Evolutionary Games with Group Selection*

Martin Kaae Jensen† **and Alexandros Rigos**‡

August 4, 2014

Abstract

This paper introduces two new concepts in evolutionary game theory: Nash equilibrium with Group Selection (NEGS) and Evolutionarily Stable Strategy with Group Selection (ESSGS). These concepts generalize Maynard Smith and Price (1973) to settings with arbitrary matching rules, in particular they reduce, respectively, to Nash equilibrium and ESS when matching is random. NEGS and ESSGS are to the canonical group selection model of evolutionary theory what Nash Equilibrium and ESS are to the standard replicator dynamics: any NEGS is a steady state, any stable steady state is a NEGS, and any ESSGS is asymptotically stable. We exploit this to prove what may be called "the second welfare theorem of evolution": Any evolutionary optimum will be a NEGS under some matching rule. Our results are illustrated in Hawk-Dove, Prisoners' dilemma, and Stag Hunt games.

Keywords: Evolutionary Game Theory, Evolutionarily Stable Strategy, ESS, Group Selection, Non-random Matching, Trait-group Model, Haystack Model.

JEL Classification Codes: C72, C73.

*We would like to thank Heinrich Nax, and participants at the 2012 UECE Lisbon Meeting and the 2014 Norms, Actions, Games Conference in London. All remaining errors are our own.

†Department of Economics, University of Leicester. Email: mj182@le.ac.uk.

‡Department of Economics, University of Leicester. Email: ar374@le.ac.uk.

1 Introduction

The canonical evolutionary game theory model of Maynard Smith and Price (1973) plays an important role in biology, economics, political science, and other fields. Its equilibrium concept, an *evolutionarily stable strategy* (*ESS*) describes evolutionary outcomes in environments where populations are *large* and matching is typically *random*.[1] Since an ESS is a refinement of Nash equilibrium, it obviously *cannot* explain any behavioral departure from purely self-serving behavior in the one-shot Nash sense. In particular it cannot account for cooperative behavior in say, a prisoners' dilemma, or shed light on altruism more generally, nor can it account for any other non-Nash behaviors such as spite (Hamilton, 1970; Alger and Weibull, 2012) or costly punishment (Fehr and Gächter, 2000).

In order to explain such deviations from Nash behavior, evolutionary game theory turned to models with a finite number of agents hence departing from the first of the mentioned conditions of Maynard Smith and Price (1973). Thus in Schaffer (1988), the finite set of individuals have "market power" and can influence average fitness, while in the model preferred by Maynard Smith (1982) – namely repeated games – a few agents, usually just two, can perfectly monitor and record each others' past actions and condition their strategies hereupon (in evolutionary theory, the repeated games approach is usually referred to as *direct reciprocity*). Both of these frameworks have led to an enormous body of research in economics and game theory (see *e.g.* Alós-Ferrer and Ania, 2005; Leininger, 2006; Samuelson, 2002; Vega-Redondo, 1997, and references therein).

While evolutionary *game theorists* turned to finite populations, evolutionary biologists more broadly devoted as much – if not more – attention to a departure from the second basic condition of Maynard Smith and Price (1973), namely the assumption that matching is random. When matching is non-random — possibly indirectly so due to prolonged interaction of individuals in separated groups (Maynard Smith, 1964) — the fitness of an individual will depend on the group he is assigned to, and so different groups will on average meet with varying reproductive success (Bergström, 2002; Kerr and Godfrey-Smith, 2002). Thus non-random matching invariably leads to *group selection* whereby one can trace the evolutionary success of certain types of groups and not just their constituent individuals (we return to this topic in a moment). Take the prisoners' dilemma. Matching is assortative if after each round of play cooperators have higher probability of being matched to other cooperators than to defectors. This is often a highly realistic assumption corresponding for example to situations where a large group of individuals cannot perfectly monitor each others' past behaviors but receive some "revealing signals" about opponents' types and exert some influence on who they are matched to (Maynard Smith, 1964; Wilson, 1975, 1977). Non-random matching also results if matching depends on the geographical location of individuals (Eshel, Samuelson, and Shaked, 1998; Nowak and May, 1992; Skyrms, 2004); or if (genetically) similar individuals match assortatively as in models of kin selection (Hamilton, 1964; Alger and Weibull, 2010). When matching is non-random a variety of different *groups* will gen-

[1] Intuitively, random matching means that an individuals' type has no influence on what type of individual he is likely to be matched to.

erally coexist at any given moment in time. For example in the prisoners' dilemma, some groups will consist of defectors only, some of cooperators only, and some will be mixed. Thus the average fitness will differ across groups, as will the fitness a specific type of individual obtains if he is placed into different groups. It follows that evolutionary pressure takes place not just at the individual level but also at the group level even though individuals are ultimately the fitness bearing entities.[2]

Now, the existing literature on non-random matching is usually informal and/or deals only with special cases (typically two types who are matched pairwise and assortatively). As a basis for this paper's main results, we begin in section 2 by laying out a unified model in a general and self-contained manner. Compared to existing literature, we add value by setting up a model that allows for arbitrary matching rules (ways to match populations into groups), any number of strategies, arbitrary group sizes, and any possible payoff structure in the group stages (*i.e.*, any possible underlying symmetric normal form game, see section 2.1). For any reader who is unfamiliar with – or confused by – the existing literature, it is our hope that section 2 will provide an easily accessible point of entry. The most substantial of the mentioned generalizations is that we allow for arbitrary matching rules. Indeed, this is what allows us to show (section 2.4) that group selection models based on prolonged interaction in what Maynard Smith (1964) calls "haystacks" (see also Wilson, 1975 and Cooper and Wallace, 2004), can be recast as models of group selection based directly on non-random matching (*e*.g. Bergström, 2003; Kerr and Godfrey-Smith, 2002). Intuitively, this is not all that surprising from the individualist perspective described above: What at the end of the day matters is whether individuals are matched randomly with each other or not. Precisely how any departure from random matching comes about is secondary to the fact that as soon as matching is non-random, different groups with different average fitness levels will exist and this is ultimately what group selection is all about. In fact, we shall from now on make this viewpoint explicit by not separating models of non-random matching from other models of group selection.

A key thing to notice about the model of section 2 is that it is *not* a game theoretic model. In the terminology of biologists it is a model of evolutionary theory, *not* a model of evolutionary game theory. More specifically, it is a dynamical model of selection where attention is devoted to steady states of the associated replicator dynamics. This is of course in sharp contrast to the random matching case where Nash equilibrium and ESS play central roles and allows the powerful machinery of game theory to be applied.[3] The first substantial contribution of the present paper is to fill the resulting gap in the literature. Specifically, we are going to ask what game theoretic equilibrium concepts form group selection's

[2]As shown by Kerr and Godfrey-Smith (2002), one may with equal formal correctness think of selection taking place at the individual or the group level. This difference in perspective has been (and is) the topic of a heated debate in evolutionary biology, a key reference here being the book "Unto Others" by Sober and Wilson (1999). As explained in section 2 we are going to take a so-called "individualist" perspective in this paper, and will not go into the more philosophical aspects of the levels of selection controversy.

[3]For example an ESS is a refinement of Nash equilibrium with the crucial property that any asymptotically stable state of the evolutionary (replicator) dynamics is an ESS (Weibull, 1995, chapter 3). Thus when one studies the set of ESSs, "bad" equilibria have been removed which not surprisingly leads to stronger results. We return to the precise relationship between steady states, Nash equilibrium, and ESS in greater detail, in section 4.

natural parallels to Nash equilibrium and ESS. This leads to two new equilibrium concepts, namely a *Nash equilibrium with group selection* (NEGS) and an *evolutionarily stable strategy with group selection* (ESSGS). These concepts turn out to be intuitive once the underlying evolutionary game, which we call a *group selection game*, is understood. Interestingly, this game turns out to be novel even from a game theoretic perspective: As in standard imperfect information games, agents make decisions without knowing with certainty the strategies pursued by opponents – all they know is the distribution of the opponents' strategies, or to put it in the evolutionary terminology, the probabilities of ending up in any of the different kinds of groups. Crucially, these probabilities depend on the actual strategies pursued by the agents. For simplicity, imagine a large group of individuals, all of whom has a choice between two strategies, "honesty" (H) or "deception" (D). Agents must commit to a strategy before being allocated into equal-sized groups where they execute these strategies (equivalently, they choose their actions with imperfect knowledge about opponents' actions). Given a specific *matching rule* (a given way to divide a population with a given fraction of H- and D-types into groups of equal size) and given that agents *know* the population-wide composition into H and D types, each agent can calculate the probabilities of ending up in any specific kind of group as a function of the specific strategy chosen (H or D).[4] In a NEGS, individuals' optimal choices precisely lead to the population-wide composition into H and D types which formed the basis of their decisions in the first place. The concept of an ESSGS simply adds a "non-invasion" criterion to this Nash/fixed point criterion precisely as is the case with random matching (Maynard Smith and Price, 1973). Note that a NEGS interchangeably can be viewed as a mixed strategy pursued by all individuals or as a vector that gives the fractions of each type in equilibrium. In the previous case, a NEGS or an ESSGS may be, say $(0.9, 0.1)$ meaning that each individual in the population will be honest with probability 90 % and deceptive with probability 10 %. With an infinite population size, this of course implies that 90 % of the population will be honest, and 10 % deceptive at any given moment in time. Intuitively, in a NEGS the deceptive individuals' purpose is to keep the honest individuals in check (and vice versa): without a sufficiently large population of deceivers, the benefit of choosing to deceive will outweigh that of being honest because deceptive individuals will face a relatively small chance of being matched to another deceiver *even though* matching is assortative.

After defining group selection games and proving equilibrium existence, we turn to the relationship with the dynamic evolutionary model of section 2. Thus in theorem 6 – which together with the "second welfare theorem of evolution" described below forms this paper's main contribution – we prove that any NEGS is a steady state for the replicator dynamics, that any (Lyapunov) stable steady state for the replicator dynamics is a NEGS, and that any ESSGS is an asymptotically stable state of the replicator dynamics. These results extend existing results on Nash equilibrium and ESS (Hofbauer and Sigmund, 1998; Maynard Smith and Price, 1973; Weibull, 1995) to settings with non-random matching, and show

[4] Obviously, the number of possible group compositions depends on the group size as well as the number of strategies. With two strategies and groups of size two, any individual can end up in precisely two different kinds of groups – one where the opponent is of the same type and one where he is not.

that NEGS and ESSGS are important new evolutionary game theory concepts. Immediately, a long list of research questions report themselves in that one could attempt to "transfer" over to group selection models all of the existing results from evolutionary game theory. We shall leave the bulk of this for future research, for example we are not going to go into topics related to neutrally stable strategies, asymptotically stable sets, doubly symmetric games or the fundamental theorem of natural selection (for these "textbook" issues see the monographs of Hofbauer and Sigmund, 1998 or Weibull, 1995). Instead we are going to focus in section 6 on a question which in some sense "ignited" this whole literature. The point of the prisoners' dilemma is that Nash equilibrium – and with it evolutionary models based on random matching – may easily fail to produce outcomes that maximize average payoff/welfare in the population.[5] The question from the group selection point of view then becomes: What types of (non-random) matching *will*, if any, lead to optimality? Our main result in this regard (theorem 7) might be called the "second welfare theorem of evolutionary theory" telling us that *any* outcome that is optimal will in fact be a NEGS under *some* matching rule. In a number of standard games (hawk-dove, stag hunt, prisoners' dilemma) we proceed to characterize these matching rules and in doing so gain an understanding of *when* evolution in a specific situation (for a fixed matching rule and payoff structure) is likely to lead to an evolutionary optimum or not. To give an example, we show that in the Hawk-Dove case even low levels of assortativity (in the "constant index" sense of Bergström, 2003) may still lead to the evolutionary optimum if a Dove who meets a Dove receives only a "modest" gain from changing to the Hawk strategy.

The structure of the paper is as follows: Section 2 describes the general group selection model and section 3 defines group selection games, NEGS and ESSGS. Section 3 also contains some some basic results on existence and the relationship between NEGS and ESSGS. Section 4 contains our main theoretical results discussed above. Section 5 contains a number of examples, and section 6 discusses the fitness/welfare issues with basis in the aforementioned "second welfare theorem". Finally, section 7 concludes.

2 Group Selection in Evolutionary Theory

In this section we present a unified model of non-random matching based group selection. The model is closely related to Kerr and Godfrey-Smith (2002) and Bergström (2003) both of which consider non-random matching as the primus motor of group selection, and both of which adopt an "individualist perspective" that assigns fitness to individuals rather than the groups they form.[6] While their analysis restricts attention to two strategies/types and certain relatively restrictive types of closed form group

[5] In the language of welfare analysis, the outcome does not maximize *utilitarian* social welfare. This, of course, also implies a break-down of *Pareto optimality*.

[6] As shown by Kerr and Godfrey-Smith (2002), one can formally recast such models so that groups become the fitness bearing entities (so the two frameworks are formally equivalent). The group-based fitness perspective is strongly advocated in the famous book "Unto Others" by Sober and Wilson (1999). See also Maynard Smith (1998) and Okasha (2005) for more on this issue.

formation rules, we allow for any number of strategies and, more importantly, arbitrary rules of group formation (called *matching rules* in what follows). The latter is crucial when in Section 2.4 we go on to show that any trait-group model — where matching is random but groups are isolated for prolonged spells in what Maynard-Smith calls "haystacks" — can be recast within our setting with non-random matching in such a way that the equilibria/steady states and dynamics, remain the same. This observation substantially extends the scope of our general results, and it also dispels the notion that group selection models based on non-random matching are somehow not "true" models of group selection (certainly, any difference will be at most a question of interpretation and terminology — any result about observables, *i.e.*, dynamics and equilibria will remain the same).[7]

Briefly, the model can be summarized as follows: At each date there is a large set of individuals, formally the continuum $I = [0,1]$. At the beginning of each period, the agents are allocated into groups of the same finite size $n \in \mathbb{N}$. This happens in accordance with what we call a *matching rule* (formally defined in subsection 2.2) which is a function that maps the type frequency of the set of agents into the distribution of group types.[8] After the n-sized groups are formed, the individuals in each group face a symmetric normal-form game (section 2.1). In accordance with the basic premise of evolutionary game theory, agents are hard-wired to follow the same strategy as the parent ('like begets like'). Thus an individual who is fathered by a parent who executed strategy j, say, in the previous round will mechanically execute strategy j in his group game, regardless of the resulting payoff/composition of individuals in the specific group he is drawn into. The payoff determines the *fitness*, *i.e.*, the (expected) number of children the agent will send on to the next round.[9] Finally, after the group game stage, a new generation is born with the relative proportion of each type determined by the success (fitness) this type's strategy enjoyed across the different groups. The above process then repeats itself leading to a new generation and so on. The evolutionary outcome of this group selection process is a steady state of the resulting replicator dynamical system as described in section 2.3.

Note that – apart from our insistence that the group games can be seen as normal form games (a perspective that is alien to the existing literature) – the model is entirely non-game theoretic.

2.1 The Underlying Normal Form Group Games

Our description begins with the underlying normal form game that agents face in the group stages. Although in evolutionary models, individuals act purely mechanically and play the strategy inherited from the parent, they nonetheless participate in a standard normal form game and receive pay-

[7] It is also worth mentioning that from a more technical perspective, our model is crafted so that the main structure of the traditional evolutionary game theory model (*e.g.* Weibull, 1995) is retained. This both makes the connection with standard replicator dynamics transparent and paves the way for our analysis in subsequent sections.

[8] Our concept of a matching rule is closely related to a construction due to Kerr and Godfrey-Smith (2002, p.484) who, however, consider only the case of two strategies (the extension to any number of strategies is non-trivial as will become clear).

[9] A different explanation of fitness that is more plausible in economic contexts is to think of it as the number of agents copying one's behavior because it is more successful: More successful behaviors will have more followers in the next round of play.

offs/fitnesses accordingly. We need to make this game theoretic aspect clear to set the stage for this paper's main results.

Let $n \in \{2,3,\ldots\}$ denote the *group size* so that $N = \{1,\ldots,n\}$ is the *set of players* in a group. A group game is a symmetric normal form game $G =< N, M, A >$ where $M = \{1,\ldots,m\}$ is the set of pure strategies and $A : M \times M^{n-1} \to \mathbb{R}$ is the payoff function over pure strategies that all players share (since the game is symmetric). The set of all n-player, m-strategy symmetric normal form games is denoted by $\mathfrak{G}_{n,m}$. Note that by symmetry, $A(y^i, y^{-i}) = A(y^i, \tilde{y}^{-i})$ where $y^i \in M$ is any pure strategy for player i, and $y^{-i}, \tilde{y}^{-i} \in M^{n-1}$ are pure strategies of i's opponents where $\tilde{y}^{-i}$ is any permutation of y^{-i}. A (symmetric) *Nash equilibrium* for G is defined in the usual way as a vector $\sigma^* \in S_m \equiv \{\sigma \in \mathbb{R}^m_+ : \sum_{j=1}^m \sigma_j = 1\}$ such that $A(\sigma^*, (\sigma^*, \ldots, \sigma^*)) \geq A(\sigma, (\sigma^*, \ldots, \sigma^*))$ for all $\sigma \in S_m$.[10]

It is convenient to write the previous payoff structure in a way that makes explicit reference to the group structure. Call an individual who executes pure strategy $j \in M$ a *type j individual*. Due to symmetry, the payoff to such a type j individual depends only on the *number* of opponents in his group who play each of the m strategies (as opposed to *which* opponents follow what strategies). Next imagine that this type j individual finds himself in a *group*, group i say, consisting of n^i_1 individuals of type 1, n^i_2 individuals of type 2, and so on up to n^i_m.[11] In this situation, the individual's payoff will be equal to $A(j, j^{\text{opp}})$ where $j^{\text{opp}} \in M^{n-1}$ is any vector of opponents' strategies which contains n^i_1 strategy 1 entries, ..., n^i_{j-1} strategy $j-1$ entries, $n^i_j - 1$ strategy j entries, n^i_{j+1} strategy $j+1$ entries, ..., n^i_m strategy m entries. Crucially, we can write the payoff $A(j, j^{\text{opp}})$ simply as $A_j^{(n^i_1,\ldots,n^i_m)}$ or even as A^i_j where i is the index of the specific group the individual finds himself in (as long as we keep record of the group composition $n^i = (n^i_1, \ldots, n^i_m)$ of group i).

In this way, we can capture all of the information we need about the normal form game in a sequence (A^i_j) where $j = 1,\ldots,m$ and $i = 1,\ldots,\gamma_{n,m}$. Here $\gamma_{n,m}$ is the *number of different n-sized groups that can be formed with m different pure strategies*.[12] From combinatorics we know that $\gamma_{n,m}$ precisely equals the number of multisets of cardinality n with elements taken from a set with cardinality m (see Aigner, 2007, p. 15), *i.e.*

$$\gamma_{n,m} = \frac{(n+m-1)!}{n!(m-1)!}. \tag{1}$$

For example, $\gamma_{2,2} = 3$ since three different groups can be formed if the group size equals 2 and there are 2 possible strategies (these groups are, respectively, one where both are of type 1, one where both are of type 2, and one where the individuals follow different strategies).

[10] Letting $\sigma^i \in S_m \equiv \{\sigma \in \mathbb{R}^m_+ : \sum_{j=1}^m \sigma_j = 1\}$ denote a *mixed strategy* for player i and $\sigma^{-i} \in S_m^{n-1}$ denote a mixed strategy profile of player i's opponents, it is easy to see that $A(\sigma^i, \sigma^{-i}) = \sum_{y \in M^n} A(y^i, y^{-i}) \prod_{k \in N} \sigma^k_{y^k}$.

