Chapter 11

NATURAL KANTIAN OR *ZOO ECONOMICUS*? EVOLUTIONARY THEORIES OF SELFISHNESS AND ALTRUISM AMONG MEN AND BEASTS

THEODORE C. BERGSTROM*

University of California at Santa Barbara, USA

Contents

* Theodore C. Bergstrom is the Aaron and Cherie Raznick Professor of Economics, University of California at Santa Barbara, Santa Barbara, California. A subset of the contents of this paper can be found in the *Journal of Economic Perspectives* under the title: "Evolution of social behavior" [Bergstrom (2002)]. The current paper includes a lot of discussion cut out of the JEP paper to meet that journal's standards for brevity. Readers who seek a terse discussion of the main issues are likely to prefer the JEP paper. I hope that some may enjoy the more leisurely and thorough discussion that is attempted here. This paper also includes discussion of some interesting work that has appeared since the earlier paper was written.

Handbook of the Economics of Giving, Altruism and Reciprocity, Volume 1
Edited by Serge-Christophe Kolm and Jean Mercier Ythier
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1574-0714(06)01011-6

Abstract

This paper addresses the question of whether our evolutionary history suggests that humans are likely to be individually selected selfish maximizers or group selected altruists. It surveys models from the literature of evolutionary biology in which groups are formed and dissolved and where the reproductive success of individuals is determined by their payoffs in a game played within groups. We show that if groups are formed "randomly" and reproductive success of group founders is determined by a multi-person prisoners' dilemma game, then selfish behavior will prevail over maximization of group payoffs. However, interesting models can be found for which "group selection" sustains cooperative behavior. Forces that support cooperative behavior include assortative matching in groups, group longevity, and punishment-based group norms.

Keywords

group selection, altruism, selfishness, evolutionary games, cooperation, biology, haystack model, punishment, reciprocity, linear public goods game, generalized prisoners' dilemma, assortative matching

JEL classification: C70, C73, D60, Dl

1. Selfishness and group selection

What can our evolutionary history tell us about human motivations and social behavior? The genes that influence our own behavior are inherited from ancestors who themselves managed to reproduce successfully. Could it be that there are evolutionary foundations for the selfishness that economists assume is characteristic of *homo economicus*?

Richard Dawkins (1989), a leading expositor of evolutionary theory, believes this is the case:

> If we were told that a man lived a long and prosperous life in the world of Chicago gangsters, we would be entitled to make some guesses as to the sort of man he was. . . . Like successful Chicago gangsters, our genes have survived, in some cases for millions of years, in a highly competitive world. . . . If you look at the way natural selection works, it seems to follow that anything that has evolved by natural selection should be selfish. (pp. 2–4)

Another well-known biologist, Michael Ghiselin (1974), states this view even more emphatically:

> Where it is in his own interest, every organism may reasonably be expected to aid his fellows . . . Yet given a full chance to act in his own interest, nothing but expediency will restrain him . . . Scratch an "altruist" and watch a "hypocrite" bleed.

But the view that evolution must lead to selfishness is not universally shared. Alexander Carr-Saunders (1922), a sociologist and pioneer in the study of demography and social evolution, observed that prehistoric humans were clustered into groups who inhabited well-defined areas, and that migration between groups was infrequent. These groups, he maintained, typically managed to avoid overpopulation and the attendant scourges of war, famine, and disease. Carr-Saunders argued that ethnographic evidence from existing primitive societies indicates that fertility is deliberately restrained by means of abortion, infanticide, and long-term sexual abstinence. Thus, he claims, these societies managed to maintain population at roughly constant levels close to those that would maximize per capita food consumption. He argued that this outcome is inconsistent with reproductive selfishness and must somehow be explained by "group selection".

Carr-Saunders believed that group selection operates for humans "who have achieved sufficient social organization", but not for more primitive animals. He was an early proponent of the view that "group selection" operates on the evolution of social norms toward those that serve the group interest.

> Now men and groups of men are naturally selected on account of the customs they practise just as they are selected on account of their mental and physical characters. Those groups practising the most advantageous customs will have an advantage in the constant struggle between adjacent groups over those that practise less advantageous customs. Few customs would be more advantageous than those which limit the number of a group to the desirable number . . . There would grow up an idea

> that it was the right thing to bring up a certain limited number of children and the limitation of the family would be enforced by convention. [Carr-Saunders (1922, p. 223)]

V.C. Wynne-Edwards, a leading ecologist of a generation ago, proposed that group selection has far more ancient roots, and applies to much of the animal kingdom. His book, *Animal Dispersion in Relation to Social Behavior* [Wynne-Edwards (1962)] includes an encyclopedic collection of data and descriptions of group behavior and territoriality among a huge variety of species of insects, fish, birds, and mammals. Wynne-Edwards maintained that the Darwinian tradition emphasized natural selection at either the level of individuals or the level of species as a whole, while paying insufficient attention to powerful selective forces that act at the level of the social group. Like Carr-Saunders, he further suggests that not only genetic material, but social norms or conventions may be subject to natural selection. Wynne-Edwards documents the importance of relatively stable localized social groups for the evolution of animal behavior and suggests that for many species, groups have evolved systems of hereditary property rights that strongly shape individual behavior.

According to Wynne-Edwards:

> It has become increasingly clear in recent years, not only that animal (and plant) species tend to be grouped into more or less isolated populations ... but that this is a very important feature from an evolutionary standpoint ... The local stock of any given animal species, exploiting its resources, consequently tends to adopt many conventions of a strictly localized or topographical character – for example the traditional sites of breeding places. Other conventions rely equally strongly on a procession of mutual relationships among the individual local inhabitants. Above all the local stock conserves its resources and thereby safeguards the future survival of its descendants; and no such conventional adaptation could have evolved if the descendants did not naturally fall heirs to the same ground. Thrifty exploitation today for the benefit of some randomly chosen and possibly prodigal generation of strangers tomorrow would make slow headway under natural selection. ... it is of the greatest importance in the long-term exploitation of resources that local populations should be self-perpetuating. If confirmation were needed of this conclusion, it could be found in the almost incredible facilities of precise navigation developed in all long-distance two-way migrants whether they are birds, bats, fish, or insects, to enjoy the advantages of two worlds, and still retain their life-long membership in the same select local stock. Ideally, localisation does not entail complete reproductive isolation however; we have to consider later the pioneering element also – in most species relatively small – that looks after colonisation and disseminates genes. [Wynne-Edwards (1962, pp. 19 and 20)]

Wynne-Edwards also believed that animals do not reproduce as rapidly as they would if individuals were attempting to maximize their own fertility. He cited examples of species in which large gatherings assemble just before breeding time. These gatherings, he claimed, allow individuals to determine the existing population density and to adjust their reproductive decisions in such a way as to maintain a relatively constant population. In Wynne-Edwards view, animal species are able to solve the "tragedy of the commons" and to maintain population densities at an "optimal level for each habitat that they occupy". In Wynne-Edwards (1962), he asserts that:

> Where we can still find nature undisturbed by human influence . . . there is generally no indication whatever that the habitat is run down or destructively overtaxed. On the contrary the whole trend of ecological evolution seems to be in the very opposite direction, leading towards the highest state of productivity that can possibly be built up within the limitations of the inorganic environment. Judging by appearances, chronic over-exploitation and mass poverty intrude themselves only as a kind of adventitious disease, almost certain to be swiftly suppressed by natural selection. (p. 8)

In the opinion of many biologists, Wynne-Edwards' conclusions represented a profound misunderstanding of evolutionary theory. According to Richard Dawkins (1989),

> being wrong in an unequivocal way, Wynne-Edwards is widely credited with having provoked people into thinking more clearly about selection. (p. 297)

An eminent ornithologist, David Lack (1966) and a leading evolutionary biologist, George C. Williams (1966) presented trenchant rebuttals of Wynne-Edwards group selectionist views. Lack presented detailed explanations of how the observations that Wynne-Edwards claimed as support for group-selection could be as well explained by the theory that individuals maximize their own reproductive interests or those of close relatives. Lack pointed out evidence from field studies and experiments that indicates that "in the Starling, the Swift, and usually the Great Tit the most frequent clutch-size is the same as that brood size from which, on average, most young are raised per brood". Lack disputed Wynne-Edwards' claim that reproductive self-regulation eliminates "over-exploitation and mass poverty". He cites one of his own studies that finds that "in many species of song-birds, nearly half of the adults and more of the juveniles die each year, probably mainly from starvation". Williams (1966) went on to list many examples of animal behavior that contribute to individual survival at the expense of the survival prospects of the group.

Williams and Lack both argued that in a fluctuating environment, reproductive restraint in times of abundance is likely to be in an individual's long term reproductive interest. Individuals who have fewer offspring and take better care of each are more likely to have descendants strong enough to survive when hard times arrive. Similarly, it may be in an individual's reproductive self-interest to defend territory that is larger than the minimum territory necessary for successful reproduction in good years because this territory will be essential for success in bad years.

Wynne-Edwards did not present a coherent theory to support his view that social groups rather than individuals might be the units of evolutionary selection. More recently, evolutionary theorists have attempted to provide formal underpinnings for group selection. Sober and Wilson (1999) offer a stimulating and detailed account of these efforts. John Maynard Smith (1976), the father of evolutionary game theory and a key contributor to this effort, concludes that:

> the argument is quantitative, not qualitative. Group selection will have evolutionary consequences: the only question is how important these consequences have been.

2. Games and social interactions

2.1. What is the game and who is playing?

To understand the conflict between the individual and group selection views, it is useful to model social interaction as a game in which the players and the payoffs are explicitly specified. In the language of game theory, the two polar positions can be stated as:

- *Individual selection theory:* To predict social outcomes, we need to examine the game in which the players are individual animals and the payoff to each animal is its expected number of progeny. The outcomes that we expect to see are the Nash equilibria for this game.
- *Group selection theory:* To predict social outcomes, we need to examine the game in which the players are geographically semi-isolated communities of individuals and the payoff is the community's expected reproductive rate. The outcomes we expect to see are Nash equilibria where the players are communities.

A third alternative game formulation is suggested by the work of William G. Hamilton (1964) on *kin selection theory*. As Dawkins (1989) suggests, individuals can be thought of as *survival machines* programmed to make copies of their programmers, the genes. The organisms that we observe are machines that were built by those genes that have in the past been most successful in getting themselves reproduced. Selfish organisms are not typically the best vehicle for genes to use in their own reproduction. Machines that are designed to care for their offspring and to help their close relatives (who are likely to carry the same genes as their own) will typically do better.

- *Kin selection theory:* To predict social outcomes, we need to examine the game in which the players are genes that operate according to Mendelian rules of replication and that carry specific instructions to the organisms that carry them. The payoffs to these genes are their replication rates.

We shall return to the discussion of kin selection theory later in this paper. In the next sections, we examine the competing models of individual and group selection theory and points between. Taken at face value, these theories have radically different implications for the evolutionary nature of men and beasts.

Individual selection theory suggests a world populated by resolutely selfish *homo economicus* and his zoological (and botanical) counterparts. By contrast, in a world shaped by group selection we would expect to see impeccable socialists with an instinctive "Kantian" morality toward other members of their group. Of course the localism that leads to group selection would also be likely to produce some unsavory impulses towards xenophobia and intertribal warfare.

When the game being played within communities is prisoners' dilemma, the contrasting predictions of the two theories are particularly stark and simple. Since the payoff from playing *defect* is always higher than that of playing *cooperate*, individual selection theory predicts a population of defectors. But since every member of a community of cooperators gets a higher payoff than any member of a community of defectors, group selection theory predicts a population of cooperators.

Using prisoners' dilemma as a research vehicle, biologists, game theorists, and anthropologists have found much interesting territory between the two poles of individual selection and group selection. Although neither of the polar theories would be supported by modern research, the tension between the forces of individual and group selection continues to be the focus of interesting research. The use of prisoners' dilemma to explore this tension has been very instructive and will play an important part in this survey. However, as we argue in later discussion, most of the really important (and problematic) social interactions in the world are probably not games with unique Nash equilibria, let alone dominant strategies, but games that have many distinct Nash equilibria among which societies somehow select.

2.2. Prisoners' dilemma games

2.2.1. Multi-player prisoners' dilemma

A multi-player prisoners' dilemma is a game in which individuals may take actions that are, in the words of J.B.S. Haldane (1932), "socially valuable but individually disadvantageous". Specifically, we consider a game that has two possible strategies for each player, *cooperate* and *defect*, where the payoff to each player depends on her own strategy and the number of other players who play cooperate. In a game with N players, where K of the *other* players cooperate, let $\Pi_{\mathrm{C}}(K, N)$ and $\Pi_{\mathrm{D}}(K, N)$ denote the payoffs to a cooperator and a defector, respectively.

DEFINITION 1 (N-player Prisoners' Dilemma Game). A game is an N-player prisoners' dilemma game if the payoff functions satisfy the following:

- All players are better off if all play *cooperate* than if all play *defect*; that is, $\Pi_{\mathrm{C}}(N-1, N) > \Pi_{\mathrm{D}}(0, N)$.
- Regardless of what other players do, an individual gets a higher payoff from playing defect than from playing cooperate; that is, $\Pi_{\mathrm{D}}(K, N) > \Pi_{\mathrm{C}}(K, N)$, for all K between 0 and $N-1$.

2.2.2. The linear public goods game

It is customary to credit game theorists, Merrill Flood and Melvin Dresher of the Rand Corporation, with inventing the prisoners' dilemma game in about 1950. But this game has an earlier history. In 1932, J.B.S. Haldane, one of the founders of modern population biology, introduced and analyzed an N person generalized prisoners' dilemma game in which each player's payoff depends linearly on the number of players in the game who cooperate. Economists will recognize Haldane's game as formally equivalent to the linear "voluntary contribution to public goods" game, much studied in experimental economics [see Ledyard (1995) for a good survey of this work]. Thus we will refer to Haldane's linear N-player prisoners' dilemma as the *linear public goods game*.[1]

DEFINITION 2 (The Linear Public Goods Game). The linear public goods game is an N player game in which each player can play either cooperate or defect. Where x is the fraction of all players who cooperate, the payoff to each cooperator is $bx - c$ and the payoff to each defector is bx.

