Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong

Matthijs van Veelen

Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong

Matthijs van Veelen

PII: S0022-5193(09)00189-1
DOI: doi:10.1016/j.jtbi.2009.04.019
Reference: YJTBI 5543

To appear in: *Journal of Theoretical Biology*

Received date: 12 January 2009
Revised date: 19 March 2009
Accepted date: 21 April 2009

Cite this article as: Matthijs van Veelen, Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.04.019

Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong

Matthijs van Veelen
CREED, Universiteit van Amsterdam
Roetersstraat 11, 1018 WB Amsterdam
the Netherlands
phone: +31 20 5255293
fax: +31 20 5255283
C.M.vanVeelen@uva.nl

March 19, 2009

Abstract Group selection theory has a history of controversy. After a period of being in disrepute, models of group selection have regained some ground, but not without a renewed debate over their importance as a theoretical tool. In this paper I offer a simple framework for models of the evolution of altruism and cooperation that allows us to see how and to what extent both a classification with and one without group selection terminology are insightful ways of looking at the same models. Apart from this dualistic view, this paper contains a result that states that inclusive fitness correctly predicts the direction of selection for one class of models, represented by linear public goods games. Equally important is that this result has a flip side: there is a more general, but still very realistic class of models, including models with synergies, for which it is *not* possible to summarize their predictions on the basis of an evaluation of inclusive fitness.

Keywords cooperation, altruism, group selection, inclusive fitness, linear and non-linear public goods games.

1 Introduction

It is safe to say that there is no consensus concerning the value of group selection models for the explanation of the evolution of altruism and cooperation. A history of disagreement has made the question evolve from whether group selection is probable or even possible (Allee, 1951, Wynne-Edwards, 1962, Williams, 1966) to whether group selection models help us understand things one would not understand without them (Sober & Wilson, 1995, Wilson & Wilson, 2007, Traulsen & Nowak, 2006, Lehmann, Keller, West & Roze, 2007, Killingback, Bieri & Flatt, 2006, Grafen, 2007, and West, Griffin & Gardner, 2007a,b, 2008). In order to show that different views need not be incompatible, I will begin with a simple but very general framework for models of the evolution of altruism and cooperation. This general framework allows us to see how and to what extent both an approach with and an approach without group selection terminology are insightful ways of looking at the same models. It also allows for a formal proof of a theorem that states that the sign of the inclusive fitness determines the direction of selection, if the model translates to a linear public goods game. The requirement of linearity turns out to a necessity; a simple example is given of a non-linear public goods game for which inclusive fitness points in the wrong direction. While a two-player situation still allows for (adjusted) formula's that do use relatedness, a slightly less simple example shows that with groups larger than two, relatedness can be the wrong population characteristic to look at. This implies that the prediction of the model cannot be given in a formula with costs, benefits and relatedness only.

There are at least three reasons why this formalism is useful. First of all it gives a formal framework for a dualistic view. This can help avoid unnecessary disagreements and helps bring out the value of both views. Second, although the first counterexample for Hamilton's rule failing is not new (see for similar counterexamples Wenseleers, 2006, and Gardner, West & Barton, 2007, which in turn relate to work by Grafen, 1979, and Day & Taylor, 1998), we should realise that the results in the literature concern 2 by 2 games. When we think of group selection, we tend to think of groups of any size, not just size 2. Also when we for instance think of the transition from single-celled to multicellular life, we tend to think of multicellular life as organisms typically consisting of more than 2 cells. An extension from groups of 2 to groups of n - or from 2 by 2 to n by n games - and a formal proof for when Hamilton's rule does and when it does not work, therefore are quite useful here. Because this goes against the intuition provided in Hamilton (1975) for why inclusive fitness should work, this paper also provides an

intuition for why it only does so for models that translate to linear public goods games, and not for models that translate to non-linear ones. The proof of the theorem also provides a general recipe for determining the direction of selection if Hamilton's rule fails due to non-linearity in the public goods game.