[11] Note that since the individual himself is counted here, we necessarily have $n^i_j \geq 1$ (there is at least one of the individual's own type). Of course we must also have $\sum_k n^i_k = n$ and each n^i_k must be non-negative.

[12] Of course, we must be a little careful here because some of these are not really properly defined. Specifically, A^i_j is not well-defined unless $n^i_j \geq 1$. But building this explicitly into the notation leads to unwarranted complications.

2.2 Group Formation

We now turn to the question of how groups are formed out of each generation's individuals. The key concept is that of a *matching rule* which generalizes what Kerr and Godfrey-Smith (2002, p. 484) call a "rule of group assembly" to more than 2 pure strategies (the concept is also related to Bergström, 2003, as returned to below).

A *population state* is the frequency distribution of the different types in the population, *i.e.*, a vector $\mathbf{x} = (x_1, \ldots, x_m) \in S_m$ where x_1 is the fraction of 1-strategists in the population, x_2 is the fraction of 2-strategists in the population, and so on.[13] A *group state* similarly represents the group frequencies and so is a vector $\mathbf{g} = (g_1, g_2, \ldots, g_{\gamma_{n,m}}) \in S_{\gamma_{n,m}}$ where g_1 is the fraction of all groups that is of type 1, g_2 the fraction of type 2 groups, and so on up to group $\gamma_{n,m}$ which it is recalled is the number of different n-sized groups it is possible to form when there are m different strategies (section 2.1).

A *matching rule* is simply a function that maps a population state $\mathbf{x} \in S_m$ into a group state $\mathbf{g} \in S_{\gamma_{n,m}}$. So, intuitively, a matching rule describes how any given population is allocated into groups.

Definition 1. (Matching Rules) *A matching rule is a function* $\mathbf{f} : S_m \to S_{\gamma_{n,m}}$ *that maps any population state* $\mathbf{x} \in S_m$ *into a group state* $\mathbf{f}(\mathbf{x}) \in S_{\gamma_{n,m}}$. *The set of matching rules in a population with* m *strategies and group-size* n *is denoted by* $\mathfrak{F}_{n,m}$.

It is natural — but not necessary for any of our results — to impose consistency on matching rules by demanding that the fraction of j-type individuals allocated into the different groups equals the fraction x_j of individuals of type j that are actually present in the population. Since the proportion of j-type individuals in an i-type group by definition is $\frac{n^i_j}{n}$, the fraction of individuals that are of type j and in i-type groups will, for any given group state $\mathbf{g}$, be $\frac{n^i_j}{n} g_i$. Hence, *across all groups* the fraction of the population that is of type j is $\sum_{i=1}^{\gamma_{n,m}} \frac{n^i_j}{n} g_i$. This number must then equal x_j for every individual to be allocated to one (and only one) group:

$$\sum_{i=1}^{\gamma_{n,m}} \frac{n^i_j}{n} f_i(\mathbf{x}) = x_j, \text{ for } j = 1, \ldots, m \tag{2}$$

If we define $\text{supp}(j)$ to be the set of group types that contain at least one j-strategist, (2) can be expressed equivalently as:

$$\sum_{i \in \text{supp}(j)} \frac{n^i_j}{n} f_i(\mathbf{x}) = x_j, \text{ for } j = 1, \ldots, m \tag{3}$$

When a matching rule satisfies (3), we say that it is *consistent*. While our main examples of matching rules below are consistent, a very important special case of our setting, namely haystack/trait-group models (Cooper and Wallace, 2004; Maynard Smith, 1964; Wilson, 1975) generally does not lead to consistent matching rules (see Section 2.4).

[13] Here S_m is the unit $(m-1)$-simplex *i.e.* $S_m = \left\{ \mathbf{x} \in \mathbb{R}^m_+ | \sum_{j \in M} x_j = 1 \right\}$.

Recall that $\frac{n_j^i}{n} f_i(\mathbf{x})$ is the fraction of the total population that is of type j and is allocated to a group of type i under the matching rule $\mathbf{f}$. When $x_j > 0$ we may divide this by the fraction x_j of the population that is of type j in order to get the fraction of j-type individuals that is allocated to a group of type i:

$$w_j^i(\mathbf{x}) \equiv \frac{n_j^i}{n x_j} f_i(\mathbf{x}) \tag{4}$$

This may be compared with Bergström (2003) who studies group selection (again in the special case $n = m = 2$), and who takes the w_j^i's as fundamentals instead of the matching rule. More specifically, Bergström (2003) considers the difference $w_1^1 - w_1^2$ and calls this difference the 'index of assortativity'. We return to the index of assortativity in example 1.3 below where we also show how one gets from a model based on a constant index of assortativity our formulation with matching rules.

We finish this subsection by presenting a number of concrete examples of matching rules. We shall be calling on these repeatedly throughout the rest of this paper.

Example 1.

1. ***Complete segregation.*** *Different strategies* do not mix. *All individuals are allocated into groups with only individuals of the same type and thus all groups contain a single type of individuals each (n individuals that follow the same strategy). The group types that have n individuals of the same type get a non-negative frequency whereas all other kinds of groups get a frequency of zero. Due to the consistency requirements for matching rules, we get that the group type that contains n j-types should get a frequency of x_j. So, formally, the matching rule for complete segregation is the following.*

$$\begin{array}{rcl} f_i(\mathbf{x}) = x_j & , & \text{if } n_j^i = n \\ f_i(\mathbf{x}) = 0 & , & \text{otherwise.} \end{array} \tag{5}$$

e.g. When $n = m = 2$ the matching rule for complete segregation take the form:

$$f_1(x_1, x_2) = x_1 \qquad f_2(x_1, x_2) = 0 \qquad f_3(x_1, x_2) = x_2.$$

2. ***Random matching.*** *Let us define the* opponent profile *of a type j individual in a type i group to be the vector $v_j^i = (v_1^i, \ldots, v_j^i, \ldots, v_m^i) \equiv (n_1^i, \ldots, n_j^i - 1, \ldots, n_m^i)$ that shows how many opponents of each type a type j individual faces when she is drawn into a group of type i. Obviously, individuals of different types that face the same opponent profile will be in groups of different types. We will say that* matching is random *when the (ex ante) probability of an individual (conditional on her type) to end up facing a specific opponent profile is independent of her type. If this is the case, then the frequencies of group types will follow a multinomial distribution (see for example Lefebvre, 2007, p. 22):*[14]

[14] To show that the property described above holds for the matching rule of equation (6), let us consider a group of type i with $n_j^i \geq 1$ for some $j \in M$. Notice that a j-type in that group has n_1^i type 1 opponents,..., $n_j^i - 1$ type j opponents, ..., n_m^i

$$f_i(\mathbf{x}) = \frac{n!}{\prod_{j\in M} n_j^i!} \prod_{j\in M} x_j^{n_j^i}. \tag{6}$$

Notice that for $m = 2$, the random matching rule becomes

$$f_i(x_1, x_2) = \frac{n!}{n_1^i!(n-n_1^i)!} x_1^{n_1^i} x_2^{n-n_1^i}.$$

That is it boils down to the binomial distribution (see Kerr and Godfrey-Smith, 2002, p. 484).

3. ***Constant Index of Assortativity.***

Bergström (2003) studies 2-person prisoner's dilemma population games by using the 'index of assortativity' which he defines as "the difference between the probability that a C-strategist meets a C-strategist and the probability that a D-strategist meets a C-strategist". In terms of notation used in this paper (with x_1 and x_2 denoting the proportion of cooperators and defectors in the population respectively), this means that the index of assortativity at a state (x_1, x_2) will be:

$$\alpha(x_1, x_2) = w_1^1(x_1, x_2) - w_2^2(x_1, x_2) = \frac{f_1(x_1, x_2)}{x_1} - \frac{f_2(x_1, x_2)}{2x_2}.$$

Bergström goes on to analyze prisoners' dilemma games under "assortative matching" rules that have a constant index of assortativity α for all values of $\mathbf{x}$. As one easily verifies, the matching rule corresponding to a constant index of assortativity α is:

$$\begin{aligned} f_1(\mathbf{x}) &= x_1(1-(1-\alpha)x_2) \\ f_2(\mathbf{x}) &= 2(1-\alpha)x_1x_2 \\ f_3(\mathbf{x}) &= x_2(1-(1-\alpha)x_1). \end{aligned}$$

In the case of $\alpha = 0$ the rule coincides with the random matching rule and in the case of $\alpha = 1$ it coincides with the complete segregation rule (for both of these statements we of course need $n = m = 2$, i.e., two players and two strategies).

4. ***"Almost" Constant Index of Dissociation.***

It is not possible to extend the previous constant index of assortativity rule to $\alpha < 0$, i.e., to dissociative matching without violating consistency of matching rules (i.e., condition (3) *of Section 2.2). Indeed, if such a "constant index of dissociation" rule is imposed, the matching rule would necessarily violate* (3) *when $\mathbf{x}$ is close to 0 or to 1.*[15] *So to consider constant index of dissociation matching rules, one must either consider matching rules that are not consistent (which, as mentioned already in Section 2.2 will not upset any of our results), or else one must "tweak" the*

type m opponents. So the opponent profile for a j strategist in a type i group will be $\nu = (n_1^i, \ldots, n_j^i - 1, \ldots, n_m^i)$. Indeed, the probability of a type j individual (conditional on her type) to end up in group with opponent profile $\nu = (\nu_1, \ldots, \nu_m)$ is given by: $w_j^i(\mathbf{x}) = \frac{n_j^i}{nx_j} \frac{n!}{\prod_{k\in M} n_k^i!} \prod_{k\in M} x_j^{n_k^i} = \frac{(n-1)!}{\prod_{k\in M} \nu_k!} \prod_{k\in M} x_k^{\nu_k}$. *i.e.* it is independent of the individual's strategy j.

[15] More specifically, this happens for $x \in (0, \frac{-\alpha}{1-\alpha}) \cup (\frac{1}{1-\alpha}, 1)$ when $\alpha \in [-1, 0)$.

construction slightly near the boundary. An example of the latter is the following matching rule with $\beta \in [0,1]$ *being the* 'index of dissociation'

$$\begin{aligned}\mathbf{f}(\mathbf{x}) &= (f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})) = \\ &= \begin{cases} (0, 2x_1, 1-2x_1) & , x_1 \in \left[0, \frac{\beta}{1+\beta}\right] \\ (x_1(1-(1+\beta)x_2), 2(1+\beta)x_1x_2, x_2(1-(1+\beta)x_1)) & , x_1 \in \left(\frac{\beta}{1+\beta}, \frac{1}{1+\beta}\right) \\ (1-2x_2, 2x_2, 0) & , x_1 \in \left[\frac{1}{1+\beta}, 1\right] \end{cases}\end{aligned}$$

2.3 Group Selection Models

At this point we have defined all of the key ingredients of a group selection model: A set of agents $I = [0,1]$, the normal form group game $G =< N, M, A >$ (here $N = \{1, \ldots, n\}$ and $M = \{1, \ldots, m\}$ where n is the group size and m the number of strategies/types), and the matching rule $\mathbf{f} : S_m \to S_{\gamma_{n,m}}$ which in each period allocates the newborn generation into groups (recall from section 2.1 that $\gamma_{n,m}$ is the number of different n-sized groups that can be formed from m different strategies).

Given the tuple $< I.G.\mathbf{f} >$, we are now in a position to describe the dynamical system that constitutes the evolutionary model of group selection. The standard solution concept in group selection models as defined is that of a steady state which we now proceed to discuss.

Let $\mathbf{x}^t \in S_m$ denote the (population) state at date t (the vector of frequencies of the different types at the given date, see subsection 2.2). At date t, the population is allocated into groups according to the matching rule $\mathbf{f}$, hence $\mathbf{f}(\mathbf{x}^t) \in S_{\gamma_{n,m}}$ is the resulting group frequency distribution. Regardless of which group an individual of type j ends up in, he will mechanically follow the strategy of his type (as inherited from the parent) and fitness will be distributed accordingly. Now recall from equation (4) of section 2.2 that $w_j^i(\mathbf{x}) = \frac{n_j^i x_j}{n} f_i(\mathbf{x})$ is the fraction of j-type individuals that is allocated to groups of type i under the matching rule $\mathbf{f}$ when the population state if $\mathbf{x}$ and $x_j > 0$. From section 2.1 we know that the payoff/fitness of a j-type who finds himself in a group of type i is A_j^i. The *average fitness of a type j individual at date t* is consequently $\sum_{i \in \text{supp}(j)} w_j^i(\mathbf{x}^t) A_j^i$. This average fitness will be denoted by $\pi_j(\mathbf{x}^t)$, and if we substitute for $w_j^i(\mathbf{x}^t)$ it is clear that this is given by:

$$\pi_j(\mathbf{x}) \equiv \sum_{i \in \text{supp}(j)} \frac{n_j^i}{n x_j} f_i(\mathbf{x}^t) A_j^i \tag{7}$$

Since $\pi_j(\mathbf{x})$ is the average fitness of a j-type, the *average fitness of all types in the population* will be:

$$\bar{\pi}(\mathbf{x}) = \sum_{j=1}^{m} x_j \pi_j(\mathbf{x}) \tag{8}$$

All that now remains is to describe how these fitnesses determine the next generation. At this point we have deliberately avoided saying whether time is to be thought of as discrete or continuous. In fact,

we are going to describe both, since both play important roles in the existing literature.

Beginning with the discrete time version, the well-known replicator dynamics equations (Hammerstein and Selten, 1994; Taylor and Jonker, 1978; Weibull, 1995, pp. 122-4), formalize the (sensible) notion that at time $t+1$ the proportion of the population that is of type j must equal the proportion of type j individuals at date t times the *relative fitness* of a type j individual.

Definition 2. *The discrete time replicator dynamics of the group selection model* $< I, G, \mathbf{f} >$ *is given by the equations:*

$$x_j^{t+1} = x_j^t \frac{\pi_j(\mathbf{x}^t)}{\bar{\pi}(\mathbf{x}^t)} \quad \textit{for all } j \in M. \tag{9}$$

where π_j *and* $\bar{\pi}$ *were defined in equations* (7) *and* (8)*, respectively.*

Turning next to the continuous-time case, the definition becomes (see Hofbauer and Sigmund, 1998, p. 67; Weibull, 1995, p. 72):

Definition 3. *The continuous time replicator dynamics of the group selection model* $< I, G, \mathbf{f} >$ *is given by the equations:*

$$\dot{x}_j = x_j(\pi_j(\mathbf{x}) - \bar{\pi}(\mathbf{x})) \quad \textit{for all } j \in M. \tag{10}$$

where π_j *and* $\bar{\pi}$ *were defined in equations* (7) *and* (8)*, respectively.*

Definition 4. *A steady state of a group selection model* $< I, G, f >$ *is a rest point of any of the dynamical systems* (9) *or* (10)

Clearly the steady states are the same whether time is continuous or discrete. Different notions of stability such as Lyapunov and asymptotic stability are defined as usual in either case, and the associated steady states (if any) are said to be Lyapunov stable, asymptotically stable, and so on. Since any uniform population state – *i.e.*, any state where all individuals are of the same type – will be a steady state, it is clear that stability must be considered or else the model will have no predictive power. Since stability analysis is very difficult, especially in cases with more than two strategies, the group selection model as presented is generally quite difficult to analyze.

2.4 Relationship with Trait-group Models

In this section, we briefly consider the group selection models in Maynard Smith (1964), Wilson (1975) and Cooper and Wallace (2004). As in the model described in the previous sections, the population of individuals is split into groups in these models and interaction in each group determines the number of offspring of different types that will enter the population of individuals in the next period, etc. The

difference is that assortativity does not stem from the matching process, but from prolonged interaction *within* groups (called "haystacks" in Maynard Smith, 1964, and trait-groups in Wilson, 1975).[16,17] Following from now on closely Cooper and Wallace (2004), consider a population of individuals who can be of two types. To facilitate comparison with this paper's main model, we assume a continuum population, so at any moment in time a proportion $x_1 \in [0,1]$ will be of type 1 and a proportion $x_2 = 1 - x_1 \in [0,1]$ of type 2. At the "dispersion phase", the population is split into *trait-groups* consisting of two individuals each. These groups are formed randomly (using the random matching rule of example 1.2). The two individuals in a group proceed to execute the strategy that their type dictates and get payoffs in fitness terms according to a symmetric normal-form matrix (A^i_j) (see section 2.1). The fitness of each individual determines the number of children it will send to the next generation. The trait-group's offspring is at this point "pooled" and dispersed but, crucially, the offspring are not pooled with the offspring of all the other groups as in the model of section 2.3. Instead, the trait-groups remain separated for $T > 1$ generations, so the second generation of a specific trait group is split into subgroups consisting of offspring of that trait group only (again the dispersion is pairwise and random). This second generation of subgroups proceed precisely as before to execute their type strategies, produce offspring according to the matrix (A^i_j), and in this way the trait-group's *third* generation is born. The process repeats itself until, after T generations the trait-group's combined offspring is finally returned to the aggregate population. The aggregate population is then again randomly matched into new trait-groups, and so on.

The described model is called a *T-period trait-group model*. A *steady state* is defined in the usual way as a population state $\mathbf{x} = (x_1, x_2)$ with the property that if the initial proportions of the types are x_1 and $x_2 = 1 - x_1$, respectively, then at any future date these will be the proportions of the two types in

[16] Maynard Smith (1964) studies the evolution of an altruistic gene in a structured population environment. In his model, each group (called a haystack) originates from a single fertilized female who gives birth to her children in the haystack. A proportion r of females mate with males from their own haystack while the rest $1-r$ mate at random. The species under consideration is assumed to be a diploid (each individual carries two genes, one from the mother and one from the father) and so, it obeys Mendel's laws of inheritance. As the altruistic gene a is assumed to be recessive and the egoistic gene A dominant, haystacks will be consisting solely of altruistic individuals only if the female that gave birth to its population was of type a/a and was fertilized by a father of the same type. In all other cases, haystacks will have non-altruistic members (of either type A/A or A/a). It is also assumed that after the prolonged interaction that takes place within the haystacks, any haystacks that start off with any non-altruistic individuals will lose the a gene through selection. So, just before the dispersal phase, all haystacks will be either of type a or of type A *i.*e. there will be no haystacks with mixed populations. The result is that A haystacks end up with smaller populations than those a haystacks. Thus, the haystack structure together with the parameter r lead to increased assortativity and altruistic behavior can evolve under some conditions depending on the number of haystacks and the factor by which an a type haystack population outnumbers an A-type one.

[17] Wilson (1975) derives a condition under which a trait would be selected for. His model assumes a population that spends most of its life separated in trait-groups that cannot affect one another. All individuals are assumed to have the same "baseline" fitness. Now, there are some individuals (donors) that have a specific trait (it can be altruism, aggressiveness, or any other kind of trait), the rest of them do not have the trait but are assumed to be otherwise identical to donors. Each donor gets an amount of fitness f_d (which can be negative) because he has the trait and also gives an amount of fitness f_r (again: it can be negative) to *all* individuals in his trait-group independently of their type.

The condition Wilson gets includes f_d, f_r, the trait-group size (which is assumed to be uniform across trait-groups although Wilson mentions that this is not what drives the result) and the distribution of the donors in the trait groups. The main result is that as assortativity (variance of the proportion of donors in each trait-group) increases, more altruistic traits are being selected for.

the aggregate population also.

As we now proceed to show, whether assortativity stems from matching (as in our preferred model), or from prolonged interaction in groups (as in the models of Maynard Smith, 1964, Wilson, 1975, and Cooper and Wallace, 2004) is of no consequence in the sense that for any model of the second variety one can reconstruct the steady states and dynamics with a model from the former.

