Punishment and the Potency of Group Selection

The final publication is available at http://link.springer.com/article/10.1007/s00191-014-0375-3

> It is an important fact of human life that individuals can often greatly increase (reward) or decrease (punish) the fitness of others at trivial costs to themselves...Secondary behaviours evolve *more easily* by group selection than primary behaviours because they are less strongly opposed by within-group selection, but they still evolve by group selection. The package of primary and secondary behaviours therefore remains a group-level adaptation.
>
> (Sober & Wilson, 1999)

> The indirect evolutionary approach is based on the assumption that players behave rationally for given preferences but that their preferences change through an evolutionary process...While preferences might be inherited literally in a genetical sense, one could also think of it in terms of social evolution since preferences and value judgements of children are shaped by taking parents or peers as role models.
>
> (Huck & Oechssler, 1999)

1.1 *Overview*

It is a well-established result in evolutionary theory that altruism can in principle be sustained by a process of group selection if a population is split into groups whose members interact disproportionately with one another, provided that there is migration between groups. The level of altruism which can be sustained depends upon the relative strength of the evolutionary forces benefiting the more selfish individuals at the expense of altruists within groups, and that favouring the more altruistic groups over the less altruistic ones. There is an evolutionary "tug-of-war" between individual-level selection and group-level selection (Sober & Wilson, 1999).

This paper embeds a simple model of a punishment system within an indirect cultural evolution framework. The result is that the availability of punishment as a device for social control is shown to drastically reduce the potency of the group selection mechanism, and thus the average equilibrium level of altruism. A normative analysis of the outcome shows that the use of the punishment system can sometimes increase social welfare at the evolutionary equilibrium, by inducing selfish individuals to behave better. However, by weakening the group selection mechanism, it can also under some conditions cause a reduction in social welfare at the equilibrium, by causing less altruism to evolve.

The idea of group selection originates with Darwin, but the contemporary formulation was developed in the twentieth century literature on evolutionary biology, most famously in the work of W. D. Hamilton (Hamilton, 1963) (Hamilton, 1972). The mathematical framework was originally devised by Price (Price, 1970). Still controversial among some biologists (but more widely accepted as a useful practical theory in social science fields), the multilevel selection paradigm has recently been popularised within and beyond the biological field by Sober and Wilson. They have provided a survey article (Sober & Wilson, 1994) and a book-length treatment of the subject (Sober & Wilson, 1999).

Group selection has also become a popular framework in theoretical anthropology, from which the fruitful suggestion that we may see cultural as well as genetic characteristics as evolving through natural selection has been developed and employed (Boyd & Richerson, 1982) (Soltis et al., 1995) (Blackmore, 1999). This idea has a pedigree going back to Darwin in biology, but arguably he was heavily influenced (Hirshleifer, 1977) (Hayek, 1988) by the application of the same principle to social institutions by the philosophers of the Scottish Enlightenment, most famously Adam Smith (Smith, 1976). Economists have also made important contributions to the modern theory of group selection, particularly in clarifying issues regarding the mathematical analysis of the different types of group structure that can enable this phenomenon to arise (Bergstrom, 2002) (Cooper & Wallace, 2004).

There is also an existing literature on the role that punishment, such as in the form of informal sanctions or a legal system, can play as an "altruism amplification device" that allows selfish individuals to be induced to behave more like altruistic ones. It has been shown that, because punishing others is often "cheaper" in terms of cost to onesself than benefiting them, the emergence of the ability to carry out altruistic punishment (which Sober and Wilson refer to as secondary behaviours) can explain how the evolution of primary altruism is made possible in a much wider variety of cases. This hypothesis fits

the empirically-observed phenomenon that the ability to punish transgressors in simple experimental games such as the public goods game results in more co-operation being sustained (usually in models where the standard Nash equilibrium with self-interested individuals leads to a complete break-down of co-operation) (Fehr & Gachter, 2000b) (Fehr & Gachter, 2000a) (Fehr & Gachter, 2002a) (Fehr & Fischbacher, 2003) (Fehr & Gachter, 2002b). There are two dimensions to this impact. Firstly, altruistic punishment improves "static" outcomes by making selfish individuals behave better, because they are afraid of being punished. Secondly, the evolution of altruistic punishment can also make it easier for altruism to evolve as a primary behaviour, by reducing the gain in fitness by selfish individuals relative to the altruists in the group (Sober & Wilson, 1999) (Boyd et al., 2003).

This paper aims to make a contribution to the theoretical understanding of the connection between group selection and punishment by applying a third conceptual strand; that of indirect evolution. Most models of the cultural evolution of altruism model cultural norms in a "mechanical" way in the sense that individuals blindly carry out their "programmed" behaviour, whereas economic theory seeks to explain phenomena from a wide variety of cultural scenarios as caused by the same underlying human rationality. The alternative is to assume that it is the weightings that individuals place on the felicity of others that form the evolving phenotype, rather than specific altruistic behaviours directly. In other words, *preferences* evolve but behaviour within the games being played is rational and forward looking, and therefore modelled in the standard manner in which game theory is applied in economic theory.

The indirect evolution approach was first proposed by Güth and Yaari (Guth & Yaari, 1992) following a suggestion originally made by Becker (Becker, 1976). It has been profitably applied to explaining the evolution of preferences for fairness in the ultimatum game (Huck & Oechssler, 1999), in which context it has been shown that "vengeful" individuals who gain utility from reducing the payoffs of others even at cost to themselves can survive and spread in an evolutionary context because, provided the damage they inflict by punishing is high enough relative to the loss to the punisher, they will still have a relatively higher fitness than the "non-vengeful" types.

The application of the indirect evolution methodology to modelling the relationship between punishment and group selection enables an original contribution to be made to an already vast literature. In the standard direct evolution group selection models, there is no distinction between "altruistic preferences" and "altruistic behaviour"; both are modelled as a simple "programmed"

phenotype. The presence of altruistic punishers in the population can only lead to more altruistic behaviour in the evolutionary equilibrium by causing the evolved proportion of selfish types to reduce, by weakening the fitness differential between selfish and altruistic individuals (see section 1.2). By contrast, an indirect evolution approach allows an analysis of punishment mechanisms that can alter the static equilibrium in the game being played at each stage of the evolutionary process. Since the use of punishment improves the static outcome for selfish phenotypes by making them behave better towards one another, it can, paradoxically, potentially result in more selfish individuals evolving.

The sequential punishment model which forms the workhorse model in this paper has the property that *only* the outcome for the selfish phenotypes is improved, because the altruistic phenotypes are unwilling to carry out punishment. Thus the result is unambiguous that fewer altruists are able to survive. The normative consequences are, however, ambiguous, because it may be that despite having more selfish individuals, the gain in static efficiency for any given proportion of selfish individuals outweighs this. The balance of normative effects can also, however, go in the opposite direction. Although the result that the evolution of altruism is unambiguously weakened is a strong one, and dependent on the specific form of model used, the phenomenon described is arguably quite general.

The analysis of the sequential punishment model uses standard conceptual tools from economics such as social welfare, subgame-perfect Nash equilibria and utility functions for individuals which are weighted sums of the felicities of other individuals (a common way to model altruism). This enables normative analysis to be carried out in the usual way. The unconventional aspect of the model is the fact that the level of altruism exhibited by the players (the weighting they place upon the consumption of others in their utility function) is, instead of being exogenously determined, endogenously evolving. The key result that punishment reduces the potency of group selection is demonstrated analytically for any population structure, but specific simulations are also carried out for illustrative purposes.

Evolutionary game theory provides the theoretical tools we can use to analyse the selection pressures that will lead to the pattern of behavioural phenotypes at an evolutionarily stable equilibrium. In general, the level of altruism which would evolve would not be expected to be that which is socially optimal, because selection at the individual level creates pressure upon individual phenotypes in the direction of those which are economically non-altruistic. With a population structure split into groups, however, there can be selection pressure at the group-level as well as the individual level. The resulting

phenotype pattern then represents a kind of "compromise" between the two. This is the essential conceptual scheme provided by multilevel selection theory (Sober & Wilson, 1999).

1.2 *The Standard Direct Evolution Model*

There are many ways to set up an evolutionary model where the evolution of a phenotype which engages in altruistic punishment makes it easier for primary altruism to evolve. A common set-up is a two-stage prisoners' dilemma, where individuals have a chance to punish their opponent at cost to themselves if they cheat in the first phase or, in the context of a group public goods game, where individuals have a chance, after having observed the contribution of other individuals, to carry out costly punishment on under-contributors. In general these models have the property that altruistic punishers do equally well against each other as against pure altruists, and that, although, in a static sense, they do worse relative to selfish individuals than pure altruists (because they carry out costly punishment), the long-run dynamic evolutionary impact of their presence is to make it much harder for selfish individuals to survive.

The intuition for this result is that the fitness reduction for the selfish types is the dominant effect on relative fitness. It is therefore possible, in a group with a sufficient number of altruistic punishers (which can be maintained by "genetic drift" through mutations), for the selfish types to be rapidly driven out of the population. It has been generally found that a three-phenotype model (selfish types, altruists, altruistic punishers) enables the population to be dominated by a mixture of altruists and altruistic punishers in a much wider variety of cases than the two-phenotype model (Boyd et al., 2003).

An indirect evolution approach, however, can provide a radically different perspective on this issue, because it can recognise the distinction between altruistic preferences and altruistic behaviour. Punishment is not carried out "blindly" but when the evolved preferences of the punisher make it rational for them to carry out the punishment. This means that there is no longer a simple connection between altruistic preferences and altruistic behaviour. More individuals with altruistic preferences in a population will not necessarily lead to more altruistic behaviour, because altruistic individuals who care about others may not be willing to go through with punishment. On the other hand, more altruistic behaviour may occur without an increase in altruistic preferences, because selfish individuals may be incentivized to behave better by the credible threat of punishment.

The results of this approach support the existing view that pure altruism is unlikely to survive evolutionarily. However, punishment is modelled not as a phenotype but as a potential equilibrium in the game being played in an evolutionary context. This means that punishment, although improving static outcomes, may worsen the dynamic outcome by helping more selfish individuals to evolve. Even if punishment is dynamically beneficial, this approach can also help explain why only imperfect forms of altruism appear to evolve in the real world. (See "The Limits to Altruism - A Survey" (Chapter 1) for a summary of the empirical results in this area.)

1.3 *The Sequential Punishment Model*

This paper uses a simplified two-move three-player version of the sequential punishment model which has been analysed extensively it its infinite-player form in "The Socially Optimal Level of Altruism" (Chapter 2). The analysis of the sequential punishment model there shows that there is a complex relationship between the altruism embodied in individual preferences and the social efficiency of the resulting outcomes. Sometimes these interactions can perversely result in too much altruism making it *harder* to support a socially efficient outcome. More frequently, the use of a self-supporting system of punishment means that, beyond a certain level, greater altruism is not necessary, as a socially efficient outcome can already be supported. These results were driven by the combination of the temptation effect (more altruistic individuals are less tempted to do harm to others), the willingness effect (more altruistic individuals are less willing to inflict punishment), and the severity effect (punishments, such as a fine where the revenue is redistributed, are less severe for more altruistic individuals, because they value the contribution of the revenue to the welfare of others).

In the version of the sequential punishment model analysed here, there are three players. Players 1 and 2 each get a opportunity in sequence to inflict harm upon[1] another individual. If they take the opportunity, they gain a benefit $\hat{\pi}$ (where $0 < \hat{\pi} < 1$) and the individual they harm suffers a felicity loss of 1. Player 1 first chooses whether or not to harm player 2. Player 2, observing player 1's action, can either choose not to inflict harm, to harm player 3, or to harm player 1. Player 2 is assumed to be indifferent as to whom they harm.[2] Player 2's ability to focus harm onto player 1 *if* they inflict harm

[1] Throughout the paper, we will use "harm" to refer to the infliction of a negative externality and "punish" to refer to the specific use of such harm opportunities to construct punishment equilibria.

[2] If individuals are indifferent between inflicting harm and not inflicting harm, they are assumed not to inflict harm.

creates the potential for a simple punishment scheme to be used to support a subgame perfect Nash equilibrium where a selfish player 1 is deterred from inflicting harm.

Imagine there is a large population of individuals, who differ in their level of altruism. This is designated by the coefficient of altruism, θ_i, which is the weighting placed on the felicity of other individuals in individual i's utility function. Since the benefit from inflicting harm always takes the value $\hat{\pi}$, we only need two distinct phenotypes, H and L, which correspond to $\hat{\pi} \leq \theta_H < 1$ and $\theta_L < \hat{\pi}$ respectively.[3] Suppose that the proportion of individuals in the population with phenotype H is q, so that $(1-q)$ have phenotype L.