In a linear public goods game with N players, if K other players cooperate, a cooperator will get

$$\Pi_{\mathrm{C}}(K, N) = b\frac{K+1}{N} - c = b\frac{K}{N} - c' \tag{1}$$

where $c' = c - \frac{b}{N}$ and a defector will get

$$\Pi_{\mathrm{D}}(K, N) = b\frac{K}{N}. \tag{2}$$

The linear public goods game can be seen to be an N-player prisoners' dilemma if $b > c > \frac{b}{N}$. If all players cooperate, each gets a payoff of $b - c$; if all defect, each gets a payoff of 0. Therefore when $b > c$, all players are better off if all cooperate than if all defect. For all K, $\Pi_{\mathrm{D}}(K, N) - \Pi_{\mathrm{C}}(K, N) = c - \frac{b}{N} = c'$. Thus if $c > \frac{b}{N}$, an individual always gets a higher payoff by defecting rather than cooperating.

In a linear public goods game with N players, a cooperator confers a benefit of $\frac{b}{N}$ on every player, including himself, so that the net cost of cooperating is $c - \frac{b}{N}$. Some writers, such as David S. Wilson (1975), analyze a variant of this game in which a cooperator confers expected benefits of $\frac{b}{N}$ on every player *other than himself* at a cost of c to himself. Results for either of these two games translate easily into corresponding results for the other, since Wilson's formulation of the game with costs c is isomorphic to a linear public goods game with costs $c + \frac{b}{N}$.

2.3. *Stag hunt games*

In one-shot prisoners' dilemma games, the socially optimal action is never a best response for selfish individuals. But in many social interactions, the action that best serves one's self-interest depends on the actions taken by others. This suggests the usefulness of a second exploratory vehicle, a simple two-person game, known as the stag hunt. This

[1] Haldane (1932, pp. 207–210 of the Appendix) presents this model in an early discussion of group selection. The notation used here is that of Cohen and Eshel (1976) rather than that of Haldane.

Table 1
A stag hunt game

	Cooperate	Defect
Cooperate	4	0
Defect	3	3

game formalizes a story told by Jean Jacques Rousseau (1950, p. 428), of two hunters who could cooperate by jointly hunting a stag or defect by individually hunting hare.[2] Table 1 is a game matrix for a stag hunt game, where entries represent payoffs to the row player.

The stag hunt has two equilibria, one where both players cooperate and one where both defect. In later discussion, we consider the question of when one or the other equilibrium is likely to be reached.

2.4. Evolutionary dynamics and altruism

2.4.1. Prisoners' dilemma in a freely mingling population

Let us consider the evolutionary dynamics of a population in which all individuals are "programmed" (perhaps genetically, perhaps by cultural experience) to play one of two strategies, *cooperate* or *defect* in a symmetric multi-person prisoners' dilemma game played by the entire population. We will assume that the dynamics satisfy *payoff-monotonicity* [Weibull (1995)], which means simply that the proportion of the population that plays the strategy with the higher payoff will increase.[3] If the game is prisoners' dilemma, the payoff to cooperators will necessarily be lower than to defectors, so the proportion of cooperators in the population must decline over time and eventually converge to zero.[4]

Gorret Hardin, in *The Limits of Altruism* (1977) explained this result and suggested that the replacement of tribalism and parochialism by a more cosmopolitan society is not likely to enhance cooperation.

[2] An engaging paper by Brian Skyrms (2001) makes a strong case that social thinkers should pay more attention to the stag hunt game.

[3] A much-studied special case of payoff monotone dynamics is *replicator dynamics* in which the growth rate of the population share using a strategy is proportional to the difference between the average payoff to that strategy and the average payoff in the entire population [Weibull (1995)]. The results found in this paper do not require the special structure of replicator dynamics.

[4] The result that the proportion of cooperators will decline monotonically is obvious. The result that it must converge to zero is less obvious. A proof can be found in Weibull (1995). Weibull credits this result to John Nachbar (1990).

> Competition is severe and total whenever members of the same species are brought together in One World . . . Conceivably some conscientious members of the community might eat less than their share of the food, but the resources they thereby released would soon be absorbed by others with less conscience. Some animals might refrain from reproducing, but the space so freed would soon be occupied by those who were less conscientious. . . . Adapting a phrase of the economist David Ricardo, we can speak of the Iron Law of the Overwhelming Minority. It is silly to dream dreams of a heaven on earth that presume a value of zero for the size of the disruptive minority.

3. Haystack models

Two of the founders of modern population biology, J.B.S. Haldane (1932) and Sewall Wright (1945), proposed that altruistic behavior is more likely to evolve in a population where group interaction takes place within relatively small subpopulations, (sometimes called *demes*) between which there is occasional, but relatively infrequent migration.

3.1. Maynard Smith's mice

John Maynard Smith (1964) produced the first formal model of group selection in which seemingly altruistic behavior prevails, even without assortative matching. Maynard Smith motivates this model with a charming story of "a species of mouse who lives in a hayfield".

The setting for Maynard Smith's haystack model is a meadow. In early summer, a farmer builds several haystacks, each of which is colonized by exactly two mice. These two mice and their descendants interact and reproduce asexually for the entire season, until the haystacks are removed.[5] When the haystacks are cleared, the dislodged mice scramble out into the meadow, mingling freely with the mice displaced from other haystacks. In the next year, when new haystacks are built, exactly two mice from the population resident in the meadow are randomly selected to colonize each new haystack. If the number of surviving mice is more than twice the number of haystacks, the extra mice are consumed by predators.

There are two types of mice in the population at large, timid mice and aggressive mice. Descendants of either type of mouse will be of the same type as their ancestors. Timid mice play the role of "altruists" by pursuing a strategy that is socially valuable but individually disadvantageous. Thus, within any haystack, the timid mice reproduce less rapidly than the aggressive mice, but at the end of the season, haystacks that are made

[5] Maynard Smith presented his model as one with sexual diploid reproduction. But he contrived special assumptions that make his model mathematically equivalent to a model with asexual reproduction. To simplify exposition and to make this model directly comparable with the later extensions by Cohen and Eshel, I present an asexual haystack model that is formally equivalent model to his sexual diploid model.

Table 2
The haystack game

	Timid	Aggressive
Timid	$r(1+K)/2$	0
Aggressive	r	$r/2$

up entirely of timid mice will have more inhabitants than those that include aggressive mice.

In a haystack settled by two timid mice, all descendants are timid and in a haystack settled by two aggressive mice, all descendants are aggressive. In a haystack settled by one mouse of each type, the descendants of the aggressive mouse eliminate the descendants of the timid mouse, and the number of its descendants at harvest time is the same as the number in a haystack colonized by two aggressive mice.

Although timid mice do poorly when matched with aggressive mice, haystacks inhabited entirely by timid mice produce more surviving offspring at harvest time than haystacks inhabited by aggressive mice. Thus a haystack colonized by two timid mice produces $1 + K$ times as many descendants as a haystack with aggressive mice.

Since the reproduction rate enjoyed by a founding mouse depends on its own type and that of its co-founder, these rates can be represented as the payoffs in a game between the two mice who colonize each haystack. If two aggressive mice colonize a haystack, they will have a total of r descendants, half of whom are descended from each founder. Thus each mouse has $r/2$ descendants. If an aggressive mouse and a timid mouse colonize a haystack, the timid mouse will have no descendants and the aggressive mouse will have r descendants. If two timid mice colonize a haystack, they will have a total of $r(1+K)$ descendants and each will have $r(1+K)/2$ descendants. In the game played by cofounders, payoffs to the row player are shown in Table 2.

If $0 < K < 1$, the haystack game is a prisoners' dilemma, since regardless of its cofounder's type, an aggressive mouse will have more offspring than a timid mouse. If $K > 1$, the haystack game is not a prisoners' dilemma but a stag hunt. If matched with a timid mouse, a mouse will have more offspring if it is timid than if it is aggressive. But if matched with an aggressive mouse, a mouse will have more offspring if it is aggressive than if it is timid.

For the prisoners' dilemma case with $K < 1$, the only equilibrium is a population made up entirely of defectors. For the stag hunt case, with $K > 1$, there are two distinct stable equilibria, one in which all mice are timid and one in which all are aggressive. We demonstrate this as follows. Let the proportion of timid mice in the population at time t be x_t. Since matching is random, any mouse is matched with a timid co-founder with probability x_t and with an aggressive co-founder with probability $1-x_t$. Given the payoffs in Table 2, the expected reproduction rate of an aggressive mouse is $x_t r + (1 - x_t)r/2$, and the expected reproduction rate of a timid mouse is $x_t r(1+K)/2$. Subtracting the latter expression from the former, we find that the difference between the expected

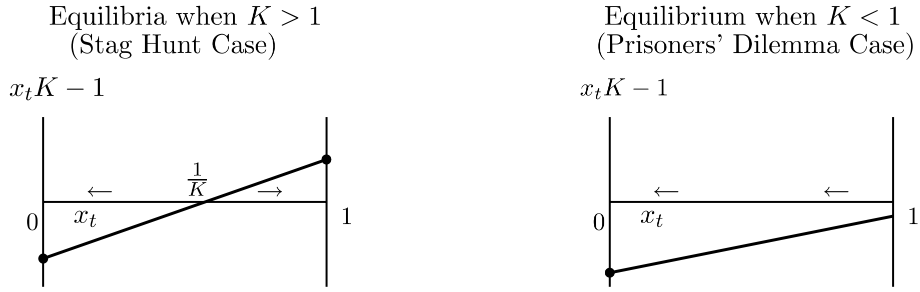

Figure 1. Dynamics of the haystack model.

reproduction rates of timid mice and of aggressive mice is proportional to $x_tK - 1$. Therefore timid mice reproduce more rapidly than aggressive mice if $x_tK > 1$ and aggressive mice reproduce more rapidly if $x_tK < 1$. These dynamics are illustrated by Figure 1. The graph on the left shows that where $K > 1$, there are two stable equilibria, one in which all mice are timid and one in which all are aggressive. (There is also an unstable equilibrium where the fraction $1/K$ of mice are timid.) The graph on the right shows that where $K < 1$, there is a unique equilibrium in which all mice are aggressive.

THEOREM 1 (Haystack Dynamics). *In Maynard Smith's haystack model with random mating:*

- *If haystacks of timid mice produce more than twice as many mice as haystacks of aggressive mice, there will be two stable monomorphic equilibria, one in which $x = 0$ (all mice are aggressive) and one in which $x = 1$ (all mice are timid), as well as one unstable polymorphic equilibrium where $x = 1/K$.*
- *If haystacks of timid mice produce fewer than twice as many mice as haystacks of aggressive mice, the only stable equilibrium is a monomorphic population of aggressive mice.*

3.2. General haystack models and assortative matching

Let us define a *generalized haystack model* to be a model with a large population of individuals, who are programmed for one of two strategies, altruist or selfish. At the beginning of each time period, these individuals are randomly partitioned into groups (possibly of different sizes). Each member produces (asexually) a number of offspring equal to her payoff in a game that she plays with other members of her own group. Offspring are programmed to use the same strategy as their parent. At the end of each time period, all groups are disbanded and new groups are randomly selected from the combined population of the disbanded groups.

Let $p_{\mathrm{A}}(K, N)$ be the probability, conditional on being an altruist, that a player is assigned to a group of size N in which K of the *other* group members are altruists. Let $p_{\mathrm{S}}(K, N)$ be the probability, conditional on being selfish, that one is assigned to a group of size N in which K of the other members are altruists. We define group formation to be *non-assortative with respect to strategy* if when new groups are assigned from the offspring of the previous groups, altruists and selfish individual offspring have equal chances to be selected to join one of the new groups, and the probability distribution of group size and the number of other group members who are altruists is the same whether one is programmed to use the altruist strategy or the selfish strategy.

DEFINITION 3 (Non-assortative Matching Process). A matching process is *non-assortative* between types if

- In each period, the number of individuals of each type who are selected to join the new groups is proportional to the total number of offspring of that type who were produced in the previous period.
- In each period, for all K and N, $p_{\mathrm{A}}(K,N) = p_{\mathrm{S}}(K,N)$.

A simple example of a non-assortative matching process is an urn model in which there is a fixed number of locations, each with a given capacity, and where each location is populated by independent random draws from the total population.

If group formation is assortative, so that an altruist is more likely to have altruist neighbors than a selfish individual, then it is easy to see that altruism can be maintained in the population. For example, in the extreme case where group formation is perfectly assortative so that groups consist either entirely of altruists or entirely of selfish, altruists will always receive higher payoffs than selfish individuals and altruists would eventually constitute the entire population.

But is it possible for altruism to be sustained in a haystack model if new groups are formed at random from the population? When groups are formed by independent random draws, the proportions in each group will not mirror the proportions in the population at large. Random selection results in some groups that have disproportionately many altruists and some that have disproportionately many selfish individuals. Within each group, the altruists get lower payoffs and hence reproduce less rapidly than the selfish. But there is a countervailing effect. Groups that contain more altruists grow more rapidly. Can this between-group effect overwhelm the within-group effect and cause the proportion of altruists *in the overall population* to increase over time? Or does Hardin's "Iron Law" extend to populations randomly selected into groups? The next sections offer a partial answer to this question.

3.3. Cohen and Eshel's generalized haystack models

Dan Cohen and Ilan Eshel (1976) produced a series of interesting generalized haystack models. In these models, there are two types of asexually reproducing individuals, "altruists" and "selfish". As in the haystack model, individuals group into distinct colonies where they live and breed. After some fixed length of time, all colonies are disbanded and new colonies are formed by individuals randomly selected from the population at large. In the Cohen–Eshel model, the number of individuals in the founding population is N. They assume that reproduction takes place continuously over time and that within any haystack, the reproduction rate of both types of individuals is an increasing function of the proportion who are altruists. However, the reproduction rate of altruists is lower than that of the selfish.