Theorem 1. *Consider a T-period trait-group model with a symmetric payoff matrix (A^i_j). Consider also the normal form game G with payoff matrix (A^i_j). Then there is a matching rule $\mathbf{f}$ such that the dynamics and steady states of the group selection model $<I,G,\mathbf{f}>$ coincide with the dynamics and steady states of the trait-group model.*

A detailed proof is provided in the Appendix. Here we provide a sketch in the 2-player, 2-strategy case stressing what the matching rule associated with a specific trait-group model actually looks like.

Consider a T-period trait-group model with payoff matrix (A^i_j). We remind the reader that A^i_j indicates the payoff that a j-type individual receives when found in an i-type group. As this is a 2-player, 2-strategy model, there are three group-types: group-type 1 which contains two individuals of type 1, group-type 2 which contains one individual of type 1 and one individual of type 2, and group-type 3 which contains two individuals of type 2.

We are tracking the evolution of the population between two consecutive dispersion phases. In order to do that we need to calculate the expected fitness (number of descendants) that an individual of each type will get at the end of the T periods whereby the trait-groups remain separated from each other. To that effect, we use a law-of-large-numbers argument and we calculate the distribution of groups across all trait-groups at the T-th period. This makes us able to calculate the expected fitness for starting individuals (individuals in the original population at the dispersion phase) of each of the types in trait-group model. We also calculate the expected payoffs for the group selection model with a matching rule given by

$$f_i(\mathbf{x}) = \frac{\sum_{k=1}^{3} r_k(\mathbf{x}) g_i^k}{\sum_{l=1}^{3}\sum_{k=1}^{3} r_k(\mathbf{x}) g_l^k}$$

where $r_1(\mathbf{x}) = x_1^2$, $r_2(\mathbf{x}) = 2x_1x_2$ and $r_3(\mathbf{x}) = x_2^2$ are the components of the random matching rule $\mathbf{r}(\mathbf{x})$ (see example 1.2). The various g_i^k are only dependent on the particular payoff matrix (A^i_j) and the number of generations T that the individuals spend in their respective trait-groups isolated from the rest of the population and are, thus, given for any trait-group model. They express the (expected) proportion of i-type groups that are found in a trait-group whose first-generation parents were a pair of type j at the pair-matching stage of generation T.

Then, we take advantage of the symmetry of the replicator dynamics *i*.e. that two processes $\mathscr{A}$ and $\mathscr{B}$ have the same dynamics iff

$$\frac{\pi_j^{\mathscr{A}}(\mathbf{x})}{\pi_k^{\mathscr{A}}(\mathbf{x})} = \frac{\pi_j^{\mathscr{B}}(\mathbf{x})}{\pi_k^{\mathscr{B}}(\mathbf{x})} \text{ for all } \mathbf{x} \in S_m \text{ and } j,k \in M.$$

Finally, we show that the two payoff structures satisfy the condition mentioned above. It follows that the dynamics of the trait-group model will be the same as those of the group selection model $< I, \mathbf{f}, G >$. And, so, the steady states of the two models will coincide too.

3 Group Selection Games

In the previous section we described in full detail what we think is a natural unified model of group selection capturing both group selection based directly on non-random matching (*e*.g. Bergström, 2003; Kerr and Godfrey-Smith, 2002), and haystack/trait-group models (*e*.g. Cooper and Wallace, 2004; Maynard Smith, 1964; Wilson, 1975). Except when we described the group games in section 2.1, we made no mention of game theory — in fact the only reason we did mention this was because we need it in this section. In this section we are going to shift the perspective entirely to a game theoretic one. The basic underlying object of study will remain the same: We have a continuum $I = [0,1]$ that is now referred to as *the set of players*, we have an underlying normal form game $G \in \mathfrak{G}_{n,m}$ as described in section 2.1, and we have a matching rule $\mathbf{f} \in \mathfrak{F}_{n,m}$ as described in section 2.2. But the "story" will be very different. All three together will define a game which we call a *group selection game*:

Definition 5. (Group Selection Games) *A group selection game is a tuple* $< I, G, \mathbf{f} >$ *where* I *is a continuum of players,* $G \in \mathfrak{G}_{n,m}$ *is a symmetric normal form game, and* $\mathbf{f} \in \mathfrak{F}_{n,m}$ *is a matching rule.*

Here is the structure of the game: As mentioned, there is a continuum $I = [0,1]$ of agents. These are identical, in particular they have the same finite set of pure strategies $M = \{1, \ldots, m\}$ given from the normal form game G. The game is symmetric, so we can conveniently summarize a *(pure) strategy profile* by its frequency distribution $\mathbf{x} = (x_1, \ldots, x_m) \in S_m$ where the j'th coordinate is the *fraction* of the players whose strategy is $j \in \{1, \ldots, m\}$. The individual player takes $\mathbf{x}$ as given and being infinitesimally small his own choice of strategy will not affect the relative proportions expressed in $\mathbf{x}$. Now, in one description, the game has two *stages*: In the *first* stage, players choose their strategies and in the *second* stage they are allocated into groups of the same finite size $n \in \{2, 3, \ldots\}$ where they execute their strategies.[18] What is crucial here is that agents do not know with certainty which group they will end up in when they choose their strategies. However, because the structure of the game is known (common knowledge), an agent *will* know the *rule* according to which agents are allocated into groups, and so will be able to calculate the probability of ending up in any particular type of group after a specific strategy is chosen. This brings us back to (4) of section 2.2. Recall from that section that if $x_j > 0$ then $w_j^i(\mathbf{x})$ is the fraction of type j individuals that are allocated into groups of type i under the matching rule $\mathbf{f}$ (and the state $\mathbf{x}$):

[18] From a game theoretic perspective, it is much more natural to think of this as a situation involving uncertainty (a type of Bayesian game). But the imperfect information perspective actually turns out to be non-standard because probabilities are endogenously determined.

$$w_j^i(\mathbf{x}) = \frac{n_j^i f_i(\mathbf{x})}{n x_j} \tag{11}$$

The case where $x_j = 0$ is returned to in a moment. It is clear that from an expected payoff point-of-view, $w_j^i(\mathbf{x})$ is the *ex-ante probability a j-strategist has of being "drawn" into group i*. It follows that the *expected payoff* to strategy j will equal,

$$\pi_j(\mathbf{x}) = \sum_{i \in \text{supp}(j)} w_j^i(\mathbf{x}) A_j^i \,, \tag{12}$$

where we remind the reader that A_j^i is the payoff received from playing strategy j in a group of type i (section 2.1); and $\text{supp}(j)$ is the set of groups that contain at least one j-strategist (section 2.2). Comparing with section 2.3, this expected payoff precisely coincides with the average fitness to a type j individual in the (deterministic) evolutionary group selection model.

Now, for the previous two definitions it is required that $x_j > 0$. The definition of the w_j^i's in (11) and so the definition of the π_j's in (12) are extended to the boundary of S_m ($\text{bd}_j(S_m) = \{x \in S_m : x_j = 0\}$) by taking $w_j^i(\mathbf{x}) = \lim_{\tilde{x}_j \downarrow 0} \frac{n_j^i f_i(\tilde{\mathbf{x}})}{n \tilde{x}_j}$ whenever $\mathbf{x} \in \text{bd}_j(S_m)$. Evidently, we need to assume that these limits exist for these extensions to be well-defined (featured in theorem 2 below).[19] Note that the limit $\lim_{x_j \downarrow 0} \frac{f_i(\tilde{\mathbf{x}})}{\tilde{x}_j}$, if it exists, is precisely the j'th partial (upper) derivative of f_i, $\partial_j^+ f_i(\mathbf{x})$. Hence $w_j^i(\mathbf{x}) = \frac{n_j^i}{n} \partial_j^+ f_i(\mathbf{x})$ when $x_j = 0$.

Finishing now the description of the game, players are allowed to choose mixed strategies, *i.e.*, their strategy set is $S_m = \{y \in \mathbb{R}_+^m : \sum_{j \in M} y_j = 1\}$. The expected payoff to a mixed strategy $\mathbf{y} \in S_m$ is then $\mathbf{y} \cdot \pi(\mathbf{x})$ where $\pi(\mathbf{x}) \equiv (\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x}))$. Note that if all players choose the *same* mixed strategy $\mathbf{y}$, the state $\mathbf{x}$ will necessarily be equal to $\mathbf{y}$. Thus the definition of an equilibrium follows naturally:

Definition 6. (NEGS) *Let $< I, G, \mathbf{f} >$ be a group selection game. A strategy $\mathbf{x}^* \in S_m$ is a* Nash Equilibrium with Group Selection (NEGS) *if:*

$$\mathbf{x}^* \cdot \pi(\mathbf{x}^*) \geq \mathbf{y} \cdot \pi(\mathbf{x}^*) \quad \textit{for all } \mathbf{y} \in S_m. \tag{13}$$

The average payoff (the welfare) *at a NEGS $\mathbf{x}^*$ is denoted by $\bar{\pi}(\mathbf{x}^*) = \mathbf{x}^* \cdot \pi(\mathbf{x}^*)$.*

Intuitively, agents in a group selection game take the matching rule's payoff effects into account, and integrate into their optimal choices the fact that different choices of strategies are associated with different probability distributions over opponents' strategies. In a NEGS, these probabilities are "self-fulfilling" in the sense that agents' *ex-post* decisions lead to the *ex-ante* probabilities upon which the

[19] Note that from a formal point of view, this is actually not acceptable because the game will not be well-defined if the expected payoffs are not well-defined. In a previous version of this paper we took instead $w_j^i(\mathbf{x}) = \limsup_{\tilde{x}_j \downarrow 0} \frac{n_j^i f_i(\tilde{\mathbf{x}})}{n \tilde{x}_j}$ which is always well-defined. This, however, tends to lead to confusion. Hence the present slight violation of mathematical rigor.

decisions are based. In the following sections we shall see that this concept has a very close relationship with the steady states of the canonical group selection model (section 2.3). The remainder of this section is devoted to showing that the NEGS concept is well-founded, and strengthening the equilibrium concept.

Our first result states that any group selection game has an equilibrium when certain regularity conditions are satisfied by the matching rule. Note that the differentiability requirement trivially will be satisfied if the matching rule is differentiable at the boundary of S_m:

Theorem 2. *Let $< I, G, \mathbf{f} >$ be a group selection game and assume that $\mathbf{f}$ is continuous and that the (upper) partial derivatives $\partial_j^+ f_i(\mathbf{x})$ exist whenever $x_j = 0$ (for all $j \in M$ and $i \in \text{supp}(j)$). Then $< I, G, \mathbf{f} >$ has a NEGS.*

Proof. See Appendix. □

Notice that all matching rules in Example 1 satisfy the conditions of Theorem 2.

Probably the most commonly used solution concept in evolutionary game theory is that of an *evolutionarily stable strategy* or *ESS* (Maynard Smith and Price, 1973). As Maynard Smith (1982, p. 14) puts it: "*If I is a stable strategy, it must have the property that, if almost all members of the population adopt I, then the fitness of these typical members is greater than that of any possible mutant; otherwise, the mutant could invade the population and I would not be stable.*"

In the literature, ESS is usually defined in games with random matching and in the special case when $n = 2$ (see Hofbauer and Sigmund, 1998, p. 63). An appropriate generalization of the ESS concept to include non-random matching and any number of strategies is the following.

Definition 7. (ESSGS) *Let $< I, G, \mathbf{f} >$ be a group selection game. A strategy $\hat{\mathbf{x}} \in S_m$ is an* Evolutionarily Stable Strategy with Group Selection (ESSGS) *if for each $\mathbf{y} \in S_m \setminus \{\hat{\mathbf{x}}\}$, there exists $\bar{\epsilon}_{\mathbf{y}} > 0$ such that*

$$\hat{\mathbf{x}} \cdot \pi(\epsilon \mathbf{y} + (1-\epsilon)\hat{\mathbf{x}}) > \mathbf{y} \cdot \pi(\epsilon \mathbf{y} + (1-\epsilon)\hat{\mathbf{x}}) \qquad \text{for all } \epsilon \in (0, \bar{\epsilon}_{\mathbf{y}}). \tag{14}$$

As Maynard Smith's quote suggests, the central idea behind the ESS (and therefore the ESSGS) concept is that of *non-invasion*. This means that a (monomorphic) population where all individuals use an ESSGS $\hat{\mathbf{x}}$ cannot be successfully invaded by a small but measurable (of measure ϵ up to $\bar{\epsilon}_{\mathbf{y}}$) group of individuals using any other strategy $\mathbf{y} \in S_m$ in the sense that in the new population – composed of $1-\epsilon$ $\hat{\mathbf{x}}$-strategists and ϵ $\mathbf{y}$-strategists – the individuals using the ESS will get higher expected payoff than the invaders (the $\mathbf{y}$-strategists). This is exactly what condition (14) expresses.

The ESSGS concept is a strengthening of the NEGS concept, just as the traditional notion of an ESS is a strengthening of Nash equilibrium:

Theorem 3. *Let $< I, G, \mathbf{f} >$ be a group selection game with $\mathbf{f}$ satisfying the assumptions of Theorem 2. Then any ESSGS is a NEGS.*

Proof. By way of contradiction, let us assume that some $\hat{\mathbf{x}} \in S_m$ is an ESSGS but *not* a NEGS. Then, there exists some $\mathbf{y} \in S_m$ such that $(\mathbf{y}-\hat{\mathbf{x}}) \cdot \pi(\hat{\mathbf{x}}) > 0$. But from the definition of an ESSGS, there must exist some $\bar{\epsilon}_{\mathbf{y}} \in (0,1)$ such that for all $\epsilon \in (0, \bar{\epsilon}_{\mathbf{y}})$, $(\mathbf{y}-\hat{\mathbf{x}}) \cdot \pi(\epsilon\mathbf{y}+(1-\epsilon)\hat{\mathbf{x}}) < 0$. By continuity therefore $(\mathbf{y}-\hat{\mathbf{x}}) \cdot \pi(\hat{\mathbf{x}}) \leq 0$. A contradiction. □

We finish this section with an alternative characterization of an ESSGS which invokes the notion of *local superiority* (see for example Weibull, 1995, p. 45) defined as follows.

Definition 8. (Local Superiority) *A strategy $\hat{\mathbf{x}} \in S_m$ is called locally superior if there exists a neighborhood U of $\hat{\mathbf{x}}$ such that for all $\mathbf{y} \in U \setminus \{\hat{\mathbf{x}}\}$*

$$\hat{\mathbf{x}} \cdot \pi(\mathbf{y}) > \mathbf{y} \cdot \pi(\mathbf{y}). \tag{15}$$

Proposition 4. *A strategy $\hat{\mathbf{x}} \in S_m$ is an ESSGS if and only if it is locally superior.*

Proof. The proof is essentially identical to that of Proposition 2.6 in Weibull (1995, pp. 45–46). There are only two changes that need to be made: *(i)* The score function now is $f(\epsilon, \mathbf{y}) = (\hat{\mathbf{x}}-\mathbf{y}) \cdot \pi(\epsilon\mathbf{y}+(1-\epsilon)\hat{\mathbf{x}})$ and *(ii)* now there is *not necessarily* "at most one ϵ" for which $f(\epsilon, \mathbf{y}) = 0$. This is because the payoff function is not necessarily bilinear (that is π is not necessarily linear). In the case where there is more than one such ϵ, we can set $\epsilon_0 = \min\{\epsilon \in (0,1] | f(\epsilon, \mathbf{y}) = 0\}$. Now everything is in place and the result carries through. □

4 Group Selection and Evolutionary Game Theory

In evolutionary models with random matching, there is a clear and well-known connection between dynamic models of the *replicator* type and game theoretic concepts such as Nash equilibrium and evolutionarily stable strategies (Hofbauer and Sigmund, 1998). The main purpose of this section, and indeed the main theoretical contribution of this paper, is to show that the previous section's notions of a Nash equilibrium and evolutionarily stable strategy with group selection (NEGS and ESSGS, respectively) are for evolutionary models of group selection what Nash equilibria and ESS are for evolutionary models with random matching. Precisely, we are going to show that any Nash equilibrium under group selection (NEGS) is a steady state of the corresponding evolutionary dynamical system (the replicator dynamics).[20] Furthermore, we are going to prove that any *stable steady state* of the replicator dynamics (be it Lyapunov or in the ω-limit sense) will be a NEGS. These results directly parallel known results on models with random matching (see *e.g.* Theorem 7.2.1. in Hofbauer and Sigmund, 1998). Finally, we will prove that any ESSGS is asymptotically stable for the associated replicator dynamics.[21] Again, this result transfers a well-known result from the random matching case over to models with non-random matching/group selection (see *e.g.* Proposition 3.10. in Weibull, 1995).

[20] Since the evolutionary dynamics have the same steady states in discrete and continuous time, this statement obviously applies to either.

[21] Note that, just as in the standard case with random matching, the stability statements refer to continuous time replicator dynamics only.

It should be noted that all of this section's results also apply to trait-group models since by Theorem 1, such models can be recast as non-random matching models *with the same dynamics* and therefore, of course, the same set of (stable) steady states.

Before turning to the main results, the following observation clarifies the exact relationship between our results and the mentioned results on random matching. Precisely, our results generalize existing ones, since as we now proceed to show, if matching is assumed to be random (example 1.2) in a group selection game, one precisely recoups the traditional Nash Equilibrium concept:

Theorem 5. *Let $< I, G, \mathbf{f} >$ be a group selection game under random matching. Then the set of Nash equilibria with group selection coincides with the set of symmetric Nash equilibria in the underlying normal form game G. Likewise, when matching is random the set of evolutionarily stable strategies with group selection coincides with the set of evolutionarily stable strategies.*

Proof. See Appendix.

Theorem 5 shows that NEGS unifies the treatment of models with or without random matching in evolutionary game theory. In particular, theorem 6 implies as special cases Theorem 7.2.1. in Hofbauer and Sigmund (1998) and Proposition 3.10. in Weibull (1995). But of course, the more interesting cases arise when matching is not random.

Theorem 6. *Let $< I, G, \mathbf{f} >$ be a group selection game and assume that* $\mathbf{f}$ *satisfies the assumptions of Theorem 2 and consider the evolutionary steady states of the associated dynamical systems* (9)-(10)*. Then,*

1. *Any NEGS is a steady state of the discrete time replicator dynamics* (9) *as well as the continuous time replicator dynamics* (10)*.*

2. *If $\mathbf{x}^*$ is the ω-limit of an orbit $x(t)$ of the replicator dynamics* (10) *that lies everywhere in the interior of S_m, then $\mathbf{x}^*$ is a NEGS.*

3. *If $\mathbf{x}^*$ is Lyapunov stable for the replicator dynamics* (10)*, then $\mathbf{x}^*$ is a NEGS.*

4. *Assume that $\mathbf{f}$ is of class C^1. Then if $\mathbf{x}^*$ is an ESSGS, it is asymptotically stable under the replicator dynamics* (10)*.*

Proof.