Each period, individuals are randomly chosen to play the sequential punishment game. Individuals are formed into triplets, where two of the individuals are able to actually make a move whilst a third individual is randomly selected to be player 3. This third individual does not play any role except to act as a passive receptacle for the harm inflicted by player 2 if player 1 co-operates by not inflicting harm. Nature randomly determines, with equal probability, which individuals will receive their harm opportunity first and second. All individuals are assumed to have full knowledge of the coefficient of altruism of the others with whom they interact. We will begin by assuming that the most socially efficient available equilibrium is played, and consider the consequences of dropping it later on.

1.4 *Derivation of Payoff Matrices*

We can think of the sequential punishment game as a sub-game nested within a supergame in which the coefficients of altruism chosen by individuals A and B[4] are a simultaneous move made before the sequential punishment game is played. The choice of the coefficients of altruism by the players then determines the payoffs, and therefore the outcome, of the sequential punishment game nested within.[5] It is, of course, not really appropriate to think of the coefficient of altruism as a strategy chosen, but rather as a phenotype which can be altered via mutations.[6] Also, whereas it is the social utility payoff that determines each player's behaviour in the nested game, it is the felicity payoff that determines

[3] We assume that individuals are always only "partially altruistic".

[4] We refer to the two individuals who are chosen to be players 1 and 2 as A and B before they know who will go first. Player A and player B both have a probability of 0.5 of being in each position.

[5] Additional assumptions about the properties of the equilibrium that will be played in the sequential punishment game are also required to make the outcome determinate. Initially we are assuming that it will be the most socially efficient one, where player 2 harms player 1 only if player 1 harms player 2 (otherwise player 2 either harms player 3, or does not inflict harm).

[6] In the context of cultural selection theory, mutations are not genetic but represent a kind of cultural random drift as behaviours are imperfectly mimicked or new ways of doing things are tried out.

the evolutionary stable equilibrium in the supergame, and the more complex evolutionarily dynamics involved in group selection that we analyse later.

Since there are two possible phenotypes for each individual, there are four possibilities when three individuals meet and interact.[7] Firstly, if both individuals have high altruism[8] then they both behave efficiently by never inflicting harm. Therefore whichever individual goes first, the felicity payoff to each individual is zero. The value of the social welfare function is also zero because no harm is inflicted, and therefore all three individuals get a felicity payoff of 0.[9] This can be seen in the upper payoff matrix in figure 1.1 in which the top left square shows the zero payoffs of individuals 1 and 2, and the resulting zero social welfare in the box in the middle of the square. Similarly, the corresponding square in the lower matrix[10] looks identical, because the payoffs are still zero for each player regardless of who gets to move first.

Suppose instead that player 2 has phenotype L and player 1 has phenotype H. Since there is no future in which they can be punished, player 2 will inefficiently inflict harm. Player 1 will still not inflict harm because he is sufficiently altruistic not to do this in a single-move game anyway. Therefore player 2 will get a felicity payoff of $\hat{\pi}$ and player 1 will get a felicity payoff of 0, because he co-operates and so player 2 follows her default behaviour and harms player 3. Total social welfare is therefore $\hat{\pi} - 1$.

On the other hand, supposing that player 1 has phenotype L and player 2 has phenotype H, player 2 will not inflict harm, and so there is then no credible threat to punish player 1 for inflicting harm, and so player 1 will do so. In this case, player 1 gets a felicity payoff of $\hat{\pi}$ and player 2 gets -1 because she is punished by player 1. Again, social welfare is $\hat{\pi} - 1$.

In the lower matrix, the payoffs for individuals A and B in the bottom left and top right squares are found by averaging the payoffs in the corresponding squares from the first matrix to produce a new symmetric matrix, because players A and B have an equal chance of being player 1 or 2.

[7] The phenotype of the individual selected to be player 3 is unimportant because they do not have any opportunity to act.

[8] Meaning that they both "play" strategy H in the supergame.

[9] The per-period social welfare function sums the felicity of the two individuals who get a harm opportunity along with the felicity of the third individual who acts as a "passive receptacle".

[10] Which shows the felicity payoffs for players A and B, once the chance of being player 1 or 2 has been randomized, and is therefore symmetric.

1's Phenotype / ***2's Phenotype***	**H** $\theta_1 \geq \hat{\pi}$	**L** $\theta_1 < \hat{\pi}$
H $\theta_2 \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\hat{\pi}$ $\boxed{\hat{\pi}-1}$ -1
L $\theta_2 < \hat{\pi}$	0 $\boxed{\hat{\pi}-1}$ $\hat{\pi}$	0 $\boxed{\hat{\pi}-1}$ $\hat{\pi}$

A's Phenotype / ***B's Phenotype***	**H** $\theta_A \geq \hat{\pi}$	**L** $\theta_A < \hat{\pi}$
H $\theta_B \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\underline{\hat{\pi}}$ $\boxed{\hat{\pi}-1}$ $-\frac{1}{2}$
L $\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\boxed{\hat{\pi}-1}$ $\underline{\hat{\pi}}$	$\frac{\hat{\pi}}{2}$ $\boxed{\hat{\pi}-1}$ $\frac{\hat{\pi}}{2}$

Figure 1.1: Sequential-move game

Finally, we have the case where both individuals have phenotype L. Here, player 2 will definitely inflict harm because there is no future. However, this allows a credible threat to be made to player 1 that if he harms socially inefficiently, the harm inflicted by player 2 will be switched from player 3 onto him. If this occurs, player 1 loses social utility of $1 - \theta_1$.[11] However, the gain in social utility he gets by inflicting harm is only $\hat{\pi} - \theta_1$. Player 1 will therefore be effectively deterred from inflicting harm. Player 1's felicity payoff is therefore 0 and player 2's is $\hat{\pi}$. Social welfare will be $\hat{\pi} - 1$. The payoffs for the second matrix are again found by averaging, in order to take into account the equal chance of players A and B being player 1 or 2 in the first matrix.

[11]This is assuming, for simplicity, no discounting. Permitting discounting would be problematic because we would then have to decide whether or not to discount felicity payoffs as well as social utility payoffs. It would also not really add anything insightful to the analysis of a finite-move sequential game.

Here we see an example of the interplay between the willingness effect and the temptation effect. If we examine the social impact of changing person 2 from a low altruism individual to a high altruism individual (with person 1 remaining a low altruism individual), we see that although the temptation effect leads person 2 not to inflict harm when she would have done so before, the willingness effect completely counteracts this by leading person 1 to defect, because he no longer faces the threat of being punished by person 2. The overall impact upon social welfare is therefore neutral.

The best response payoffs in the second matrix are underlined, and the pure strategy Nash equilibrium[12] is for both individuals A and B to have phenotype L. Since each player is always better off in felicity terms by having low altruism, regardless of whether the other individual has high or low altruism, the individual level selection pressure in this simple model leads to a socially inefficient evolutionarily stable equilibrium, in a similar manner to the standard prisoners' dilemma. This comes about because individuals with low altruism receive higher felicity payoffs and therefore reproduce faster than high altruism individuals, thus coming to dominate the population.

Before further analysing the properties of this evolutionary equilibrium, it is instructive to compare it to that of an identical model, except that rather than having a two-move sequential punishment model nested within the supergame, there is instead a game where each individual chooses whether or not to inflict harm in a single-move game simultaneously.[13] So, person A inflicts harm if and only if $\theta_A < \hat{\pi}$ and person B if and only if $\theta_B < \hat{\pi}$. (We continue to assume that person 1 will harm person 2 by default and that person 2 will harm person 3 by default. Thus individuals A and B only take the felicity loss of -1 if they turn out to be person 2, with probability $\frac{1}{2}$.) The payoff matrix for this model is shown in figure 1.2. Although the evolutionarily stable equilibrium is again for all individuals to have phenotype L, the important difference compared to the case where the nested game is sequential is that in the evolutionary equilibrium for this model, both individuals will inflict harm, whereas in the case of the two-move sequential game, although all individuals have low altruism in the evolutionary equilibrium, the individual who has a chance to inflict harm first does not inflict harm, due to the threat of having the harm inflicted by player 2 focused on to him if he defects by inflicting harm. This difference between the two models will turn out to be of crucial importance in determining the nature of their evolutionarily stable equilibria when group selection effects can occur.

[12] Which is also a dominant strategy equilibrium and therefore the unique Nash equilibrium.

[13] When we analyse this model is more detail later on, we will see that in terms of the evolutionary pressures, this model is essentially the same as the standard prisoners' dilemma set-up.

	H $\theta_A \geq \hat{\pi}$	**L** $\theta_A < \hat{\pi}$
H $\theta_B \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\underline{\hat{\pi}}$ $\boxed{\hat{\pi}-1}$ $-\frac{1}{2}$
L $\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\boxed{\hat{\pi}-1}$ $\underline{\hat{\pi}}$	$\underline{\hat{\pi}-\frac{1}{2}}$ $\boxed{2\hat{\pi}-2}$ $\underline{\hat{\pi}-\frac{1}{2}}$

Figure 1.2: Simultaneous-move game

The relevant difference between the sequential-move and simultaneous-move versions of the model can be brought out if we consider the effect on social welfare of a marginal increase in the proportion of the population with high altruism (phenotype H) from the evolutionarily stable equilibrium in a single homogeneous population. The expected value of the social welfare function, $E(W)$, depends upon the proportion of each phenotype in the population. In the case of the nested sequential-move punishment model, in a finite population of size n, this will be given by:

$$E\left(W\right)=\frac{q\left(nq-1\right)0}{n-1}+2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1}+\frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(\hat{\pi}-1\right)}{n-1}$$

In the case of the nested simultaneous-move punishment model, this will be:

$$E\left(W\right)=\frac{q\left(nq-1\right)0}{n-1}+2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1}+\frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(2\,\hat{\pi}-2\right)}{n-1}$$

If we now differentiate these expressions with respect to q, we can find an expression for the gains in social welfare from a marginal increase in the proportion of altruists. For the nested sequential punishment model, we get:

$$\frac{d}{dq}E\left(W\right)\left(q\right)=\frac{\left(1-\hat{\pi}\right)\left(2\,nq-1\right)}{n-1} \tag{1.1}$$

For the simultaneous-move model, we get:

$$\frac{d}{dq}E\left(W\right)\left(q\right)=2\,\left(1-\hat{\pi}\right) \tag{1.2}$$

As $n\longrightarrow\infty$, (1.1) goes to:

$$\frac{d}{dq}E\left(W\right)\left(q\right)=2\,\left(1-\hat{\pi}\right)q \tag{1.3}$$

The diagram below shows social welfare as a function of q for both types of nested model, letting $\hat{\pi} = \frac{1}{2}$ and taking the limit as $n \longrightarrow \infty$. We see that at the evolutionary equilibrium where $q = 0$, the marginal increase in social welfare when q increases is positive for the simultaneous-move model but 0 for the sequential-move model. This is because introducing a small number of high altruism individuals into a population of low altruism individuals means that they are almost certain to interact with low altruism individuals. In the nested sequential move game, however, this means that if the new high altruism individual inflicts harm first, they do not change their behaviour, whereas if they go second, although they do not inflict harm, this causes the low altruism individual to defect and inflict harm, whereas they would not do so if the second individual had low altruism instead of high altruism.

So, altruism is only socially beneficial in the sequential punishment model when altruists encounter each other rather than low altruism individuals. In the simultaneous-move game, by contrast, the presence of even a small number of high altruism individuals is socially beneficial because even if they do interact with a low altruism individual, their behaviour is changed because they now do not inflict harm, and this increases social efficiency, even though the low altruism individual they interact with still defects and inflicts harm.

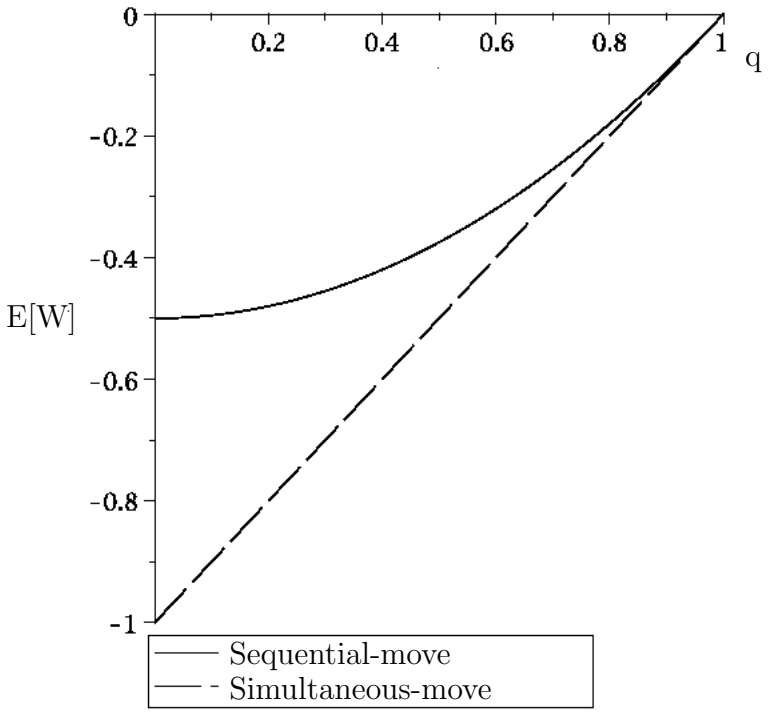

Figure 1.3: Comparison of evolutionary models

What this analysis shows us is that by introducing a change in the way social interactions are modelled so that there are sequential moves enabling conditional punishment, the properties of the evolved equilibrium pattern of individual altruism are altered in the sense that marginal injections of additional altruism into the population are not as socially beneficial. This gives us the intuition for the result that the group selection mechanism is weaker in the sequential-move model.