Cohen and Eshel focus on determining the stability of each the two possible monomorphic populations, all altruists and all selfish. This investigation is simplified by the following observation. With random group formation, when invaders are rare, almost all invaders will be selected into groups in which all other members are normal. Thus a monomorphic population of either type can be invaded by an initially small influx of the other type only if the reproduction rate of a single invader who joins $N - 1$ normal individuals in founding a colony is larger than that of a normal individual among a group made up entirely of the normal type.

3.3.1. Linear public goods games in haystacks

One model that Cohen and Eshel analyze is Haldane's linear public goods game. In the Cohen–Eshel formulation, if $x(t)$ is the fraction of a group that are altruists at time t, then the reproduction rate of selfish group members is $a + bx(t)$, while that of altruists in the same group is $a + bx(t) - c$. Cohen and Eshel find the ranges of parameter values in the linear public goods game for which each kind of monomorphic equilibrium is stable.[6] The length of time T for which communities remain intact before dispersing is of critical importance.

THEOREM 2 (Cohen–Eshel). *In the Cohen–Eshel haystack model, where reproduction rates are determined by the linear public goods game and where T is the length of time for which groups remain intact:*

- *For small T, if $b/N < c$, the only stable equilibrium is a population of selfish individuals and if $b/N > c$, the only stable equilibrium is a monomorphic population of cooperators.*
- *If T is sufficiently large, and $b > c > 0$, there exist two distinct stable monomorphic equilibria; one with selfish players only and one with altruists only.*

The most surprising result is that if T is large enough, there exists a stable equilibrium with a population made up of altruists, even though groups are formed by an independent random matching process, and even though the game that determines instantaneous reproductive rates is an N-player prisoners' dilemma game. To see why this happens, recall that a population of altruists will be stable if the expected number of descendants of a single selfish individual who joins $N - 1$ altruists in founding a community is higher than the expected number of descendants of an altruist who is among a founding group consisting entirely of altruists. The number of altruists in a group consisting

[6] They are able to find closed-form solutions for the reproduction rates of a mutant cooperator in a population of defectors and of a mutant defector in a population of cooperators.

entirely of altruists grows at the rate $a + b - c > a$. The descendants of the selfish invader will reproduce more rapidly than the *altruist members of the group which she joins*. But to invade the population, her descendants must reproduce more rapidly than *altruists who live exclusively among altruists*. As T is large, the descendants of a selfish invader will eventually comprise almost the entire group to which they belong. Hence the growth rate of the invader population will approach a. Thus when T is large enough, the growth rate of normal altruists is higher than that of the invading selfish. Moreover, this difference in growth rates does not diminish over time. It follows that there exists some survival period T such that if groups persist for longer than T, a monomorphic population of altruists is a stable equilibrium.

3.3.2. Mutation in the haystacks

Cohen and Eshel's Theorem 2 assumes the absence of mutation within haystacks. Indeed, their conclusion that a population of altruists will be stable if the time between haystack dispersals is sufficiently long is not in general true if there is a non-zero probability of mutation at each moment in time. Eshel and Cohen's result depended on the observation that a haystack that initially consists entirely of cooperators will remain uninfected by defectors and will if the haystack remains intact long enough reproduce faster than groups of defectors. But with mutation and very long-lasting haystacks, the haystacks that start out with cooperators only are likely to be infected at some time by mutant defectors. Once infected, their growth will slow as the defectors within the group reproduce more rapidly than the cooperators.

A recent paper by Ben Cooper and Chris Wallace (2001) runs simulations of a haystack model with mutations. In their simulations, altruism does not survive either if haystacks are dispersed after a very short time or after a very long time. However in their simulations, altruism can prevail if the number of generations from the formation of haystack groups until their dispersal is of intermediate size.

3.3.3. Nonlinearity and polymorphic equilibria

The Haldane linear public goods model assumes that a community's growth rate depends linearly on its proportion of altruists. This implies constant returns to altruism in the sense that an additional altruist makes the same contribution to growth regardless of the number of other altruists. Cohen and Eshel show that without this linearity, monomorphic equilibrium do not always exist. They define a "generalized Haldane model" in which the reproduction rate of selfish individuals is $a + b\psi(x)$ when x is the proportion of altruists in their community; where $\psi(\cdot)$ is an increasing function such that $\psi(0) = 0$ and $\psi(1) = 1$. They show that if there is diminishing returns to the addition of altruists to the community, it can happen that the only equilibria are polymorphic, with both types being present in equilibrium.

3.3.4. Congested resources

Cohen and Eshel (1976) also study a version of the haystack model in which growth within each community is constrained by the amount of resources available. There are "selfish" individuals who reproduce more rapidly than "altruists", but consume more resources. At the end of a fixed period of time, T, the original communities are dispersed and new communities are founded by groups who are randomly selected from the entire population. In this model, a community whose founders are mainly altruists will produce more offspring because each uses less resources. On the other hand, the selfish members of a community produce more offspring than an altruistic member. For fixed growth rates and resource exhaustion parameters, if founding populations are small enough, there will be a stable equilibrium with altruists only, if populations are large enough, there will be a stable equilibrium with selfish only, and for intermediate sizes of population, there will be two distinct stable equilibria; one with altruists only and one with selfish only.

3.4. The iron rule of selfishness

David S. Wilson (1975), in a pioneering study of group selection, showed that for his particular model, "random" formation of groups must result in the elimination of altruism. In a survey article called "Natural, kin and group selection" (1984), Alan Grafen states that "with random grouping there is no selection for altruism". However, Maynard Smith (1964, 1976), Eshel (1972), Cohen and Eshel (1976), and Matessi and Jayakar (1976) seem to have contrary results. Although mating is random in Maynard Smith's haystack model, for some parameter values, there is a stable equilibrium in which the entire population consists of altruists. Eshel (1972) asserts that "for any altruistic trait, there is a critical level of demographic mobility under which selection would always operate for the establishment of the altruist". In Cohen and Eshel's models (1976), there is "random distribution of altruist and selfish in small founder groups" and it turns out that if groups remain together long enough before being dispersed, there exists a stable equilibrium consisting entirely of altruists (as well as another stable equilibrium consisting entirely of selfish.)

To establish the circumstances under which Grafen's claim of no-altruism-with-random-sorting is correct, we need to specify the reproductive dynamics that we have in mind, as well as what we mean by altruism, and by random mating. In this section altruism is defined as playing altruist in an N-person prisoners' dilemma game in which a player's payoff is her reproduction rate. As we will later discuss, this does not exhaust the forms of behavior that might reasonably be called altruistic.

THEOREM 3 (Iron Rule of Selfishness). *In a generalized haystack model, if groups are formed by a matching process that is non-assortative and if the game that determines reproduction rates is an N-player prisoners' dilemma played with members of one's* *own group,then the proportion of altruists (cooperators) in the population will approach zero as the number of periods gets large.*

PROOF. In each period at the time when new groups are formed, the expected numbers of offspring produced by each selfish individual and each altruist of the previous generation are, respectively:

$$\sum_N \sum_{K=1}^{N-1} p_{\mathrm{S}}(K,N)\Pi_{\mathrm{S}}(K,N) \quad \text{and} \quad \sum_N \sum_{K=1}^{N-1} p_{\mathrm{A}}(K,N)\Pi_{\mathrm{A}}(K,N). \tag{3}$$

The difference between the growth rate of the number of altruists and the growth rate of the number of selfish individuals is proportional to the difference between these two rates. Since matching is non-assortative, $p_{\mathrm{A}}(K,N) = p_{\mathrm{S}}(K,N)$. Therefore the difference between the two reproduction rates in (3) is

$$\sum_N \sum_{K=1}^{N-1} p_{\mathrm{S}}(K,N)\big(\Pi_{\mathrm{S}}(K,N) - \Pi_{\mathrm{A}}(K,N)\big). \tag{4}$$

Since the game is an N-player prisoners' dilemma game, it must be that $\Pi_{\mathrm{S}}(K,N) - \Pi_{\mathrm{A}}(K,N) > 0$ for all K and N, and hence the expression in (4) must be positive. It follows that the growth rate of the population of selfish individuals exceeds that of the population of altruists at all times. Therefore, the limiting value of the proportion of altruists in the population is zero. □

3.4.1. *Where* not *to look*

It is important to understand that this "Iron Rule" does not tell us that evolutionary selection must eliminate altruistic behavior. The usefulness of Theorem 3 is that it tells us where *not* to look for the evolutionary foundations of such behavior. If we are looking for environments in which cooperative behavior is sustained by group selection, we should expect that at least one of the following is NOT true.

- The game that determines long term reproduction rates is an N-person prisoners' dilemma.
- The matching process that forms groups is "random".

3.5. *Haystacks and the iron rule*

In the haystack models of Maynard Smith and of Cohen and Eshel, communities are formed by independent random draws and the game played by individuals within each community seems to be an N-person prisoners' dilemma. Nevertheless, we have seen that in these models a population of altruists can survive evolutionary selection. How do these populations escape the reach of the "Iron Rule of Selfishness?"

The game that is specified in the Iron Rule is the game played between community founders, in which the payoffs are measured by the number of descendants at the time when the community is dissolved. For Maynard Smith's mice, the payoff matrix for this game is given in Table 2 above. In the case where $K < 1$, this game is a prisoners' dilemma and, as we have seen, cooperators will eventually disappear. If $K > 1$, then the game is not a prisoners' dilemma, but a stag hunt. Thus the conditions of the Iron Rule are not satisfied, and indeed its conclusion does not follow.

The way in which Cohen and Eshel's linear public goods model eludes the strictures of the Iron Rule is especially instructive. In this model, group formation is non-assortative. Furthermore, the number of offspring that any individual produces is the payoff in a multi-player prisoners' dilemma game played among contemporaries in the same group. Nevertheless, cooperative behavior can be sustained if groups spend sufficient time together before dispersal.

The reason that the Iron Rule is not violated is that if time to dispersal is long, the relevant game is not a prisoners' dilemma. In a group where all others are cooperators, a single defecting founder's defecting descendants would slow the growth of the group so that it would eventually be overtaken by a group consisting only of cooperators.

But why can't the Iron Rule be applied at times shortly before dispersal to individuals within a group? If the time to dispersal is short, then the game in which payoffs are descendants at dispersal time will be a prisoners' dilemma. But this game is played between individuals in the same group who are all descendants of the initial population. Matching among these individuals is decidedly not non-assortative. Thus, for the players who are matched non-assortatively, the game is not a prisoners' dilemma, while for the players for whom the game is a prisoners' dilemma, matching is not non-assortative.

Another instructive way of looking at the Cohen–Eshel game is to note that if we measure payoffs of each individual by the number of her own offspring, then the game is a multi-person prisoners' dilemma. But current rates of reproduction are not a proper measure of reproductive success. One's long run reproductive success depends not only on the number of one's own offspring, but on the rate at which these offspring, in turn, will reproduce. In the Cohen–Eshel model, the long-term reproductive value of an additional offspring depends on the proportion of altruists that are expected to be in one's group for the duration of survival of this group. In a population of altruists, an individual could increase her current reproduction by switching to the selfish strategy. But over time, her selfish descendants will slow the rate of reproduction for each other and if groups are sufficiently long-lived, the number of her descendants at the time her group disperses will be lower than it would have been had she remained an altruist.

3.6. Migration and stochastic extinction

Haystack models are artificial in that they assume that groups persist in perfect isolation until they are simultaneously disbanded. More realistic models would allow some migration between groups and would have asynchronous extinctions and resettlement. Such models have been studied, with results that are qualitatively similar to those of the haystack models. Ilan Eshel (1972), R. Levins (1970), Bruce Levin and William Kilmer (1974) and Scott Boorman and Paul Levitt (1980) consider stochastic dynamic models of group selection, in which selfish individuals reproduce more rapidly than altruists within their own group, but where groups face a probability of extinction that increases with the proportion of their members who are selfish. Locations in which extinction has occurred are reoccupied by the descendants of a random selection from the population at large. In the Levins and Boorman–Levitt models, monomorphic populations of altruists are not stable, but polymorphism is favored if the difference in extinction rates between altruistic and selfish groups is large enough relative to the selective pressure within groups. Eshel adds random migration between groups to his model and finds that if the migration rate is sufficiently small, then with probability one, the population will fix at a monomorphic population of altruists, and for larger migration rates the population will fix at a monomorphic selfish population. Levin and Kilmer (1974) conducted Monte Carlo simulations of a model similar to that proposed by Eshel[7] and found that altruism emerged when founding populations were no larger than 25 individuals and migration rates no larger than 5% per generation.

3.7. *Relative and absolute payoffs*

Some confusion in the debate on group selection has resulted from the fact that there exist games in which, paradoxically, *cooperate* is a dominant strategy, even though *defectors* always receive higher payoffs than cooperators. For example, consider N-player linear public goods game in which x is the fraction of cooperators in the population, the payoffs are bx for each defector and $bx - c$ for each cooperator. Thus defectors always get higher payoffs than cooperators. But suppose that $b > \frac{b}{N} > c > 0$. If this is the case, then given the action of other players, any player will get a higher payoff if she cooperates than if she defects. To see this, notice that if K other players cooperate, a player will get $\Pi_{\mathrm{C}}(K, N) = b\frac{K+1}{N} - c$ if she cooperates and $\Pi_{\mathrm{D}}(K, N) = b\frac{K}{N}$ if she defects. Thus we have $\Pi_{\mathrm{C}}(K, N) - \Pi_{\mathrm{D}}(K, N) = \frac{b}{N} - c > c$ and so *cooperate* is a dominant strategy.