The third reason why this formalism is useful is at first perhaps a bit more difficult to see. In the literature, relatedness is regularly defined as a statistical property. In modelling, that would in principle be inappropriate; in a theoretical model, relatedness should be a probabilistic property, while statistics is only involved in testing of models or estimation of parameters using actual data. In the formal setup here, relatedness is a proper difference in conditional probabilities that is to follow from model assumptions. It fortunately does match with what we think relatedness should be in most models, and therefore one could see it as a formal justification for those cases. The formal setup on the other hand also helps understand why in some models with groups larger than 2 relatedness is the wrong population characteristic to look at. It thereby helps us formalise and sharpen our interpretation of relatedness.

2 Public goods games

Public goods games can be seen as the mother of all cooperation models.[1] Therefore it is useful to first properly define and picture how different situations in which selection takes place translate to different public goods games. In a selection process concerning a trait that has an effect on the carrier itself as well as on other members of the group it is in, we can write these effects as payoffs in a game. If the effects of different group members having the trait simply add up, then this results in a linear public goods game, in which the payoffs, or (expected) numbers of offspring, can be described as follows. In a group that consists of n individuals, i of which have the trait, payoffs for bearers (T) and for non-bearers (N) of the trait are, respectively

$$\begin{cases} \pi(T,i,f) = 1 + b(f)\cdot i - c(f) \\ \pi(N,i,f) = 1 + b(f)\cdot i \end{cases} \tag{1}$$

Here, $f \in [0,1]$ represents the frequency of the trait in the entire population. This description matches models in for instance Hamilton (1975), Nunney (1985) and Wilson

[1] In an e-mail discussion group on the topic of multilevel selection theory, Michael Doebeli described public goods games as the mother of all cooperation models. I thought that was a nice description, so I borrowed it here.

& Dugatkin (1997) and is only a little more general in that it allows for $b(f)$ and $c(f)$ to depend on the frequency of the trait in the entire population. One could also make them depend on other overall population characteristics without changing the analysis. The restriction that (1) imposes on the payoff function π can also be seen as a natural generalization of "equal gains from switching" as used in Traulsen, Shoresh & Nowak (2008), Wild & Traulsen (2007) and defined in Nowak & Sigmund (1990); see also Section 5 for a discussion.

Figure 1 graphically describes behaviours for this class of models. This figure is perhaps not that easy to read at first, but I firmly believe it is very much worth the effort, as it embraces a wide variety of models.