1. Let $\mathbf{x}^* \in S_m$ be a NEGS, $I(\mathbf{x}) \equiv \{j \in M | x_j > 0\}$ and $O(\mathbf{x}) \equiv \{j \in M | x_j = 0\}$. Then, from (13) we get for all $\mathbf{y} \in S_m$: $\sum_{j \in M} y_j \pi_j(\mathbf{x}^*) \leq \sum_{l \in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) + \sum_{q \in O(\mathbf{x}^*)} x_q^* \pi_q(\mathbf{x}^*)$. Hence:

$$\sum_{j \in M} y_j \pi_j(\mathbf{x}^*) \leq \sum_{l \in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) \tag{16}$$

Now let $p = \arg\max_{j \in M} \pi_j(\mathbf{x}^*)$ and $r = \arg\max_{l \in I(\mathbf{x}^*)} \pi_l(\mathbf{x}^*)$. Clearly, $\sum_{l \in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) \leq \pi_r(\mathbf{x}^*) \leq \pi_p(\mathbf{x}^*)$ where the second inequality holds because $I(\mathbf{x}^*) \subseteq M$. Hence for all $\mathbf{y} \in S_m$:

$$\sum_{j\in M} y_j \pi_j(\mathbf{x}^*) \leq \sum_{l\in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) \leq \pi_r(\mathbf{x}^*) \leq \pi_p(\mathbf{x}^*) \tag{17}$$

Taking $\mathbf{y} = (0,\ldots,0,\underbrace{1}_{p\text{-th}},0\ldots,0)$, we get $\pi_p(\mathbf{x}^*) \leq \sum_{l\in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) \leq \pi_r(\mathbf{x}^*) \leq \pi_p(\mathbf{x}^*)$ which obviously implies that $\sum_{l\in I(\mathbf{x}^*)} x_l^* \pi_l(\mathbf{x}^*) = \pi_r(\mathbf{x}^*)$. But this is only possible if $\pi_j(\mathbf{x}^*) = \pi_k(\mathbf{x}^*)$ for all $j,k \in I(\mathbf{x}^*)$, and this in turn implies that $\pi_j(\mathbf{x}^*) = \mathbf{x}^* \cdot \pi(\mathbf{x}^*)$ for all $j \in I(\mathbf{x}^*)$. From equation (10), we therefore get $\dot{x}_j^* = 0$ for all $j \in M$, *i.e.*, $\mathbf{x}^*$ is a steady state.

2. Assume that $x(t) \in \operatorname{int} S_m$ converges to $\mathbf{x}^*$ and that $\mathbf{x}^*$ is not a NEGS. That $\mathbf{x}^*$ is not a NEGS means that there exists a j with $\mathbf{e}_j \cdot \pi(\mathbf{x}^*) = \pi_j(\mathbf{x}^*) > \mathbf{x}^* \cdot \pi(\mathbf{x}^*)$. Hence $(\pi_j(\mathbf{x}^*) - \mathbf{x}^* \cdot \pi(\mathbf{x}^*)) \geq \epsilon > 0$ for some $\epsilon > 0$. Since $x(t)$ converges and π_j is continuous on the interior, $(\pi_j(x(t)) - \mathbf{x}^* \cdot \pi(x(t)) \to 0$ as $t \to \infty$. This is a contradiction. Note that at the boundary, this holds because we have defined the π_j's so that they are continuous onto the boundary. If we had not done that, the claim would in general be false for a vector $\mathbf{x}^*$ on the boundary. This same problem does not arise with random matching/in the usual replicator dynamics setting because the payoff functions trivially are continuous. This shows exactly why our "continuous extension to the boundary" is the right thing to do.

3. Precisely as in the previous proof and by continuity of the π's we get that if $\mathbf{x}^*$ is not an NEGS then there exists an $\epsilon > 0$ such that for all x in a neighborhood of $\mathbf{x}^*$: $(\pi_j(x) - x \cdot \pi(x)) \geq \epsilon > 0$. For such x, the component x_i increases exponentially which contradicts Lyapunov stability.

4. Following Weibull (1995, pp. 95–100), we will use Lyapunov's direct method to prove the proposition. What we need is to find a scalar function H that is defined on a neighborhood Q of $\mathbf{x}^*$ which has the following properties: *(i)* H is continuously differentiable on Q, *(ii)* $H(\mathbf{x}^*) = 0$, *(iii)* $H(\mathbf{y}) > 0$ for all $\mathbf{y} \in Q \setminus \{\mathbf{x}^*\}$ and *(iv)* $\dot{H}(\mathbf{y}) = \frac{d}{dt} H(\mathbf{y}) < 0$ for all $\mathbf{y} \in Q \setminus \{\mathbf{x}^*\}$.

Let us consider the set $Q_{\mathbf{x}^*} \equiv \{\mathbf{y} \in S_m | I(\mathbf{x}^*) \subseteq I(\mathbf{y})\}$ *i.*e. the set of all states that assign positive weights to all the pure strategies that $\mathbf{x}^*$ assigns positive weights. Obviously, $\mathbf{x}^* \in Q_{\mathbf{x}^*}$ and $Q_{\mathbf{x}^*}$ is an open set (in the topology induced from $\mathbb{R}^m$). So, $Q_{\mathbf{x}^*}$ is a neighborhood of $\mathbf{x}^*$. We will show that the function $H_{\mathbf{x}^*} : Q_{\mathbf{x}^*} \to \mathbb{R}$ defined by

$$H_{\mathbf{x}^*}(\mathbf{y}) = \sum_{j\in I(\mathbf{x}^*)} x_j^* \log\left(\frac{x_j^*}{y_j}\right)$$

satisfies all of the above conditions (*i–iv*) *i.*e. that it is a strict local Lyapunov function on $Q_{\mathbf{x}^*}$.

First of all, it is easy to verify that *(i)* $H_{\mathbf{x}^*}$ is continuously differentiable on $Q_{\mathbf{x}^*}$ and that *(ii)* $H_{\mathbf{x}^*}(\mathbf{x}^*) = 0$. Now, as $\mathbf{x}^*$ is an ESSGS, we know that there exists a neighborhood U of $\mathbf{x}^*$ such that condition (14) holds for all $\mathbf{y} \in U$. We will consider the restriction of $H_{\mathbf{x}^*}$ on the set $U \cap Q_{\mathbf{x}^*}$, a neighborhood of $\mathbf{x}$. The next step is to show that $H_{\mathbf{x}^*}$ is strictly positive on $U \cap Q_{\mathbf{x}^*}$. As the function $-\log(\cdot)$ is convex, we get from Jensen's inequality:

$$H_{\mathbf{x}^*}(\mathbf{y}) = \sum_{j\in I(\mathbf{x}^*)} x_j^* \left(-\log\left(\frac{y_j}{x_j^*}\right)\right) \geq -\log\left(\sum_{j\in I(\mathbf{x}^*)} x_j^* \left(\frac{y_j}{x_j^*}\right)\right) \geq -\log\left(\sum_{j\in M} y_j\right) = 0$$

Now, in the case where $I(\mathbf{x}^*) = I(\mathbf{y})$, the first inequality is strict (because of the log's strict concavity) and in the case where $I(\mathbf{x}^*) \subsetneq I(\mathbf{y})$, the second inequality is strict. In any case, we will always have that *(iii)* $H_{\mathbf{x}^*}(\mathbf{y}) > 0$ for all $\mathbf{y} \in U \cap Q_{\mathbf{x}^*} \setminus \{\mathbf{x}^*\}$.

The last step is to show that $\dot{H}_{\mathbf{x}^*}$ is negative for all $\mathbf{y} \in U \cap Q_{\mathbf{x}^*} \setminus \{\mathbf{x}^*\}$. Indeed:

$$\dot{H}_{\mathbf{x}^*}(\mathbf{y}) = \sum_{j \in I(\mathbf{x}^*)} \partial_j H_{\mathbf{x}^*}(\mathbf{y}) \dot{y}_j = - \sum_{j \in I(\mathbf{x}^*)} \frac{x_j^*}{y_j} \dot{y}_j$$

and using equation (10), we get:

$$\dot{H}_{\mathbf{x}^*}(\mathbf{y}) = -\mathbf{x}^* \cdot \pi(\mathbf{y}) + \bar{\pi}(\mathbf{y}) = (\mathbf{y} - \mathbf{x}^*)\pi(\mathbf{y})$$

which we know is negative because of (14). So, the final condition *(iv)* is satisfied. □

Remember from the discussion at the end of section 2.3, that the evolutionary/replicator dynamics model has limited predictive power without accompanying stability analysis (*e.g.*, *any* uniform population state will be a steady state). This places high demands on the analyst because models with non-random matching lead to complicated non-linear dynamical systems whose stability properties are non-trivial to analyze. As we shall see in the following sections, our new game theoretic concepts (NEGS and ESSGS) to a large extent overcome these problems. The reason is the previous theorem which ensures that when we look at the set of NEGS we select all steady states for the replicator dynamics that are (Lyapunov) stable, in particular we capture any steady state for the replicator dynamics that is also an ESSGS.

5 Some Examples

In this section we analyze a number of group selection games with 2 players and 2 strategies under different matching rules. We apply a method that allows us to graphically portray matching rules and makes the process of finding NEGSs and ESSGSs as simple as finding the intersections of two curves.[22] We also provide comparative statics results for the class of matching rules with a constant index of assortativity. We restrict ourselves to analysis of consistent matching rules (see section 2.2) throughout this section.

5.1 Hawk-Dove/Chicken

A game often analyzed in the literature of both economics and biology is the Hawk-Dove (HD) game.[23] Players in this game have two available pure strategies: Hawk (H) and Dove (D). In our formalization, a Hawk-Dove game is a 2×2 game with $A_2^2 > A_1^1 > A_1^2 > A_2^3$.[24] The payoff matrices of three Hawk-Dove games are depicted in Table 1.

[22] The method is described in the Appendix in detail.

[23] Economists usually refer to this game as Chicken rather than Hawk-Dove.

[24] As a convention in what follows and without loss of generality we will assume that $A_1^1 \geq A_2^3$.

	D	H
D	50,50	40,80
H	80,40	0,0

(a) $A_1^2 + A_2^2 > 2A_1^1$

	D	H
D	50,50	40,60
H	60,40	0,0

(b) $A_1^2 + A_2^2 = 2A_1^1$

	D	H
D	50,50	20,60
H	60,20	0,0

(c) $A_1^2 + A_2^2 < 2A_1^1$

Table 1: The payoff matrices of three Hawk-Dove games.

In this game, there are three Nash Equilibria: Two asymmetric ones in pure strategies (H,D) and (D,H) and a symmetric one in mixed strategies where both players play Dove with probability $p_D = \frac{A_1^2 - A_2^3}{A_1^2 + A_2^2 - A_1^1 - A_2^3}$ and Hawk with probability $p_H = \frac{A_2^2 - A_1^1}{A_1^2 + A_2^2 - A_1^1 - A_2^3}$. In the group selection game the state will be summarized by x which indicates the proportion of the population that follows D.

Equilibria of the Group Selection Game

Now, in order to find the NEGS and ESSGSs of the PD game, we follow the methodology proposed in the Appendix. The equilibrium curves of the games in Table 1 are shown in Figure 1.

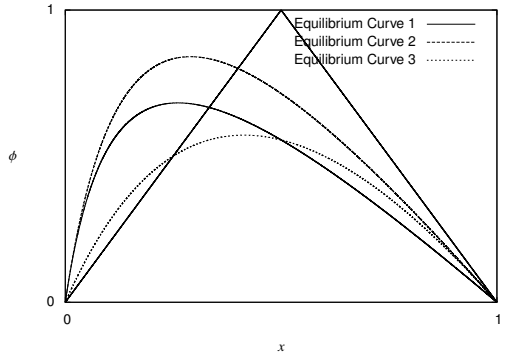


Figure 1: Equilibrium curves of the HD games in Table 1.

Random Matching As expected, the unique equilibrium of the group selection game under the Random Matching rule yields the unique symmetric Nash equilibrium of the game where a proportion $x^* = \frac{A_1^2 - A_2^3}{A_1^2 + A_2^2 - A_1^1 - A_2^3}$ of the population play D.

Complete Segregation Under complete segregation, there is a unique equilibrium of the group selection game $x^* = 1$ where the whole population follows D.

Constant Index of Assortativity Under a constant index of assortativity rule (see Appendix), the group selection game has a unique equilibrium given by:

$$x^* = \begin{cases} \frac{A_1^1 - A_2^3}{1-\alpha} + A_1^2 - A_1^1 \\ \overline{A_1^2 + A_2^2 - A_1^1 - A_2^3} \\ 1 \end{cases} \quad \begin{matrix} \text{if } 0 \leq \alpha < \frac{A_2^2 - A_1^1}{A_2^2 - A_2^3} \\ \\ \text{if } \frac{A_2^2 - A_1^1}{A_2^2 - A_2^3} \leq \alpha \leq 1 \end{matrix} \tag{18}$$

The equilibrium-finding process is shown in Figure 2 for constant index of assortativity rules for different values of α. The comparative statics results are summarized in Figure 3.

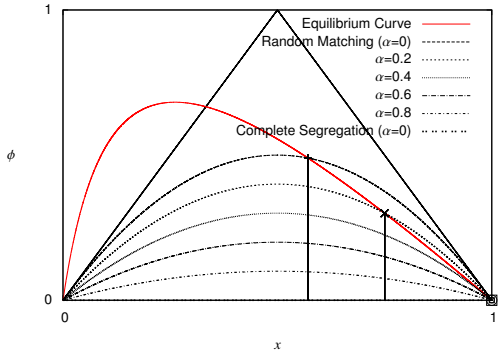


Figure 2: NEGS with a constant index of assortativity in a Hawk-Dove game.

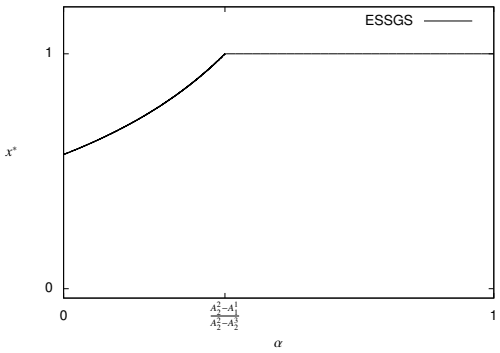


Figure 3: NEGS as a function of the index of assortativity in a Hawk-Dove game.

In the HD game, strategies $x \in [0, \frac{A_1^2 - A_2^3}{A_2^2 + A_1^2 - 2A_2^3})$ cannot be equilibria of the group selection game *under any (consistent) matching rule* due to constraint (31) on ϕ.

Welfare

In order to conduct welfare analysis, we use the methodology described in the Appendix. The isogrowth diagram of a Hawk-Dove game is shown in Figure 4. The comparison of equilibrium welfare in the group selection game and the normal form game is shown in Figure 5. Notice that the equilibrium welfare curve is not defined for $x \in [0, \frac{A_1^2 - A_2^3}{A_2^2 + A_1^2 - 2A_2^3})$ as these states can never be attained as equilibria of the group selection game. In all HD games, the level of equilibrium welfare is strictly increasing with the proportion of Doves in the population and thus, maximum equilibrium welfare is obtained when the equilibrium state is $x = 1$ *i.e.* when the whole population follows D.

Now, in the case where $A_1^2 + A_2^2 \leq 2A_1^1$, maximum equilibrium welfare coincides with the maximum expected payoff players using symmetric strategies can get in the normal form game (which is attained when both players play D with certainty).

In the case where $A_1^2 + A_2^2 > 2A_1^1$, the normal form game maximum expected payoff (under symmetric strategies) is obtained if both players play D with probability $p_D^* = \frac{A_1^2 + A_2^2 - 2A_2^3}{2(A_1^2 + A_2^2 - A_1^1 - A_2^3)}$. However, when a matching rule that makes $x = p_D^*$ an equilibrium is implemented, equilibrium welfare is reduced below A_1^1. This is because the proportion of Hawk-Dove pairs – which are efficient in the utilitarian sense – is reduced in favor of more Hawk-Hawk and Dove-Dove pairs which are not as efficient.

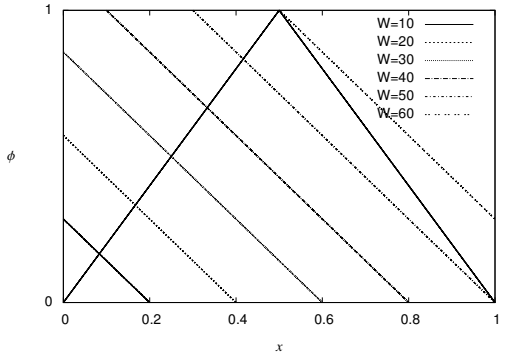


Figure 4: Isogrowth diagram for a HD game.

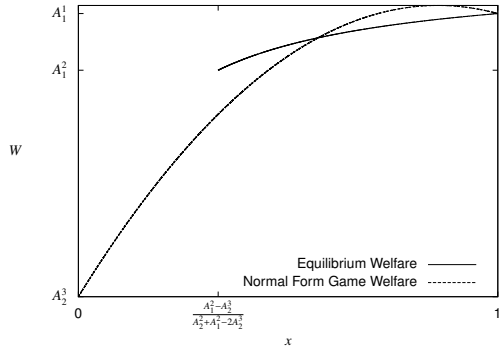


Figure 5: Equilibrium welfare and normal form payoff for a HD game.

5.2 Stag Hunt

Another game with interesting insights on social behavior is the Stag Hunt.[25] In our notation a SH game will have values $A_1^1 > A_2^2 \geq A_2^3 > A_1^2$. The payoff matrices of three SH game are depicted in Table 2.

	S	H
S	100, 100	0, 70
H	70, 0	60, 60

	S	H
S	100, 100	0, 70
H	70, 0	70, 70

	S	H
S	100, 100	0, 80
H	80, 0	70, 70

Table 2: The payoff matrices of three SH games.

The game has three Nash equilibria, all symmetric. Two of them are in pure strategies (S,S) and (H,H) and one in mixed strategies where both players play S with probability $p_S = \frac{A_2^3 - A_1^2}{A_1^1 + A_2^3 - A_1^2 - A_2^2}$ and H with probability $p_H = \frac{A_1^1 - A_2^2}{A_1^1 + A_2^3 - A_1^2 - A_2^2}$. Also, we require that $A_2^2 + A_2^3 > A_1^1 + A_1^2$ so that even though the pure strategy equilibrium (S,S) is payoff dominant (*i.e.* it yields higher payoffs for both players), the pure strategy equilibrium (H,H) is risk dominant (*i.e.* if we assume that players are not sure which strategy their opponent will follow and assign equal probabilities to the two strategies, then the expected payoff from playing H exceeds the expected payoff from playing S).[26]

The importance of the Stag Hunt is that it shows that although the efficient outcome (S,S) is a Nash equilibrium, it may not always be selected. More than that, it has been shown that in some stochastic evolutionary models the risk dominant outcome occurs with probability 1 (Young, 1993) and that in global games, the risk dominant outcome is the only one that survives iterative elimination of dominated strategies when noise tends to vanish (Carlsson and Van Damme, 1993). So the literature suggests that in several environments it is the risk dominant rather than the payoff dominant outcome that prevails. We show that in our model this inefficiency can be amended under matching rules with high enough assortativity.

[25] For an extensive analysis see Skyrms (2004).
[26] See Carlsson and Van Damme (1993).

Equilibria of the Group Selection Game

The equilibrium curves of the games in Table 2 are shown in Figure 6.

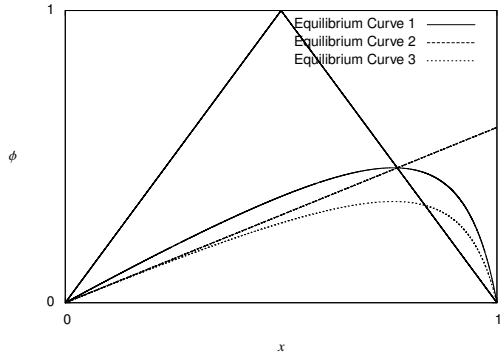


Figure 6: Equilibrium curves of the SH games in Table 2.