1.5 *Price Equations*

The conditions under which group-level selection pressures will dominate, and altruism will evolve, can be described using Price equations. Price was an evolutionary biologist and the first person to outline the mathematical conditions required for altruism to evolve by group selection (Price, 1970). It is clear that altruists will always be wiped out in the long run in a single isolated group because the selfish individuals always get better felicity payoffs from their interactions on average, and thus breed more rapidly. However, if altruists are sufficiently concentrated together in sub-groups within the population, whose members interact only (or mainly) with one another, then altruists can get better average payoffs in the population as a whole, and thus outbreed selfish types, because the benefits of altruism will be focused (mainly) upon other altruists. For this to work in the long run, there must be a dispersal mechanism which has the property that it allows altruists to migrate between groups whilst maintaining sufficient inter-group variance of the altruistic phenotype relative to the intra-group variance that altruists are fitter on average than selfish types. The dispersal mechanism determines the manner in which migration of individuals between groups occurs.[14] The Price Equation for a particular model establishes the minimum variance ratio required to enable altruism to survive.

In this section, we will show that the sequential-move game described above leads to a more stringent requirement on the variance ratio achieved by the dispersal mechanism than the simultaneous-move game. This means that altruism will, under general conditions, evolve to a higher degree if the social control mechanism provided by person 2's threat to punish person 1 is removed. Although the presence of this threat prevents person 1 from inflicting harm, and therefore causes a social welfare gain, ceteris paribus, the potential for the use of punishment also weakens the group selection mechanism, and thus the ability to achieve a socially efficient high altruism equilibrium "anarchically". Potentially,

[14]There must always be some migration of altruists in order for altruism to survive, since otherwise they would always be extinguished within each isolated group.

	H	**L**
H	$b-c$ $b-c$	$\underline{b}$ $-c$
L	$-c$ $\underline{b}$	$\underline{0}$ $\underline{0}$

Figure 1.4: Standard prisoners' dilemma

therefore, removing the social control mechanism might improve social welfare by giving a boost to altruism sufficient to cause a net rise in social welfare. After deriving analytic results regarding the Price equation which apply to any dispersal mechanism, we will proceed to show, using some simple specific simulations, that this can indeed be the case.

We will proceed by first showing that the simultaneous-move version of the punishment game has essentially the same Price equation as the standard prisoners' dilemma, which is the classic example of a Price equation in the literature. We will then derive the Price equation for the sequential-move punishment game, and proceed to prove that it always involves a more stringent requirement upon the inter-group variance. Finally, we will illustrate the analysis using a computer simulation of a specific dispersal mechanism.

1.6 *The Standard Prisoners' Dilemma*

The standard model involves a population split into m groups with average size n, so that $n = \frac{1}{m}\sum_{i=1}^{m} n_i$, where n_i is the number of individuals in group i. Individuals in each group only interact with members of their own group, and do so by playing a 2-person prisoners' dilemma game with the payoff matrix shown below (where $b > c$). Now, since the only Nash equilibrium is the dominant strategy equilibrium where both individuals play L (low altruism, equivalent to "defect" in the usual parlance), the only evolutionarily stable equilibrium if all individuals are in a single group is for all individuals to have phenotype L.[15]

[15]If there are random mutations, there may be some type Hs but they are always evolutionarily less fit than the type Ls and so are always in the process of dying out.

When there is more than one group, however, we can show that conditions exist under which the altruistic phenotype can spread through the population. We do this by deriving the change which will occur in the number of high altruism individuals and the total population, thus enabling the derivation of the condition for the proportion of high altruism individuals to increase, assuming that the number of offspring is equal to the felicity payoff. The payoffs for a member of a particular group depend upon the proportion in the group of each type, q_i being the proportion of high altruism types in group i and $q = \frac{1}{mn}\sum_{i=1}^{m} n_i q_i$ being the proportion of altruists in the population.

The payoffs of high and low altruism individuals in group i will be:[16]

$$U_i{}^H = f + \frac{(q_i n_i - 1)\, b}{n_i - 1} - c \tag{1.4}$$

$$U_i{}^L = f + \frac{q_i n_i b}{n_i - 1} \tag{1.5}$$

From the above, we can see that, once interactions have occurred and breeding has taken place, the new proportion of altruists in the overall population after one phase of interactions will be given by the following expression, derived by dividing the new number of high altruism individuals in the population by the new total population:

$$q' = \frac{\sum_{i=1}^{m}\left(\left(f + \frac{(q_i n_i - 1)b}{n_i - 1} - c\right) q_i n_i\right)}{\sum_{i=1}^{m}\left(\left(f + \frac{(q_i n_i - 1)b}{n_i - 1} - c\right) q_i n_i + \left(f + \frac{q_i n_i b}{n_i - 1}\right)(1 - q_i)\, n_i\right)} \tag{1.6}$$

Dividing the numerator and denominator through by n and collecting like terms gives us:

$$q' = \frac{\left(\sum_{i=1}^{m} \frac{q_i n_i f}{n} + \sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2 b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i c}{n}\right)}{\left(\sum_{i=1}^{m} \frac{f n_i}{n} - \sum_{i=1}^{m} \frac{q_i n_i b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i c}{n} + \sum_{i=1}^{m} \frac{q_i {n_i}^2 b}{n(n_i - 1)}\right)} \tag{1.7}$$

The following expressions are used now to simplify the above expression, and also subsequently for the same purpose later in this section:

$$\begin{aligned} \textstyle\sum_{i=1}^{m} \frac{q_i {n_i}^2}{n_i - 1} &= m\,Cov\left(\tfrac{q_i n_i}{n_i - 1}, n_i\right) + \textstyle\sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} n \\ \textstyle\sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2}{n_i - 1} &= m\,Cov\left(\tfrac{q_i n_i}{n_i - 1}, q_i n_i\right) + \textstyle\sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} nq \\ \textstyle\sum_{i=1}^{m} \frac{{n_i}^2}{n_i - 1} &= m\,Cov\left(\tfrac{n_i}{n_i - 1}, n_i\right) + n \textstyle\sum_{i=1}^{m} \frac{n_i}{n_i - 1} \end{aligned} \tag{1.8}$$

$$\begin{aligned} \textstyle\sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} &= m E\left(\tfrac{q_i n_i}{n_i - 1}\right) & \textstyle\sum_{i=1}^{m} q_i n_i &= qnm \\ \textstyle\sum_{i=1}^{m} n_i &= mn & \textstyle\sum_{i=1}^{m} \frac{n_i}{n_i - 1} &= m E\left(\tfrac{n_i}{n_i - 1}\right) \end{aligned} \tag{1.9}$$

[16] U_i^H and U_i^L are the expected felicity payoffs in group i. Note also the introduction of a fixed payoff f. This is the same for both phenotypes and thus has no effect on relative fitness, but is needed to ensure that both types always gain a strictly positive payoff. It will be set to a fixed value in the simulations later on.

In order for the proportion of high altruism individuals to grow in the population, we require that $q' - q > 0$, where $q' - q$ can be derived to be the following:

$$q' - q = \frac{\left(-qbCov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + bCov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) - b\left(1-q\right)E\left(\frac{q_i n_i}{n_i - 1}\right) - qcn\left(1-q\right)\right)}{\left(fn + bCov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + b\left(n-1\right)E\left(\frac{q_i n_i}{n_i - 1}\right) - cqn\right)} \tag{1.10}$$

Provided f is set high enough to ensure a positive payoff for both phenotypes, the denominator of (1.10) will be positive. Therefore the sign of the numerator will determine whether $q' - q > 0$ is positive or negative. It will therefore be the case that $q' - q > 0$ if and only if the following is fulfilled:

$$\frac{c}{b} < \frac{Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right)}{qn\left(1-q\right)} - \frac{Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right)}{n\left(1-q\right)} - \frac{E\left(\frac{q_i n_i}{n_i - 1}\right)}{nq} \tag{1.11}$$

This result can be most easily interpreted in the situation where all groups are of equal size, so that $E\left(\frac{q_i n_i}{n_i - 1}\right) = \frac{qn}{n-1}$, $Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) = \frac{n^2}{n-1}Var(q_i)$ and $Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) = 0$. In this case, expressions (1.10) and (1.11) simplify respectively to give:

$$q' - q = \frac{\left(n\,Var\left(q_i\right) - q(1-q)\right)b - q\left(n-1\right)\left(1-q\right)c}{\left(n-1\right)\left(f + q(b-c)\right)} \tag{1.12}$$

$$\frac{c}{b} < \frac{n\,Var\left(q_i\right)}{\left(n-1\right)q\left(1-q\right)} - \frac{1}{\left(n-1\right)} \tag{1.13}$$

The intuition for this result is that altruism is able to survive if altruists are sufficiently concentrated together that they have a higher average fitness level than the selfish types. Within a particular group, selfish individuals still do better than altruistic individuals, but across the population, altruists are able to do better than selfish individuals because the altruistic groups spread more rapidly. The $Var(q_i)$ part of the above condition is the inter-group variance of the level of altruism. The $q(1-q)$ part is the intra-group variance: the variance of the random variable formed by taking a single individual from the population and assigning a value of 1 if they have phenotype H and 0 if they have phenotype L. As $n \longrightarrow \infty$, (1.13) simplifies even further to give $\frac{c}{b} < \frac{Var(q_i)}{q(1-q)}$; the variance ratio must be greater than the ratio of the cost of co-operating to the benefit bestowed upon the other individual by doing so. The lower the cost relative to the benefit, the easier it is for altruism to evolve, because the individual-level selection pressure in favour of the selfish types is weakened relative to group-level selection.

If we now take the example of the simultaneous-move punishment game, we can see that a high altruism individual *refraining* from inflicting harm and imposing the cost of 1 on the other individual at benefit $\hat{\pi}$ to herself is logically equivalent to bestowing a benefit of 1 upon the other individual at

a cost of $\hat{\pi}$ to herself. The simultaneous-move punishment game therefore has almost the same payoff matrix as the standard prisoners' dilemma, with $b = 1$ and $c = \hat{\pi}$, except that person 2, if she has phenotype L, inflicts harm upon person 3 rather than person 1. We will see when we come to derive the Price equation that this is an insignificant difference. We will also see that the simultaneous-move game can be analysed as an instance of the sequential-move game but with player 2's strategy being not to condition her actions upon those of player 1. In this context, therefore, the standard prisoners' dilemma situation can essentially be viewed as *one* of the possible equilibrium outcomes of the sequential punishment model, and thus as a subcase of this more general model.

1.7 *The Sequential-Move Game*

Before deriving the Price equation for the sequential-move game, we need to further consider the possible subgame-perfect equilibria in this game, and justify why we pay particular attention to certain of these. Player 1's moves are restricted to either inflicting harm upon player 2 or not inflicting harm at all.[17] This means that player 1 has only 2 available strategies. If player 2 has the high altruism phenotype then she only has one credible strategy available, which is not to inflict harm. If player 2 has low altruism then, after the elimination of strictly dominated strategies, she has only 4 possible strategies that could be played in a subgame-perfect Nash equilibrium.[18] These restrictions enable us to fully characterize all of the equilibria of the embedded subgame. The discussion and payoff matrices below describe the subgame-perfect Nash equilibria for the four combinations of phenotype: (H,H), (L,H), (H,L) and (L,L)[19], where L corresponds to $\theta < \hat{\pi}$ and H corresponds to $\theta \geq \hat{\pi}$.

Taking first the two cases when player 2 has high altruism, it is clear that here player 2 will choose not to harm either player 1 or player 3. This is turn means that player 1 will face no future punishment when deciding whether or not to harm player 2, and so will do so if he has low altruism, but not if he has high altruism. Taking instead the two cases where player 2 has low altruism, it is clear that player 2 will choose to inflict harm, but she will be indifferent between inflicting harm upon player 1 and player 3. This makes the strategic possibilities more interesting.

[17] We could justify this by assuming that individuals can only harm those adjacent to them.