David S. Wilson (1979) noticed this interesting case and argues for its significance. Wilson suggested that someone who cooperates when $b > c > \frac{b}{N}$ be called a *strong altruist* and someone who cooperates $\frac{b}{N} > c$ but not when $\frac{b}{N} < c$ be called a *weak altruist*.[8] Thus, in Wilson's terms, a strong altruist will cooperate even if doing so reduces both his *absolute* payoff and his relative payoff. A weak altruist will cooperate if doing so increases his absolute payoff, even if doing so reduces his payoff relative to that of other members of his group. Wilson maintains that "many, perhaps most, group-advantageous traits such as population regulation, predation defense, and role differentiation" may be explained by weak altruism. Wilson argues that individual selection models will incorrectly predict that weak altruistic behavior will be selected

[7] Eshel's model has asexual reproduction. The Levin–Kilmer model, like those of Levins and of Boorman–Levitt models has sexual diploid reproduction.

[8] As remarked in Section 2.2.2, Wilson formulates the game slightly differently. The conditions stated here are equivalent to his when Wilson's game is recast as an equivalent linear public goods game.

against, while properly constructed group selection models will predict selection *for* such behavior.

Alan Grafen (1984) suggests that Wilson's use of the term weak altruism for behavior that is consistent with maximization of absolute payoffs is misleading. According to Grafen,

> Another source of misunderstanding arises from the use of the word 'altruism'. As we noted earlier, altruism will not evolve in simple one-generation groups that are formed at random from the population . . . (Wilson, Cohen and Eshel and others) . . . redefined altruism to refer to relative success within the group rather than absolute success. . . . Under the 'relative' definition, 'altruism' can spread. Wilson calls the acts that are altruistic under the relative definition, but not under the 'absolute' definition, 'weakly altruistic'. An alternative I prefer is 'a self-interested refusal to be spiteful'.

The question of what to call the pursuit of absolute benefit at the expense of relative benefit is of some interest, but a more important question is whether such behavior will prevail under evolutionary dynamics. Cohen and Eshel (1976) answered this question for the case of haystack models. The answer is that in haystack models, where relative and absolute benefit are in conflict, absolute benefit tends to prevail. A more careful statement can be found as part of Cohen and Eshel's Theorem 2 above. In a generalized haystack model in which the linear public goods game is played within localities, if $\frac{b}{N} > c$ then if the length of time T between founding and dispersal groups is short, there will be a unique stable equilibrium and it is a population of cooperators only. If, however, T is sufficiently large, then there will be two distinct stable equilibria, one populated by cooperators only and one by defectors only.[9] Thus Cohen and Eshel's result as applied to "one generation groups formed at random from the population" is in full agreement with Grafen's statement. In equilibrium, individuals will "cooperate" if and only if the direct benefits that they get for themselves exceed the cost. In Wilson's language, strong altruism will be driven out, but weak altruism will prevail. In Grafen's language, altruism will not survive, but the surviving population will show a self-interested refusal to be spiteful. Somewhat more surprisingly, Cohen and Eshel also find that if groups have long persistence, there will exist two equilibria, one where all cooperate and one where all defect, even where cooperation is a dominant strategy in terms of absolute payoffs in the single-shot game.

3.8. *"Too stringent to be realistic?"*

There seems to be broad agreement with Maynard Smith's (1976) statement that the argument about the significance of group selection for altruism is "not quantitative, but

[9] Wilson (1979) claims that theoreticians, including Cohen and Eshel, "tend to lump" the cases of weak altruism and strong altruism since neither is selected for in standard population models. In the case of Cohen and Eshel, I believe that Wilson is mistaken. As we see from Theorem 2, Cohen and Eshel find sharply divergent results for the cases of "weak" and "strong" altruism.

qualitative". At least for some time, there also appeared to be agreement that conditions under which group selection could work were not plausible.

David S. Wilson 1975 said that

> recent models ... make it plausible that (group selection) can occur – the main question is to what extent the conditions for its operation (small group size, high isolation, high extinction rates) are met in nature. The current consensus is that the proper conditions are infrequent or at least limited to special circumstances ...

In his survey of the theory of group selection and kin selection, Alan Grafen (1984) asserted that

> the final consensus on these models was that the conditions for (them) to be successful were too stringent to be realistic.

Even the beleaguered V.C. Wynne-Edwards called it quits, at least temporarily.[10] In a 1978 article Wynne-Edwards reports that

> in the last 15 years, many theoreticians have wrestled with it and in particular with the specific problem of the evolution of altruism. The general consensus of theoretical biologists at present is that credible models cannot be devised by which the slow march of group selection could overtake the much faster spread of selfish genes that bring gains in individual fitness. I therefore accept their opinion. [Wynne-Edwards (1978)]

Levin and Kilmer (1974) seem to have been the first to explore the plausibility of the parameter values under which models of group selection with random matching can lead to altruism. They conducted Monte Carlo simulations of a model similar to Eshel's extinction model (1972) and report that

> Interdemic selection favoring an allele was able to override the effects of Mendelian selection operating against it and led to maintenance of this allele in either fixed or polymorphic states. However, with potentially realistic deme survival functions and significant levels of Mendelian selection, restrictive conditions were necessary for this result to be obtained. In our simulated population, genetically effective deme sizes of less than 25 and usually closer to 10 were required, and the rate of gene exchange, through flow migration could not be much greater than 5% per generation.

Wilson (1987) ran Monte Carlo simulations of a model similar to Maynard Smith's haystack model, with founding populations of two individuals, and with dispersal and rematching of the population at the end of a fixed length of time. Wilson drops Maynard Smith 's assumption that in populations with some genes for aggressive behavior,

[10] In a (1986) book, Wynne-Edwards attempts to reestablish his group-selectionist arguments on firmer theoretical ground.

all carriers of the gene for timidity are eliminated before the haystack population is dispersed. In Wilson's simulation, in each generation, an altruist reduces its own reproduction rate by c and contributes $b > c$ to the reproduction rate of a randomly selected other member of the group. As in the Eshel–Cohen model, a group stays together for a fixed, finite number of periods before dispersing and mating at random. But while reproduction is asexual in the Eshel–Cohen model, Wilson has sexual diploid reproduction. Wilson points out that if communities disperse after a single period, then the model is the same as Hamilton's model of kin-selection (1964), and Hamilton's rule applies; there will be a unique stable equilibrium, which will be populated by altruists if $b > 2c$ and by selfish individuals if $b < 2c$. When the number of generations is 5, his simulation results that indicate that with $b/c = 2.2$, there are two distinct equilibria, a stable polymorphic equilibrium with a population of 80% altruists and a stable monomorphic equilibrium where the entire population is selfish.

Theoretical considerations may help us to recognize circumstances under which Maynard Smith's haystack model and its generalizations would plausibly support a population of altruists. In the Maynard Smith model where each haystack population gets genetic material from just individuals, we find that a monomorphic population of altruists will be a stable equilibrium if at season's end, the number of mice produced in haystacks of altruists is more than twice the number of mice produced in haystacks of selfish mice. In the Cohen–Eshel extension, with N co-founders, in order for a monomorphic population of altruists to be a stable equilibrium, it must be that a single selfish individual in a community of altruists will have fewer descendants within that community at the time of dispersal than the *per capita* number of descendants of a community consisting entirely of altruists. Thus, if at the time the group disperses, the descendants of the selfish individual constitute the fraction s of its community, then it must be that groups consisting entirely of altruists have more than sN times as many inhabitants as groups that included a selfish individual among their founders. If, much as in Maynard Smith's model, descendants of a selfish individual dominate the population of their community quickly and thoroughly, then the purely altruistic groups would have to produce more than N times as many descendants as groups that included a selfish cofounder.

In haystack models, with durable groups, we have seen that when there is a stable equilibrium of cooperators, there typically exists another equilibrium comprised entirely of defectors. We need to be concerned about whether and how the system could move into the basin of attraction of an equilibrium of cooperators. One possibility is that payoffs to particular actions are likely to shift across time and space. As Wilson (1979) suggested, actions that are "strongly altruistic" in the current environment may have emerged as equilibrium actions in an environment where costs were smaller or benefits were greater so that these actions were once individually rational in single shot games. These equilibria might survive changes in costs or benefits such that self-interested play in single shot games no longer supports cooperation.

4. Assortative matching

In prisoners' dilemma games, everyone gets a higher payoff from playing with a cooperator than with a defector, but in any encounter, playing *defect* yields a higher payoff than playing *cooperate*. In a population where both types are equally likely to play with cooperators, defectors will enjoy higher expected payoffs. But if matching is assortative, so that cooperators have better chances of meeting cooperators than do defectors, the cost of cooperation may be repaid by a higher probability of playing a cooperative opponent.

4.1. Measures of assortativity

Suppose that a population is made up of two types of individuals and each of these individuals is matched with a partner. Let $x = (x_1, x_2)$ where x_1 is fraction of the population that is of type 1 and x_2 the fraction that is of type 2. Let $p_{ij}(x)$ be the conditional probability that an individual is matched with a type j, given that she, herself, is of type i. Since an individual is matched either to its own type or to the other type, it must be that $p_{11}(x) + p_{12}(x) = 1$ and $p_{21}(x) + p_{22}(x) = 1$. These equations imply that $p_{22}(x) - p_{12}(x) = p_{11}(x) - p_{21}(x)$. This equality allows us to define a useful measure of assortativity.

DEFINITION 4 ((Pairwise) Index of Assortativity). Let there be two types of individuals i and j and let $x = (x_1, x_2)$ where x_i is the proportion of type i's in the population. If individuals are matched in pairs, the index of assortativity $a(x)$ is the difference between the probability that an individual of type i is matched with its own type and the probability that an individual of type j is matched with a type i. That is, $a(x) = p_{11}(x) - p_{21}(x) = p_{22}(x) - p_{12}(x)$.

Sewall Wright (1921) defined assortativeness of mating with respect to a given trait as "the coefficient of correlation m between the two mates with respect to their possession of the trait". Cavalli-Sforza and Feldman (1981) interpret Wright's correlation as follows. "The population is conceived of as containing a fraction $(1 - m)$ that mates at random and a complementary fraction m which mates assortatively". With this interpretation, if the population frequency of a type is x, then the probability that an individual of that type mates an individual of its own type is $p(x) = m + x(1 - m)$. Wright's definition and that of Cavalli-Sforza and Feldman are seen to be equivalent where we take Wright to mean that m is the coefficient of correlation between indicator random variables for possession of the trait by mates.[11] It turns out that the definition of assor-

[11] Let I_i be an indicator variable that takes on value 1 if mate i has the trait and 0 otherwise. Wright's definition of the correlation coefficient between mates 1 and 2, is the correlation between the random variables I_1 and I_2. Thus we have $m = (E(I_1 I_2) - E(I_1)E(I_2))/(\sigma_1\sigma_2)$ where σ_i is the standard deviation of I_i. Now $E(I_1 I_2) = xp(x)$, and for $i = 1, 2$, $E(I_i) = x$ and $\sigma_i = \sqrt{x(1-x)}$. Therefore $m = (xp(x) - x^2)/x(1 - x)$. Rearranging terms, we find that this expression is equivalent to $p(x) = m + x(1 - m)$.

tativeness proposed by Wright and by Cavalli-Sforza and Feldman is equivalent to the special case of our definition where $a(x)$ is constant.

REMARK 1. Where there are two types of individuals and $a(x)$ is the index of assortativity,

- $p_{ii}(x) = a(x) + (1 - a(x))\, x_i$ for each i.
- $p_{ji} = a(x)(1 - x_i)$.

PROOF. The fraction of all partnerships in which a type i is matched to a j is $x_i p_{ij}(x)$ and the fraction of all partnerships in which a type j is matched to a type i is $x_j p_{ji}(x)$. Since these are just two different ways of counting the same encounters it must be that $x_i p_{ij}(x) = x_j p_{ji}(x)$. From the definition of $a(x)$, we have $p_{ii}(x) = a(x) + p_{ji}(x)$. We also have $p_{ij}(x) = 1 - p_{ii}(x)$ and $x_1 + x_2 = 1$. Simple manipulations of these equations yields the claimed results. □

The simplest, and perhaps most useful, way to generalize the index of assortativity from groups of two members to groups of arbitrary size is to simply restate the pairwise definition in terms of expected proportions. Thus for an individual of type i, let p_{ij} to be the *expected proportion of other group members* who are of type j. Where group size is two, this expected proportion is simply the conditional probability that one's partner is a type j, given that one's own type is i. It remains the case, as with pairwise matching that $p_{11}(x) - p_{21}(x) = p_{22}(x) - p_{12}(x)$.

DEFINITION 5 ('Generalized' Index of Assortativity). Where there are two types of individuals and groups are of size N, for an individual of type i, let $p_{ij}(x)$ be the expected proportion of the $N - 1$ other group members who are of type j. The index of assortativity is defined as $a(x) = p_{11}(x) - p_{21}(x) = p_{22}(x) - p_{12}(x)$.

Where there are more than two possible types, we could define an index of assortativity between any two types as previously. In general, the index of assortativity between one pair of types need not be the same as that between another.

4.2. Hamilton's kin selection theory

Families are among the most conspicuous examples of non-randomly formed groups. William G. Hamilton (1964) developed a theory that predicts the strength of benevolent interactions between relatives based on their degree of relatedness. Maynard Smith (1964) conferred the name *kin selection theory* on this theory, while Dawkins (1989) gave it the evocative name *theory of the selfish gene*.

Biologists define the coefficient of relatedness between two individuals to be the probability that the alleles found in a randomly selected genetic locus in the two individuals are inherited from the same ancestor. In a population without inbreeding, the coefficient of relatedness is one half for full siblings, one fourth for half siblings, and

Table 3
Hamilton's help game

		Player 2	
		C	D
Player 1	C	$b - c$	$-c$
	D	b	0

one eighth for first cousins. According to Hamilton's theory, evolutionary dynamics selects for individuals who are willing to help a genetic relative if (and only if) if the increase in reproductive value, b gained by the beneficiary, multiplied by the *coefficient of relatedness* r between the two relatives exceeds the cost in reproductive value c to the helper. The following "maxim" has come to be known as *Hamilton's rule*.

DEFINITION 6 (Hamilton's Rule). Help someone whose coefficient of relatedness to you is r if and only if $br > c$.