Random Matching As before, under the Random Matching rule, as expected, we get that there are three NEGS in the group selection game that coincide with the three Nash equilibria of the normal form game: two stable ones (ESSGSs) at $x = 0$ and $x = 1$ (where the whole population follows H and S respectively) and a NEGS which is not an ESSGS where a fraction of the population $x = \frac{A_2^3 - A_1^2}{A_1^1 - A_1^2 + A_2^3 - A_2^2}$ follows S.

Complete Segregation Under the complete segregation matching rule, there is only one NEGS where the whole population follows S ($x = 1$) and it is also an ESSGS.

Constant Index of Assortativity Under a matching rule with a constant index of assortativity α we have two cases depending on the value of α:

- if $\alpha \leq \frac{A_2^3 - A_1^2}{A_1^1 - A_1^2}$ we have three NEGS: two NEGS that are also ESSGSs where everybody follows H ($x = 0$) or S ($x = 1$) and a NEGS which is not an ESSGS where a proportion of the population $x = \frac{\frac{A_2^3 - A_1^1}{1-\alpha} + A_1^1 - A_1^2}{A_1^1 + A_2^3 - A_1^2 - A_2^2}$ follows S
- if $\alpha > \frac{A_2^3 - A_1^2}{A_1^1 - A_1^2}$ there is only one NEGS that is also an ESSGS where the whole population follows S ($x = 1$).

The equilibrium-finding process is shown in Figure 7 for constant index of assortativity rules for different values of α. The comparative statics results are summarized in Figure 8.

As in the case of the Hawk-Dove game, in the Stag Hunt there are some states that cannot be attained as equilibria under any matching rule. At these states, namely $x \in \left(\frac{A_1^1 - A_1^2}{2A_1^1 - A_1^2 - A_2^2}, 1 \right)$, the dynamics will tend to lead the population towards $x = 1$ where they all follow S under any matching rule. So, if

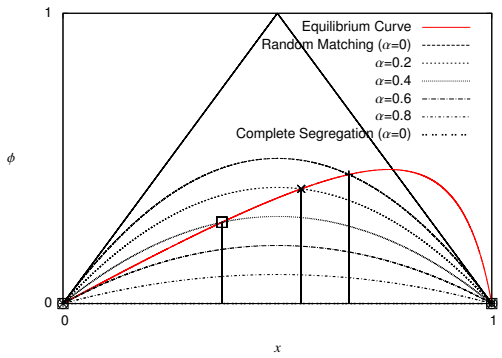


Figure 7: NEGS with a constant index of assortativity in a SH game.

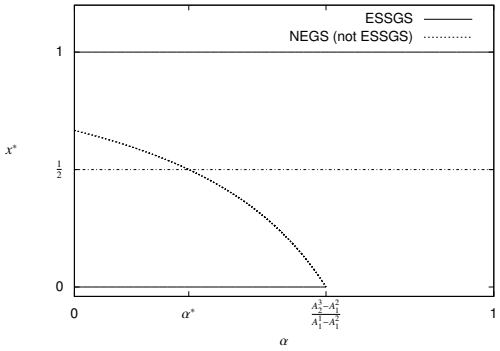


Figure 8: NEGS as a function of the index of assortativity.

it happens that the system reaches one of these states, then it will be eventually brought to the state where the whole population uses the efficient strategy S.

Risk Dominance Notice that there is a value $\alpha^* = \frac{(A_2^2-A_1^2)-(A_1^1-A_2^3)}{(A_2^2-A_1^2)+(A_1^1-A_2^3)}$ for which the basin of attraction of the ESSGS at $x=1$ is greater than that of the ESSGS at $x=0$ iff $\alpha \in (\alpha^*, 1]$. We can interpret that as follows: Assume that players in the population do not know whether each of the other players is going to play S or H and so, using the principle of insufficient reason, they ascribe equal probabilities (equal to 0.5 each) to each other player following S and H.[27] Then, if $\alpha \in (\alpha^*, 1]$ the expected payoff for a player following S is higher than his expected payoff when he follows H and so, given the aforementioned beliefs, it is a best response for all of them to follow H, leading to the state being $x=1$. Conversely when $\alpha \in [0, \alpha^*)$.

So, in the terms described above, we can have a notion of *risk dominance* in the group selection game. Of course – having assumed that $A_2^2+A_2^3 > A_1^1+A_1^2$ as is usually done in Stag Hunt games – in the case where $\alpha = 0$, it is always the case that the risk dominant equilibrium is the one where the whole population follows H ($x=0$).

Welfare

The isogrowth diagram of a Stag Hunt game is shown in Figure 9. The comparison of equilibrium welfare in the group selection game and the normal form game is shown in Figure 10. Notice that the equilibrium welfare curve is not defined for $x \in \left(\frac{A_1^1-A_1^2}{2A_1^1-A_1^2-A_2^2}, 1\right)$ as these states can never be attained as equilibria of the group selection game. The maximum level of welfare is obtained when the equilibrium state is the one where everybody follows S ($x=1$) and it coincides with the maximum expected payoff players using symmetric strategies can get in the normal form game.

[27] See also Carlsson and Van Damme (1993).

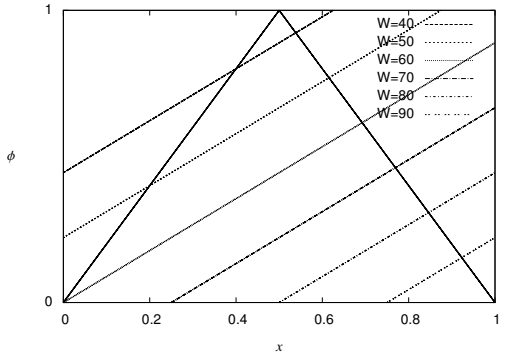


Figure 9: Isogrowth diagram for a SH game.

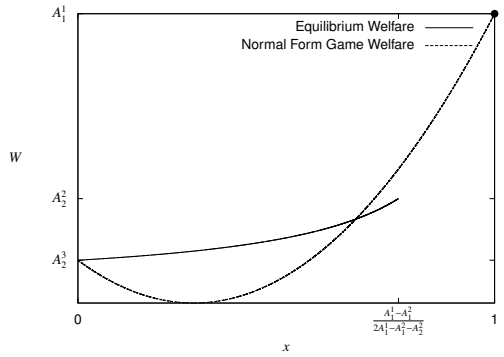


Figure 10: Equilibrium welfare and normal form payoff for a SH game.

5.3 Prisoner's Dilemma

As a final application, we leave arguably the most analyzed game in the literature and which has served as the canonical way to model altruistic behavior: The Prisoner's Dilemma (PD). The two players involved in the game have two possible (pure) strategies each: Cooperate (C) or Defect (D). In our notation, a PD game is a game with $A_2^2 > A_1^1 > A_2^3 > A_1^2$. The payoff matrices of three PD games are shown in Table 3.

	C	D
C	40, 40	0, 100
D	100, 0	20, 20

(a) $A_1^2 + A_2^2 > A_1^1 + A_2^3$

	C	D
C	60, 60	0, 70
D	70, 0	40, 40

(b) $A_1^2 + A_2^2 < A_1^1 + A_2^3$

	C	D
C	60, 60	0, 80
D	80, 0	20, 20

(c) $A_1^2 + A_2^2 = A_1^1 + A_2^3$

Table 3: The payoff matrices of three Prisoner's Dilemma games.

In any PD game, there exists a unique pure strategy Nash equilibrium (D,D) as defection strictly dominates cooperation. The outcome is far from optimal as there is an obvious Pareto improvement if we move to (C,C).

Equilibria of the Group Selection Game

The equilibrium curves the Prisoner's Dilemma games of Table 3 are shown in Figure 11.

Random Matching Under random matching, $\phi(x) = 2x(1-x)$. Out of the three conditions (32), (33) and (34), only (33) is satisfied for any PD game and this agrees with what we expected as the NEGS under random matching should coincide with the Nash Equilibrium *i.e.* all follow D ($x = 0$). It's easy to check using condition (37) that the NEGS at $x = 0$ is also an ESSGS.

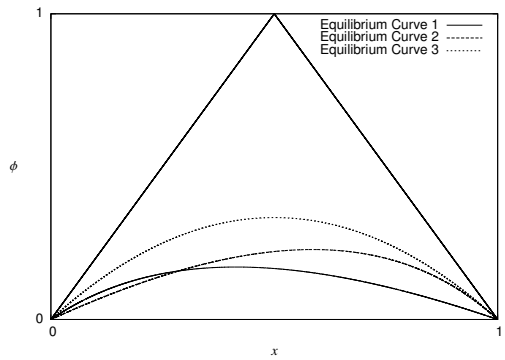


Figure 11: Equilibrium curves of the PD games in Table 3.

Complete Segregation Under complete segregation, $\phi(x)=0$ and so, only condition (34) is satisfied. Thus, the unique NEGS is $x=1$, *i.*e. pure cooperation. This state is also an ESSGS.

Constant Index of Assortativity In the case of a 'constant index of assortativity' rule, $\phi(x)=2(1-\alpha)x(1-x)$ (see example 3 in section 2.2). Depending on the value of α, we get all three cases. As intuition would suggest, the higher the assortativity, the higher the level of cooperation in equilibrium.

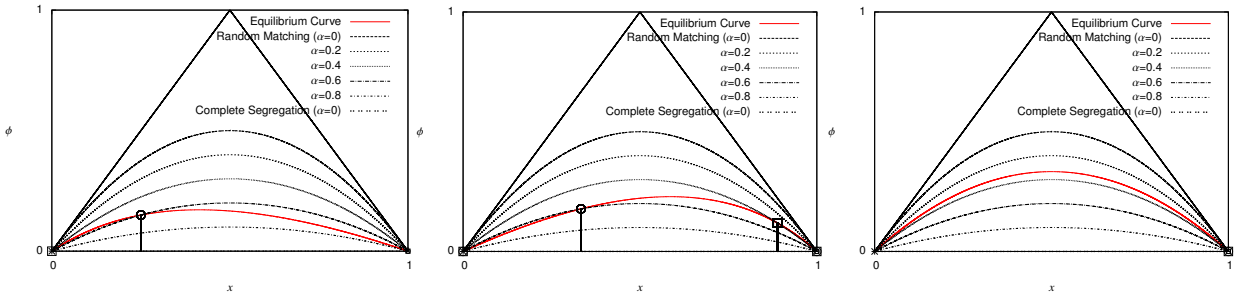
1. If $A_1^2+A_2^2>A_1^1+A_2^3$, then there is a *unique equilibrium* given by:

$$x^*(\alpha)=\begin{cases} 0 & \text{if } \alpha \le \frac{A_2^3-A_1^2}{A_1^1-A_1^2} \\ \frac{\frac{A_1^1-A_2^3}{1-\alpha}+A_1^2-A_1^1}{A_2^2-A_2^3+A_1^2-A_1^1} & \text{if } \frac{A_2^3-A_1^2}{A_1^1-A_1^2}<\alpha<\frac{A_2^2-A_1^1}{A_2^2-A_2^3} \\ 1 & \text{if } \alpha \ge \frac{A_2^2-A_1^1}{A_2^2-A_2^3} \end{cases} .$$

2. If $A_1^2+A_2^2<A_1^1+A_2^3$ then
   (a) if $\alpha<\frac{A_2^2-A_1^1}{A_2^2-A_2^3}$, there is a *unique equilibrium* at $x^*=0$ (all play D),
   (b) if $\alpha>\frac{A_2^3-A_1^2}{A_1^1-A_1^2}$, there is a *unique equilibrium* at $x^*=1$ (all play C),
   (c) if $\alpha=\frac{A_2^2-A_1^1}{A_2^2-A_2^3}$ or $\alpha=\frac{A_2^3-A_1^2}{A_1^1-A_1^2}$, there are *two equilibria*: one at $x_1^*=0$ and one at $x_2^*=1$ and
   (d) if $\frac{A_2^2-A_1^1}{A_2^2-A_2^3}<\alpha<\frac{A_2^3-A_1^2}{A_1^1-A_1^2}$, there are *three equilibria*: one at $x_1^*=0$, one at $x_2^*=\frac{\frac{A_2^3-A_1^1}{1-\alpha}+A_1^1-A_1^2}{A_1^1+A_2^3-A_1^2-A_2^2}$ and one at $x_3^*=1$.

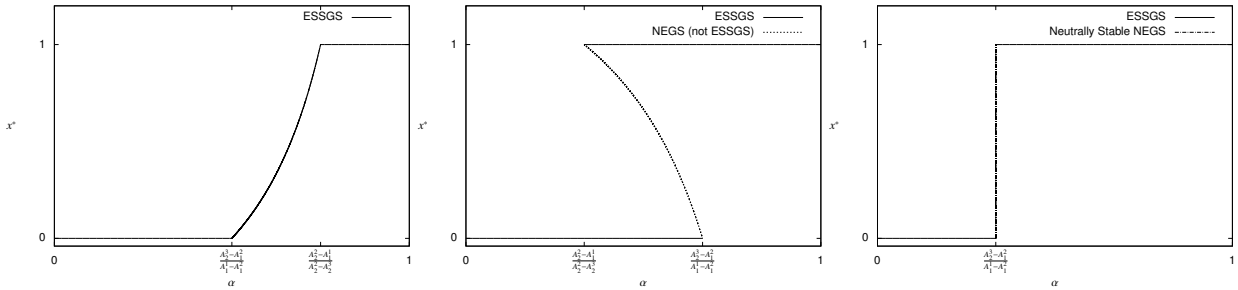
3. If $A_2^2+A_1^2=A_2^3+A_1^1$ then
   (a) if $\alpha<\frac{A_2^2-A_1^1}{A_2^2-A_2^3}$, there is a *unique equilibrium* at $x^*=0$ (all play D),
   (b) if $\alpha>\frac{A_2^2-A_1^1}{A_2^2-A_2^3}$, there is a *unique equilibrium* at $x^*=1$ (all play C) and
   (c) if $\alpha=\frac{A_2^2-A_1^1}{A_2^2-A_2^3}$, there is a *continuum of equilibria*. Actually, *any* $x\in[0,1]$ is an equilibrium.

The equilibrium-finding process for all three cases is shown in Figure 12 for constant index of assortativity rules with different values of α. The comparative statics results are summarized in Figure 13.



(a) $A_1^2+A_2^2>A_1^1+A_2^3$ (b) $A_1^2+A_2^2<A_1^1+A_2^3$ (c) $A_1^2+A_2^2=A_1^1+A_2^3$

Figure 12: NEGS with a constant index of assortativity in three different cases of PD games.



(a) $A_1^2+A_2^2>A_1^1+A_2^3$ (b) $A_1^2+A_2^2<A_1^1+A_2^3$ (c) $A_1^2+A_2^2=A_1^1+A_2^3$

Figure 13: NEGS as a function of the index of assortativity for three different cases of PD games.

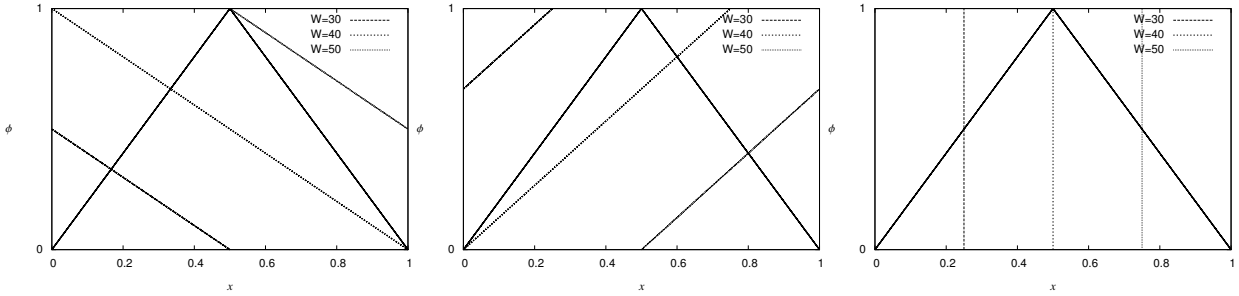
Risk Dominance In the case where $A_1^2+A_2^2<A_1^1+A_2^3$ (where two ESSGSs exist for certain values of α), the risk dominant equilibrium (in the sense introduced in 5.2) is the one where all play D ($x=0$ when $\alpha<\alpha^*=\frac{(A_2^2-A_1^2)-(A_1^1-A_2^3)}{(A_2^2-A_1^2)+(A_1^1-A_2^3)}$ and the one where all play C ($x=1$) when $\alpha>\alpha^*$, as was the case in the SH game.

Notice that unlike the HD and the SH games, in a PD game, *all states can be attained as equilibria* if an appropriate matching rule is selected.

Welfare

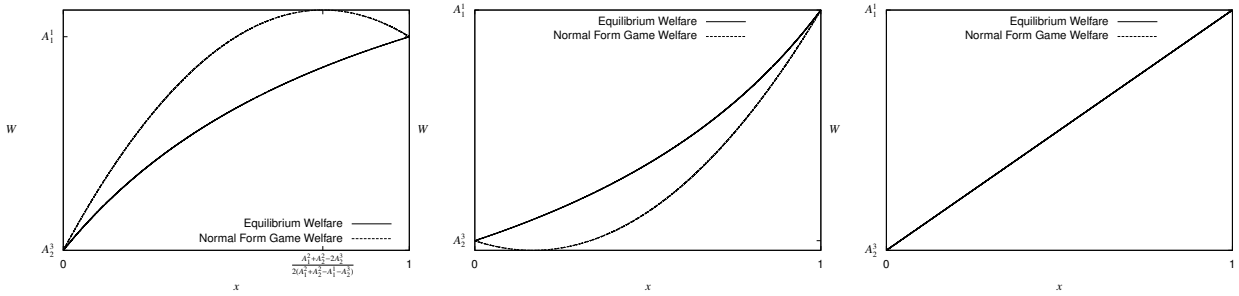
The isogrowth diagrams of three Prisoner's Dilemma games are shown in Figure 14. The comparison of equilibrium welfare in the group selection game and the normal form game for each of the three cases is shown in figure 5. The maximum level of welfare is obtained when the equilibrium state is the one where all cooperate ($x=1$) and it coincides with the maximum expected payoff players using

symmetric strategies can get in the normal form game when $A_1^2 + A_2^2 > 2A_1^1 > A_1^1 + A_2^3$. In the case where $A_1^2 + A_2^2 > 2A_1^1$ the maximum value of welfare in the normal form game is obtained when both players play C with probability $p_C = \frac{A_1^2+A_2^2-2A_2^3}{2(A_1^2+A_2^2-A_1^1-A_2^3)}$. However, when this state is implemented as an equilibrium in the group selection game, it does not grant the players such high expected payoffs as the frequency of (C,D) or (D,C) pairs is not high enough. The implementation of an assortative matching rule can make the state an equilibrium but this happens at the expense of obtained payoff at that state. Also, if we restrict ourselves to equilibrium payoffs, then the payoff obtained at $x = \frac{A_1^2+A_2^2-2A_2^3}{2(A_1^2+A_2^2-A_1^1-A_2^3)}$ is no longer the optimal payoff. Once again, utilitarian optimality is achieved when $x = 1$ (all cooperate) is implemented as an equilibrium.



(a) $A_1^2+A_2^2 > A_1^1+A_2^3$ (b) $A_1^2+A_2^2 < A_1^1+A_2^3$ (c) $A_1^2+A_2^2 = A_1^1+A_2^3$

Figure 14: Isogrowth diagrams for three PD games.