[18] This is because player 2 can either harm player 1 or harm player 3 in response to each of player 1's possible moves. She cannot credibly threaten to refrain from inflicting harm, or to harm herself.

[19] (H,L) and (L,H) are distinct because players 1 and 2 do not have symmetric moves or information sets.

	Punish (A)	**Don't Punish (B)**
A → 1 B → 1	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - 1$ $\underline{\hat{\pi} - \theta_2}$
A → 1 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 1	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - 1$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 3	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - \theta_1$ $\underline{\hat{\pi} - \theta_2}$

Figure 1.5: Player 1 has phenotype L and player 2 has phenotype L

The first payoff matrix in figure 1.5 illustrates the case where both individuals have low altruism. The underlined payoffs indicate potential best responses for each player,[20] so that squares with both payoffs underlined represent subgame-perfect Nash equilibria.[21] In three such equilibria, both individuals inflict harm, but there is also one where player 2 (the row player) makes a credible threat to switch harm onto player 1 (the column player) if he chooses to inflict harm, thus resulting in player 1's best response being not to inflict harm. This is a reasonably plausible equilibrium because player 2 can gain by making the threat (if it is credible), but is indifferent as to who she inflicts harm upon and so never incurs a cost from carrying out the threat, thus rendering it credible.

The second payoff matrix, in figure 1.6, illustrates the case where player 1 has phenotype H and player 2 has phenotype L. Provided $\theta_1 > \frac{1+\hat{\pi}}{2}$, all the sub-game perfect Nash equilibria involve player 1

[20] When a player is indifferent between payoffs, all of the equally preferred payoffs are underlined.

[21] Note that although, in general, a best response to a best response is a necessary, but not sufficient condition for a subgame-perfect Nash equilibrium, in this simple game *all* Nash equilibria are subgame-perfect.

	Punish (A)	**Don't Punish (B)**
A → 1 B → 1	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - 1}$ $\underline{\hat{\pi} - \theta_2}$
A → 1 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 1	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - 1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$

Figure 1.6: Player 1 has phenotype H and player 2 has phenotype L, assuming $\theta_1 > \frac{1+\hat{\pi}}{2}$.

not inflicting harm and player 2 inflicting harm. If $\frac{1+\hat{\pi}}{2} \geq \theta_1 \geq \hat{\pi}$ then there can be a Nash equilibrium where player 2 threatens to harm player 1 *unless* he inflicts harm, and so player 1 does indeed inflict harm even though he loses utility from doing so as a standalone act. However, this equilibrium is not very believable as it involves player 2 making a threat which it is not in her interest to make. She would prefer player 1 not to inflict harm, and so it is not intuitively plausible that she would make such a threat.

Although it is not likely that player 2 would make a threat which influences player 1's behaviour in a manner which is against her own interest, it is more plausible that for some reason player 2 might not be able to effectively make the threat to induce player 1 not to inflict harm when he has low altruism. If player 1 does not believe that player 2 will carry out the threat, then, once it comes to enforcing it, player 2 is indifferent as to whether she does in fact do so, since "bygones are bygones". If

the punishment equilibrium breaks down, then the sequential-move game essentially collapses into the simultaneous-move game. This can occur if the subgame-perfect equilibrium that is played involves player 2 playing either the strategy $A \to 1, B \to 1$ or the strategy $A \to 3, B \to 3$. So, in this sense, the sequential-move move model contains the simultaneous-move model as a special case.

1.8 *Deriving the Price Equations*

We can now derive the expected payoffs of high altruism and low altruism types in the sequential-move game. We assume that each individual has a $\frac{1}{3}$ chance of being player 1, 2 or 3 respectively in each interaction. We will first assume that player 2 plays either strategy $A \to 1, B \to 1$ or strategy $A \to 3, B \to 3$. In both cases, we will see that the resultant Price equation is basically the same as the standard prisoners' dilemma. We will then move on to the subgame-perfect equilibrium where strategy $A \to 1, B \to 3$ is played, and where the resultant Price equation places a more stringent condition upon the variance ratio, thus reducing the potency of the group selection mechanism.

Suppose first of all that a selfish player 2 (a player 2 with a low altruism phenotype, L) chooses to play the strategy $A \to 1, B \to 1$, so that she always inflicts harm upon player 1 regardless of whether player 1 inflicts harm upon her or not[22]. A selfish player 1 will therefore definitely choose to harm player 2, because he will be punished anyway and so will optimally wish to take his opportunity to inflict harm for a gain in his social utility. This is a subgame-perfect Nash equilibrium because once player 1 has made his choice, player 2 will be indifferent over whether she inflicts harm upon player 1 or player 3.

1.8.1 Selfish player 2 plays strategy $A \to 1, B \to 1$

We derive the Price equation condition by finding the expected felicity payoff of an altruistic individual with phenotype H and a selfish individual with phenotype L. Altruistic individuals have a $\frac{1}{3}$ chance of being player 1, 2 or 3 respectively. If they are player 1, then if player 2 is selfish (with probability $\frac{(1-q_i)n_i}{n_i-1}$), they will receive a felicity payoff of -1. Otherwise they will receive a felicity payoff of 0. If they are player 2, they also will receive a felicity payoff of -1, if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$), and 0 otherwise. If they are player 3, they will always receive a felicity payoff

[22] As we have already seen, since there is no future an altruistic player 2 with phenotype L will never inflict harm.

of 0, because they are never punished. The expected felicity payoff for an altruist will thus be:

$$U_i{}^H = f - \frac{2}{3}\frac{(1-q_i)\,n_i}{n_i-1} \tag{1.14}$$

If players are selfish, then if they turn out to be player 1, they will definitely choose to harm player 2, who will harm them in turn if they too are selfish. The expected payoff if they are player 1 would therefore be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. If they turn out to be player 2, they will again definitely inflict harm, and player 1 will harm them if they are selfish. Again, the expected payoff would be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. As before, if they turn out to be player 3, their expected payoff is definitely 0. So:

$$U_i{}^L = f + \frac{2}{3}\hat{\pi} - \frac{2}{3}\frac{(1-q_i)\,n_i-1}{n_i-1} \tag{1.15}$$

The new proportion of altruists in the population after one stage of interaction will therefore be:

$$q' = \frac{\sum_{i=1}^{m}\left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right)q_i n_i}{\sum_{i=1}^{m}\left(\left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right)q_i n_i + \left(f + \frac{2}{3}\hat{\pi} - \frac{2}{3}\frac{(1-q_i)n_i-1}{n_i-1}\right)(1-q_i)\,n_i\right)} \tag{1.16}$$

Multiplying out, dividing the numerator and denominator by n and collecting like terms gives us:

$$q' = \frac{\left(\sum_{i=1}^{m}\frac{3}{2}\frac{q_i n_i f}{n} - \sum_{i=1}^{m}\frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m}\frac{{q_i}^2 {n_i}^2}{n(n_i-1)}\right)}{\left(\sum_{i=1}^{m}\frac{3}{2}\frac{f n_i}{n} + \sum_{i=1}^{m}\frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m}\frac{\hat{\pi}\,n_i}{n} - \sum_{i=1}^{m}\frac{\hat{\pi}\,n_i q_i}{n} - \sum_{i=1}^{m}\frac{{n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m}\frac{n_i}{n(n_i-1)} - \sum_{i=1}^{m}\frac{q_i n_i}{n(n_i-1)}\right)} \tag{1.17}$$

We can now apply (1.8) and (1.9) to derive the following expression for the change in the proportion of altruists in the overall population:

$$q' - q = \frac{2\left(-(1+q)\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - (n-q)E\left(\frac{q_i n_i}{n_i-1}\right) + q(n-1)E\left(\frac{n_i}{n_i-1}\right) + q\left(Cov\left(\frac{n_i}{n_i-1}, n_i\right) - \hat{\pi}\,n(1-q)\right)\right)}{\left(3\,fn + 2\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + 2(n-1)\left(E\left(\frac{q_i n_i}{n_i-1}\right) - E\left(\frac{n_i}{n_i-1}\right)\right) - 2\,Cov\left(\frac{n_i}{n_i-1}, n_i\right) + 2\,\hat{\pi}\,n(1-q)\right)} \tag{1.18}$$

Provided f is high enough so that both types always get a positive payoff, the denominator of (1.18) will be positive, and $q' - q > 0$ if and only if the following condition is fulfilled:

$$\hat{\pi} < -\frac{(1+q)\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right)}{qn(1-q)} + \frac{Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right)}{qn(1-q)} - \frac{(n-q)E\left(\frac{q_i n_i}{n_i-1}\right)}{qn(1-q)} + \frac{(n-1)E\left(\frac{n_i}{n_i-1}\right)}{n(1-q)} + \frac{Cov\left(\frac{n_i}{n_i-1}, n_i\right)}{n(1-q)} \tag{1.19}$$

If all groups are of equal size, the conditions (1.18) and (1.19) become, respectively:

$$q' - q = \frac{2\left(n\,Var\left(q_i\right) - q\left(n-1\right)\left(1-q\right)\hat{\pi} - q(1-q)\right)}{\left(n-1\right)\left(3\,f - 2(1-q)(1-\hat{\pi})\right)} \tag{1.20}$$

$$\hat{\pi} < \frac{n\,Var\left(q_i\right)}{q\left(n-1\right)\left(1-q\right)} - \frac{1}{(n-1)} \tag{1.21}$$

If all groups are the same size, the condition on the variance ratio is therefore identical to the prisoners' dilemma, since (1.21) is identical to (1.13) with $\frac{c}{b} = \hat{\pi}$.[23]

[23]There is a slight difference between the two models when groups are of different sizes, due to the differing position of player 3 in different sized groups.

1.8.2 Selfish player 2 plays strategy $A \to 3, B \to 3$

This case will be very similar to the previous one, except that it is player 3 rather than player 1 who always receives the harm inflicted by a selfish player 2. The expected utility payoffs will therefore be the same as above, as will the Price equation.

1.8.3 Selfish player 2 plays strategy $A \to 1, B \to 3$

We will now assume that player 2 always plays strategy $A \to 1, B \to 3$, so that player 1 is always induced not to inflict harm if player 2 is of type L. Our first step will be to derive the expected felicity payoff for individuals with high and low altruism.

Taking first the expected felicity payoff of an altruistic individual, they have a $\frac{1}{3}$ chance of being player 1 in their interaction. In this case, whether or not player 2 is altruistic, the individual will not inflict harm, and so will receive a payoff of 0. If, on the other hand, they turn out to be player 2 (probability $\frac{1}{3}$), they will be punished by player 1 if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1), because they can make no credible threat to punish a selfish player 1 for doing this. The third possibility is that they will be player 3, in which case they will be punished by player 2 if player 2 turns out to be selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1). This is because even if player 1 turns out to be selfish, he will never choose to harm player 2 due to his fear of being punished by having the harm inflicted by player 2 focused onto him. Hence, a selfish player 2 will always harm player 3. So, the expected felicity payoff of an altruistic individual in this model is:

$$U_i{}^H = f - \frac{2}{3}\frac{(1-q_i)\,n_i}{n_i-1} \tag{1.22}$$

Now we take the case of selfish individuals. If they turn out to be player 1, they will choose to harm player 2 if and only if player 2 is altruistic (the unconditional probability of this scenario is $\frac{1}{3}\frac{q_i n_i}{n_i-1}$ and the felicity payoff would be $\hat{\pi}$).[24] If selfish individuals turn out to be player 2 (probability $\frac{1}{3}$), then they will definitely harm either player 1 or player 3, gaining a felicity payoff of $\hat{\pi}$. If they turn out to be player 3, they are in the same situation as they would be if they were altruistic, except that the probability that player 2 is selfish and inflicts harm upon them is now $\frac{(1-q_i)n_i-1}{n_i-1}$. The expected utility payoff of an altruistic individual will therefore be:

$$U_i{}^L = f + \frac{1}{3}\frac{\hat{\pi}\,q_i n_i}{n_i-1} + \frac{1}{3}\hat{\pi} - \frac{1}{3}\frac{(1-q_i)\,n_i-1}{n_i-1} \tag{1.23}$$

[24] The benefit received by inflicting harm inefficiently is equivalent to the cost that must be incurred in order to behave efficiently in the prisoners' dilemma model.