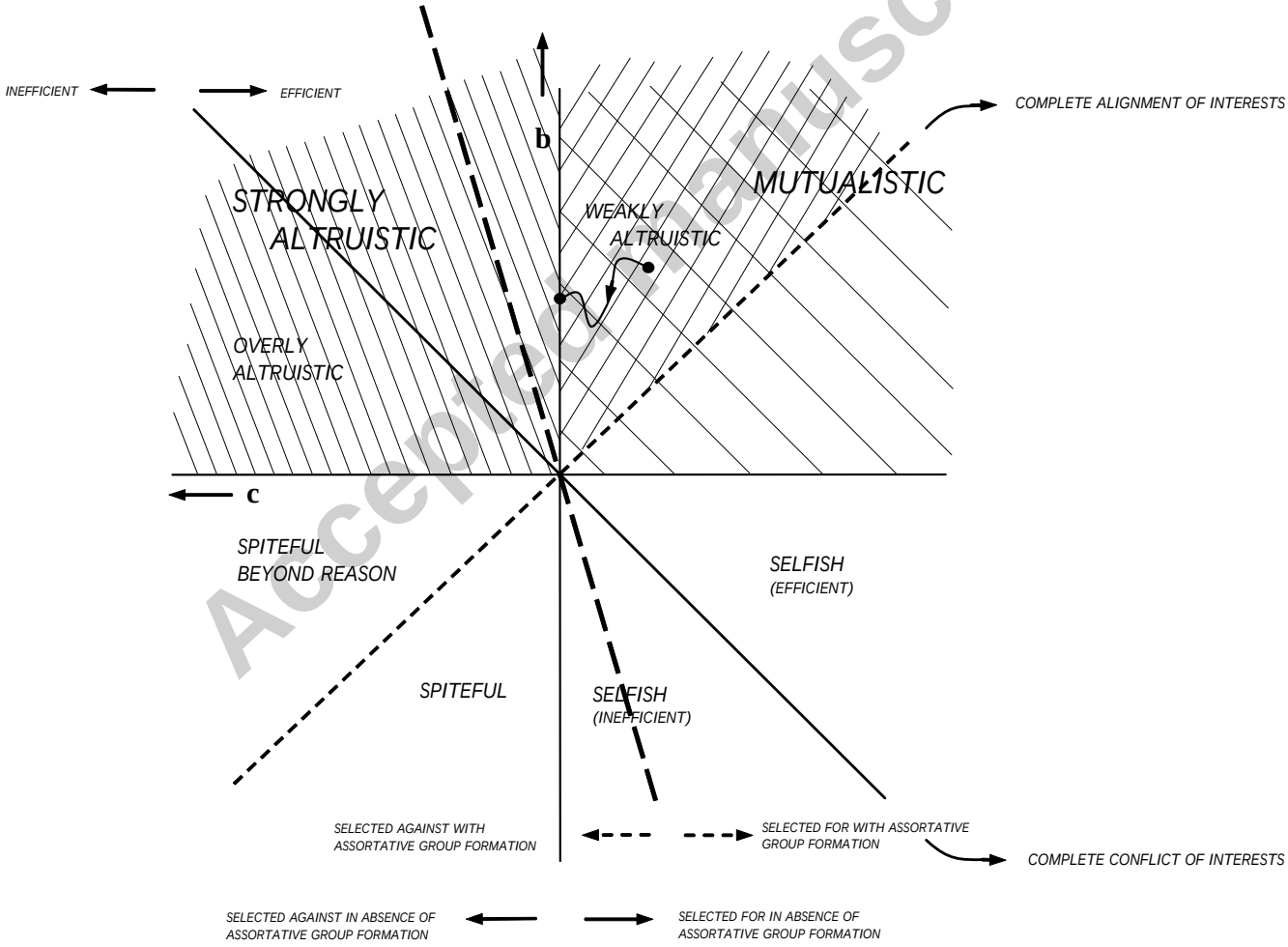

Figure 1: Fitness effects are represented by net costs **c** to the acting individual on the

horizontal axis and aggregate benefits to the other group members **b** on the vertical axis. Please note that net costs to the acting individual are positive to the left and negative to the right, so that the first quadrant consists of traits that have a positive fitness effect both on the acting individual itself and on the other group members.

Any choice for $b(f)$ and $c(f)$ can be associated with a point in Figure 1, which represents the fitness effects of the behaviour. The horizontal axis represents the net effect on the fitness of the individual itself, while the vertical axis represents the aggregate effect on all other group members. From equation (1) it therefore follows that $\mathbf{c} = c(f) - b(f)$ and $\mathbf{b} = (n-1) \cdot b(f)$, which makes **c** the net cost of the behaviour to the acting individual, and **b** the aggregated benefits to the others. Behaviour in individual interactions is subsumed in this setting, because groups of any size are allowed, including groups of size 2. The origin in Figure 1 is naturally associated with not displaying the behaviour – which can be seen as a status quo.

The setting does not restrict the behaviour to whole-group or others-only traits; all one has to do in order to translate a whole group trait to an others-only setting is shift the benefits that accrue to the actor as a benefiting member of the group from the aggregate group benefit to the actor itself, as we did above (see also Pepper, 2000). The figure also allows for frequency dependence; if fitness effects on the actor and on the rest of the group change with the frequency of the trait, then the point that depicts the difference between having the trait and not having it – or performing the behaviour and not performing it – simply shifts during selection as illustrated in Figure 1.