(a) $A_1^2+A_2^2 > A_1^1+A_2^3$ (b) $A_1^2+A_2^2 < A_1^1+A_2^3$ (c) $A_1^2+A_2^2 = A_1^1+A_2^3$

Figure 15: Equilibrium welfare and normal form payoff in three PD games.

6 Group Selection and the Fitness of Populations

Group selection can explain behavioral traits such as altruism or cooperation which cannot arise in Nash equilibrium and so cannot be favored by natural selection if matching is random (see theorem 5). Importantly, such departures from egoism may be superior to the outcomes under random matching in the sense that the *average fitness* may be higher. The classical example here is of course the prisoners'

dilemma where the outcome of random matching yields lower average fitness than outcomes with assortative matching (see section 5.3 and also Bergström, 2002). In this section we are going to discuss these issues drawing on both the abstract results and the concrete examples of the previous sections. As will become clear, our new concepts (NEGS and ESSGS) allow us to push the discussion substantially forward in comparison with existing literature.

First, we need to define the concepts involved. Recall from section 2 that the *average fitness* $\bar{\pi}(\mathbf{x})$ at a population state $\mathbf{x} \in S_m$ is given by $\bar{\pi}(\mathbf{x}) = \sum_{j=1}^{m} x_j \pi_j(\mathbf{x})$ (equation (8)). In the context of a group selection game $< I, G, \mathbf{f} >$, we referred instead to this as the *average payoff* or the *welfare* (see equation (12)). Since average fitness in the evolutionary model is obviously equal to average payoff in the (evolutionary) game theory model, and since by theorem 6 we know how the various equilibrium/steady state concepts relate to each other, we need not differentiate between them in what follows. Accordingly, we use the term *average fitness* exclusively from now on. Average fitness at a population state $\mathbf{x}$ will from now on be denoted by $\bar{\pi}_{\mathbf{f}}(\mathbf{x})$ so as to explicitly mention the matching rule. This allows us to easily compare average fitnesses under different matching rules for a fixed underlying payoff structure/normal form game G (*e.g.*, prisoners' dilemma or hawk-dove).

Now as was already mentioned, random matching – or for that matter any other specifically given matching rule $\mathbf{f}$ – may not maximize average fitness in a NEGS $\mathbf{x}^*$. The prisoners' dilemma was already mentioned above, but in the previous section we saw that the observation remains valid in other standard 2 by 2 games such as hawk-dove and stag-hunt; and it also remains valid if instead of NEGS we focus on ESSGS. Thus, evolution under non-random matching certainly does not imply fitness maximization. The interesting next question therefore is whether for a *fixed* underlying normal form game there exists *some* matching rule given which average fitness will be maximized in NEGS; and if the answer is yes, to characterize these matching rules in concrete situations. Thus in the prisoners' dilemma, random matching is inferior in average fitness terms, but as we saw in section 5.3, a rule such as complete segregation will lead to equilibria where everybody cooperates and so to average fitness maximization. When discussing this topic it is important to understand that when $\mathbf{f}$ is varied, not only does the set of NEGS (and ESSGS and also, the set of steady states of the replicator dynamics) change – the average fitness $\bar{\pi}_{\mathbf{f}}(\mathbf{x})$ will also change at any given population state $\mathbf{x}$. So if some population state maximizes welfare but is not a NEGS at some matching rule $\mathbf{f}^1$, it could be a NEGS at another matching rule $\mathbf{f}^2$ but no longer maximize welfare! Any sensible discussion must therefore consider the *joint* selection of a population state and matching rule as captured by the following definition.

Definition 9. (Evolutionary Optimum) *Let G be a normal form game. A population state $\mathbf{x}^* \in S_m$ together with a matching rule $\mathbf{f}^* \in \mathfrak{F}_{n,m}$ is said to be an evolutionary optimum if $\bar{\pi}_{\mathbf{f}^*}(\mathbf{x}^*) \geq \bar{\pi}_{\mathbf{f}}(\mathbf{x})$ for all $(\mathbf{x}, \mathbf{f}) \in \mathfrak{E} = \{(\mathbf{x}, \mathbf{f}) \in S_m \times \mathfrak{F}_{n,m} : \mathbf{x}$ is a steady state of $\langle I, G, \mathbf{f} \rangle\}$.*

Intuitively, a population state $\mathbf{x}^*$ and a matching rule $\mathbf{f}^*$ form an optimum if they lead to maximum average *fitness of the population* among all population state/matching rule combinations that satisfy the steady state restriction. Note that the restriction to steady states is entirely natural here: Any pop-

ulation state that is *not* a steady state under some matching rule would immediately be "destroyed" by natural selection.[28] Given these definitions, we can now answer the previous question:

Theorem 7. *Let* $(\mathbf{x}^*, \mathbf{f}^*)$ *be an evolutionary optimum. Then there exists a matching rule* $\mathbf{h} \in \mathfrak{F}_{n,m}$ *which satisfies the assumptions of theorem 2, such that* $\mathbf{x}^*$ *is a NEGS under* $\mathbf{h}$*, and such that* $(\mathbf{x}^*, \mathbf{h})$ *is an evolutionary optimum (in particular,* $\bar{\pi}_{\mathbf{h}}(\mathbf{x}^*) = \bar{\pi}_{\mathbf{f}^*}(\mathbf{x}^*)$*).*

Proof. See Appendix. □

Theorem 7 can be thought of as the "second welfare theorem of evolution" telling us that *any* evolutionary optimum can be "decentralized" in the evolutionary environment through *some* matching rule.[29] That this should be so is easy to see in simple cases, but it is in general a surprising result. In most standard games (including the ones considered in this paper), there is a premium on coordination/uniformity, and so what is needed in order to reach an evolutionary optimum is a sufficiently high level of assortativity. In games where there is a premium on agents in a group being *different* – *e.g.*, due to specialization – it will instead be a sufficiently high degree of dissociation that leads to evolutionary optimality. It is not obvious that Theorem 7 should hold in the latter case, to say nothing of cases that are neither assortative or dissociative. To illustrate with some concrete examples, consider first the Hawk-Dove model of section 5.1. As was shown in that section, the Hawk-Dove model has a unique evolutionary optimum, namely the state where all individuals are doves (since this state is uniform, it will be a steady state of the replicator dynamics under *any* matching rule). As was also shown in section 5.1, the doves only outcome is *not* a NEGS for all matching rules, however.[30] Specifically, (18) shows that only when matching is *sufficiently* assortative will a uniform population of doves be a NEGS.[31] Intuitively, what happens when this "assortativity threshold" is crossed is that hawks become so likely to end up with other hawks that it is not worthwhile playing hawk even if the population of hawks is infinitely small. What all of this shows is that as predicted by theorem 7, the evolutionary optimum (doves only) will be a NEGS for *some* matching rule. But unless the environment is such that matching is sufficiently assortative, evolution will not lead to the evolutionary optimum. In a real-world situation where a specific matching rule and a specific payoff structure is in effect, this of course implies that evolution can easily produce a mixed population of hawks and doves. But it is interesting that evolution *may* in fact lead to the evolutionary optimum even without recourse to the extremities of either complete segregation or direct reciprocity. Furthermore, equation (18) tells us exactly which pa-

[28] Note in this connection that *any* uniform population state is a steady state (in fact, any uniform population state is a steady state under *any* matching rule).

[29] If in Definition 9 matching rules are required to be consistent, one can show that the "decentralizing" matching rule of Theorem 7 can be chosen to be consistent also.

[30] Compare with *direct reciprocity* where the doves only outcome is always supported – along with any other payoff in the maximin set of the associated normal form game – as a subgame perfect Nash equilibrium in the infinitely repeated game (see *e.g.* Rubinstein, 1979).

[31] In fact, this population state will be an ESSGS for such levels of assortativity, and so cannot be invaded by hawks even in the highly demanding sense of an ESSGS (in particular, it will be asymptotically stable for the replicator dynamics), see section 5.1 for details.

rameters account for the relationship between the level of assortativity in matching and evolutionary optimality. For example, less assortativity is needed if $A_2^2 - A_1^1$ is "small" which simply means that a Dove facing a Dove will gain relatively less from switching to Hawk (A_1^1 is the payoff to a Dove facing a Dove, and A_2^2 is the payoff to a Hawk facing a Dove).

Another illustration is provided by the Stag Hunt model of section 5.2. As we saw in that section, the Stag Hunt model has multiple NEGS for "low" levels of assortativity: Two uniform population states (everyone hunts for hare, everyone hunts for stag), and a mixed population state. For "high" levels of assortativity, only the state where everyone hunts for stag is a NEGS. The evolutionary optimum is for everyone to hunt for stag (and again this evolutionary optimum is supported by *any* matching rule since it is uniform). Thus the evolutionary optimum is a NEGS for *all* levels of assortativity which means that the prediction of theorem 7 bears out in a particularly strong way. But only if assortativity is sufficiently high will the evolutionary optimum be the unique NEGS, and so – just as in the Hawk-Doves model – evolution may not lead to the evolutionary optimum in a real-world situation with a specific matching rule and specific payoffs.[32]

7 Conclusion

This paper had two main purposes. Firstly, to extend the existing machinery of evolutionary game theory to include models of group selection; and secondly, to use the new concepts developed to discuss the relationship between different kinds of selection and the fitness of populations. Two new equilibrium concepts were proposed, *Nash equilibrium with group selection* (NEGS) and *evolutionarily stable strategy with group selection* (ESSGS). These equilibrium concepts contain as special cases the standard ones; indeed when matching is random, the set of NEGS is just the symmetric Nash equilibria and the set of ESSGS is the evolutionarily stable strategies (theorem 5). We proceeded to show in our main theoretical result (theorem 6) that NEGS and ESSGS are for models with arbitrary matching rules what Nash equilibrium and ESS are for models with random matching. In particular, any stable steady state of the replicator dynamics is a NEGS and any ESSGS is an asymptotically stable steady state. As in the standard random matching setting, these results form the theoretical foundation upon which evolutionary game theory rests; hence our concepts extend the traditional game theoretic framework to models with group selection. As for the fitness of populations, our main result is the "second welfare theorem" of evolution (theorem 7) which states that *any* evolutionary optimum will be a NEGS under some matching rule.

[32] Concerning the situation with multiple NEGSs, the standard way to "resolve" multiplicity in the evolutionary setting would be to think of this as a situation with path-dependence so that, depending on initial conditions, a society may end up either as one where everyone hunts for hares or everyone hunts for stags (as mentioned, the mixed NEGS is not an ESSGS). From our game theoretic perspective, it is however more natural to employ a suitable selection criterion (see the discussion in section 5.2, where we saw that a global games approach will actually favor the hares only outcome for "low" levels of assortativity because this is the *risk-dominant* outcome (Carlsson and Van Damme, 1993).

We also showed (theorem 1) that models with structured populations, such as the haystack and trait-group models, can be captured by appropriately defined matching rules. This makes the dynamics and equilibrium analysis of such complicated models considerably easier as one can then simply apply the concepts of NEGS and ESSGS in a straightforward manner.

From an applied point of view, the great advantage of the game theoretic approach is the additional structure it imposes compared to dynamic models of the replicator type. In particular, the analysis becomes simpler and the results become more powerful. Recall that *all* uniform population states (all individuals employing the same strategy) are steady states for the replicator dynamics. In fact, the set of steady states includes *everything* that is "evolutionarily feasible" (and a good way to think of this set is in fact as evolutionary models' parallel to the feasible set of an exchange economy). This of course makes stability analysis absolutely critical in the dynamic setting – the problem being that such stability analysis is *not* straight-forward in group selection models where the replicator dynamics forms a complex non-linear dynamical system.[33] In contrast, we saw in section 5 that the set of NEGS and ESSGS can be computed with great ease in group selection games, and equally importantly, the game theoretic formulation allows for abstract analysis and the derivation of general results. An example of such a general result is theorem 5 which states that with random matching, the set of NEGS coincides with the symmetric Nash equilibria in the underlying normal form game which intuitively means that random matching precisely corresponds to "self-serving" behavior in general. Such a result would be impossible to establish within the traditional group selection framework of section 2. The "second welfare theorem" of evolution (theorem 7) is another example of this.

Often, matching is a geographical phenomenon (think of viruses, neighborhood imitation amongst humans, or trait-group models as studied in section 2.4), or a reflection of individuals' limited ability to monitor other individuals (see the introduction for further details). But when matching rules correspond to institutions or conventions, not explaining how they come about misses half the story. A clear weakness of existing group selection models – including the results in this paper – is in this connection that the matching rules are taken as given. An obvious topic for future research would be to model the evolution of the matching rules (*i.e.*, to endogenize them). Consider monitoring: If individuals gain an advantage by increasing their ability to monitor (by increasing their intelligence and memory), we can see how matching rules will over time evolve to be less and less random (typically more and more assortative). This then would be a true endogenous description of matching (institutions, conventions). The simplicity of the game theoretic framework presented in this paper should definitely put such a theory of matching rules within reach.

[33] Thus, consider for example the discrete time replicator dynamics of equation (9) in the often-studied case with two strategies. Unlike in models with random matching where the π's are linear, in models with non-random matching these coefficients will depend on the population state through the matching rules in an often very complicated way. This of course makes even local stability analysis a daunting task.

Appendix

Proofs

Proof of Theorem 1

Proof. We will restrict our attention to 2×2 games but similar extensions will hold for games with more strategies and/or players.

As the population is evolving in two different time modes (one related to dispersion phases and another one related to generations within trait-groups), we choose to use t to denote dispersion phases and τ to denote generations within the trait-groups. Our aim is to identify how the population evolves from one dispersion phase to the next. Intuitively, that would relate more to discrete-time dynamics but one can extend that to continuous time if the time scale of that evolutionary changes need to occur is large enough. In what follows, we calculate the fitness of each type of individual. The relevant dynamic equations (either discrete- or continuous-time) will determine the evolution of the population thereafter.

At dispersion phase t, the original population which comprises of a proportion of x_1 1-type individuals and x_2 2-type individuals is being randomly drawn to form trait-groups of initial size 2. The outcome is that there will be a proportion x_1^2 of type-1 trait-groups ($\{11\}$), a proportion $2x_1x_2$ of type-2 trait-groups ($\{12\}$) and a proportion x_2^2 of type-3 trait-groups ($\{22\}$). Each of these trait-groups will evolve independently and in isolation of the rest of the trait-groups for T generations.

Now at each generation τ a (isolated) trait-group will have a population that comprises of N_1 1-type and N_2 2-type individuals. These individuals are going to be drawn into pairs at each generation where they will act according to their types and get payoffs. Obviously, out of N_1+N_2 individuals $\frac{N_1+N_2}{2}$ groups (pairs) can be formed. Let's call κ the random variable that indicates how many of these groups are of type 2. Then the number of groups of type 1 will be given by $\frac{N_1-\kappa}{2}$ and the number of groups of type 2 will be given by $\frac{N_2-\kappa}{2}$. The probability that κ type 2 groups will be formed by a population of N_1 1-type and N_2 2-type individuals is given by

$$F(\kappa;N_1,N_2)=\begin{cases}\frac{N_1!N_2!(N_1-\kappa-1)!!(N_2-\kappa-1)!!}{(N_1-\kappa)!(N_2-\kappa)!\kappa!(N_1+N_2-1)!!} & \kappa\in\{0,1,\ldots,\min\{N_1,N_2\}\}\\ 0 & \text{otherwise}\end{cases}$$

where

$$(2l-1)!!=\begin{cases}\prod_{i=1}^{l}(2i-1) & \text{if } l\in\mathbb{N}_+\\ 1 & \text{if } l=0\\ 0 & \text{otherwise}\end{cases}$$

is an appropriate extension of the odd factorial.

Each of the pairs formed in the first stage (when drawn from the population) will be the first generation of a *trait-group* that will evolve separately from all other groups (pairs) for T generations.

The evolution process of each of the *trait-groups* follows a stationary Markov chain. The state of the trait-group at time t is a vector $\omega^t = (\omega_1^t, \omega_2^t, \omega_3^t)$ that represents the number of pairs of each type in the given *trait-group*. Let $A_{\max} = \max_{i\in\{1,2,3\}} \sum_j n_j^i A_j^i$ be the maximum number of children that can be obtained by a matched pair in a trait-group. Then, after T generations, a trait-group that began with 2 individuals cannot exceed a population of $K = \left(\frac{A_{\max}}{2}\right)^T$. So, the number of groups after T periods cannot exceed $\frac{K}{2}$. This means that the state space is finite and is $\Omega = \left\{(\omega_1, \omega_2, \omega_3) \in \mathbb{N}^3 : 0 < \sum_i \omega_i \leq \frac{K}{2}\right\}$. We will let μ denote the cardinality of Φ. We can also impose an ordering $\succ$ such that:

$$\omega \succ \omega' \Leftrightarrow \begin{cases} \sum_i \omega_i < \sum_j \omega'_j & \text{or} \\ \sum_i \omega_i = \sum_j \omega'_j \text{ and } \omega_1 > \omega'_1 & \text{or} \\ \sum_i \omega_i = \sum_j \omega'_j \text{ and } \omega_1 = \omega'_1 \text{ and } \omega_2 > \omega'_2 \end{cases}$$

This is a *total order* over Ω and thus it induces a ranking # of the elements of Ω. By $\#\omega$ we will denote the rank of state ω under $\succ$. Likewise $\#^{-1} n$ will denote the n-th state according to the ranking induced by $\succ$.

The only element of the Markov chain that we need to determine is the transition probabilities. So, the probability that state ω' will occur at period $t+1$ when we know that the trait-group was at state ω at period t is given by:

$$P(\omega'|\omega) = \begin{cases} F\left(\omega'_2; N_1(\omega), N_2(\omega)\right) & \text{if } (\omega'_1, \omega'_3) = \left(\frac{N_1(\omega)-\omega'_2}{2}, \frac{N_2(\omega)-\omega'_2}{2}\right) \\ 0 & \text{otherwise} \end{cases}$$

where $N_1(\omega) = 2\omega_1 A_1^1 + \omega_2 A_1^2$ and $N_2(\omega) = 2\omega_3 A_2^3 + \omega_2 A_2^2$ give the population of 1-type and 2-type individuals in the trait-group at state ω respectively. We will call $\mathscr{P}$ the matrix that is defined as follows: $\mathscr{P}_{ij} = P(\#^{-1} j | \#^{-1} i)$.

At each period τ, let $\mathbf{P}(\tau) \in S_\mu$ denote the vector whose i-th entry gives the probability that the trait-group is in state $\#^{-1} i$. As the Markov process is stationary, $\mathbf{P}(\tau)$ will be given by $\mathbf{P}(\tau) = \mathscr{P}^t \mathbf{P}(0)$. Where $\mathbf{P}(0)$ is the initial state of the trait-group *i.e.* $\mathbf{P}(0) \in \{(1,0,\ldots,0),(0,1,0,\ldots,0),(0,0,1,0,\ldots,0)\}$ as there's exactly one pair of individuals in each of the trait-groups in the beginning. In the interest of brevity, we will call these vectors $\mathbf{P}^1(0)$, $\mathbf{P}^2(0)$ and $\mathbf{P}^3(0)$ respectively. So, at the end of the T periods we will have $\mathbf{P}^i(T) = \mathscr{P}^T \mathbf{P}^i(0)$ for $i = 1,2,3$. So after T periods have gone by, the expected number of type-i groups that will be at a trait-group that contained one type-k group at time 0 will be:

$$g_i^k = \sum_{l=1}^{\mu} P_l^k(T)(\#^{-1} l)_i$$

Actually, as we have a continuum of trait-groups, by using a law of large numbers we can say that the distribution of group types in trait-groups will be (almost surely) exactly the one given by the above formula.