Hamilton's work on kin selection came almost 10 years before Smith and Price (1973) introduced formal game theory to biologists. Therefore he did not think of the interaction between relatives as a game, but it is instructive to model Hamilton's interactions as a two-person game. In Hamilton's model, each player can choose whether to "cooperate" by helping the other or to "defect" by not helping. A player who helps the other player reduces her own reproductive success by an amount $c > 0$, but increases that of the other player by $b > c$. The payoff matrix for this game is as in Table 3. When $b > c > 0$, we see that Hamilton's help game satisfies the conditions for a two-person prisoner's dilemma. Since $b - c > 0$, both players are better off when both cooperate than when both defect. Given the other player's action, a player is always better off defecting than cooperating, since $b > b - c$ and $0 > -c$. As later discussion will show, Hamilton's help games are *special cases* of a prisoners' dilemma. There is a large class of prisoners' dilemma games which have quite different evolutionary dynamics from this special class.

A two-person linear public goods might appear to differ from a Hamilton's help game. In a linear public goods game a cooperator incurs a cost to produce "public benefits that help the other player *and also benefit herself*", while in the Hamilton game the other player is the only beneficiary of a helper's efforts. But simple linear transformations of payoffs allow one to express every Hamilton game as a linear public goods game and *vice versa*. For example, consider a Hamilton game in which a helper can confer a benefit b on the other player at cost c to herself. This game has the same payoff matrix as a linear public goods game in which *each* the payoff to a defector is $b'x$ and the payoff to a cooperator is $b'x - c'$, where x is the fraction of contributors and where $b' = 2b$ and $c' = b + c$. Similarly a two-player linear public goods game, with parameters b' and c' will be equivalent to a Hamilton help game with parameters $b = b'/2$ and

$c = c' - (b'/2)$. This game is a prisoners' dilemma if and only if $b > c > 0$, which is equivalent to $b' > c' > b'/2$.

4.2.1. Index of assortativity for relatives

In later work, Hamilton (1975) recognized that his theory of kin selection could usefully be understood as a special case of assortative matching of partners in social interactions. It is helpful to see just how this is done by calculating the index of assortativity between prisoners' dilemma playing siblings who inherit their type by copying one of their parents.

We follow Hamilton (1964) in considering a simplified version of genetics, known to biologists as *sexual haploidy*. Most animals, including humans, are sexual diploids. A sexual diploid carries two alleles in each genetic locus, one of which is inherited from its mother and one from its father. These two alleles jointly determine those individual characteristics governed by this locus. A sexual haploid has only one allele at each locus. This allele is a copy of the allele in the corresponding locus of one of its parents, chosen at random. Sexual haploidy occurs as a genetic process among some living organisms, but is of special interest in the theory of cultural transmission since it is formally identical to a theory in which for a specified behavior, a child randomly selects one of its parents to copy.[12]

Suppose that individuals can adopt one of two possible strategies, cooperate or defect, in games played with their siblings. Each child is able to observe the type of its father and of its mother and copies one or the other with probability $1/2$; independently of the choice made by its siblings. Suppose further that parents mate monogamously and independently of their strategy in games with siblings.

Let x be the proportion of cooperators in the entire population. If a child is a cooperator, then with probability $1/2$ its sibling will have copied the same parent. In this case, the sibling must be a cooperator. With probability $1/2$, the sibling role will have copied the other parent. Since parents are assumed to mate independently of their strategies, the probability that the other parent is a cooperator is x. Therefore the probability that a randomly chosen sibling of a cooperator is also a cooperator is

$$p_{cc}(x) = \frac{1}{2} + \frac{1}{2}x. \tag{5}$$

If a child is a defector, then its sibling will be a cooperator only if the sibling's role model is different from the defector's. With probability $1/2$, the two siblings will have different role models, and given that they have different role models, the probability that the other parent is a cooperator is x. Therefore the probability that a randomly chosen

[12] Similar techniques can be applied and similar results obtained in the study of monomorphic equilibria in kin selection models with diploid sexual reproduction. For details, see Bergstrom (1995) or Boorman and Levitt (1980).

sibling of a defector is a cooperator is

$$p_{dc}(x) = \frac{1}{2}x. \tag{6}$$

Notice that in a family of N siblings, $p_{cc}(x)$ and $p_{dc}(x)$ are equal to the expected proportion of an individual's siblings who are cooperators, conditional on that individual being a cooperator or a defector, respectively. Therefore the index of assortativity between full siblings is

$$a(x) = p_{cc}(x) - p_{dc}(x) = \frac{1}{2}. \tag{7}$$

Thus we see find that with non-assortative monogamous mating, the index of assortativity between siblings is constant and equal to their coefficient of relatedness, $r = 1/2$.

Similar calculations show that the index of assortativity between other related individuals is equal to their degree of relatedness. For example, the index of assortativity between half-siblings is $1/4$ and the index of assortativity between first cousins is $1/8$. Bergstrom (2001) calculates the index of assortativity for siblings under a variety of more general assumptions. For example, if parents mate assortatively, with an index of assortativity of mating m, then the index of assortativity between full siblings is $(1+m)/2$. If with some probability v a child copies neither of its parents, but a randomly chosen stranger, the index of assortativity is $v(1 + m)/2$. That paper also calculates indexes of assortativity for children of polygamous marriages, and for cases where children preferentially copy the mother or the father.

4.3. Evolutionary dynamics with assortative mating

4.3.1. The linear public goods game

We can now investigate the evolutionary dynamics of populations of prisoners' dilemma players under assortative mating. The effect of assortative mating on expected payoffs is particularly easy to calculate when payoffs depend linearly on the proportion of cooperators in the group as in Haldane's N-player linear public goods game. Let x be the fraction of cooperators and $1 - x$ the fraction of defectors in the entire population. Define $p_{cc}(x)$ as the expected proportion of cooperators that a cooperator finds among other members of her group and $p_{dc}(x)$ as the expected proportion of cooperators that a defector finds in her group. Recalling Equations (1) and (2), the expected payoff of a cooperator is $p_{cc}(x)b - c'$ and the expected payoff of a defector is $p_{dc}(x)b$. Therefore the difference between the expected payoff of cooperators and that of defectors is just

$$p_{cc}(x)b - c' - p_{dc}(x)b = a(x)b - c' \tag{8}$$

where $a(x)$ is the index of assortativity.

Equation (8) generalizes Hamilton's rule from linear pairwise interactions to the N player linear public goods game with voluntary provision of public goods. In this generalization, the index of assortativity plays the same formal role that the coefficient of relatedness plays in kin selection theory. In the case of kin selection theory, the index of assortativity $a(x)$ is constant for all x and equal to the coefficient of relatedness r between any two players.

If $a(x) = a$ is constant, then except for the knife-edge case where $ab = c$, there will be a unique stable equilibrium. If $a > b/c$, then so long as both types are present, the proportion of cooperators will grow relative to that of defectors. If $a < b/c$, the reverse is true. Thus the unique stable equilibrium is a population made up entirely of cooperators if $a > c/b$ and a population made up entirely of defectors if $a < b/c$.

If $a(x)$ is variable, then it is possible that there may be more than one equilibrium, or there may be a polymorphic equilibrium with some individuals of each type. In Section 4.4 we analyze an interesting example in which $a(x)$ is variable and where there is a stable polymorphic equilibrium.

4.3.2. Dynamics with nonlinear payoff functions

Alan Grafen (1979) and Gordon Hines and Maynard Smith (1979) show that Hamilton's rule is not correct in general for the wider class of games in which the costs of helping and the benefits of being helped may depend on the actions taken by both players. Bergstrom (1995) classifies two-player non-linear games according to whether there is complementarity or substitutability between actions and shows the way that equilibrium is altered from the Hamilton's rule predictions in each of these cases.

We follow Rappaport and Chammah (1965), in denoting the payoffs (Table 4) in a general prisoners' dilemma game by R (reward) for mutual cooperation, P (punishment) for mutual defection, T (temptation) to a defector whose opponent cooperates, and S (sucker's payoff) to a cooperator whose opponent defects.

This game is a prisoners' dilemma whenever $T > R > P > S$.[13] In the case of Hamilton's help game, described by Table 2 in Section 4.2, we have $T = b$, $R = b - c$, $P = 0$, $S = -c$. It follows that for Hamilton's game, $R + P = T + S = b - c$. Not every prisoners' dilemma game has this property. There are prisoners' dilemma games

Table 4
Payoff matrix

		Player 2	
		C	D
Player 1	C	*R*	*S*
	D	*T*	*P*

[13] Some writers use a definition that adds the additional restriction that $2R > T + P$ which ensures that mutual cooperation yields a higher *total* payoff than the outcome where one player cooperates and the other defects.

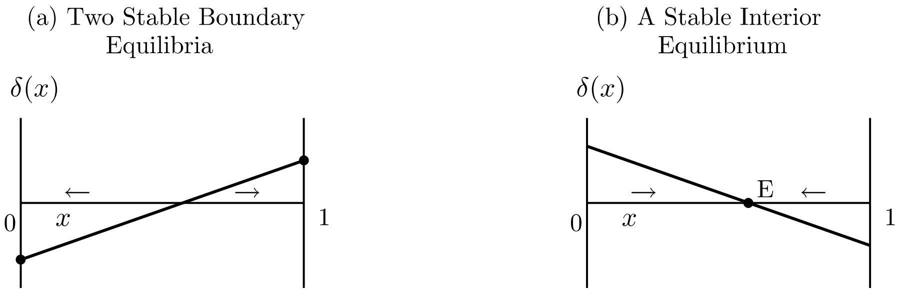
(a) Two Stable Boundary Equilibria
$\delta(x)$
←
→
0
x
1
(b) A Stable Interior Equilibrium
$\delta(x)$
→
E
←
0
x
1

Figure 2. Dynamics of prisoners' dilemma.

in which $R + P > T + S$ and some in which $R + P < T + S$. The evolutionary dynamics of each of these prisoners' dilemma games are qualitatively different from those of Hamilton's help game.

Let x be the fraction of cooperators in the population, $p_{cc}(x)$ the probability that a cooperator is matched with a cooperator and $p_{dc}(x)$, the probability that a defector is matched with a cooperator. Then the expected payoff to a cooperator is:

$$\begin{aligned} p_{cc}(x)R + \bigl(1 - p_{cc}(x)\bigr)S &= S + p_{cc}(x)(R - S) \\ &= S + a(x)(R - S) + x\bigl(1 - a(x)\bigr)(R - S) \qquad (9) \end{aligned}$$

where the latter equation follows from Remark 1.

The expected payoff to a defector is:

$$\begin{aligned} p_{dc}(x)T + (1 - p_{dc})P &= P + p_{dc}(T - P) \\ &= P + x\bigl(1 - a(x)\bigr)(T - P) \qquad (10) \end{aligned}$$

where again the latter equation follows from Remark 1. If we subtract the expression in Equation (10) from that in Equation (9), we can express the difference between the expected payoff to a cooperator and that to a defector as a function of x:

$$\delta(x) = S - P + a(x)(R - S) + x\bigl(1 - a(x)\bigr)\bigl(R + P - (S + T)\bigr). \qquad (11)$$

Equation (11) can be used to characterize the equilibria, under the assumption of monotone dynamics (see Section 2.4.1), of any symmetric two-player, two-strategy games with assortative matching.[14]

Where $a(x) = a$ is constant, we see from Equation (11) that the difference between the payoffs to the two strategies is linear in the proportion x of cooperators in the population. In this case, we see that $\delta(0) = aR + (1 - a)S - P$ and $\delta(1) = R - (aP + (1 - a)T)$. A simple calculation shows that $\delta(1) - \delta(0) =$

[14] Though most of our discussion focusses on prisoners' dilemma, this formula applies as well to games without a dominant strategy, such as *chicken*, and the *stag hunt*.

$(1 - a)(R + P - S - T)$. Thus the graph of $\delta(x)$ slopes upward if $R + P > S + T$, downward if $R + P < S + T$, and is horizontal if $R + P = S + T$. It could happen that $\delta(0)$ and $\delta(1)$ are both positive, in which case there is a unique stable equilibrium populated entirely of cooperators or both negative, in which case there is a unique stable equilibrium populated entirely by defectors. But there are also two other interesting cases. In Figure 2(a), where $\delta(0) < 0$ and $\delta(1) > 0$, there are two distinct "monomorphic" equilibria, one consisting of cooperators only and one consisting of defectors only. In Figure 2(b) where $\delta(0) > 0$ and $\delta(1) < 0$, neither monomorphic population is stable and there is a unique stable "polymorphic" equilibrium at the point E.

4.4. Assortative matching with partner choice

We can expect to see assortative matching if individuals have some evidence of each others' types and some choice about with whom they match. In a multiplayer prisoners' dilemma game, everyone would rather be matched with cooperators than with defectors. If players' types were perfectly observable and if groups are able to restrict entry, then groups of cooperators would not admit defectors, and so the two types would be strictly segregated. But suppose that detection is less than perfectly accurate.

Bergstrom (2001) presents a model in which players are labeled with an imperfect indicator of their type. The indicator might be a reputation based on partial information or a set of behavioral cues, or perhaps the result of a psychological test. Assume that with probability $\alpha > 1/2$, a cooperator is correctly labeled as a cooperator and with probability $1 - \alpha$ is mislabeled as a defector. Assume that with probability $\beta > 1/2$, a defector is correctly labeled and with probability $1 - \beta$ is mislabeled as a cooperator.

Everyone sees the same labels, so that at the time when players choose partners there are only two distinguishable types: players who appear to be cooperators and players who appear to be defectors. Although everyone realizes that the indicators are not entirely accurate, everyone prefers to match with an apparent cooperator rather than an apparent defector. Therefore, with voluntary matching, there will be two kinds of groups, those made up entirely of apparent cooperators and those made up entirely of apparent defectors.