In this figure we can discern a few characteristic situations. The horizontal axis represents traits that only have an effect on the acting individual itself, and not on other group members. The vertical axis represents traits that only have an effect on other group members, and not on the acting individual itself. The diagonal that runs from the top-left to the bottom-right corner separates the traits that increase the aggregated fitnesses of all group members (right-up from the diagonal) from the traits that decrease the aggregated fitnesses of the group (left-down). A setting in which the reproductive success of all group members coincides, makes all possible behaviours map onto the diagonal that runs from bottom-left to top-right. In Figure 1, which pictures a situation with groups of size 2, and hence represents interactions between individuals, this diagonal makes a 45° angle (or $\pi/4$) with the horizontal axis. Groups of larger size result in larger angles; because the vertical axis represents the *aggregate* fitness effect

on the other group members, a group of size n would require a line through the origin that makes an angle of $\arctan(n-1)$ with the horizontal axis in order to represent a situation in which the interests of all group members are perfectly aligned. (The other diagonal is the same for all group sizes).

We can also identify different regions in this figure with different qualifications of behaviour. The entire top-right quadrant can be qualified as mutualistic behaviour, because fitness effects on both actor and recipients are positive. Such behaviour is also regularly referred to as a by-product mutualism. Mutualistic behaviour from which every recipient gains more than the actor does, is called weakly altruistic in Wilson (1979, 1990); the fitness of the actor increases in absolute terms, but decreases relative to the other individuals in the group. The top-left quadrant represents strongly altruistic behaviour (see again Wilson, 1979, 1990), where behaviour for which the others gain less than the actor loses, could be qualified as overly altruistic. Spiteful behaviours map onto the bottom-left quadrant, where behaviour with which the actor even reduces its own fitness relative to the recipients could be called spiteful beyond reason. The selfish behaviour in the bottom-right quadrant can be divided in selfish behaviour that is efficient and selfish behaviour that is not, depending on whether or not the total aggregated fitness effects – that is, the effect on the actor plus the effects on the recipients – are positive or negative.

Whether or not we should expect a particular behaviour to be selected in a model depends on the assumptions that are made concerning the composition of the groups. If groups are formed randomly, then the vertical axis separates the behaviours that we predict will be selected (right of the vertical axis) from the behaviours that we predict will not be selected (left from it, see also Figure 3a). If groups are not formed randomly, but assortatively, then the line that separates behaviours that will be selected from behaviours that will not, will be tilted counterclockwise (see also Figure 3b. The idea of such a line being tilted by assortative matching is also present in Wilson (1975) and, in a different setting, in Rousset (2004)). If groups are formed anti-assortatively, the line will be tilted clockwise. How far it will be tilted, depends on what the assumptions of the model imply for a population characteristic that we can write as a difference in probabilities in a hypothetical chance experiment: $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$. If we randomly draw a member of the population, with all individuals having equal probability, and then randomly draw another member of the same group, with all other group members having equal probability, then $\mathbb{P}(T \mid T)$ is the probability that this individual has the trait, if the first has it too, and $\mathbb{P}(T \mid N)$ is the probability that this individual has

the trait, if the first does not. For games that fit equation (1), this difference in probabilities times the benefits on the vertical axis is the difference between the expected benefits of a carrier and the expected benefits of a non-carriers (see Theorem 1 below, which implies that this holds). The expression $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ can be seen as a generalized measure of relatedness, but it is helpful to first of all see it as a reflection of the assumptions of the model concerning the population- or interaction structure. It is important to stress that this expression is not specific to any model; it embraces whatever it is that is assumed to cause the distribution of carriers and non-carriers over the groups. Section 6 contains a precise interpretation, including a reason why it is appropriate to call it a generalized measure of relatedness.

If there is random group formation, then $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) = 0$, and the line will just coincide with the vertical axis. A probability exercise in the appendix shows that how much it can vary, depends on group size; complete assortment always leads to $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) = 1$, but anti-assortment cannot make this difference go below $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) = -\frac{1}{n-1}$, where n is the group size. The angle that the line then makes with the vertical axis is $\arctan[\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)]$. The two diagonals in Figure 1 therefore not only represent models with, respectively, complete alignment and complete conflict of interests, but they also give the boundaries between which this assortment-line can be tilted. This also implies that being overly altruistic or spiteful beyond reason will never be favoured by selection.

