ such that whenever $n > \bar{N}$ there exists an interior fixed point to the right of $\frac{j^*}{N}$. Fixing p_0, take $\hat{N} = \max\{\bar{N}, \frac{j^*}{p_0}\}$, and any $n > \hat{N}$. Then $p_0 \geq \frac{j^*}{n}$, and, from Theorem 3, $\tilde{D}_n^t p_0 > \frac{j^*}{n}$ and hence

$$1 - F(n, j^* - 1, \tilde{D}_n^t p_0) > 1 - F(n, j^* - 1, \frac{j^*}{n})$$

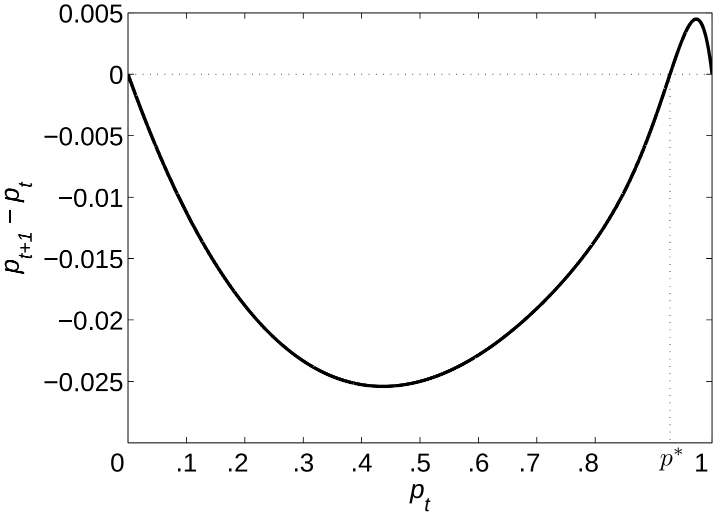
for all t. The fraction of cooperative groups is thus greater than the probability of at least j^* successes out of n tries with the binomial distribution with $p = \frac{j^*}{n}$. As $n \to \infty$, this binomial distribution converges to the Poisson distribution with expected value j^*. Adell and Jodrá (2005) show that $G_{j^*}(j^* - 1) < 0.5$, where $G_{j^*}(\cdot)$ is the cumulative density function for the Poisson distribution with expected value j^*. We conclude:

$$\lim_{n \to \infty} \left(1 - F(n, j^* - 1, \frac{j^*}{n})\right) = 1 - G_{j^*}(j^* - 1) > 0.5,$$

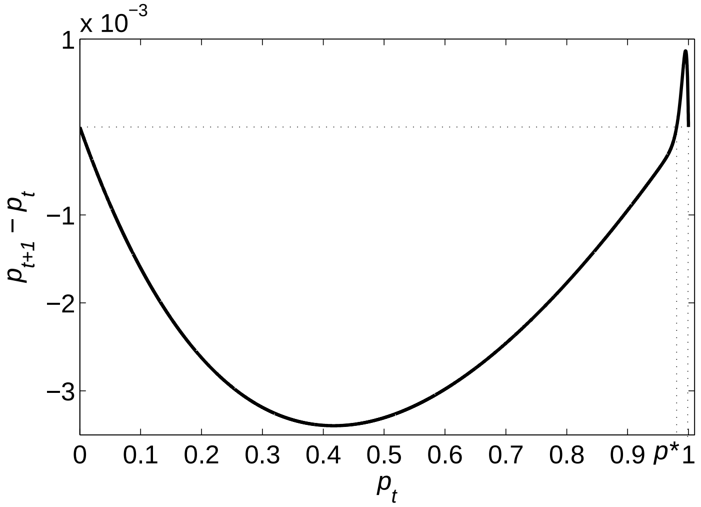
so there exists $N \geq \hat{N}$ such that $n > N$ ensures

$$1 - F(n, j^* - 1, \tilde{D}_n^t p_0) > 1 - F(n, j^* - 1, \frac{j^*}{n}) > 0.5.$$