We will calculate the average fitness that each starting j-type individual will get (*i.e.* the number of descendants a j-type is expected to have) after T periods. A j-type individual that is drawn into

a k type trait-group is expected to have $\sum_{i\in \text{gsupp}(k)} g_i^k \frac{n_j^i}{n_j^k} A_j^i$ descendants. As the distribution of trait-groups is given by the random matching rule $\mathbf{r}(\mathbf{x}) = (r_1(\mathbf{x}), r_2(\mathbf{x}), r_3(\mathbf{x})) = (x_1^2, 2x_1x_2, x_2^2)$ and as a k-type trait-group contains n_j^k first-generation i-type individuals, we can calculate the average fitness of a first-generation j-type individual by:

$$\pi_j(\mathbf{x}) = \frac{\sum_{k\in \text{supp}(j)} r_k(\mathbf{x}) n_j^k \sum_{i\in \text{gsupp}(k)} g_i^k \frac{n_j^i}{n_j^k} A_j^i}{\sum_{k\in \text{supp}(j)} r_k(\mathbf{x}) n_j^k}$$

Where $\text{gsupp}(k) = \left\{ i \in \{1,2,3\} \,|\, g_i^k > 0 \right\}$. Explicitly, for type-1 individuals, the average fitness is

$$x_1 g_1^1 A_1^1 + x_2 \left(2 g_1^2 A_1^1 + g_2^2 A_1^2 \right) \tag{19}$$

whereas for type-2 individuals, average fitness is

$$x_2 g_3^3 A_2^3 + x_1 \left(2 g_3^2 A_2^3 + g_2^2 A_2^2 \right). \tag{20}$$

The system will follow the replicator dynamics (either the discrete-time version of equation (9) or the continuous-time version of equation (10)) with fitness functions given by (19) and (20). We will show that a group selection model with a matching rule given by

$$f_i(\mathbf{x}) = \frac{\sum_{k\in \text{gsupp}^{-1}(i)} r_k(\mathbf{x}) g_i^k}{\sum_{l=1}^3 \sum_{k\in \text{gsupp}^{-1}(l)} r_k(\mathbf{x}) g_l^k} \tag{21}$$

has exactly the same dynamic behavior as the trait-group model. Actually, we can rewrite the above matching rule as

$$f_i(\mathbf{x}) = \frac{\sum_{k=1}^3 r_k(\mathbf{x}) g_i^k}{\sum_{l=1}^3 \sum_{k=1}^3 r_k(\mathbf{x}) g_l^k} \tag{22}$$

as $g_l^k = 0$ for all $k \notin \text{gsupp}^{-1}(l)$.

In the group selection model $< I, \mathbf{f}, G >$, the payoffs for type-1 individuals is given by (see equation (12))

$$\frac{f_1(\mathbf{x})}{x_1} A_1^1 + \frac{f_2(\mathbf{x})}{2x_1} A_1^2$$

while the fitness for type-2 individuals is given by

$$\frac{f_2(\mathbf{x})}{x_2} A_2^3 + \frac{f_2(\mathbf{x})}{2x_2} A_2^2.$$

The key observation that makes it easy to show the result is that two models have identical dynamics if they have identical fractions $\frac{\pi_1(\mathbf{x})}{\pi_2(\mathbf{x})}$ for all $\mathbf{x} \in S_2$. So, in order for the trait-group model to have the same dynamics as the group selection model, it is sufficient for $\mathbf{f}(\mathbf{x})$ to satisfy:

$$\frac{x_1 g_1^1 A_1^1 + x_2 \left(2 g_1^2 A_1^1 + g_2^2 A_1^2 \right)}{x_2 g_3^3 A_2^3 + x_1 \left(2 g_3^2 A_2^3 + g_2^2 A_2^2 \right)} = \frac{\frac{f_1(\mathbf{x})}{x_1} A_1^1 + \frac{f_2(\mathbf{x})}{2x_1} A_1^2}{\frac{f_2(\mathbf{x})}{x_2} A_2^3 + \frac{f_2(\mathbf{x})}{2x_2} A_2^2}$$

It is easy to confirm that the above condition is satisfied for the matching rule given by (21). Also notice that **f** satisfies $f_1(\mathbf{x})+f_2(\mathbf{x})+f_3(\mathbf{x})=1$ and $f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})\geq 0$ as well as the conditions of theorem 2.

Notice that **f** as calculated above would not necessarily be consistent as it may fail to satisfy condition (3). As the trait-group model and the group selection model under **f** share the same dynamics, they will also have the same steady states. It is also interesting to point out that the matching rule **f** reduces to the random matching rule when $T=1$ (in this case $g_k^k=1$ and $g_i^k=0$ for $k\neq i$). □

Proof of Theorem 2

Best reply correspondence The best reply correspondence (BRC) is a correspondence $B: S_m \rightrightarrows S_m$ defined by:

$$B(\mathbf{x})=\left\{\mathbf{y}\in S_m : \left(\forall \tilde{\mathbf{y}}\in S_m\right)\Pi(\mathbf{y},\mathbf{x})\geq\Pi(\tilde{\mathbf{y}},\mathbf{x})\right\}$$

and gives the mixed strategies an agent can follow so as to maximize his/her expected payoff given that the state is **x**.

We also define the *value function* $V: S_m\to\mathbb{R}$ that gives the maximum payoff an agent can achieve at any given state. Formally: $V(\mathbf{x})=\max_{\mathbf{y}\in S_m}\Pi(\mathbf{y},\mathbf{x})$.

Equilibrium

We intend to show that under some assumptions on **f**, an equilibrium state always exists. We will prove the existence result by using Kakutani's fixed point theorem. In order to do that, we need to show that the BRC is convex-valued , nonempty-valued and upper hemicontinuous. These prerequisites are proven in Lemmata 8 and 9.

Lemma 8 (Convex-valued BRC)**.** *For any group selection game under a matching rule* $\mathcal{G}=<I,G,\mathbf{f}>$ *the best reply correspondence B is convex-valued.*

Proof. We can identify three different cases for $B(\mathbf{x})$:

- $B(\mathbf{x})=\emptyset$ and thus B is convex-valued at **x**.
- $B(\mathbf{x})=\{\mathbf{y}^*\}$ *i.e.* the best reply correspondence contains only one element at **x** and thus B is convex-valued at **x**.
- $B(\mathbf{x})$ contains at least two elements at **x** *i.e.* there exist $\mathbf{y}_1^*,\mathbf{y}_2^*\in S_m$ such that

$$\Pi(\mathbf{y}_1^*,\mathbf{x})\geq\Pi(\mathbf{y},\mathbf{x}) \text{ for all } \mathbf{y}\in S_m$$

$$\Pi(\mathbf{y}_2^*,\mathbf{x})\geq\Pi(\mathbf{y},\mathbf{x}) \text{ for all } \mathbf{y}\in S_m$$

which is possible only if $\Pi(\mathbf{y}_1^*, \mathbf{x}) = \Pi(\mathbf{y}_2^*, \mathbf{x}) = L$. Now, for all $\lambda \in [0,1]$ we have the following series of equalities:

$$\begin{aligned} \Pi(\lambda \mathbf{y}_1^* + (1-\lambda)\mathbf{y}_2^*, \mathbf{x}) &= (\lambda \mathbf{y}_1^* + (1-\lambda)\mathbf{y}_2^*) \cdot \pi(\mathbf{x}) = \\ = \lambda \mathbf{y}_1^* \cdot \pi(\mathbf{x}) + (1-\lambda)\mathbf{y}_2^* \cdot \pi(\mathbf{x}) &= \lambda \Pi(\mathbf{y}_1^*, \mathbf{x}) + (1-\lambda)\Pi(\mathbf{y}_2^*, \mathbf{x}) = \\ = \Pi(\mathbf{y}_1^*, \mathbf{x}) &= \Pi(\mathbf{y}_2^*, \mathbf{x}) = L \end{aligned}$$

So, for any $\mathbf{y}_1^*, \mathbf{y}_2^* \in B(\mathbf{x})$ we get that $\lambda \mathbf{y}_1^* + (1-\lambda)\mathbf{y}_2^* \in B(\mathbf{x})$ for all $\lambda \in [0,1]$ and thus B is convex-valued at $\mathbf{x}$.

Since these are the only possible cases, we can conclude that B is convex-valued in S_m. □

Lemma 9 (BRC: Nonempty-valued and upper hemicontinuous)**.** *For a group selection game under a matching rule $\mathscr{G} =< I, G, \mathbf{f} >$, if*

1. $\mathbf{f}$ *is continuous on S_m and*
2. *the partial derivatives $\partial_j f_i$ for all $j \in M$ and all $i \in \text{supp}(j)$ exist on $\text{bd}_j(S_m)$*

then the best reply correspondence B is non-empty valued and upper hemicontinuous.

Proof. From assumption 2 of the lemma, we get that the limits $\lim_{\tilde{\mathbf{x}} \to \mathbf{x}} \frac{f_i(\tilde{\mathbf{x}})}{\tilde{x}_j} = \partial_j f_i$ for all $j \in M$ and all $i \in \text{supp}(j)$ exist on $\text{bd}_j(S_m)$ and from the definition of π_j (7), we get that

$$\lim_{\tilde{\mathbf{x}} \to \mathbf{x}} \pi_j(\tilde{\mathbf{x}}) = \pi_j(\mathbf{x}) \quad \text{on } \text{bd}_j(S_m).$$

So, π_j are continuous on $\text{bd}_j(S_m)$ and since all f_i are continuous on S_m, π_j are continuous on $S_m \setminus \text{bd}_j(S_m)$ as sums of quotients of continuous functions. So, π is continuous on S_m and therefore, Π is continuous on S_m^2.

Now we can see that the conditions for Berge's maximum theorem are satisfied: (i) S_m is compact and (ii) Π is continuous. So, using Berge's theorem, we get that the value function V is continuous on S_m and that the best reply correspondence B is nonempty-valued, compact-valued, upper hemicontinuous and has a closed graph on S_m.

The results needed are the nonempty-valuedness and upper hemicontinuity of B. □

Now we have all that is needed in prove the theorem. From the results of Lemmata 8 and 9, we know that $B : S_m \to S_m$ is a nonempty-valued, convex-valued, upper hemicontinuous correspondence defined on the nonempty, compact and convex set S_m. So, the conditions for the application of Kakutani's fixed point theorem are satisfied. From Kakutani's fixed point theorem, we get that there exists a $\mathbf{x}^* \in S_m$ such that $\mathbf{x}^* \in B(\mathbf{x}^*)$ which means that there exists a $\mathbf{x}^* \in S_m$ such that

$$\Pi(\mathbf{x}^*, \mathbf{x}^*) \geq \Pi(\mathbf{x}, \mathbf{x}^*) \quad \text{for all } \mathbf{x} \in S_m.$$

That is, $\mathscr{G}$ has an equilibrium. □

Proof of Theorem 5

Let us denote by $\mathbf{y}^i \in S_m$ the (mixed) strategy used by player i in the normal-form game G and by $\mathbf{x}^{-i} \in S_m$ the strategy used in the normal-form game G by *all* player i's opponents. Let also $P_i(\mathbf{y}^i|\mathbf{x}^{-i})$ be the expected payoff of player i in the normal-form game when he/she is using strategy $\mathbf{y}^i$ and *all* of his opponents use strategy $\mathbf{x}^{-i}$. Since G is symmetric, we have $P_i(\mathbf{y}^i|\mathbf{x}^{-i}) = P_j(\mathbf{y}^j|\mathbf{x}^{-j})$ for all $i, j \in N$. So we can write $P(\mathbf{y}|\mathbf{x})$ to express the expected payoff in the normal-form game of any player using strategy $\mathbf{y}$ when all his opponents use the same strategy $\mathbf{x}$.

A symmetric Nash equilibrium of game G is a strategy $\mathbf{x}^* \in S_m$ such that:

$$P(\mathbf{x}^*|\mathbf{x}^*) \geq P(\mathbf{y}|\mathbf{x}^*) \quad \text{for all } \mathbf{y} \in S_m.$$

So, for $\mathbf{x}^*$ to be a symmetric Nash equilibrium, if every opponent of any given player i is using strategy $\mathbf{x}^*$, it must be a best response for player i to use the same strategy $\mathbf{x}^*$ as well.

On the other hand, a strategy $\mathbf{x}^*$ will be an equilibrium in game $< I, G, \mathbf{f} >$ iff:

$$\Pi(\mathbf{x}^*, \mathbf{x}^*) \geq \Pi(\mathbf{y}, \mathbf{x}^*) \quad \text{for all } \mathbf{y} \in S_m$$

Where $\Pi(\mathbf{y}, \mathbf{x}^*)$ expresses the expected payoff of an individual using strategy $\mathbf{y}$ while the rest of the population is using strategy $\mathbf{x}^*$. In order to prove the proposition, all we need to show is that

$$\Pi(\mathbf{y}, \mathbf{x}) = P(\mathbf{y}|\mathbf{x}) \quad \text{for all } \mathbf{y} \in S_m \tag{23}$$

under the random matching rule. If we let $\mathbf{e}_j$ be the probability vector that corresponds to pure strategy j, then (23) boils down to

$$\pi_j(\mathbf{x}) = P(\mathbf{e}_j|\mathbf{x}) \quad \text{for all } j \in M. \tag{24}$$

Calculating $\pi_j(\mathbf{x})$. Let us denote by M^i_{-j} the set of all strategies other than j represented in group i and by Γ^1_j the set of all groups that contain exactly one individual following strategy j. Formally $M^i_{-j} = \{k \in M \setminus \{j\} | i \in \text{supp}(j)\}$ and $\Gamma^1_j = \{i \in \Gamma_{n,m} | n^i_j = 1\}$. Calculating $\pi_j(\mathbf{x})$ under $\mathbf{r}^{n,m}$ yields:

$$\pi_j(\mathbf{x}) = \sum_{i \in \text{supp}(j)} \frac{(n-1)! x_j^{n^i_j - 1}}{(n^i_j - 1)!} \prod_{k \in M^i_{-j}} \frac{x_k^{n^i_k}}{n^i_k!} A^i_j \qquad , \mathbf{x} \in S_m \setminus \text{bd}_j(S_m) \tag{25}$$

$$\pi_j(\mathbf{x}) = \sum_{i \in \Gamma^1_j} (n-1)! \prod_{k \in M^i_{-j}} \frac{x_k^{n^i_k}}{n^i_k!} A^i_j \qquad , \mathbf{x} \in \text{bd}_j(S_m) \tag{26}$$

Calculating $P(\mathbf{e}_j|\mathbf{x})$. In general, all players use mixed strategies *i.e.* a randomization over the set of pure strategies M. We will denote the pure strategy a player l ends up using after the randomization process has taken place – *i.e.* the realization of player l's mixed strategy – as $\mathbf{s}^l$. The probability of

a player ending up in a situation where his/her opponents follow (pure) strategies $\mathbf{s}^{-l} \in M^{n-1}$ with $\mathbf{s}^{-l} = (s^1, \ldots, s^{l-1}, s^{l+1}, \ldots, s^n)$ will be denoted by $p(\mathbf{s}^{-l})$. When all player l's opponents use the same strategy $\mathbf{x}$, those probabilities can be calculated to be:

$$p(\mathbf{s}^{-l}) = \prod_{k \in M} (x_k)^{\nu_k(\mathbf{s}^{-l})}$$

where $\nu_k(\mathbf{s}^{-l}) \in \{0, 1, \ldots, n-1\}$ is the number of player l's opponents using strategy k in the ordered set $\mathbf{s}_{-l}$.

Let us fix player l's strategy (realization) to be $\mathbf{s}^l = \mathbf{e}_j$. Since the game G is symmetric, the payoff of player l will not depend on the exact ordering in $\mathbf{s}^{-l}$ but on the vector $\nu(\mathbf{s}^{-l}) = \left(\nu_1(\mathbf{s}^{-l}), \ldots, \nu_m(\mathbf{s}^{-l})\right)$. This means that different $\mathbf{s}^{-l}$s with the same $\nu(\mathbf{s}^{-l})$ will yield the same payoff for player l. The number of the different ν outcomes are elements is $\gamma_{n-1,m}$. Let us use $\kappa \in \Gamma_{n-1,m}$ to index the different ν. By abusing notation, we can calculate the probability of a specific ν^κ to occur as

$$p(\nu^\kappa) = \frac{(n-1)!}{\prod_{k \in M} \nu_k^\kappa!} \prod_{k \in M} (x_k)^{\nu_k^\kappa}. \tag{27}$$

As player l is using strategy j, if he ends up in a situation where his/her opponents' realizations are κ, it is as if he ends up in a group i where $n_k^i = \nu_k^\kappa$ for $k \neq j$ and $n_j^i = \nu_j^\kappa + 1$. This group will be in $\text{supp}(j)$ and we will write $i = j \rhd \kappa$ and read: "i is the group that we get if we add an individual who uses strategy j to a set of opponents whose realizations are κ. Notice that the probabilities in (27) are *independent* of player l's choice of strategy. So, the probability of player l ending up in situation i conditional on him using strategy j will be the same as the probability realization κ occurring. Using the i- rather than the κ- indexing, we can rewrite (27) (abusing the notation once again) as:

$$p(i|j) = p(j \rhd \kappa | j) = p(\nu^\kappa) = \frac{(n-1)! x_j^{n_j^i - 1}}{(n_j^i - 1)!} \prod_{k \in M_{-j}^i} \frac{(x_k)^{n_k^i}}{n_k^i!}.$$

Now, in each of these cases i, player l gets a payoff of A_j^i and his expected payoff is:

$$P(\mathbf{e}_j|\mathbf{x}) = \sum_{i \in \text{supp}(j)} p(i|j) A_j^i = \sum_{i \in \text{supp}(j)} \frac{(n-1)! x_j^{n_j^i - 1}}{(n_j^i - 1)!} \prod_{k \in M_{-j}^i} \frac{(x_k)^{n_k^i}}{n_k^i!} A_j^i. \tag{28}$$

In the special case where $\mathbf{x} \in \text{bd}_j S_m$, player l can be sure that he is the only one using strategy j and thus, the only groups that get positive probability are the ones in Γ_j^1 which have $n_j^i = 1$. So his/her expected payoff is:

$$P(\mathbf{e}_j|\mathbf{x}) = \sum_{i \in \Gamma_j^1} p(i|j) A_j^i = \sum_{i \in \Gamma_j^1} (n-1)! \prod_{k \in M_{-j}^i} \frac{(x_k)^{n_k^i}}{n_k^i!} A_j^i. \tag{29}$$

By comparing equation (25) to (28) and equation (26) to (29), we can see that

$$\pi_j(\mathbf{x}) = P(\mathbf{e}_j|\mathbf{x})$$

and as we showed that for an arbitrary j, it holds for all $j \in M$. □

Proof of Theorem 7

Let us define the following sets of group types:

$$\mathscr{E}(\mathbf{x}^*) = \{i \in \Gamma_{n,m} : \text{supp}^{-1}(i) \subseteq I(\mathbf{x}^*)\}$$

$$[M] = \{i \in \Gamma_{n,m} : \text{supp}^{-1}(i) = \{j\} \text{ for some } j \in M\}$$

$\mathscr{E}(\mathbf{x}^*)$ consists of the group types that contain only individuals of types that are present in the population at $\mathbf{x}^*$. $\mathscr{E}'(\mathbf{x}^*)$ will denote its complement *i*.e. group types that contain at least one individual of one of the types that are not present at $\mathbf{x}^*$. $[M]$ consists of the groups types that contain only one type of individuals. We will denote the group type that contains only individuals of type j by $[j]$. Now we can separate all group types in the following four categories:

- $SP(\mathbf{x}^*) = \mathscr{E}(\mathbf{x}^*) \cap [M]$ is the set of all group types that contain a single type of individuals that are present at $\mathbf{x}^*$.
- $SA(\mathbf{x}^*) = \mathscr{E}'(\mathbf{x}^*) \cap [M]$ is the set of all group types that contain a single type of individuals that are absent at $\mathbf{x}^*$.
- $MP(\mathbf{x}^*) = \mathscr{E}(\mathbf{x}^*) \setminus [M]$ is the set of all group types that contain more than one types of individuals that are present at $\mathbf{x}^*$.
- $MA(\mathbf{x}^*) = \mathscr{E}'(\mathbf{x}^*) \setminus [M]$ is the set of all group types that contain more than one types of individuals and at least one of them is absent at $\mathbf{x}^*$.