3 A dualistic view on group selection models

Group selection of altruistic traits is described, for instance in Sober & Wilson (1998), as a process where selective forces work at different levels and in opposite directions. Within groups, individuals that have the trait typically have a lower fitness than individuals that do not, so that within groups, selection is said to work against the trait. Groups with a larger share of individuals that have the trait however typically grow larger (or faster) than groups with a smaller share of individuals that have the trait. Selection between groups therefore is said to work in favour of the trait. Or, in the words of Wilson & Wilson (2007): "Selfishness beats altruism within groups. Altruistic groups beat selfish groups. Everything else is commentary."

In Figure 2 these two opposing forces are visualized. The first characteristic – within groups, carriers of the trait do worse – implies that the fitness effects lie up-left from the "complete alignment of interests" line. The second characteristic – groups with many

carriers of the trait do better – implies that the fitness effects lay up-right from the line that separates efficient from inefficient behaviours. This implies that such models lie in the area north of the V-shaped boundary that consists of the two diagonals.

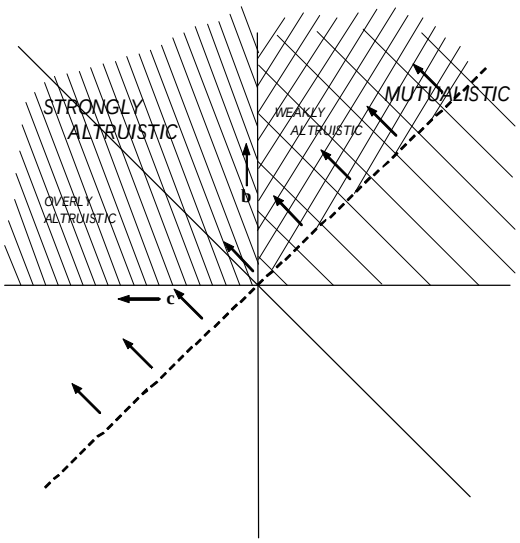

Figure 2a: "Selfishness beats altruism within groups."

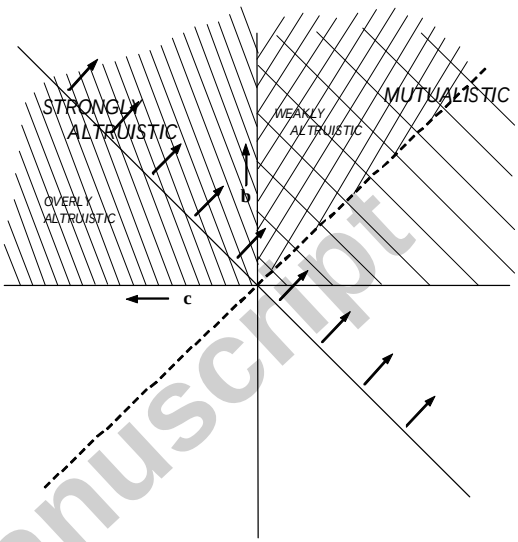

Figure 2b: "Altruistic groups beat selfish groups"

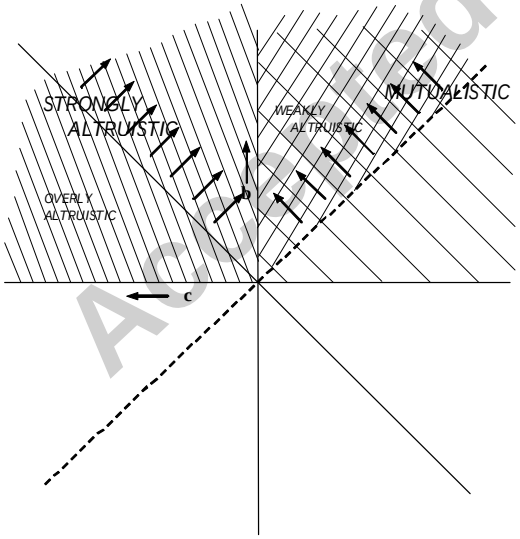

Figure 2c: Group selection models map onto points in this area.