In this model, in contrast to the case of kin selection, the index of assortativity varies with the proportion of cooperators in the population. If we graph $a(\cdot)$ as a function of x, the graph looks qualitatively like Figure 3.[15]

There is a simple intuitive explanation for the fact that $a(0) = a(1) = 0$. In general, a cooperator is more likely to be matched with a cooperator than is a defector because a cooperator is more likely to be labeled a cooperator than is a defector. But if x is small, so that actual cooperators are rare, the advantage of being matched with an apparent cooperator is small because almost all apparent cooperators are actually defectors who

[15] In Bergstrom (2001), I exhibit a closed form expression for $a(x)$ and show that $a(0) = a(1) = 0$, $a'(0) > 0$, $a'(1) < 0$ and $a''(x) < 0$ for all $x \in [0, 1]$.

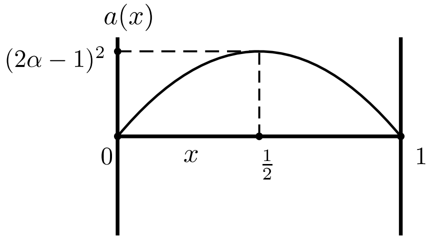

Figure 3. Graph of $a(x)$ where $\alpha = \beta$.

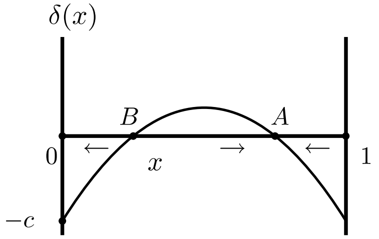

Figure 4. Graph of $\delta(x)$ for additive prisoner's dilemma.

have been mislabeled. Similarly, when x is close to one, defectors are rare, so that most apparent defectors are actually cooperators who have been mislabeled. In the latter case, even if a defector is labeled a defector, his chance of getting matched with a cooperator are good. Thus in the two extreme cases, where x approaches zero and where x approaches one, the chances of being matched with a cooperator are nearly the same for a defector as for a cooperator.

Recall from Equation (8), that in the Haldane linear multiperson prisoners' dilemma game, the difference between the expected payoff of cooperators and that of defectors is simply $\delta(x) = a(x)b - c$ where x is the fraction of cooperators in the population and $a(x)$ is the index of assortativity. Figure 4 shows the graph of $\delta(x)$ for a case in which $\delta(x)$ takes some positive values. As we see from the graph, under monotone dynamics there are two locally stable equilibria. One of these equilibria occurs where $x = 0$ and the other is at the point marked A. For any level of x to the left of the point B or to the right of the point A, $\delta(x) < 0$ and so x, the proportion of cooperators in the population, would decline. For any level of x between the points A and B, $\delta(x) > 0$ and so in this region x would increase.

For Prisoners' Dilemma games with additive payoffs, $\delta(x) = a(x)b - c$. We have shown that $a(0) = a(1) = 0$, $a'(0) > 0$, $a'(1) < 0$, and $a''(x) < 0$ for all x between 0 and 1. It follows that $\delta(0) = \delta(1) < 0$, $\delta'(0) > 0$, and $\delta'(1) < 0$, and $\delta''(x) < 0$ for all x between 0 and 1. The fact that $\delta''(x) < 0$ on the interval [0, 1] implies that the graph of $\delta(x)$ is "single-peaked" as in Figure 4. Where this is the case, and if $\delta(x) > 0$

for some x, there must be exactly one stable polymorphic equilibrium and one stable monomorphic equilibrium with defectors only.

An earlier model by Robert Frank (1987) also explores the evolutionary dynamics in a population of cooperators and defectors.[16] In Frank's model, each member of each type projects a "signal of trustworthiness" that is a random draw from a continuous signal distribution. The two types draw from different distributions, whose supports overlap, but where the higher one's signal is the more likely it is that one is a cooperator. Each individual has the option of matching with a partner or of having no partner. Partners play a game of prisoners' dilemma. Those who choose to have no partner are assumed to receive the same payoff as that received by a defector matched with a defector. Players understand the game, including the payoff values and statistical distributions of payoffs and can rationally calculate their own optimal responses. Since each player prefers those who project higher signals, every individual will be matched with someone who projects approximately the same signal. In equilibrium, cooperators who project a signal lower than some critical value realize that the partners that they can attract are so likely to be defectors that it is better to stay unmatched. Frank shows that for this model there is a unique stable equilibrium and it occurs with a polymorphic population that includes both cooperators and defectors.

Skyrms and Pemantle (2000) explicitly model the dynamic formation of group structure by reinforcement learning. Individuals begin to interact at random to play a game. The game payoffs determine which interactions are reinforced and a social network emerges. They report that social interaction groups that tend to form in their model consist of small interaction groups within which there is partial coordination of strategies.

4.5. Assortative matching induced by spatial structure

The reason that evolution selects for individuals who value their siblings' well-being is that two siblings have a high probability of carrying the same genetic program. Hence an individual who is programmed to be kind to his brother is likely to be the beneficiary of a kind brother. Similarly, if neighbors have a significant probability of sharing the same role model, those who cooperate with neighbors may enjoy a higher likelihood of benefiting from neighborly cooperation than those who act selfishly.

Evolutionary biologists have stressed the importance of spatial structure on the spread of mutations, genetic variation and the formation of species. Wright (1943) studied the degree of inbreeding in a model in which a population is distributed uniformly over a large area, but individuals are more likely to find mates who live nearby. Kimura and Weiss (1964) studied genetic correlations in a one dimensional "stepping stone model" in which there is an array of colonies along a line and where "in each generation an individual can migrate at most 'one step' in either direction" and extended this model to colonies located on two and three dimensional lattices.

[16] Frank calls them "honest" and "dishonest" types.

More recent authors have explored the dynamics of a population of agents located on a spatial grid, who repeatedly play a game with their neighbors and who may switch their strategies either deterministically or stochastically in response to their observations of the payoffs realized by themselves and their neighbors. Nowak and May (1993) ran computer simulations with a deterministic model of prisoners' dilemma playing agents located on a two-dimensional grid. The grid is initially populated with some assortment of cooperators and defectors. In each round, each individual uses its preassigned strategy in a game of prisoners' dilemma with each of its immediate neighbors. After this round, each site is occupied by its original owner or by one of its neighbors, depending on who had the highest score in the previous round. Their simulations show that this process can generate chaotically changing spatial patterns in which the proportions of cooperators and defectors fluctuate about long-term averages.

Bergstrom and Oded Stark (1993) model a population of farmers located on a road that loops around a lake. Each farmer plays prisoners' dilemma with his two adjacent neighbors, using one of the two strategies cooperate or defect. The farmers' sons observe the strategies and payoffs of their fathers and their immediate neighbors and imitate the most successful of these individuals. For this setup, it turns out that any arrangement of cooperators and defectors will be stable if cooperators appear in clusters of three or more and defectors in clusters of two or more. Bergstrom and Stark show that if the sons do not pay attention to their fathers, but copy the more successful of their father's neighbors, then some patterns of behavior will "move in a circle" around the lake. For example, if there are at least eight farmers on the road, a pattern of the form *CDCCC* would move clockwise around the lake, moving by one farm in each generation. Thus a long-lived chronicler, who observed behavior at a single farm would see "cyclic behavior" in which spells of cooperation are interrupted by defection according to a regular temporal pattern.

Eshel, Larry Samuelson and Avner Shaked (1998) present a thorough analysis of the circular setup considered by Bergstrom and Stark. For the nonstochastic case, they show that in addition to an equilibrium with defectors only, there are stable equilibria in which some cooperators and some defectors survive and that in all such equilibria, at least 60 per cent of the population must be cooperators. They also show that if the initial distribution of cooperators and defectors is determined by independent random draws, then as the size of the population gets large, the probability that the initial distribution is in the basin of attraction of one of the equilibria that includes cooperators approaches unity.

Surprisingly, Eshel et al. were able to show that when there is a positive probability of mutations, in the limit as the mutation rate becomes small, the only stationary states that have positive probability are the ones in which at least 60 percent of the population are cooperators. As the authors explain:

> One's initial impression might be that mutation should be inimical to Altruists because a mutant Egoist will thrive and grow when introduced into a collection of Altruists, while a lone Altruist will whither and die when introduced into a collection of Egoists. ... Altruists can thus invade a world of Egoists with only a local

burst of mutation that creates a small string of Altruists, which will then subsequently grow to a large number of Altruists. Mutations can create small pockets of egoism, but these pockets destroy one another if they are placed too close together, placing an upper bound on the number of Egoists that can appear.

Although the structure of equilibrium sets in the Bergstrom–Stark model and in the Eshel–Samuelson–Shaked model seem too complicated and diverse for a simple measure of assortativity to be of any use, Eshel, Emilia Sansone and Shaked (1999) constructed a model of prisoners' dilemma players on a line in which, quite remarkably, the dynamics depend on the index of assortativity for a specific critical configuration of cooperators and defectors. The model starts with an initial configuration of cooperators and defectors. In each period, each individual plays a prisoners' dilemma game with each of her k nearest neighbors. A player will not change strategies from one period to the next if her two nearest neighbors use the same strategy that she uses. But one of these neighbors uses a different strategy, she will observe the average realized payoffs of cooperators and of defectors who are within n positions of herself. She will randomly adopt a strategy for the next period, where the probability that a strategy is adopted is proportional to the average payoff of those whom she observes using that strategy. The authors show that the long run fate of this system depends entirely on what happens at a frontier between long strings of individuals of each type. From this configuration, one can calculate the probability that a defector situated at the boundary will switch to cooperation and the probability that a cooperator situated at the boundary will switch to defection. These two probabilities depend on comparisons of the average payoffs of cooperators and of defectors who are located within n positions of the boundary between a long strong of cooperators and a long string of defectors. The dynamics is a simple random walk in which the limiting outcome is a population of cooperators or of defectors, depending on whether defectors are more likely to switch than cooperators or *vice versa*.

In the Eshel, Shaked, Sansone model the critical observers on the frontier see their own payoffs and the payoffs to their n neighbors. Each observed individual plays prisoners' dilemma with her k nearest neighbors. Since the observed defectors are located somewhere in a string of defectors and the observed cooperators are located somewhere in a string of cooperators, the cooperators enjoy the advantage of a larger proportion of encounters with cooperators than that experienced by defectors. If k, the number of opponents played in each direction is large and if n the distance over which the boundary individuals observe payoffs, this advantage will be slight since both the cooperators and defectors that are observed will be near the boundary and will play nearly equal numbers of cooperators and defectors. On the other hand, if n is large relative to k, then the average payoff of the observed cooperators will be close to the payoff in a community of cooperators only and the average payoff of the observed defectors will be close to the payoff in a community of defectors only.

The defectors would of course get higher payoffs if they played against the same number of cooperators as did the cooperators, but in this spatial setup, the defectors will be matched with more defectors than the cooperators and individuals living further from the frontier will have larger proportions of their neighbors being of their own type. The larger n is relative to k, the greater the proportion of observed neighbors who play their own type. The authors find expressions for the proportions of cooperators and of defectors encountered by those members of each type who can be observed by the frontier individual. From these calculations they produce an explicit function $r(k, n)$ that corresponds exactly to the *index of assortativity* as we have defined it. In the special case where the prisoners' dilemma game has the linear payoffs that we have earlier described as the linear public goods game, they observe that the outcome is exactly as would be predicted by Hamilton's rule where the coefficient of relatedness is $r(k, n)$. That is to say, cooperation will prevail if $r(k, n)b > c$ and defection will prevail if $r(k, n)b < c$.

5. Repeated games and group selection

5.1. Group selection from multiple Nash equilibria

"Punishment allows the evolution of cooperation (or anything else) in sizeable groups" by Robert Boyd and Peter Richerson (1992) is one of those rare scholarly titles that nudges readers' minds toward a productive line of thought.[17] In an earlier paper, Boyd and Richerson (1990) noticed that group selection is a highly plausible evolutionary mechanism where games with more than one Nash equilibrium are played within relatively distinct subpopulations. They suggested that group selection is likely to be effective "if processes increasing the frequency of successful strategies *within* groups are strong compared to rate of migration among groups" and if "individuals drawn from a single group make up a sufficiently large fraction of newly formed groups". In (1990), Boyd and Richerson succinctly explain the essence of group selection among alternative local Nash equilibria in the following words. "Viewed from the within-group perspective, behavior will seem egoistic, but the egoistically enforced equilibria with the greatest group benefit will prevail". In (1992), they strengthen the case for group selection by noting that within stable groups where individuals encounter each other repeatedly and can punish defections from a group norm, an extremely diverse range of results can be sustained as Nash equilibria.

Ken Binmore (1994b) observed that "If our Game of Life were the one-shot Prisoners' Dilemma, we should never have evolved as social animals". Binmore argues that the "Game of Life" is best modeled as an indefinitely repeated game in which reciprocal rewards and punishments can be practiced. As Binmore reminds us, this idea is not new. In the seventh century before Christ, Hesiod stated [Waugh (1929)] the maxim "Give to him who gives, and do not give to him who does not". David Hume (1978) says in language that is suggestive of modern game theory:

[17] Dawkins' *The Selfish Gene* is another member of this class.

> I learn to do service to another, without bearing him any real kindness, because I foresee, that he will return my service in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me and others. And accordingly, after I have serv'd him . . . he is induc'd to do his part, as foreseeing the consequences of his refusal. (p. 521)

Several game theorists in the 1950's nearly simultaneously discovered the *folk theorem*, which informs us that in indefinitely repeated games, almost all possible patterns of individual behavior can be sustained as Nash equilibria. For example, in the simple case of repeated prisoners' dilemma between two players, almost any intertemporal pattern of cooperation and defection on the part of each players can be maintained as a Nash equilibrium. The logic of the folk theorem is that in repeated games, almost any behavior can be induced as a Nash equilibrium by the threat of punishment for deviant actions. Individuals can be coordinated on a configuration of strategies by a stable, self-policing norm. Such a norm prescribes a course of action to each player conditional on the actions of the others and it includes instructions on how to punish any deviant player who violates his prescribed course of action. The punishments for each deviation must be sufficient to ensure that each individual takes the prescribed action.