Let us define for any $\mathbf{x} \in S_m$ the following:

$$\mu = \arg\min_{j \in I(\mathbf{x}^*)} \frac{x_j}{x_j^*}$$

We construct $\mathbf{h}$ as follows:

- For all $i \in MA(\mathbf{x}^*)$ we define $h_i(\mathbf{x}) = 0$.
- For all $i \in SA(\mathbf{x}^*)$ we define $h_{[j]}(\mathbf{x}) = x_j$.
- For all $i \in MP(\mathbf{x}^*)$ we define $h_i(\mathbf{x}) = \frac{x_\mu}{x_\mu^*} f_i^*(\mathbf{x}^*)$.
- For all $i \in SP(\mathbf{x}^*)$ we define $h_{[j]}(\mathbf{x}) = \frac{x_\mu}{x_\mu^*} f_{[j]}^*(\mathbf{x}^*) + x_j - \frac{x_\mu}{x_\mu^*} x_j^*$.

It is easy to check that $\mathbf{h}$ is a matching rule as it satisfies definition 1 *i*.e. it is a function from S_m to $S_{\gamma_{n,m}}$. More than that it is also easy to see that $\mathbf{h}(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^*)$ and so $(\mathbf{x}^*, \mathbf{h})$ is an evolutionary optimum. All we have to do is to show that $\mathbf{x}^*$ is a NEGS under $\mathbf{h}$.

Now let us define $A^* = \max_{(\mathbf{x},\mathbf{f}) \in \mathfrak{C}} \bar{\pi}_{\mathbf{f}}(\mathbf{x})$. As $(\mathbf{x}^*, \mathbf{h})$ is an evolutionary optimum, it has to be that $\mathbf{x}^*$ is a steady state of the replicator dynamics under $\mathbf{h}$. So:

1. For all $j \in I(\mathbf{x}^*)$ it has to be that $\pi_{\mathbf{h}}(\mathbf{x}^*) = A^*$ which is ensured by the fact that $\mathbf{h}(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^*)$ and

2. there is no restriction for all $j \in O(\mathbf{x}^*)$.

For $\mathbf{x}^*$ to be a NEGS it must hold that:

$$\mathbf{x}^* \cdot \pi_{\mathbf{h}}(\mathbf{x}^*) \geq \mathbf{y} \cdot \pi_{\mathbf{h}}(\mathbf{x}^*) \qquad \text{for all } \mathbf{y} \in S_m.$$

Notice that from point 1. above, if $\mathbf{x}^* \in \operatorname{int} S_m$, it is a NEGS as $\mathbf{y} \cdot \pi_{\mathbf{h}}(\mathbf{x}^*) = A^*$ for all $\mathbf{y} \in S_m$ and the proposition holds.

If $\mathbf{x}^* \in \operatorname{bd} S_m$, then all we need to do is show that $\pi_{\mathbf{h}j}(\mathbf{x}^*) \leq A^*$ for all $j \in O(\mathbf{x}^*)$. By definition,

$$\pi_{\mathbf{h}j}(\mathbf{x}^*) = \sum_{i \in \operatorname{supp}(j)} \frac{n_j^i}{n} \partial_j^+ h_i(\mathbf{x}^*) A_j^i = \partial_j^+ h_{[j]}(\mathbf{x}^*) A_j^{[j]} + \sum_{i \in MA \cap \operatorname{supp}(j)} \frac{n_j^i}{n} \partial_j^+ h_i(\mathbf{x}^*) A_j^i = A_j^{[j]}$$

Finally, notice that under any matching rule the states $e_j = (0, \ldots, 0, \underbrace{1,}_{j-\text{th}} 0, \ldots, 0)$ are steady states and the payoff of all individuals on these states is simply: $\bar{\pi}_{\mathbf{h}}(\mathbf{e}_j) = A_j^{[j]}$. But as $(\mathbf{x}^*, \mathbf{h})$ is an evolutionary optimum, we know that $A_j^{[j]} \leq A^*$ for all $j \in M$. So, $\pi_{\mathbf{h}j}(\mathbf{x}^*) \leq A^*$ for all $j \in M$. □

Finding all equilibria in 2×2 games

In this section we provide a tool that makes it easy for one to find and visualize NEGSs and ESSGSs in the 2×2 case. By use of our method, we can easily identify equilibria of such games by looking for intersections between two lines: one that depends on the payoffs (the *equilibrium curve*) and one that depends on the matching rule in effect (the *matching rule curve*). An example is shown in Figure 16; the equilibrium state is at the intersection of the two lines.

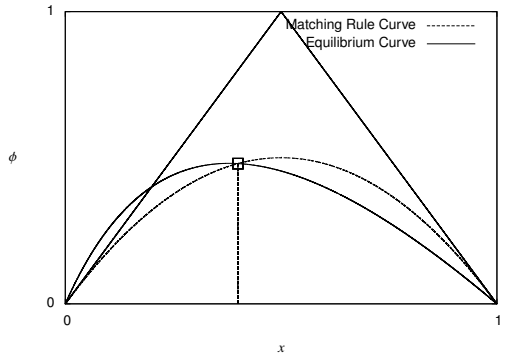


Figure 16: Example of finding an equilibrium.

In what follows, we analyze games that have a payoff bimatrix of the general form presented in Table 4. Without loss of generality, we will assume that $A_1^1 \geq A_2^3$.

	Strategy 1	Strategy 2
Strategy 1	A_1^1, A_1^1	A_1^2, A_2^2
Strategy 2	A_2^2, A_1^2	A_2^3, A_2^3

Table 4: The general form of a 2×2 game. $A_1^1 \geq A_2^3$.

The Matching Rule Curve

A matching rule for the 2×2 case, will be of the form $\mathbf{f}(\mathbf{x}) = \big(f_1(x_1,x_2), f_2(x_1,x_2), f_3(x_1,x_2)\big)$. Now notice that under consistency, it can be easily described by only defining one of the three coordinates $f_i(\mathbf{x})$. This is because in order for $\mathbf{f}$ to satisfy the equations in (3) (two linearly independent equations in our example of 2 strategies), only one degree of freedom remains.[34] We pick the value of $f_2(\mathbf{x})$ – that expresses the extent to which the two strategies get mixed with one another – to describe the matching rule. Of course, because there are only two strategies available, the state can be summarized by the proportion of individuals using Strategy 1 (the remaining individuals are clearly using Strategy 2). We will use x to denote this proportion and thus to express the state.[35] So any matching rule will be described by a function $\phi : [0,1] \to [0,1]$. Under the consistency requirement in (3), the three coordinates of $\mathbf{f}$ can be calculated to be:

$$f_1(x) = x - \frac{1}{2}\phi(x) \quad f_2(x) = \phi(x) \quad f_3(x) = 1 - x - \frac{1}{2}\phi(x). \tag{30}$$

More than that, the conditions $0 \leq f_1(x)$, $0 \leq f_2(x)$ and $0 \leq f_3(x)$ must be satisfied for all $x \in (0,1)$. From these, we get that the values ϕ can take are restricted by:

$$0 \leq \phi(x) \leq 2x \quad \text{for } x \in \left[0, \frac{1}{2}\right], \quad 0 \leq \phi(x) \leq 2(1-x) \quad \text{for } x \in \left(\frac{1}{2}, 1\right]. \tag{31}$$

So any consistent matching rule in the case of 2-strategy, 2-person normal form games can be summarized by a function ϕ that satisfies (31).

It is now possible for us to draw diagrams that show what matching rules look like. Examples of graphs of matching rules are given in Figure 17. A matching rule is summarized by a line that begins at (0,0), assumes values 'within' the triangle bounded by (31) and ends at (1,0).

Under this formalization, the random matching rule will be given by

$$\phi(x) = 2x - 2x^2$$

whereas the complete segregation rule is simply

$$\phi(x) = 0$$

[34] Equations (3) are in essence 'balancing conditions' similar to condition (2) in Alger and Weibull (2012). *i.e.* They ensure that the number of 1-strategists that are matched to 2-strategists is equal to the number of 2-strategists that are matched to 1-strategists.

[35] Obviously, $x_1 = x$ and $x_2 = 1 - x$.

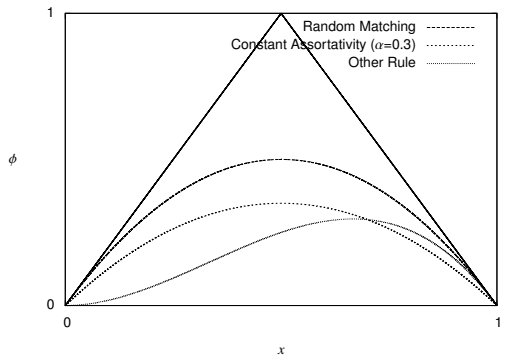


Figure 17: Examples of Matching Rule Curves.

. Another example would be the constant index of assortativity rule (Bergström, 2003) which can be summarized by

$$\phi(x) = 2(1-\alpha)x(1-x)$$

where $\alpha \in [0,1]$ is the index of assorativity.

The Equilibrium Curve

Under any matching rule,it is easy to show that an interior state x^* is an equilibrium iff:

$$\pi_1(x^*) = \pi_2(x^*) \Leftrightarrow$$

$$[(A_2^2 - A_2^3)x^* + (A_1^1 - A_1^2)(1-x^*)]\phi(x^*) = 2(A_1^1 - A_2^3)x^*(1-x^*) \tag{32}$$

and, looking for boundary equilibria, if ϕ is differentiable at 0 and at 1, for $x = 0$ to be an equilibrium, it must be the case that:

$$\pi_1(0) \leq \pi_2(0) \quad \Rightarrow \quad (A_1^1 - A_1^2)\frac{\partial \phi}{\partial x}(0) \geq 2(A_1^1 - A_2^3) \tag{33}$$

and for $x = 1$ to be an equilibrium, it must be the case that:

$$\pi_1(1) \geq \pi_2(1) \quad \Rightarrow \quad (A_2^3 - A_2^2)\frac{\partial \phi}{\partial x}(1) \leq 2(A_1^1 - A_2^3). \tag{34}$$

Now, provided that there is actually some strategic interaction occurring between the two players, *i.e.* either $A_2^2 \neq A_2^3$ or $A_1^1 \neq A_1^2$ (or both), then from condition (32) we get two cases:

- If $A_1^1 \neq A_2^3$, then an interior state will be an equilibrium iff the value of ϕ for that state is equal to the value of a function E for that given state. We will call this function the *equilibrium curve* of the game and it is given by:

$$E(x) = \frac{2(A_1^1 - A_2^3)x(1-x)}{(A_2^2 - A_2^3)x + (A_1^1 - A_1^2)(1-x)}. \tag{35}$$

- In the case where $A_1^1 = A_2^3$, then the condition for an interior state to be an NEGS is:

$$\begin{cases} \phi(x) = 0 & \text{or} \\ x = \frac{A_1^2 - A_1^1}{A_2^2 - A_2^3 + A_1^2 - A_1^1} \end{cases} \tag{36}$$

  Condition (36) says that any state for which the two strategies do not mix at all will be an equilibrium state (obviously, as no strategy gets an advantage over the other) and, more importantly, that the state $\frac{A_1^2 - A_1^1}{A_2^2 - A_2^3 + A_1^2 - A_1^1}$ will be an equilibrium *for all matching rules* (as long as this value is withing the boundaries (0,1).

Stability and the Equilibrium Curve If we assume that the matching rule is $\mathscr{C}^1$, then we can easily check that a state x will be an ESSGS iff

$$\begin{cases} \phi(x) = E(x) & \text{and} \\ \frac{\partial \phi}{\partial x}(x) > \frac{\partial E}{\partial x}(x) \end{cases} \tag{37}$$

Using the above analysis in conjunction with diagrams like the one in Figure 17 can help us spot NEGS and ESSGSs very easily. All one has to do is to plot the matching rule ϕ and the equilibrium curve E on the same diagram. If the two lines meet at an interior state, then this state is a NEGS. If along with that the equilibrium curve is above the matching rule to the left of the state and below it to the right of the state, then the state is an ESSGS as well. Finally, for the states 0 and 1, one can say that in order for one of these states to be a NEGS (ESSGS), then it has to be that the slope of the matching rule is greater than (or equal to) the slope of the equilibrium curve at that state.

Welfare in 2×2 Games

In the case of 2×2 games, by using the formalization introduced above, we can make equilibrium welfare considerations. What we are interested in is to see how the different equilibria fare in terms of welfare. For a 2×2 game, the welfare at state x when the value of the matching rule at x is ϕ is given by:

$$W(x, \phi) = A_2^3 + (A_1^1 - A_2^3)x + \frac{\left(A_1^2 + A_2^2 - A_1^1 - A_2^3\right)\phi}{2} \tag{38}$$

And as long as $A_1^2 + A_2^2 \neq A_1^1 + A_2^3$, solving for ϕ, we get:

$$\phi = \frac{2(W - A_2^3)}{A_1^2 + A_2^2 - A_1^1 - A_2^3} - \frac{2(A_1^1 - A_2^3)x}{A_1^2 + A_2^2 - A_1^1 - A_2^3} \tag{39}$$

For any value of W, the above equation gives the set of points on the (x, ϕ) plane that yield an average payoff of W for the population. We will call such lines *isogrowth lines* as all points on each of these lines leads to the same growth rate of the population (which is the same as the average payoff). Drawing such lines can help us visualize what is really happening in terms of welfare under the various matching rules. More than that, by combining the isogrowth lines with the equilibrium curves of different games,

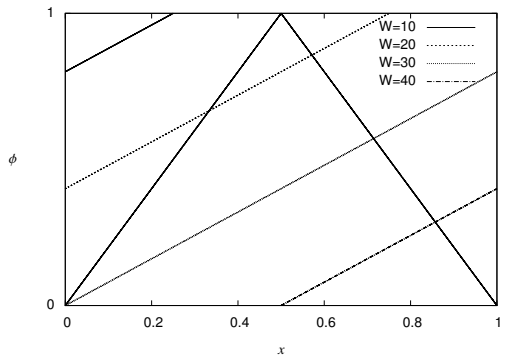


Figure 18: An example isogrowth diagram.

we can see which matching rules can lead to some (utilitarian) optimality. An example of an isogrowth diagram is depicted in Figure 18. Finally, using the welfare function (38) along with the equilibrium curve (35) we can calculate the equilibrium welfare in the group selection game and then compare that to the expected payoff of a player in the normal form game. Such comparisons are carried out in Section 5 for three classes of 2×2 games.

References

Aigner, M. (2007). *A course in enumeration*. Berlin: Springer – Verlag.

Alger, I. and J. W. Weibull (2010). "Kinship, incentives, and evolution". *The American Economic Review*, pp. 1725–1758.

Alger, I. and J. W. Weibull (2012). "A generalization of Hamilton's rule – Love others how much?" *Journal of Theoretical Biology* 299, pp. 42–54.

Alós-Ferrer, C. and A. B. Ania (2005). "The Evolutionary Stability of Perfectly Competitive Behavior". *Economic Theory* 26, pp. 497–516.

Bergström, T. C. (2002). "Evolution of Social Behavior: Individual and Group Selection". *Journal of Economic Perspectives* 2.16, pp. 67–88.

Bergström, T. C. (2003). "The algebra of assortative encounters and the evolution of cooperation". *International Game Theory Review* 5.3, pp. 211–228.

Carlsson, H. and E. Van Damme (1993). "Global games and equilibrium selection". *Econometrica* 61.5, pp. 989–1018.

Cooper, B. and C. Wallace (2004). "Group selection and the evolution of altruism". *Oxf. Econ. Pap.* 56.2, p. 307.

Eshel, I., L. Samuelson, and A. Shaked (1998). "Altruists, Egoists, and Hooligans in a Local Interaction Model". *American Economic Review* 88.1, pp. 157–179.

Fehr, E. and S. Gächter (2000). "Cooperation and Punishment in Public Goods Experiments". *American Economic Review* 90.4, pp. 980–994.

Hamilton, W. D. (1964). "The genetical evolution of social behaviour. II". *Journal of Theoretical Biology* 7.1, pp. 17–52.

Hamilton, W. D. (1970). "Selfish and spiteful behaviour in an evolutionary model". *Nature* 228, pp. 1218–1220.

Hammerstein, P. and R. Selten (1994). “Game theory and evolutionary biology”. In: *Handbook of Game Theory with Economic Applications*. Ed. by R. J. Aumann and S. Hart. Vol. 2. Amsterdam: North-Holland. Chap. 28, pp. 929–993.

Hofbauer, J. and K. Sigmund (1998). *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Kerr, B. and P. Godfrey-Smith (2002). “Individualist and multi-level perspectives on selection in structured populations”. *Biology and Philosophy* 17.4, pp. 477–517.

Lefebvre, M. (2007). *Applied stochastic processes*. New York: Springer.

Leininger, W. (2006). “Fending off one means fending off all: evolutionary stability in quasi-submodular aggregative games”. *Economic Theory* 29.3, pp. 713–719.

Maynard Smith, J. (1964). “Group selection and kin selection”. *Nature* 201.4924, pp. 1145–1147.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. (1998). “The Origin of Altruism”. *Nature* 393.5427, pp. 639–640.

Maynard Smith, J. and G. R. Price (1973). “The logic of animal conflict”. *Nature* 246.5427, pp. 15–18.

Nowak, M. A. and R. M. May (1992). “Evolutionary games and spatial chaos”. *Nature* 359.6398, pp. 826–829.

Okasha, S. (2005). “Maynard Smith on the levels of selection question”. *Biology and Philosophy* 20.5, pp. 989–1010.

Rubinstein, A. (1979). “Equilibrium in Supergames with the Overtaking Criterion”. *Journal of Economic Theory* 21.1, pp. 1–9.

Samuelson, L. (2002). “Evolution and Game Theory”. *Journal of Economic Perspectives* 16, pp. 47–66.

Schaffer, M. E. (1988). “Evolutionarily stable strategies for a finite population and a variable contest size”. *Journal of Theoretical Biology* 132, pp. 469–478.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Sober, E. and D. S. Wilson (1999). *Unto others: The evolution and psychology of unselfish behavior*. 2nd. Cambridge, Massachussetts: Harvard University Press.

Taylor, P. D. and L. B. Jonker (1978). “Evolutionary stable strategies and game dynamics”. *Mathematical Biosciences* 40.1, pp. 145–156.

Vega-Redondo, F. (1997). “The Evolution of Walrasian Behavior”. *Econometrica* 65.2, pp. 375–384.

Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge Massachussets: The MIT Press.

Wilson, D. S. (1975). “A theory of group selection”. *Proceedings of the National Academy of Science of the U.S.A.* 72.1, p. 143.

Wilson, D. S. (1977). “Structured demes and the evolution of group-advantageous traits”. *American Naturalist* 111.977, pp. 157–185.

Young, H. P. (1993). “The evolution of conventions”. *Econometrica* 61.1, pp. 57–84.