It is important to see that if a trait in this area is selected, there are two ways of understanding why it is. Both are correct, and both contain a part of the insight that

the other does not capture. The classic group selection view is that if such a trait is selected, then between group selection is stronger than within group selection; see, again, Sober & Wilson (1998) and Wilson & Wilson (2007). Whether or not it indeed is, of course has to follow, one way or another, from the assumptions of the model, but it is undeniably true that on the border between traits that do and traits that do not get selected within a certain model, these two forces must cancel each other out. These two opposing selection pressures are a characteristic of the whole region north of the V-shaped boundary.

We can, on the other hand, also make a very relevant distinction between two parts of this region. If a weakly altruistic trait is selected, as in Figure 3a, then the reason that it is selected is that the fitness effect on the actor itself is positive. If a strongly altruistic trait is selected, as in Figure 3b, then the explanation can only be that the assumptions of the model imply that groups, in expectation, are composed assortatively enough, tilting the dotted line counter-clockwise.

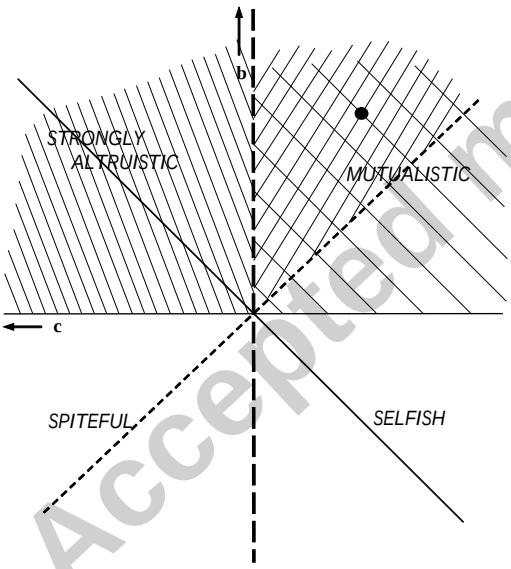

Figure 3a: Group selection by shared interests. The dashed line being vertical represents random group formation.

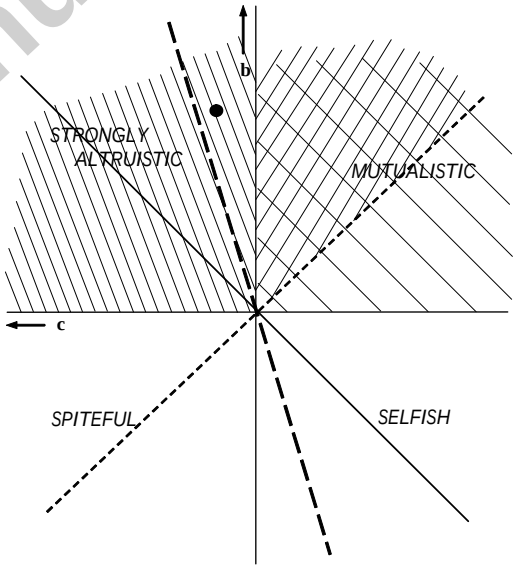

Figure 3b: Group selection by shared genes / assortment. The dashed line being tilted represents assortative group formation.

A few useful insights already follow directly from this dualistic view on group selection models. The first is that not all group selection models are the same. Group selection models with random group formation, in which weakly altruistic traits are selected, are fundamentally different from group selection models in which strongly altruistic traits

are selected. Also, some group selection models can be understood as kin selection models, but not all: only those in which strongly altruistic traits are selected. They do however all fit within an inclusive fitness setting; for weakly altruistic traits, by definition $\mathbf{c} < 0$ and $\mathbf{b} > 0$ and hence, trivially, $r\mathbf{b} - \mathbf{c} > 0$.

4 Not all group selection models can be reformulated in terms of inclusive fitness

While equation (1) implies that costs and benefits of individual behaviour do not depend on the composition of the remainder of the group, there are examples of realistic models, both with random and with assortative or non-random matching that do not share this characteristic. Examples are Avilés (1999, 2002), Avilés, Abbot & Cutter (2002), Avilés, Fletcher & Cutter (2004), Bowles, Choi & Hopfensitz (2003) - see also Van Veelen & Hopfensitz (2007) - and Hauert, Michor, Nowak & Doebeli (2006). In fact, one could argue that (1) defines only a small subset of all public goods games, excluding for instance all models that contain synergies. In order to capture all group selection models, we will therefore have to let go of the linearity in the public goods game that (1) imposes, and allow for all possible functions $\pi(T, i, f)$ and $\pi(N, i, f)$.