From the above, we can see that, once interactions have occurred and breeding has taken place, the new proportion of high altruism individuals in the population will be given by:

$$q' = \frac{\sum_{i=1}^{m} \left(f - \frac{2}{3} \frac{(1-q_i)n_i}{n_i - 1} \right) q_i n_i}{\sum_{i=1}^{m} \left(\left(f - \frac{2}{3} \frac{(1-q_i)n_i}{n_i - 1} \right) q_i n_i + \left(f + \frac{1}{3} \frac{\hat{\pi}\, n_i q_i}{n_i - 1} + \frac{1}{3}\hat{\pi} - \frac{1}{3} \frac{(1-q_i)n_i - 1}{n_i - 1} \right) (1 - q_i)\, n_i \right)} \tag{1.24}$$

Multiplying out, dividing the numerator and denominator by n and collecting like terms yields:

$$q' = \frac{\left(\sum_{i=1}^{m} \frac{3}{2} \frac{q_i n_i f}{n} - \sum_{i=1}^{m} \frac{q_i {n_i}^2}{n(n_i - 1)} + \sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2}{n(n_i - 1)} \right)}{\frac{1}{2} \left(3 \sum_{i=1}^{m} \frac{f n_i}{n} + \sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2 (1-\hat{\pi})}{n(n_i - 1)} + \sum_{i=1}^{m} \frac{\hat{\pi}\, q_i {n_i}^2}{n(n_i - 1)} + \sum_{i=1}^{m} \frac{\hat{\pi}\, n_i}{n} - \sum_{i=1}^{m} \frac{\hat{\pi}\, n_i q_i}{n} - \sum_{i=1}^{m} \frac{{n_i}^2}{n(n_i - 1)} + \sum_{i=1}^{m} \frac{n_i}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i}{n(n_i - 1)} \right)} \tag{1.25}$$

Expressions (1.8) and (1.9) can now be applied to derive the following expression for the change in the proportion of altruists in the overall population:

$$q' - q = \frac{(2 - q(1-\hat{\pi}))\, Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i \right) - (2 + \hat{\pi}\, q)\, Cov\left(\frac{q_i n_i}{n_i - 1}, n_i \right) - \left(\left((1-q)(q\hat{\pi} + 2) + q^2 \right) n - q \right) E\left(\frac{q_i n_i}{n_i - 1} \right) + (n-1) q E\left(\frac{n_i}{n_i - 1} \right) - q(1-q)\hat{\pi}\, n + q\, Cov\left(\frac{n_i}{n_i - 1}, n_i \right)}{\left(3\, f n + \hat{\pi}\, Cov\left(\frac{q_i n_i}{n_i - 1}, n_i \right) + (1-\hat{\pi})\, Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i \right) - (1 - ((1-q)\hat{\pi} + q)n) E\left(\frac{q_i n_i}{n_i - 1} \right) - (n-1) E\left(\frac{n_i}{n_i - 1} \right) + (1-q)\hat{\pi}\, n - Cov\left(\frac{n_i}{n_i - 1}, n_i \right) \right)} \tag{1.26}$$

Assuming f is high enough to make the denominator of the RHS of (1.26) positive, $q' - q$ will be positive if and only if the following condition holds:

$$\hat{\pi} < \frac{(2-q)\, Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i \right) - 2\, Cov\left(\frac{q_i n_i}{n_i - 1}, n_i \right) - \left(n(1-q)^2 + n - q \right) E\left(\frac{q_i n_i}{n_i - 1} \right) + q(n-1) E\left(\frac{n_i}{n_i - 1} \right) + q\, Cov\left(\frac{n_i}{n_i - 1}, n_i \right)}{q \left(Cov\left(\frac{q_i n_i}{n_i - 1}, n_i \right) - Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i \right) + n(1-q) \left(E\left(\frac{q_i n_i}{n_i - 1} \right) + 1 \right) \right)} \tag{1.27}$$

When all groups are of equal size, the relevant conditions become:

$$q' - q = \frac{(qn\hat{\pi} + n\,(2-q))\; Var\,(q_i) - q\,(1-q)\,(n(q+1) - 1)\,\hat{\pi} - q\,(1-q)\,(n + 1 - nq)}{3\,(n-1)\,f + n\,(1-\hat{\pi})\; Var\,(q_i) - (1-q)\,(nq + n - 1)\,(1-\hat{\pi})} \tag{1.28}$$

$$\hat{\pi} < \frac{q\,(1-q)\,(n + 1 - nq) - n\,(2-q)\; Var\,(q_i)}{q\,(\,Var\,(q_i)\,n - (1-q)\,(nq + n - 1))} \tag{1.29}$$

Condition (1.27) will be shown in Theorem 1.I to be more stringent than the equivalent condition (1.19) for the simultaneous-move game. This means that it is unambiguously more difficult for group selection to operate in the sequential-move model. The proof works by making a number of simplifying substitutions so that the RHS of (1.27) can be directly compared to the RHS of (1.19). It can then be shown that, if the RHS of (1.27) is positive then the RHS of (1.19) must also be positive and strictly greater. In other words, if group selection can survive given some values of $\hat{\pi}$ in the sequential-move model, it must be able to survive for a wider range of $\hat{\pi}$ values in the simultaneous-move model.

Theorem 1.I

If $\hat{\pi}'_{seq} > 0$, *then* $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$.

Proof: Let $\hat{\pi}'_{sim}$ be the RHS of (1.19) and $\hat{\pi}'_{seq}$ be the RHS of (1.27). The following substitutions can be used to rewrite (1.19) and (1.27) in a more easily comparable form:

$$\alpha_i = Cov\left(\frac{n_i}{n_i - 1}, n_i\right) + nE\left(\frac{n_i}{n_i - 1}\right) - Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) - nE\left(\frac{q_i n_i}{n_i - 1}\right) \tag{1.30}$$

$$\beta_i = E\left(\frac{n_i}{n_i - 1}\right) - E\left(\frac{q_i n_i}{n_i - 1}\right) \tag{1.31}$$

$$\gamma_i = Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + nE\left(\frac{q_i n_i}{n_i - 1}\right) - Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) - qnE\left(\frac{q_i n_i}{n_i - 1}\right) \tag{1.32}$$

Note that $\alpha_i > \beta_i > 0$ and that $\alpha_i > \gamma_i > 0$. Using these substitutions, (1.19) and (1.27) become:

$$\hat{\pi}'_{sim} = \frac{(\alpha_i - \beta_i)\, q - \gamma_i}{qn\,(1 - q)} \tag{1.33}$$

$$\hat{\pi}'_{seq} = \frac{(\alpha_i - \beta_i)\, q - (2 - q)\, \gamma_i}{(\gamma_i + n\,(1 - q))\, q} \tag{1.34}$$

It can now be seen clearly by observation that if $\hat{\pi}'_{seq} > 0$, then $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$.

1.9 *Illustration - "Haystacks" Model*

The "Haystacks" model is a well-known scenario in evolutionary biology which provides one possible population structure which could produce sufficient inter-group variance to allow altruism to survive via group selection. The original analogy was a population of mice which splits into haystacks, following which interaction takes place entirely within each haystack for a period of time. Periodically, however, the hay is taken away and the mice are forced out, so that the meta-population once again intermingles. If altruistic mice are sufficiently concentrated within the haystacks, they will breed more rapidly on average than selfish mice.

We are not seeking here to demonstrate the conditions under which such a population structure will enable group selection to take place. There has been controversy about the nature of the mechanism by which the haystacks are formed required to enable group selection to operate effectively. It has been shown that there must either be assortative group formation or more than one period of isolation in order for altruism to survive (Bergstrom, 2002). A thorough investigation has now been undertaken

(Cooper & Wallace, 2004). Cooper and Wallace have shown that altruism can indeed evolve by group selection, even with finite groups and when the assortment mechanism is completely random, provided the ratio of benefit to cost is high enough and the number of periods of isolation is within an intermediate "Goldilocks" band.

We will use different lengths of the isolation period to illustrate the phenomenon that the sequential-move model makes it more difficult for altruism to survive by, as we have already seen analytically, making the Price equation condition more stringent. The haystacks structure acts as a kind of amplification device for the inter-group variance. The longer groups are isolated, the longer the altruists have to benefit one another. However, counteracting this is the fact that the longer the groups are isolated, the better the selfish individuals are doing at the expense of the altruists within each haystack. This means that there is an optimal amount of time for the haystacks isolation, in terms of maximizing the success of the altruists. Isolating the haystacks for a longer or shorter period than this results in a lower average proportion of altruists evolving.

The haystacks model can only operate if there is some initial inter-group variance to be amplified. This is usually achieved by introducing randomness into the assortment process by which groups are formed. In the simulations that follow, we assume that individuals in the overall population are sorted into groups of size 6. We use the hypergeometric distribution to approximate this process. Simulations are "smoothed", in the sense that the fraction of altruists is treated as a continuous variable, even though the Price equations are based on finite group sizes. This has been found to deliver a reasonably close approximation to the discrete model, and allows for much faster simulations, and thus better quality data (Cooper & Wallace, 2004). The haystacks idea is not here being used primarily to justify the possibility of group selection working, but to provide a variable, in the number of periods of isolation g, that can be used to adjust the inter-group variance and illustrate the differences between the sequential-move model and the simultaneous-move model.

Group selection can, of course, occur by other methods aside from a Haystacks population structure. An example commonly used in the social sciences is assortative interaction, where altruists are able to disproportionately interact with one another by forming groups and excluding selfish types. It is commonly argued that group selection is likely to be stronger in human cultural evolution than in human biological evolution not only because cultural phenotypes can be transmitted more rapidly (e.g.

by imitation) than biological ones (which must be passed on genetically via biological reproduction) but also because the concentration of particular cultural traits in groups of humans does not have to rely on randomness as it does in biological models of group selection. For example, groups can expel or reject interaction with non-altruists, or bring extra pressures to bear to enforce conformity. One empirical study (Soltis et al., 1995) found sufficient empirical evidence from anthropological studies of group formation and interaction to conclude that cultural group selection may occur in this manner over a long time scale in human society.

The simulations proceed as follows. The population begins at size 100, with $\frac{1}{3}$ of the individuals having high altruism. This is split into groups of 6, approximated by a hypergeometric distribution for the proportion of groups with each different possible composition, then multiplied by $\frac{100}{6}$ to give the number of each type of group.[25] Each group then evolves in isolation for g periods. The members of the group are formed each period into triplets to play the sequential punishment game, assuming that its simultaneous-move and sequential-move equilibria are played respectively for the simulations with and without the use of punishment. At the end of each period, mutations occur where a fraction ϵ of each type change into the other type. The new population composition is then generated as a weighted average of the different group types after g periods. The number of individuals in the overall population is then normalised back to 100, but preserving the new proportion of altruists q'. (This is done in order to prevent the population from exploding to infinity, to aid the running of the simulations.)

Figures 1.7 through 1.11 overleaf illustrate the outcomes from the simulation over 500 generations, with $\pi = 0.075$ and $\epsilon = \frac{1}{500}$ (time being measured along the x-axis), in the two models, with isolation times of $g = 3$, $g = 6$, $g = 52$, $g = 117$ and $g = 204$ periods respectively.[26] The black line illustrates the simultaneous-move model and the grey line the sequential-move model. The top graph shows the proportion of altruists, and the bottom graph shows the average value of the social welfare function.[27]

In figure 1.7, $g = 3$ is not high enough for altruism to survive in the long run in either model. Since the sequential-move model results in a socially superior static outcome, social welfare can be seen to be higher with the sequential move model (the grey line in the bottom diagram) than with the simultaneous-move model (the black line in the bottom diagram).

[25] Recall that the simulations are smoothed, so there is no reason why the population size need be an integer multiple of the group size.

[26] The simulations were written in Ox. Source code is provided in the appendix.

[27] Calculated as the expected felicity payoff for each individual, and normalized so that both individuals choosing to inflict harm results in a felicity of 0.

In figure 1.8, $g = 6$ is high enough for altruism to survive in the simultaneous move model, but not the sequential-move model. Once sufficient time has elapsed for the proportion of altruists in the population to become high enough, social welfare in the simultaneous model ends up higher than social welfare in the sequential model.

Figure 1.9 shows a situation where $g = 52$ is in the range necessary for altruism to survive in both types of model. However, the analytic result from Theorem 1.I still results in the proportion of altruists oscillating around a lower average in the sequential model. This can be seen to lead to lower average social welfare in the sequential model.

Figure 1.10 shows a situation where $g = 117$ is too high for altruism to survive in the sequential model. It is still able to survive in the simultaneous model, where average social welfare is higher. In figure 1.11, however, $g = 204$ is sufficiently high that although altruism still survives in the simultaneous model, it oscillates so much that social welfare is actually higher on average in the sequential model.

Figures 1.12 and 1.13 show the average values of the proportion of altruists, q and the social welfare function over 5000 generations for different values of g on the x-axis, given two different possible values for $\hat{\pi}$. The dashed grey line in the top diagram in each figure shows the initial proportion of altruists $q = \frac{1}{3}$. The important features to note are, firstly, that there is a wider range where altruism survives in the simultaneous model than in the sequential model and, secondly, that there is a range of values of g where altruism survives in the simultaneous model and not in the sequential model, but average social welfare is nonetheless higher in the sequential model. This shows that the normative consequences of the weakening of group selection are ambiguous, despite the unambiguous result that it is harder for altruism to evolve in the sequential model.