Where the game is single-shot prisoners' dilemma, the theory of individual selection almost inevitably predicts universal defection, but in repeated games, even repeated prisoners' dilemma, individual selection theory leaves us an embarrassment of Nash equilibria and essentially no predictive power. As Boyd and Richerson (1992, 2001), Binmore (1992, 1994a, 1994b), and Sober and Wilson (1999) suggest, the stage is set for group selection to play a mighty role. Consider a population in which individuals are clustered into semi-isolated groups within which most of their social interactions occur. Within groups, these individuals play a repeated game that has many equilibria, some of which are better for all members than others. Binmore (1994b) suggests that we can expect groups using Pareto-superior equilibria to grow in size and number relative to the rest of the population and that eventually the groups that coordinate on Pareto-inferior equilibria will disappear. The transmission process may be speeded either by migrants who move to more successful groups and adopt local ways or by imitation. Boyd and Richerson (2001) propose that in geographically structured populations, imitation of behavior in successful neighboring groups is likely to greatly speed the spread of Pareto-superior equilibria.

5.2. How can costly punishment survive?

While the *folk theorem* goes a long way toward explaining the power of norms and punishment threats for maintaining a great variety of possible outcomes as Nash equilibria within communities, there remain some troubling details to be resolved in determining whether plausible evolutionary processes will sustain the punishment strategies needed to support all of the outcomes that folk theorem postulates. As Henrich and Boyd (2001) put it

> Many students of human behavior believe that large-scale human cooperation is maintained by threat of punishment. ... However, explaining cooperation in this way leads to a new problem: why do people punish noncooperators? ... Individuals who punish defectors provide a public good, and thus can be exploited by non-punishing cooperators if punishment is costly.

The standard game theoretic answer to this conundrum is that equilibrium strategies include instructions to punish others if they are "supposed to punish" and fail to do so. These instructions include a requirement to punish those who won't punish others when they are supposed to do so. In equilibrium, if you fail to perform your obligation to punish someone who doesn't do his task, you will be punished by someone else who in turn would be punished if he did not punish you, and so on *ad infinitum*. From an evolutionary point of view, this resolution seems unsatisfactory. Can we really expect that people or animals will keep track of their obligations to do nth order punishment for n greater than one or two? Moreover if the society is really in an equilibrium, deviations that require punishment will be rare and usually the result of a "mistake". Selection is likely to be very weak in such circumstances.

As Rajiv Sethi and R. Somanathan (to appear), point out in their survey paper "Understanding Reciprocity", "(The) problem of reciprocity being undermined by the gradual encroachment of unconditional cooperation is pervasive in the literature". Not only is it likely that punishment is costly in terms of direct payoffs. A strategy that involves unused punishments is, by any reasonable measure, more complex than a strategy that dictates the same actions in a world of cooperators but omits the punishment branch. Binmore and Samuelson (1992) present a model in which strategies are modeled as finite-state automata and complexity is measured by the number of states. They postulate that a mutant that gets the same payoff as the incumbents but is less complex will invade a population. This assumption eliminates the possibility that 'nice' strategies, such as tit-for-tat will be stable monomorphic equilibria.

Nowak and May (1998) introduce an evolutionary model in which individuals accumulate reputations. In each generation, a large number of pairs of individuals are selected randomly. One member of each pair is given a chance to play donor and the other is the potential recipient. Those who choose to donate reduce their wealth by c, while the recipient's wealth increases by $b > c$. Each player has an *image score* that starts out at 0 at the beginning of life and is incremented by one unit every time that she makes a donation. A strategy for any individual i takes the form of a threshold k_i, such that if given a chance to donate to a recipient with image score s, i will do so if and only if $s \geqslant k_i$. After the interactions for the current generation have taken place, members of this generation are replaced by their offspring, who inherit the strategies of their parents (but not their image scores). The number of offspring that a parent has is proportional to the wealth that she accumulates during the course of her life. Nowak and Sigmund run computer simulations of this model. They find that when the model is run for about 150 generations, almost all population members adopt a strategy of donating to everyone with an image score of 0 or higher. When these strategies are played out, this means that almost everyone donates at every opportunity. When Nowak and Sigmund add a very small rate of mutation to new strategies, the results are very different. According to Nowak and Sigmund,

> with mutation the population, long term simulations with mutation . . . show endless cycles. . . . defectors are invaded by discriminators, who only help players whose score exceeds some threshold. Next discriminators are undermined by unconditional cooperators. The prevalence of these indiscriminate altruists subsequently allows the return of defectors.

The Nowak–Sigmund model does not follow the course suggested by game theoretic constructions of punishment strategies. In their model, one's reputation improves whenever one makes a donation, regardless of whether the potential recipient has been generous or not. The kind of punishment strategy that the folk theorem suggests would be more like the following. Initially, everyone is in *good standing*. After each play, a person is in good standing if and only if she donated whenever she had a chance to donate to a person in good standing and she refused to donate whenever she had a chance to donate to a person not in good standing.

Bowles and Gintis (2000) build an evolutionary model of a population that includes some *shirkers* and some *reciprocators* who don't shirk and who, despite the fact that it is costly, will punish shirkers when they catch them shirking. Equilibrium in their model has a mixed population of workers and shirkers. However, they evade the problem of the evolutionary stability by not allowing the possibility of an invader who neither shirks nor punishes.

In "The viability of vengeance", Dan Friedman and Nirvikar Singh (1999) present a good discussion of the evolutionary stability of costly punishment. Friedman and Singh distinguish between punishment of group members and of outsiders. They suggest that within groups, one's actions are observed and remembered. A reputation for being willing to avenge actions harmful to oneself may be sufficient compensation for the costs of retribution. They propose that in dealing with outsiders, one is remembered not as an individual but as a representative of one's group. Accordingly, a willingness to avenge harm done by outsiders is a *public good* for one's own group since it deters outsiders from uncooperative behavior to group members. They propose that a failure to avenge wrongs from outsiders is punished (costlessly) by one's own group, through loss of status.

In their paper "Why punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas",[18] Henrich and Boyd (2001) present an ingenious theory of the viability of expensive vengeance. The authors suggest that "the evolution of cooperation and punishment are plausibly a side-effect of a tendency to adopt common behaviors during enculturation". They argue that since it is not possible to analyze and "solve" the complex social games that we play, imitation

[18] This paper is a contender with the earlier cited Boyd–Richerson paper for an "informative title award".

plays a large role in decision-making. Since observation of the realized payoffs of others is not always possible, much of this imitation takes the form of 'copy-the-majority' rather than 'copy-the-most-successful'.

Henrich and Boyd test this idea on a multi-stage game. The first stage of this game is a "Haldane" game in which each individual can choose whether to make a contribution to the group at a cost of c to himself and with a total benefit of b divided equally among all group members. Those who don't contribute share the benefits but don't pay the cost. With a small probability, individuals who intend to contribute mistakenly do not. The game has a second stage in which each individual decides whether or not to punish those who defected in the first stage. Punishing costs ϕ to the punisher and ρ to the punished, where $\phi < \rho < c$. There is a second punishing stage in which individuals decide, with the same cost structure, whether to punish those who have not punished the malefactors of the first stage. And a finite number of additional stages is constructed recursively. At each stage the authors suppose that there is some small probability of mistakes.

At each stage of the game, there are two possible strategies, cooperate or defect. In the first stage, cooperate means to contribute. In later stages, cooperate means to punish those who defected in the previous stage. The population evolves according to "replicator dynamics" applied separately to the strategy used in each stage. In particular the difference between the growth rate of cooperators and the growth rate of defectors for this stage is a weighted average of two differences: the difference between the average payoffs of cooperators and defectors in that stage and the difference between the fraction of the population who are cooperators and the fraction who are defectors. The latter difference reflects the force of conformism.

If the weight placed on conformism is sufficiently large, then of course any strategy, including cooperate and don't punish can be maintained, simply because an invader's payoff advantage would be overwhelmed by the conformist advantage of the incumbent strategy. But while placing some weight on copying the majority is plausibly adaptive, placing such a large weight does not seem likely to be so. The authors stress that

> ... stabilization of punishment is from the gene's point of view a maladaptive side-effect of conformist transmission. If there were genetic variability in the strength of conformist transmission and cooperative dilemmas were the *only* problem humans faced, then conformist transmission might never evolve.

The key to Henrich and Boyd's result is that it takes only a very small weight on conformity to maintain an equilibrium that supports punishment strategies. To see why, let us look at a version of the Henrich–Boyd model with only one punishment stage. Suppose that the population is initially one in which everyone tries to cooperate at the first stage and also in the punishment stage. Then the only defections observed will be mistakes (or possibly actions of a few mutants). Individuals who defect in the first stage will get lower payoffs than those who cooperate in the first stage because almost everyone is cooperating in the punishment stage by punishing first-stage defectors. Individuals who defect in the punishment stage by not punishing first stage defectors *will* get higher payoffs than those who cooperate by punishing first stage defectors, but only slightly higher since there are very few defections in the first stage. Since almost everyone is observed to cooperate in the second stage, even a very small coefficient on conformism will be sufficient to overcome this small payoff difference. Henrich and Boyd show that when higher levels of punishment are accounted for, an even smaller coefficient on conformism is sufficient to maintain cooperation at all stages.

The Henrich–Boyd argument leaves some room for skepticism. If defections on the first round are rare, isn't it likely that in realistic models few individuals would observe a defection? But if that is the case, then conformists who observe a defection might not be able to determine that first-order punishment is the social norm. Perhaps a polymorphic equilibrium that has just enough defectors to make the prevalence of punishment observable to conformists could be obtained in this setting.

There is room to question whether the visceral, seemingly irrational anger that people feel when they are cheated or otherwise violated can really be explained as a result of cultural transmission rather than as genetically hard-wired.

A recent paper by Florian Herold (2003) proposes another interesting explanation for the survival. Herold studies a "haystack model" in which individuals are randomly assembled into groups where they interact and reproduce. The number of offspring that a player has will be her payoff in an n-player prisoners' dilemma game in her group. Players can observe the play of others and are able to inflict punishment, but at a cost to themselves. Individuals have a hard-wired inclination either to punish defectors or not, but make a calculated choice of whether to cooperate or defect. All in the group will cooperate if and only if the number of punishers exceeds some threshold. Otherwise they will all defect. Herold shows that with monotone selection dynamics, there is an evolutionarily stable equilibrium in which all players are programmed to engage in costly punishment and where everyone therefore cooperates. In Herold's model, if almost everybody in the population at large is a punisher, then in almost all groups, there is a preponderance of punishers and so everybody chooses to cooperate. In this case, punishers don't have to bear the costs of punishing. The only way that a non-punisher could have a different payoff from a punisher would be if the random matching process selects a number of punishers that is below the cooperation-inducing threshold. Herold noticed the important fact that if non-punishers are rare, then conditional on the event that a group does not achieve the threshold number of punishers, the probability is very high that the number of punishers in the group is just one below threshold, so that each non-punisher in the group is "pivotal" to whether the group exhibits cooperation or defection. This implies that when they are rare, non-punishers will get lower expected payoffs than punishers.

5.3. Evidence from psychology and anthropology

Leda Cosmides, a psychologist and John Tooby, an anthropologist, offer [Cosmides (1989), Cosmides and Tooby (1989)] experimental evidence indicating that people are much better at solving logical problems that are framed as "cheater-detection" problems than at solving equivalent problems in other frameworks. In their view, this is evidence that individuals have evolved special modules in their brains for solving such problems.

There is interesting experimental evidence that cultural transmission plays an important role in determining when people get angry. Richard Nisbett and Dov Cohen (1996) conducted experiments in which male college students are subjected to rude and insulting behavior in the laboratory. Using questionnaires, behavioral responses, and checks of testosterone levels, they find that students who were raised in the American South become much angrier and more ready to fight than those who were raised in the North. The authors attribute this difference to the existence of a "culture of honor" in the South that is not present in the North.

Economists and anthropologists have recently conducted a remarkable series of experimental studies of how people in different cultures play the *ultimatum game*. In an ultimatum game, two players are matched and there is a fixed sum of money to be allocated. The first player, "the proposer" offers a portion of the total to the second player, "the responder". The responder can either accept or reject the offer. If the responder accepts, the division is that proposed. If the responder rejects, both players receive nothing. If this game is played by rational players who care only about their money payoff, then equilibrium in this behavior is for the proposer to offer the responder a very small share, which the responder will accept. In actual experiments with laboratory subjects in the United States, it was discovered that typically proposers offered a share of nearly one half, and this was accepted. When proposers attempted to capture a significantly larger share, responders would usually reject the proposal, thus acting as if they were willing to forego the small share that they were offered in order to "punish" a greedy proposer. In 1991, Alvin Roth and his coworkers (1991) did a "cross-cultural" conducted in which they compared the results from running the experiment in the U.S., and in Israel, Japan, and in Slovenia. They found very similar results in all four countries. In 2000, Joe Henrich (2000), published a study of an ultimatum game performed with the Machiguenga of Peru. The Machiguenga live in mobile, single-family units and small extended-family hamlets scattered throughout the tropical forests of the Amazon, where they practice hunting, fishing, gathering, and some horticulture. According to Henrich, among the Machiguenga, "cooperation above the family level is almost unknown". Henrich found that in sharp contrast to the results in the Western countries, where the modal offer was usual fifty percent, the modal share offered by the Machiguenga was only fifteen percent. Moreover, although the Machiguenga responders were offered a much smaller share than their counterparts in the developed world, they accepted these offer about 95 percent of the time – a higher acceptance rate than the average in the developed world. A recent study [Henrich et al. (2001)] reports on game experiments that have been conducted in a total of 15 "small-scale societies", including hunter–gathers, pastoralists, and farmers, and villagers. The studies found a great deal of divergence among these societies. In some of them results strongly suggested an equal-split norm and in others most proposers made offers much less generous than equal splitting and were not punished for doing so.