Without restrictions on the payoff functions, it is natural to ask ourselves whether or not we can still arrive at a description of costs, benefits and relatedness that makes all group selection models map onto Figure 1. More precisely, we would like to find out if the direction of selection in a group selection model can always be determined by computing inclusive fitness, which is also a question that emerges from the recent literature; see for instance Traulsen & Nowak (2006) and the kin-selection reinterpretation of that model by Lehmann, Keller, West & Roze (2007) as well as Killingback, Bieri & Flatt (2006) and a similar reinterpretation by Grafen (2007). The following theorem provides a positive answer for linear public goods games.

Theorem 1 *If the payoff function satisfies equation (1), then the direction of selection follows from Hamilton's rule, with* $\mathbf{c} = c(f) - b(f)$, $\mathbf{b} = (n-1) \cdot b(f)$ *and* $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$.

Proof. The division of the population in groups is given by values for $f_i, i = 1, ..., n$. Here f_i is the frequency of groups that have i carriers of the trait, and naturally we assume that $\sum_{i=0}^{n} f_i = 1$ and we define $p = \frac{\sum_{i=0}^{n} i \cdot f_i}{n}$, or $\sum_{i=0}^{n} i \cdot f_i = np$. The frequency of the trait goes up if

$$\frac{\sum_{i=1}^{n} i \cdot f_i \cdot \pi(T,i,f)}{np} > \frac{\sum_{i=0}^{n-1} (n-i) \cdot f_i \cdot \pi(N,i,f)}{n(1-p)}$$

If we fill in the fitness / payoff function from (1), this is

$$\begin{aligned}
\frac{\sum_{i=0}^{n} i \cdot f_i \cdot \{1 + b(f) \cdot i - c(f)\}}{np} &> \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot \{1 + b(f) \cdot i\}}{n(1-p)} \\
\frac{(1-c(f)) \sum_{i=0}^{n} i \cdot f_i + b(f) \sum_{i=0}^{n} i^2 \cdot f_i}{np} &> \frac{n - \sum_{i=0}^{n} i \cdot f_i + b(f) \sum_{i=0}^{n} (n-i) \cdot f_i \cdot i}{n(1-p)} \\
1 - c(f) + b(f) \frac{\sum_{i=0}^{n} i^2 \cdot f_i}{np} &> 1 + b(f) \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot i}{n(1-p)} \\
-c(f) + b(f) + b(f) \frac{\sum_{i=0}^{n} i \cdot f_i \cdot (i-1)}{np} &> b(f) \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot i}{n(1-p)} \\
-c(f) + b(f) + (n-1) b(f) \frac{\sum_{i=0}^{n} i \cdot f_i \cdot \frac{i-1}{n-1}}{np} &> (n-1) b(f) \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot \frac{i}{n-1}}{n(1-p)} \\
\left\{ \begin{array}{c} (n-1) b(f) \left(\frac{\sum_{i=0}^{n} i \cdot f_i \cdot \frac{i-1}{n-1}}{np} - \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot \frac{i}{n-1}}{n(1-p)} \right) \\ -c(f) + b(f) \end{array} \right\} &> 0
\end{aligned}$$

If we randomly draw a carrier of the trait from the whole population, with all carriers of the trait having equal probability, and $\mathbb{P}(T \mid T)$ is the probability that a randomly chosen other group member, with all other group members having equal probability, is a carrier, then it follows that $\mathbb{P}(T \mid T) = \frac{\sum_{i=0}^{n} i \cdot f_i \cdot \frac{i-1}{n-1}}{np}$ and that $\mathbb{P}(T \mid N) = \frac{\sum_{i=0}^{n} (n-i) \cdot f_i \cdot \frac{i}{n-1}}{n(1-p)}$. Hence, we can rewrite the inequality as follows: the frequency of carriers of the trait increases if