Figures 1.14 and 1.15 show, for two different values of ϵ, the region in (g, π) space where the simultaneous model results in a higher average value of the social welfare function in black and the region where the sequential model results in a higher average value in white. This shows most clearly the result that the use of the social control mechanism of person 2's ability to conditionally punish person 1 is a mixed blessing. In some circumstances, in which the group selection mechanism would have been too weak to operate, it will improve the social efficiency of the evolutionary outcome. However, in other circumstances, where the group selection mechanism would have been strong enough (corresponding to the black area of the diagram), the use of conditional punishment can weaken

the group selection mechanism and thus, by reducing the number of altruists in the evolutionary equilibrium, actually result in society being worse off than in the "anarchic" equilibrium, where punishment is not used.

1.10 *Conclusion*

This paper has shown that the use of punishment is a "double-edged sword" for the evolution of altruism in that it may help selfish types to do better evolutionarily, because they are more willing to make use of opportunities to harm others at benefit to themselves. In the traditional literature on altruistic punishment, altruistic punishers are modelled as a specific behavioural phenotype. This is unsatisfactory from the viewpoint of economic theory because it begs the question of how different available punishment technologies might interact with such evolved preferences. The indirect evolution methodology provides a useful way to approach this question, because it allows a separation between altruistic preferences and altruistic behaviour. The central message is that the ability of humans to punish one another may weaken the selection pressure in favour of altruistic preferences, with potentially negative dynamic welfare implications.