Captions

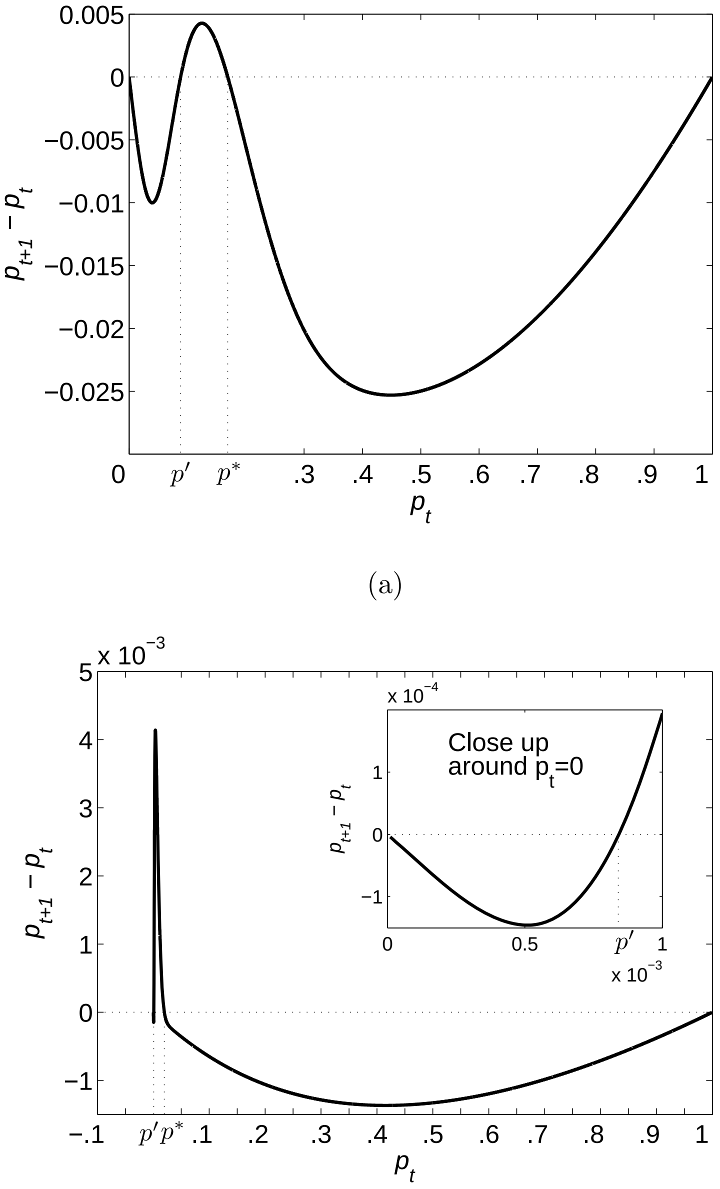


(a)



(b)

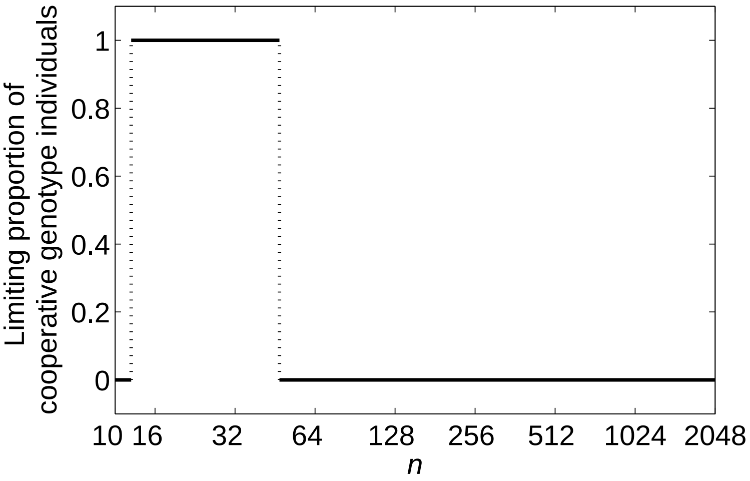
Fig. 1. Dynamics in the baseline model with $\beta = .2$, $M = 2$ for group sizes (a) $n = 25$ and (b) $n = 200$. To the left (right) of p^*, $p_{t+1} < p_t$ ($p_{t+1} > p_t$), and intrinsically cooperative T-types (uncooperative U-types) die off over time. Increasing group size move the cutoff p^* towards 1.



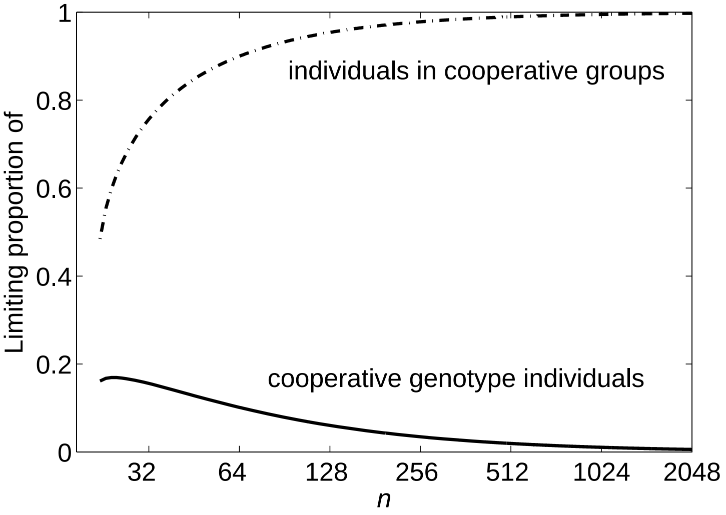
(a)

(b)

Fig. 2. Dynamics in the modified model with $\beta = .2$, $M = 2$ for group sizes (a) $n = 25$ and (b) $n = 500$. The points p^* and $\underline{p} \equiv 0$ are stable steady states with basins of attraction $(p', 1)$ and $(0, p')$, respectively, where p' is an unstable steady state. Increasing group size moves p^* towards 0, and increasing p^*'s basin of attraction $(p', 1)$.



(a)



(b)

Fig. 3. The limiting fraction of genotypically cooperative individuals (i.e., $\lim_{t\to\infty} p_t$) and cooperative groups ($\lim_{t\to\infty} 1 - F(n, j^* - 1, p_t)$) when $p_0 = .95$ for various group sizes in the baseline and modified dynamics. Note that in the baseline dynamics, the limiting fraction of cooperative groups is identical to the limiting fraction of genotypically cooperative individuals.