# Evolution of group reciprocity

*David Hugh-Jones, Moti Michaeli, Ro'i Zultan*

*31/07/2019*

## Introduction

## Theory

### A single generation

There are $G$ groups. Each group has $N$ members. The set of groups is $\mathcal{G} = \{1...G\}$.

[XXX may need group sizes to vary due to evolution, in which case N is the group size above which the group splits/establishes new groups]

A single generation consists of $T$ periods. In every period, each individual "acts" against one randomly selected member (the "target") from *each* group. That is, let $P$ be the set of permutations $(p_1, ..., p_N)$ of $(1, ..., N)$ such that $p_i \neq i$ for all $i$. For each pair of groups $i, j$, a permutation $p$ is drawn from the uniform distribution on $P$. Player $m$ of group $i$ acts against player $p_m$ of group $j$. Thus each player acts against one member of each group, and is the target of one member from each group.

The acting player unilaterally chooses whether to "help" or "harm" the target. Helping costs 1 to the player and benefits the target by $b > 1$. Harming has zero cost and zero benefit. (Alternatively and equivalently, helping has zero cost and benefit, while harming gives 1 to the player and costs the target $b$.)

There are the following types of players:

- Selfish ($s$-) types always harm.
- Generalized reciprocity ($r$-) types harm if and only if they were harmed in the previous period.
- Group reciprocity ($g$-) types harm a target from group X, if and only if the last time somebody from group X acted against them, that person harmed them.

Let $\pi^i_\sigma$ be the proportion of individuals of type $\sigma \in \{s, r, g\}$ in group $i$. Let $\pi_\sigma$ be the overall proportion of type $\sigma$ in the population.

Write $h^t_{ij}$ for the probability that an individual from group $i$ helps an individual from group $j$ in period $t$.

We can calculate this as follows:

$$h^{t+1}_{ij} = \pi^i_s 0 + \pi^i_r \frac{1}{G} \sum_{k \in \mathcal{G}} h^t_{ki} + \pi^i_g h^t_{ji} \tag{1}$$

[XXX could we write the above in matrix form?]

Here, the first term reflects the fact that selfish types in group $i$ never help. The second term gives the probability of a r-type helping. This is given by the overall probability that the r-type was helped by any one actor from the $G$ individuals (one from each group) who targeted her in the previous period. The third term gives the probability that a group reciprocal g-type was previously helped by an actor from group $j$. Setting $h^0_{ij} = 1$ for all $i, j$ ensures that the sum is correct in the first period.

Total payoffs are averaged over all periods and normalized by the number of groups $G$. When $T \to \infty$ the payoffs will be determined by the steady state such that $h^{t+1}_{ij} = h^t_{ij} \equiv h_{ij}$ for all $i, j \in \mathcal{G}$.

Denote by $X \subset \mathcal{G}$ the set of groups $i$ which contain *only* group-reciprocators: $\pi_g^i = 1$.

*Theorem.* There is a unique steady state. If $\pi_s > 0$ it is as follows:

- For all $i, j \in X$, $h_{ij} = 1$.
- For all $i \notin X$ and all $j \in \mathcal{G}$, $h_{ij} = h_{ji} = 0$.

If $\pi_s = 0$, the steady state is $h_{ij} = 1$ for all groups $i, j$.

*Proof.* To prove these values are steady states, plug them into the fixed point equations:

$$h_{ij} = \pi_s^i 0 + \pi_r^i \frac{1}{G} \sum_{k \in \mathcal{G}} h_{ki} + \pi_g^i h_{ji} \tag{2}$$

To prove they are the unique steady states: when $\pi_s = 0$, the second and third terms in (1) always sum to 1, by induction from the first period. (If nobody is selfish, then all players in all groups start by helping and never have a reason to stop.)

When $\pi_s > 0$, first note that if $i \in X$, $h_{ij} = h_{ji}$; furthermore, if $i, j \in X$, $h_{ij} = h_{ji} = 1$ again by using the initial condition.

Pick $i, j$ such that

$$h_{ij} = \bar{h} \equiv max_{k,l;k \notin X} h_{kl}. \tag{3}$$

For a contradiction, suppose that $h_{ij} > 0$.

Fix an arbitrary group $k$. If $k \in X$ then $h_{ki} = h_{ik}$ as just noted and $h_{ik} \leq h_{ij}$ by (3); if $k \notin X$ then $h_{ki} \leq h_{ij}$ again by (3). In either case, then, $h_{ki} \leq h_{ij}$, for any $k$. Since (2) is a weighted sum of $h_{ki}$'s and 0, this immediately implies:

1. $\pi_s^i = 0$;
2. $h_{ki} = h_{ij} = \bar{h}$ for all groups $k$.

Now apply the above argument to $h_{ki}$ for any $k \notin X$. This shows $\pi_s^k = 0$. But for $k \in X$, $\pi_s^k = 0$ by definition. Thus $\pi_s = 0$. Contradiction. This shows $h_{ik} = 0$ for all $i \notin X$ and all $k$. Finally, if $i \in X$ and $k \notin X$, $h_{ik} = h_{ki}$ as noted above, and this is 0. QED.

[XXX: there might be a simpler "contraction mapping" argument to immediately prove the steady states are unique.]

## Payoffs

If $T$ approaches infinity, the payoffs from the generation can be approximated by the steady state payoffs, since play is arbitrarily close to the steady state for an arbitrarily large number of periods.

The steady state payoffs for any member of group $i \notin X$ are 0 since these groups never help and are never helped. If I am a member of $i \in X$, my payoffs are $\frac{|X|}{G}(b-1)$; each period, a member of every group in $X$ helps me, and I help a member of every such group.

## Evolution of strategies

At the end of a generation, fitness is calculated according to payoffs and there is a selection process. . . .

## Extensions/ideas

- Institutional evolution - i.e. the "types" exist at group level only, avoiding any within-group heterogeneity.
- Intra-group public goods
- Direct reciprocity types
- Doing the maths for finite $T$
- (Re)doing the simulations for finite $T$
- The "blood feud" institution in e.g. Boehm suggests there is a deterrent effect - if I plan to harm another clan, my cousins may dissuade me. Selfish types could be (slightly) strategic, e.g. only harming if there is not too much "comeback" in the following period.

## Simulations

## Conclusion