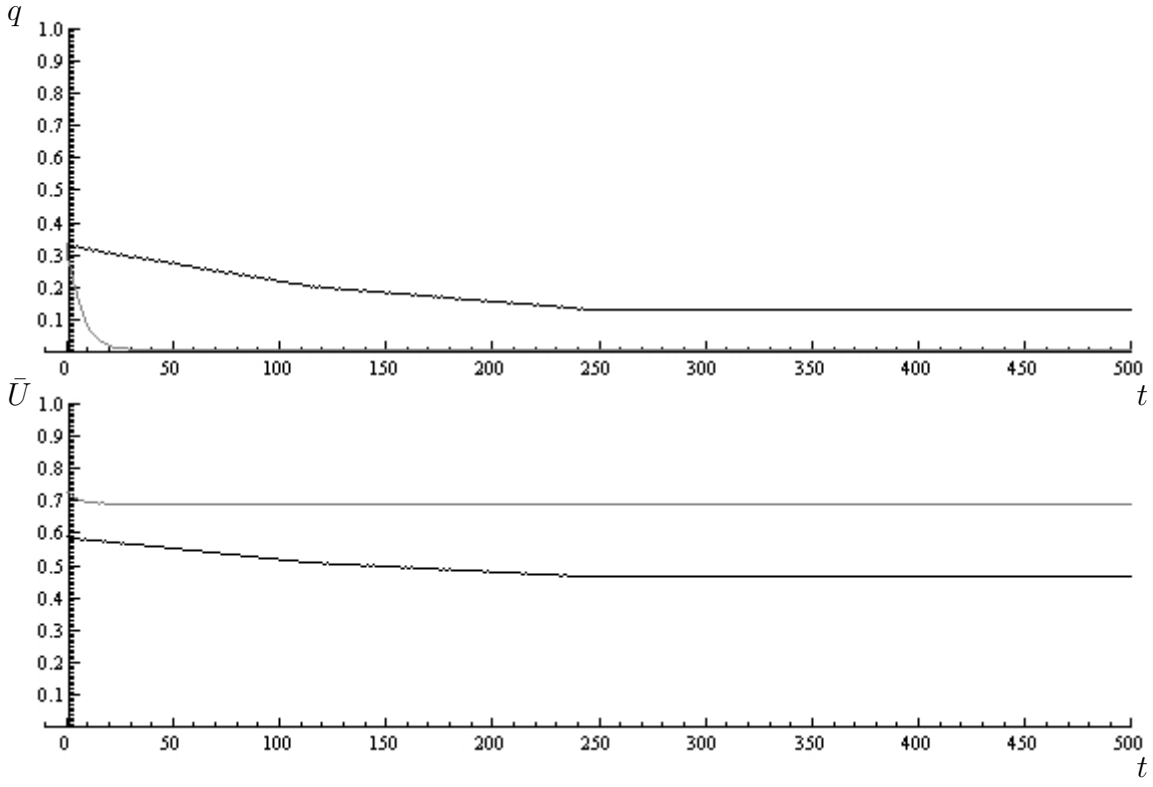

Figure 1.7: 3 periods of isolation, $\pi = 0.075$

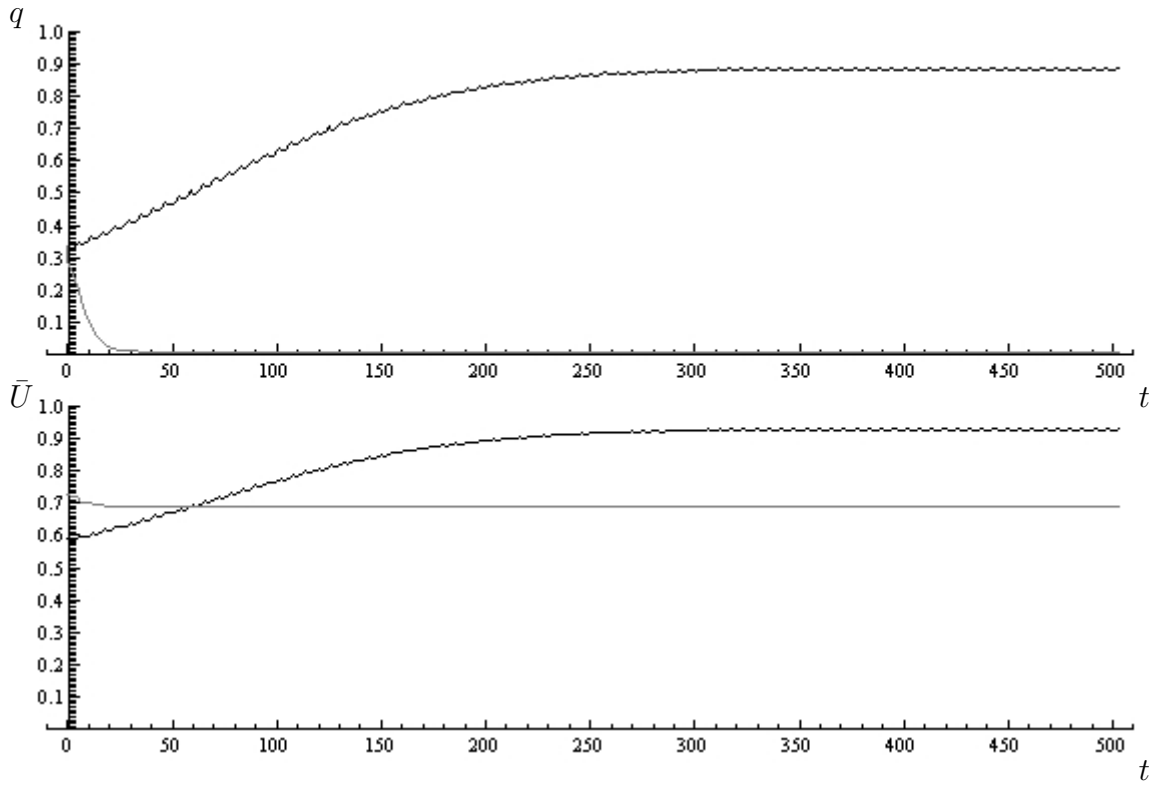

Figure 1.8: 6 periods of isolation, $\pi = 0.075$

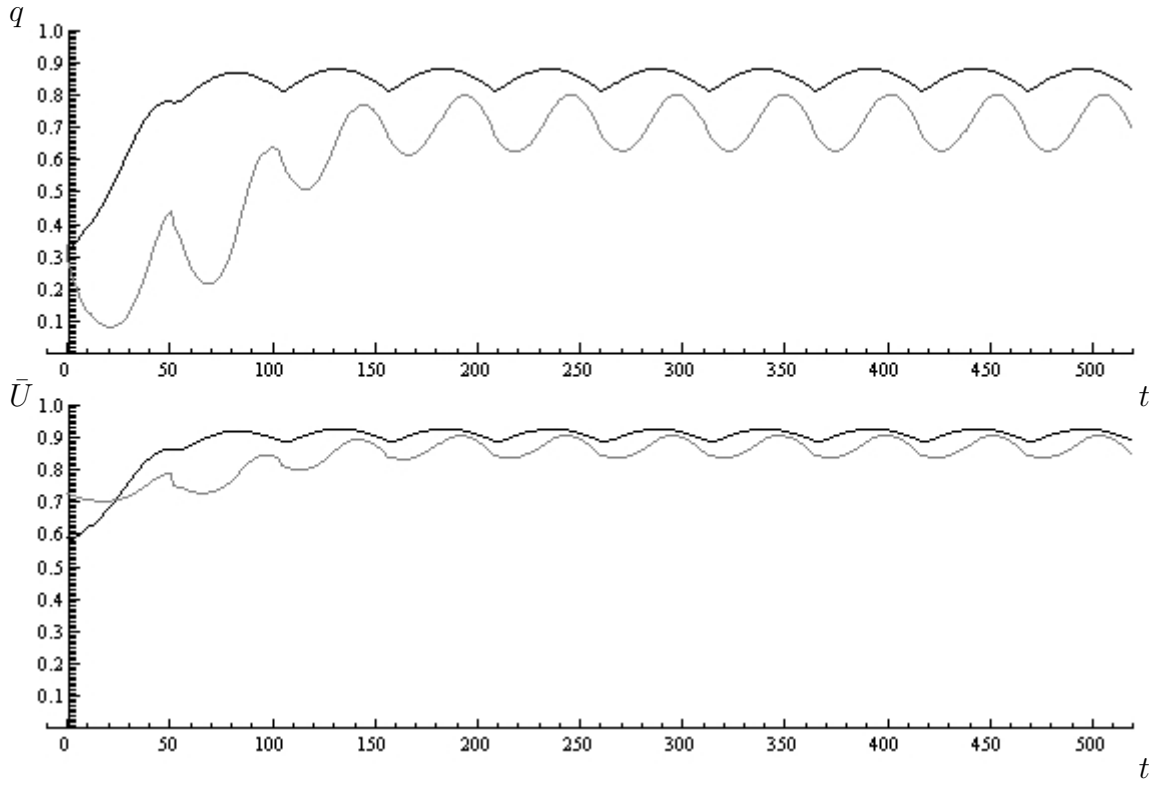

Figure 1.9: 52 periods of isolation, $\pi = 0.075$

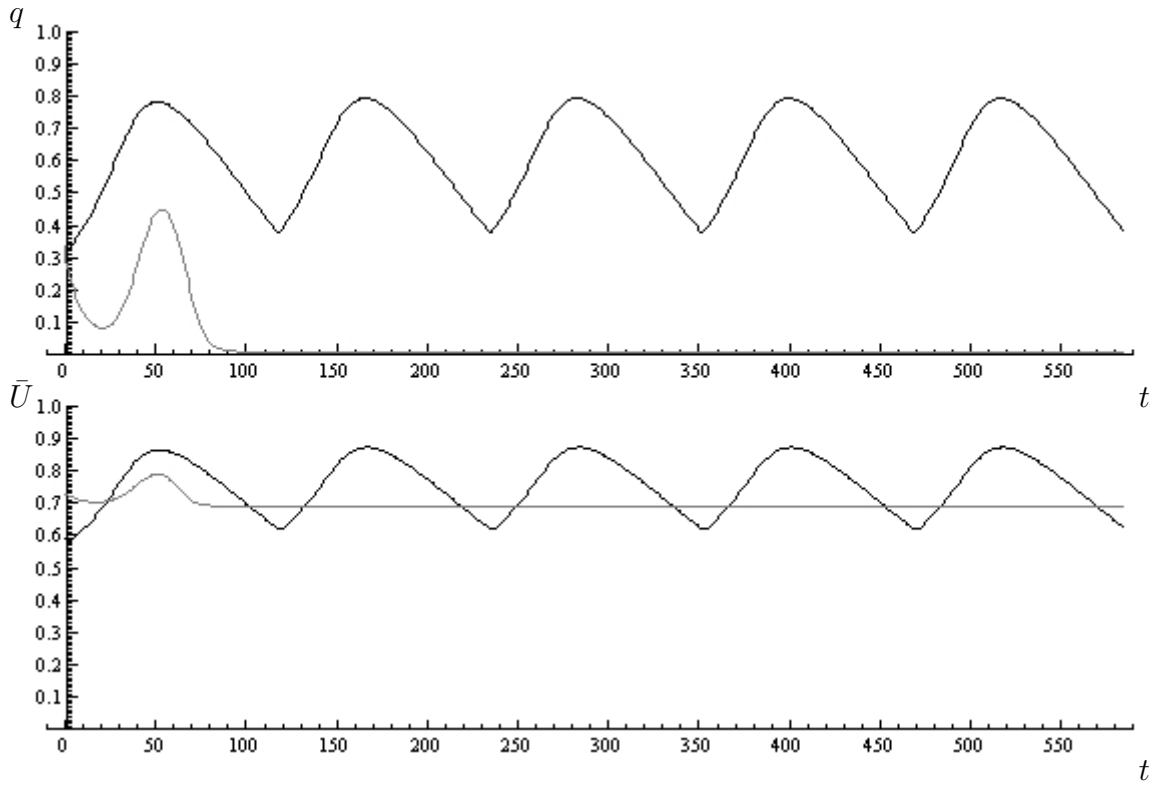

Figure 1.10: 117 periods of isolation, $\pi = 0.075$

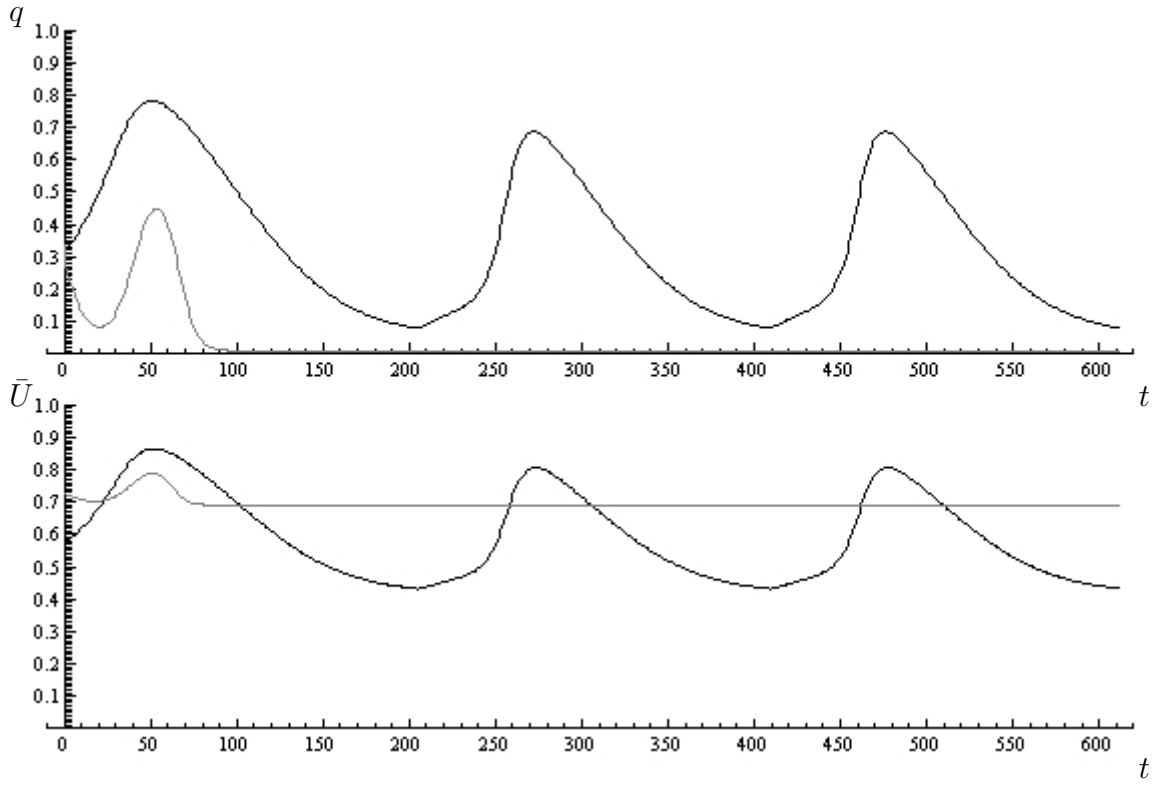

Figure 1.11: 204 periods of isolation, $\pi = 0.075$

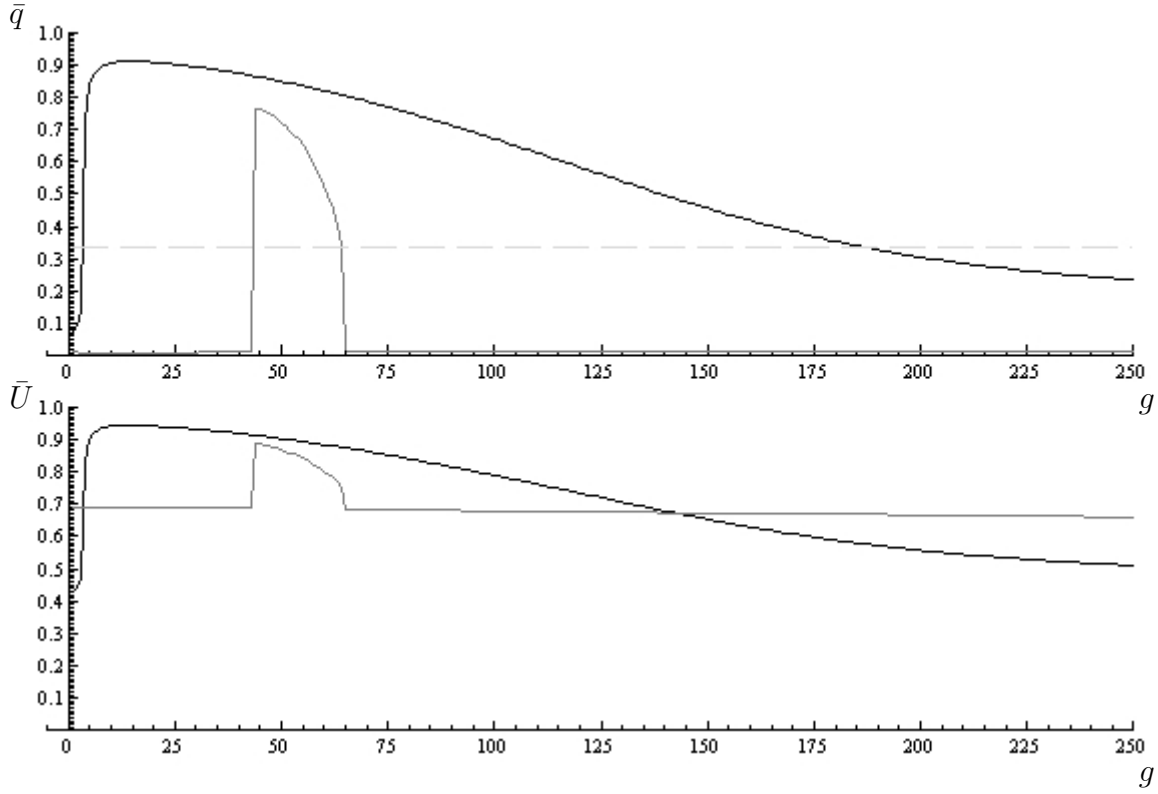

Figure 1.12: Average level of altruism and social welfare, $\pi = 0.075$

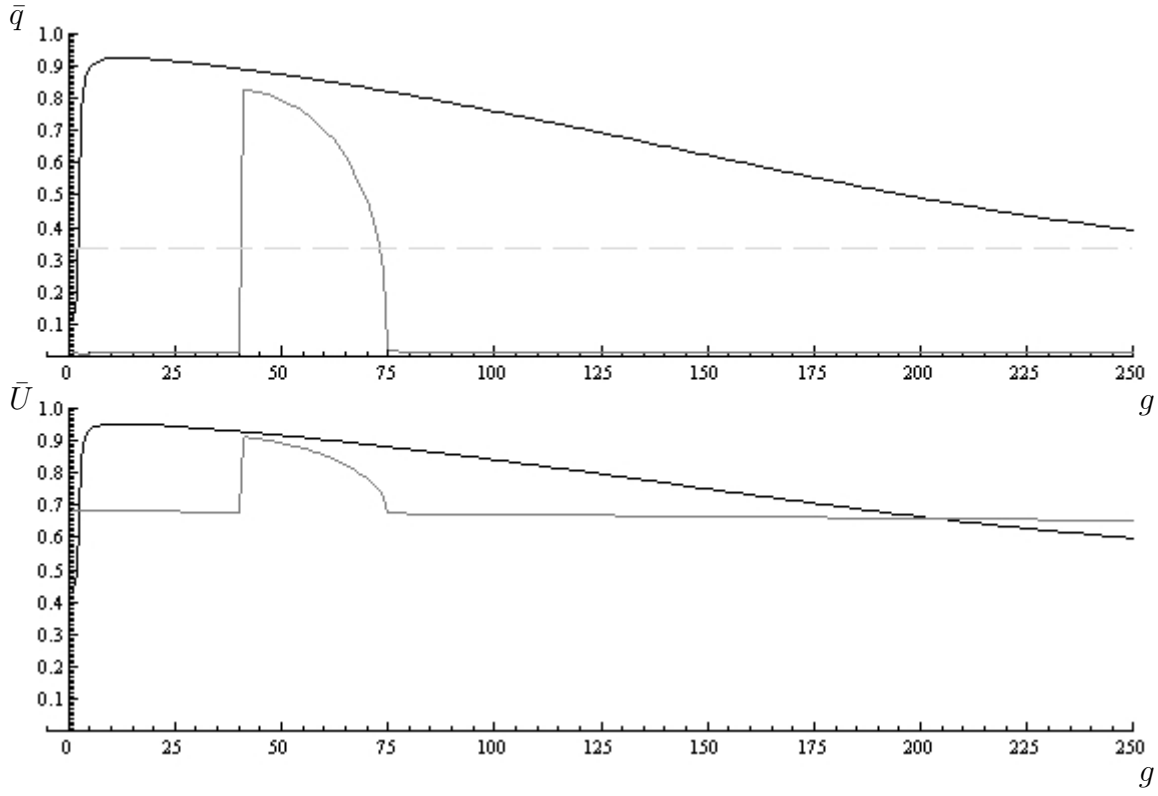

Figure 1.13: Average level of altruism and social welfare, $\pi = 0.05$

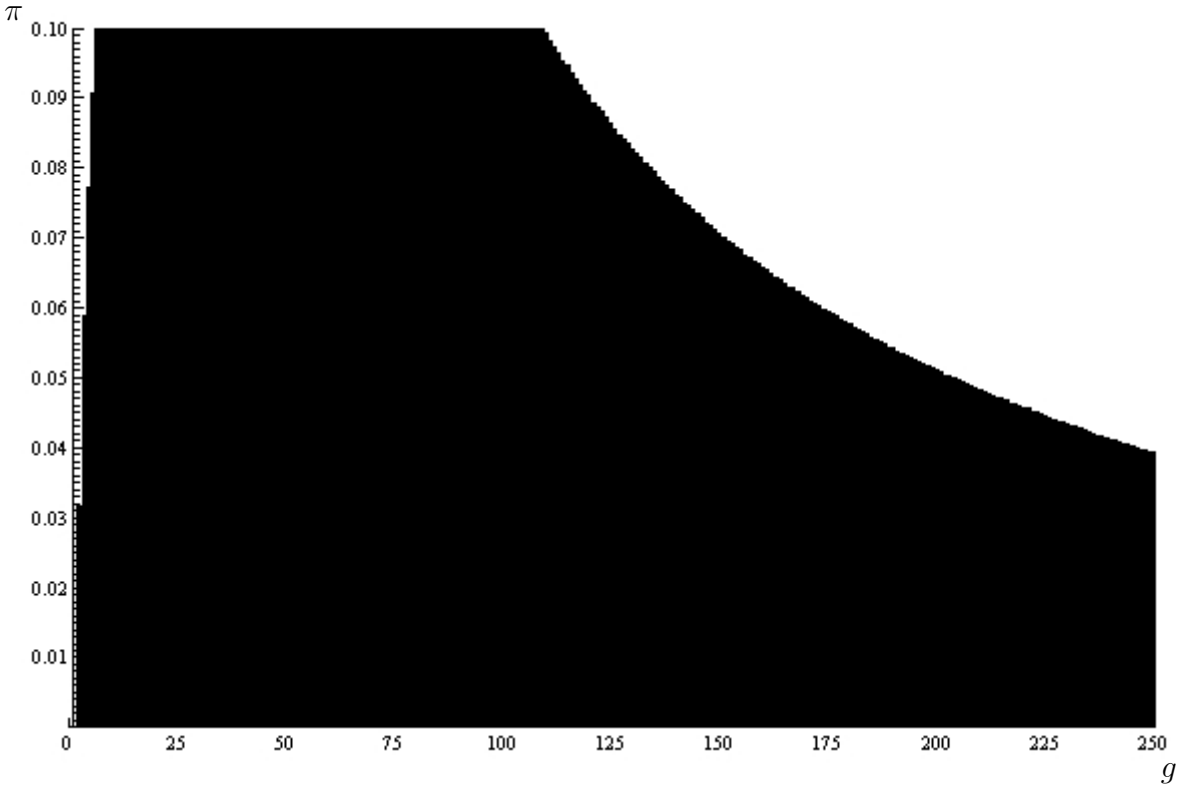

Figure 1.14: Comparison of simultaneous-move and sequential-move game, $\epsilon = \frac{1}{500}$