6. Conclusion

6.1. Further reading

The literature on social evolution is large, diverse, and multi-disciplinary. There is a great deal of good work that I have failed to discuss. Some of the omissions are simply due to my ignorance. Some work that I admire and intended to include, didn't find its way into the survey because I had to narrow my focus to limit its length. Fortunately, the seriousness of these omissions is diminished by the fact that much of the omitted work is beautifully presented in other sources.

For a survey article that partially overlaps this material, but also examines a lot of good work not covered here, I recommend Rajiv Sethi and R. Somanathan's (to appear) lucid and insightful article, "Understanding reciprocity".

There are several books that I strongly recommend to anyone interested in the subject of social evolution. These books tell their stories better than I could, so I confine my remarks to brief descriptions and hope that readers will find and enjoy them in undiluted form.

Cavalli-Sforza and Feldman's book, *Cultural Transmission and Evolution* (1981) pioneered formal modeling of this subject. Their introductory chapter is richly endowed with examples and presents a clearheaded formulation of the way that the implications of mutation, transmission, and natural selection can be extended from the study of genetically transmitted characteristics to that of culturally transmitted characteristics. Their formulation of the contrasting effects of *vertical transmission*, (from parent to child) and of *oblique* and *horizontal* transmission is insightful and provocative. They illustrate this formulation with fascinating examples such as he spread of linguistic patterns, the introduction of birth control methods, the spread of the kuru virus, which is contracted by ceremonial ingestion of dead relatives, in the Fore tribe of New Guinea. There is also a very interesting empirical study of the transmission from parents to children of such cultural behavior as religious beliefs, political affiliation, listening to classical music, reading horoscopes, and high salt usage.

Robert Trivers' book, *Social Evolution* (1985) is a stimulating and attractive treatise on the evolution of social behavior of animals (including humans) and plants. It is full of interesting examples from the natural world, thought-provoking bits of theory, and delightful photographs and drawings.

Brian Skyrms' short book, *Evolution of the Social Contract* (1996), is a beautifully written and highly accessible application of the methods of evolutionary dynamics to behavior in bargaining games and the evolution of notions of fairness and "the social contract".

My own thinking about matters related to the evolutionary foundations of social behavior has been strongly influenced by Ken Binmore's two volume work, *Game Theory and the Social Contract* (1994a, 1994b). This book combines social philosophy, political theory, evolutionary theory, anthropology, and modern game theory with great depth and subtlety.

Sober and Wilson's book *Unto Others* is written in advocacy of a modern version of the group selectionist view. It contains an extensive and interesting history of theoretical controversies between group selectionists and individual selectionists. There are also reports on interesting empirical work with group selection as well as a useful survey of group norms in a sample of twenty-five cultures that they selected *randomly* from the anthropological literature.

H. Peyton Young's *Individual Strategy and Social Structure: An Evolutionary Theory of Social Institutions* (1998) contains a remarkably accessible introduction to the mathematical theory of stochastic dynamics and to its applications in the study of the evolution of social institutions. Almost all of the work discussed in the present review uses deterministic dynamics to approximate the outcomes in a stochastic model. Heuristically, the justification for doing so is that if an equilibrium that is locally stable under deterministic dynamics receives a small, one-time stochastic shock, then as the shock wears off, equilibrium will be restored.[19] Young observes that the difficulty with this argument is that occasionally, *albeit* extremely rarely, the system may receive a sufficiently large number of shocks to knock it out of the basin of attraction of any locally stable equilibrium that is not globally stable. Thus, Young argues, a proper treatment of the very long run must directly incorporate the stochastic process into the laws of motion. He shows that in models with multiple equilibria, "long run average behavior can be predicted much more sharply than that of the corresponding determinate dynamics".

Those seeking clear, mathematical presentations of the major technical issues in evolutionary game theory will do well to look at Jörgen Weibull's *Evolutionary Game Theory* (1995) and Larry Samuelson's *Evolutionary Games and Equilibrium Selection* (1997).

Uncited references

[Bergstrom (2002)]

References

Benaim, M., Weibull, J. (September 2000). "Deterministic approximation of stochastic evolution in games". Technical Report 534, IUI Working Paper Series. Stockholm.

Bergstrom, T.C. (1995). "On the evolution of altruistic ethical rules for siblings". American Economic Review 85 (1), 58–81.

Bergstrom, T.C. (2001). "The algebra of assortative encounters and the evolution of cooperation". International Game Theory Review. (To appear).

Bergstrom, T.C. (2002). "Evolution of social behavior: Individual and group selection". Journal of Economic Perspectives 16 (2), 67–88.

[19] Michel Benaim and Jörgen Weibull (2000) have developed a careful formal treatment of the circumstances in which deterministic approximation of stochastic dynamic evolutionary processes is justified.

Bergstrom, T., Stark, O. (1993). “How altruism can prevail in an evolutionary environment”. American Economic Review 83 (2), 149–155.

Binmore, K. (1992). Fun and Games. D.C. Heath, Lexington, MA.

Binmore, K. (1994a). Game Theory and the Social Contract I: Playing Fair. MIT Press, Cambridge, MA.

Binmore, K. (1994b). Game Theory and the Social Contract II: Just Playing. MIT Press, Cambridge, MA.

Binmore, K., Samuelson, L. (1992). “Evolutionary stability in repeated games played by finite automata”. Journal of Economic Theory 57, 278–305.

Boorman, S.A., Levitt, P.R. (1980). The Genetics of Altruism. Academic Press, New York.

Bowles, S., Gintis, H. (2000). “The evolution of reciprocal preferences”. Technical Report. Santa Fe Institute, Santa Fe, NM.

Boyd, R., Richerson, P. (1990). “Group selection among alternative evolutionarily stable strategies”. Journal of Theoretical Biology 145, 331–342.

Boyd, R., Richerson, P. (1992). “Punishment allows the evolution of cooperation (or anything else) in sizeable groups”. Ethology and Sociobiology 113, 171–195.

Boyd, R., Richerson, P. (April 2001). “Group beneficial norms can spread rapidly in structured populations”. Technical Report. UCLA anthropology department. Los Angeles, CA.

Carr-Saunders, A.M. (1922). The Population Problem: A Study in Human Evolution. Clarendon Press, Oxford.

Cavalli-Sforza, L.L., Feldman, M.W. (1981). Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Cohen, D., Eshel, E. (1976). “On the founder effect and the evolution of altruistic traits”. Theoretical Population Biology 10, 276–302.

Cooper, B., Wallace, C. (2001). “Group selection and the evolution of altruism”. Technical Report. Economics Department, Oxford University.

Cosmides, L. (1989). “The logic of social exchange: Has natural selection shaped how humans reason”. Cognition 31, 187–276.

Cosmides, L., Tooby, J. (1989). “Evolutionary psychology and the generation of culture ii: A computational theory of exchange”. Ethology and Sociobiology 10, 51–97.

Dawkins, R. (1989). The Selfish Gene. (New edition.) Oxford University Press, Oxford.

Eshel, I. (1972). “On the neighbor effect and the evolution of altruistic traits”. Theoretical Population Biology 3, 258–277.

Eshel, I., Samuelson, L., Shaked, A. (1998). “Altruists, egoists, and hooligans in a local interaction structure”. American Economic Review 88, 157–179.

Eshel, I., Sansone, E., Shaked, A. (1999). “The emergence of kinship behavior in structured populations of unrelated individuals”. International Journal of Game Theory 28, 447–463.

Frank, R.H. (1987). “If homo economicus could choose his own utility function, would he want one with a conscience”. American Economic Review 77 (4), 593–604.

Friedman, D., Singh, N. (1999). “The viability of vengeance”. Technical Report. U.C. Santa Cruz, Santa Cruz, CA.

Ghiselin, M. (1974). The Economy of Nature and the Evolution of Sex. University of California Press, Berkeley, CA.

Grafen, A. (1979). “The hawk–dove game played between relatives”. Animal Behaviour 27 (3), 905–907.

Grafen, A. (1984). “Natural selection, group selection, and kin selection”. In: Kreb, J.R., Davies, N.B. (Eds.), Behavioural Ecology, Chapter 3, 2nd edn. Blackwell, London, pp. 62–80.

Haldane, J.B.S. (1932). The Causes of Evolution. Harper & Brothers, New York and London.

Hamilton, W.D. (1964). “The genetical evolution of social behavior, Parts i and ii”. Journal of Theoretical Biology 7, 1–52.

Hamilton, W.D. (1975). “Innate social aptitudes in man: An approach from evolutionary genetics”. In: Fox, R. (Ed.), Biosocial Anthropology. Malaby Press, London.

Hardin, G. (1977). The Limits of Altruism. Indiana University Press, Bloomington, IN.

Henrich, J. (2000). “Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon”. American Economic Review 90 (4), 9730979.

Henrich, J., Boyd, R. (2001). "Why people punish defectors". Journal of Theoretical Biology 208, 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., McElreath, R. (2001). "In search of homoeconomicus: Behavioral experiments in 15 small-scale societies". American Economic Review 91 (2), 73–78.

Herold, F. (2003). "Carrot or stick: Group selection and the evolution of reciprocal preferences". Technical Report. Munich University.

Hines, W.G.S., Smith, J.M. (1979). "Games between relatives". Journal of Theoretical Biology 79, 19–30.

Hume, D. (1978). A Treatise of Human Nature, 2nd edn. Clarendon Press, Oxford. (Edited by Selby-Bigge, L.A., revised by Nidditch, P.; first published 1739.)

Kimura, M., Weiss, G.H. (1964). "The stepping stone model of population structure and the decrease of genetic correlation with distance". Genetics 49, 561–576.

Lack, D. (1966). Population Studies of Birds. Clarendon Press, Oxford.

Ledyard, J.O. (1995). "Public goods: A survey of experimental research". In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics, Chapter 2. Princeton University Press, Princeton, NJ, pp. 111–181.

Levin, B.R., Kilmer, W.L. (1974). "Interdemic selection and the evolution of altruism". Evolution 28 (4), 527–545.

Levins, R. (1970). "Extinction". In: Gerstenhaber, M. (Ed.), Some Mathematical Problems in Biology. American Mathematical Society, Providence, pp. 77–107.

Matessi, C., Jayakar, S.D. (1976). "Conditions for the evolution of altruism under Darwinian selection". Theoretical Population Biology 9.

Nachbar, J. (1990). "Evolutionary selection dynamics in games: Convergence and limit properties". International Journal of Game Theory 19, 59–89.

Nisbett, R.E., Cohen, D. (1996). Culture of Honor: The Psychology of Violence in the South. Westview Press, Boulder, CO.

Nowak, M.A., May, R.M. (1993). "Evolutionary games and spatial chaos". Nature 359, 826–829.

Nowak, M.A., May, R.M. (1998). "Evolution of indirect reciprocity by image scoring". Nature 393, 573–577.

Rappaport, A., Chammah, A.M. (1965). Prisoner's Dilemma. University of Michigan Press, Ann Arbor, MI.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S. (1991). "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo, an experimental study". American Economic Review 81 (5), 1068–1095.

Rousseau, J.J. (1950). Discourses on the Origins and Foundation of Inequality Among Men (Second Discourse). Everyman's Library, Dutton, NY. (First publication 1755).

Samuelson, L. (1997). Evolutionary Games and Equilibrium Selection. MIT Press, Cambridge, MA.

Sethi, R., Somanathan, E. (to appear). "Understanding reciprocity". Journal of Economic Behavior and Organization.

Skyrms, B. (1996). Evolution of the Social Contract. Cambridge University Press, Cambridge.

Skyrms, B. (2001). "The stag hunt". Proceedings and Addresses of the American Philosophical Association 75 (2), 31–41.

Skyrms, B., Pemantle, R. (2000). "A dynamic model of social network formation". Proceedings of the National Academy of Science 97 (16), 9340–9346.

Smith, J.M. (1964). "Group selection and kin selection". Nature 201, 1145–1147.

Smith, J.M. (1976). "Group selection". Quarterly Review of Biology 51, 277–283.

Smith, J.M., Price, G.R. (1973). "The logic of animal conflict". Nature 246, 15–18.

Sober, E., Wilson, D.S. (1999). Unto Others. Harvard University Press, Cambridge, MA.

Trivers, R. (1985). Social Evolution. Benjamin Cummings, Menlo Park, CA.

Waugh, H.E. (Ed.) (1929). Hesiod: The Homeric Hymns and Homerica. Heineman, London.

Weibull, J. (1995). Evolutionary Game Theory. MIT Press, Cambridge, MA.

Williams, G.C. (1966). Adaptation and Natural Selection. A Critique of Some Current Evolutionary Thought. Princeton University Press, Princeton, NJ.

Wilson, D.S. (1975). "A theory of group selection". Proceedings of the National Academy of Sciences 72 (1), 143–146.

Wilson, D.S. (1979). "Structured demes and trait-group variation". American Naturalist 113, 157–185.

Wilson, D.S. (1987). "Altruism in Mendelian populations derived from sibling groups". Evolution 41 (5), 1059–1070.

Wright, S. (1921). "Systems of mating". Genetics 6 (2), 111–178.

Wright, S. (1943). "Isolation by distance". Genetics 28, 114–138.

Wright, S. (1945). "Tempo and modes in evolution: A critical review". Ecology 26, 415–419.

Wynne-Edwards, V.C. (1962). Animal Dispersion in Relation to Social Behaviour. Oliver and Boyd, Edinburgh and London.

Wynne-Edwards, V.C. (1978). "Intrinsic population control: An introduction". In: Ebling, F.J., Stoddart, D.M. (Eds.), Population Control by Social Behaviour. Institute of Biology, London, pp. 1–22. (Volume *Population Control by Social Behaviour*.)

Wynne-Edwards, V.C. (1986). Evolution through Group Selection. Alden Press, Oxford.

Young, H.P. (1998). Individual Strategy and Social Structure. Princeton University Press, Princeton, NJ.

Proof of Raw Subject Index