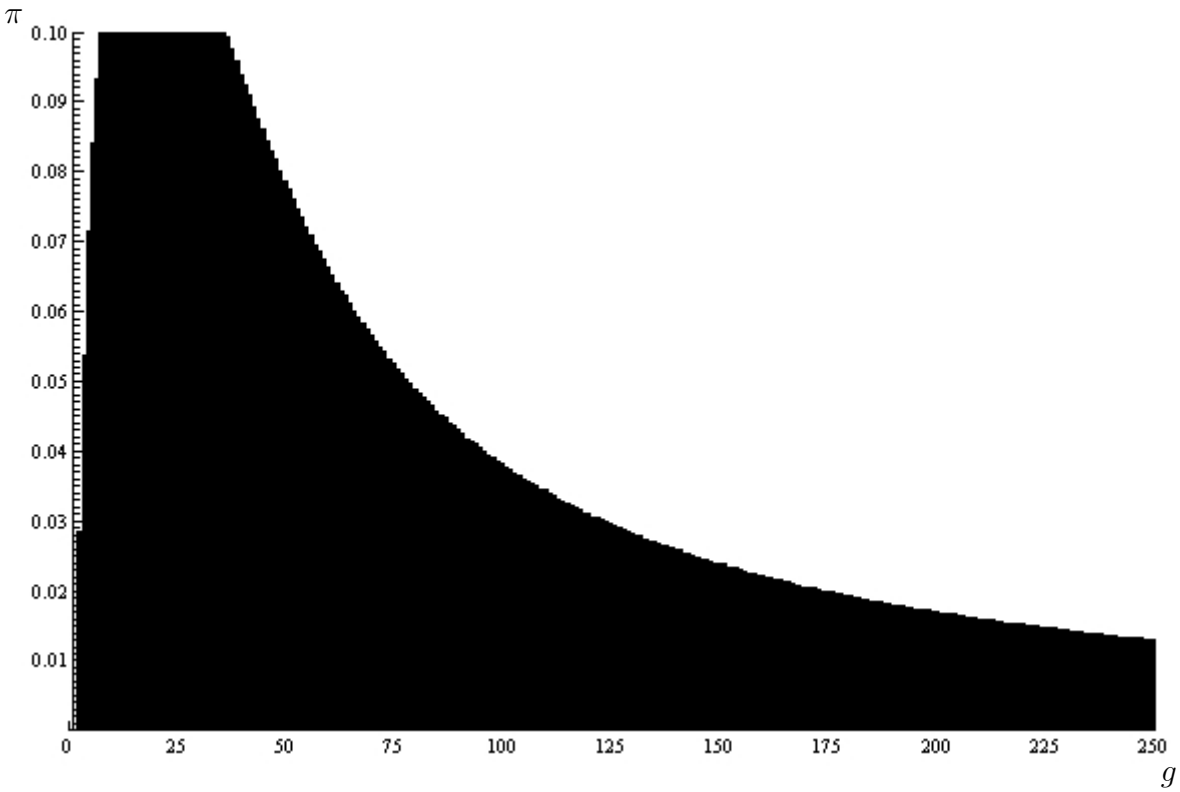

Figure 1.15: Comparison of simultaneous-move and sequential-move game, $\epsilon = \frac{1}{100}$

Appendix

This is the source code to generate figure 1.13:

```
#include <oxstd.h>
#include <oxdraw.h>
#include <oxprob.h>

decl f=2;
decl pi=0.05;
decl qs=1/3;
decl gl=250;
decl totgen=5000;
decl n=6;
decl m=100;
decl e=1/500;

hypergeom(N,p,n,x)
{
return binomial(N*p,x)*binomial(N*(1-p),n-x)/binomial(N,n);
}

class Group
{
decl n;
decl nh;
decl nq;
decl swf;
Group();
reset(nn,nnh,nnq);
nextgen();
nextgenseq();
NH();
N();
SWF();
}

Group::Group()
{
}
```

```
Group::reset(nn,nnh,nnq)
{
n=nn;
nh=nnh;
nq=nnq;
}

Group::nextgen()
{
decl nn,nnh;
nnh=nh*(f-2/3*(n-nh)/(n-1));
nn=n*f-2/3*nh*(n-nh)/(n-1)+2/3*(n-nh)*(pi-(n-nh-1)/(n-1));
swf=nn-f*n+n;
decl q=nnh/nn;
q=(1-e)*q+e*(1-q);
n=nn;
nh=q*nn;
}

Group::nextgenseq()
{
decl nn,nnh;
nnh=nh*(f-2/3*(n-nh)/(n-1));
nn=n*f-2/3*nh*(n-nh)/(n-1)+1/3*(n-nh)*(pi*nh/(n-1)+pi-(n-nh-1)/(n-1));
swf=nn-f*n+n;
decl q=nnh/nn;
q=(1-e)*q+e*(1-q);
n=nn;
nh=q*nn;
}

Group::NH()
{
return nh*nq;
}

Group::N()
{
```

```
return n*nq;
}

Group::SWF()
{
return swf*nq;
}

main()
{
SetDrawWindow("Simulation Output");
DrawAxis(0,0,0,0,1,0.1,0.1,0.01,0);
DrawAxis(1,0,0,0,1,0.1,0.1,0.01,0);
decl i;
decl j;
decl qa;
decl group=new array[n+1];
decl gen;
decl mg;
for(mg=0;mg<n+1;mg++)
{
group[mg]=new Group();
}
decl g;
decl q;
decl NH;
decl N;
decl gb;
decl oldx;
decl oldy;
decl oldSWFx;
decl oldSWFy;
decl tSWF;
decl pass;
decl oldN;
decl SWF;

oldx=0;
oldy=0;
```

```
oldSWFx=0;
oldSWFy=0;
for(j=1;j<gl+1;j++)
{
gb=j;
q=qs;
qa=0;
tSWF=0;
for(gen=0;gen*gb<totgen;gen++)
{
N=m*n;
for(mg=0;mg<n+1;mg++)
{
group[mg].reset(n,mg,N/n*hypergeom(N,q,n,mg));
}
for(g=0;g<gb;g++)
{
oldN=N;
N=0;
NH=0;
SWF=0;
for(mg=0;mg<n+1;mg++)
{
group[mg].nextgen();
N=N+group[mg].N();
NH=NH+group[mg].NH();
SWF=SWF+group[mg].SWF();
}
SWF=SWF/oldN;
tSWF=tSWF+SWF;
q=NH/N;
qa=qa+q;
}
}
qa=qa/(gen*gb+g);
tSWF=tSWF/(gen*gb+g);
if(oldx==0&&oldy==0) {oldx=j; oldy=qa;}
DrawLine(0,oldx,oldy,j,qa,1);
oldx=j;
```

```
oldy=qa;
if(oldSWFx==0&&oldSWFy==0)
{oldSWFx=j; oldSWFy=tSWF; DrawLine(1,oldSWFx,oldSWFy,j,tSWF,1);}
else
{DrawLine(1,oldSWFx,oldSWFy,j,tSWF,1); oldSWFx=j; oldSWFy=tSWF;}
}
oldx=0;
oldy=0;
oldSWFx=0;
oldSWFy=0;
for(j=1;j<gl+1;j=j+1)
{
gb=j;
q=qs;
qa=0;
tSWF=0;
for(gen=0;gen*gb<totgen;gen++)
{
N=m*n;
for(mg=0;mg<n+1;mg++)
{
group[mg].reset(n,mg,N/n*hypergeom(N,q,n,mg));
}
for(g=0;g<gb;g++)
{
oldN=N;
N=0;
NH=0;
SWF=0;
for( mg=0;mg<n+1;mg++)
{
group[mg].nextgenseq();
N=N+group[mg].N();
NH=NH+group[mg].NH();
SWF=SWF+group[mg].SWF();
}
SWF=SWF/oldN;
tSWF=tSWF+SWF;
q=NH/N;
```

```
qa=qa+q;
}
}
qa=qa/(gen*gb+g);
tSWF=tSWF/(gen*gb+g);
if(oldx==0&&oldy==0) {oldx=j; oldy=qa;}
DrawLine(0,oldx,oldy,j,qa,14);
oldx=j;
oldy=qa;
if(oldSWFx==0&&oldSWFy==0)
{oldSWFx=j; oldSWFy=tSWF; DrawLine(1,oldSWFx,oldSWFy,j,tSWF,14);}
else
{DrawLine(1,oldSWFx,oldSWFy,j,tSWF,14); oldSWFx=j; oldSWFy=tSWF;}
}
DrawLine(0,1,qs,gl,qs,10);
DrawText(0, "Periods of Isolation", 0, 0, -1, -1, TEXT_XLABEL);
DrawText(0, "Average Proportion of Altruists", 0, 0, -1, -1, TEXT_YLABEL);
DrawText(1, "Periods of Isolation", 0, 0, -1, -1, TEXT_XLABEL);
DrawText(1, "Average Social Welfare", 0, 0, -1, -1, TEXT_YLABEL);
ShowDrawWindow();
}
```

References

Becker, Gary S. (1976). **"Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology"**. *Journal of Economic Literature*, 14(3), 817–826.

Bergstrom, Theodore C. (2002). **"Evolution of Social Behavior: Individual and Group Selection"**. *Journal of Economic Perspectives*, 16(2), 67–88.

Blackmore, Susan J. (1999). **The Meme Machine**. Oxford University Press, Oxford.

Boyd, Robert and Gintis, Herbert and Bowles, Samuel and Richerson, Peter J. (2003). **"The evolution of altruistic punishment"**. *PNAS*, 100(6), 3531–3535.

Boyd, R. and Richerson, Peter J. (1982). **"Cultural transmission and the evolution of cooperative behavior"**. *Human Ecology*, 10(3), 325–351.

Cooper, Ben and Wallace, Chris (2004). **"Group Selection and the Evolution of Altruism"**. *Oxford Economic Papers*, 56, 307–330.

Fehr, Ernst and Fischbacher, Urs (2003). **"The Nature of Human Altruism"**. *Nature*, 425, 785–791.

Fehr, Ernst and Gachter, Simon (2000a). **"Cooperation and Punishment in Public Goods Experiments"**. *The American Economic Review*, 90(4), 980–994.

Fehr, Ernst and Gachter, Simon (2000b). **"Fairness and Retaliation: The Economics of Reciprocity"**. *The Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, Ernst and Gachter, Simon (2002a). **"Altruistic Punishment in Humans"**. *Nature*, 415, 137–140.

Fehr, Ernst and Gachter, Simon (2002b). **"Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms"**. *Human Nature*, 13, 1–25.

Guth, W. and Yaari, M. (1992). **An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game**. In U. Witt (Ed.), *Explaining Process and Change: Approaches to Evolutionary Economics* (pp. 23–34). University of Michigan Press.

Hamilton, W. D. (1963). **"The Evolution of Altruistic Behavior"**. *The American Naturalist*, 97(896), 354–356.

Hamilton, W D (1972). **"Altruism and Related Phenomena, Mainly in Social Insects"**. *Annual Review of Ecology and Systematics*, 3(1), 193–232.

Hayek, Friedrich A. (1988). **The Fatal Conceit: The Errors of Socialism**. Routledge.

Hirshleifer, Jack (1977). **"Economics from a Biological Viewpoint"**. *The Journal of Law and Economics*, 20(1), 1.

Huck, Steffen and Oechssler, Jorg (1999). **"The Indirect Evolutionary Approach to Explaining Fair Allocations"**. *Games and Economic Behavior*, 28(1), 13–24.

Price, George R. (1970). **"Selection and Covariance"**. *Nature*, 227, 520–521.

Smith, Adam [1776] (1976). **An Inquiry into the Nature and Causes of the Wealth of Nations**. Oxford University Press, Oxford.

Sober, E. and Wilson, D. S. (1994). **"Reintroducing group selection to the human behavioral sciences"**. *Behavioral and Brain Sciences*, 17(4), 585–654+.

Sober, Elliot and Wilson, David S. (1999). **Unto Others**. Harvard University Press, Cambridge, Massachussets.

Soltis, Joseph and Boyd, Robert and Richerson, Peter J. (1995). **"Can Group-Functional Behaviors Evolve by Cultural Group Selection?: An Empirical Test"**. *Current Anthropology*, 36(3), 473–494.