(2)

$$\begin{aligned}
(n-1) b(f) (\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)) - c(f) + b(f) &> 0 \\
r \cdot \mathbf{b}(f) - \mathbf{c}(f) &> 0
\end{aligned}$$

which is Hamilton's rule, if we define the net costs as $\mathbf{c} = c(f) - b(f)$, the total benefit conferred to the other group members as $\mathbf{b} = (n-1) b(f)$ and relatedness as $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$. ■

This theorem therefore implies that the (tilted) dotted line in Figure 1 indeed separates traits that are selected from traits that are not, if the public goods game is linear. Or, in other words, if the fitness function satisfies (1), then the direction of selection is given by (2). This linearity however is crucial. Below I will provide a simple non-linear counterexample with $n = 2$ – which turns public goods games into prisoners

dilemma's – for which Hamilton's rule does not give the correct prediction. In order to see why this counterexample is not some constructed rarity, but a general problem for non-linear public goods games, it is useful to first read the proof of Theorem 1 in reverse too. Therefore we need to realize again that what inclusive fitness does, is separate the population- or interaction structure from the fitness effects, or (the payoffs of) the game. The first implies an r, the second is reflected by $\mathbf{b}$ and $\mathbf{c}$. That means that we can do two things. We can keep the fitness effects (the game) constant, and vary the population- or interaction structure. This changes the r and separates population- or interaction structures in those for which the r is high enough, and the behaviour is selected, and those for which the r is not high enough, and the behaviour is not selected. On the other hand, Theorem 1 also shows that within the set of games that satisfy (1), we can also keep the population- or interaction structure constant, and distinguish between games for which the cooperative behaviour is selected ($\frac{\mathbf{b}}{\mathbf{c}}$ is high enough) and games for which the cooperative behaviour is not selected ($\frac{\mathbf{b}}{\mathbf{c}}$ is not high enough). Therefore, if we want to extend Theorem 1 from linear public goods games to all public goods games, this would imply that the r should still only reflect the population- or interaction structure, and should not change between games. This implies that when we generalize, Theorem 1 restricts the choice for r to this difference in probabilities that works for linear public goods games. Reading the proof in reverse, we see that if we indeed stick to $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$, we can follow every step on the way back up again, apart from the last one (or the first one on the way down), in which $1 + b(f) \cdot i - c(f)$ is replaced by $\pi(T, i, f)$ and $1 + b(f) \cdot i$ is replaced by $\pi(N, i, f)$. This means that it really is the linearity of the payoff function that ties the direction of selection to Hamilton's rule. In other words, one could say that linearity is the only real ingredient of the proof; the rest is just rewriting of the inequality. Any divergence from linearity therefore means that a wedge is driven between the direction of selection and Hamilton's rule.

Therefore it is generally the case that as soon as a group selection model implies a public goods game that is not linear, inclusive fitness can give the wrong prediction. If costs and benefits of the trait do indeed depend on how many other bearers of the trait the group contains, and hence the functions $\pi(T, i, f)$ and $\pi(N, i, f)$ no longer fit in the structure of equation (1), then one can also no longer distill measures of (expected) costs and benefits – neither marginal nor aggregated or averaged – that combine with some measure of relatedness or assortativity to a concise prediction of the form: the trait will be selected if and only if $r\mathbf{b} - \mathbf{c} > 0$, where $\mathbf{b}$ and $\mathbf{c}$ characterize a fitness transfer, and r

characterizes the composition of the groups. This is an important conclusion, because it shows that not all group selection models can be translated to a prediction in the form of an expression of inclusive fitness, as is sometimes suggested.

4.1 Counterexample I

The first, simple counterexample is similar to examples given in Wenseleers (2006) and Gardner, West & Barton (2007) and is related to examples given in Day & Taylor (1998). Here, the example is given in a way that directly fits Theorem 1. Section 7 discusses how Theorem 1 and the counterexample relate to existing results.

With groups of size 2, we can represent the fitnesses as payoffs in a 2 x 2 game. We will also assume that $\mathbf{T} > \mathbf{R} > \mathbf{P} > \mathbf{S}$, which makes it a prisoners' dilemma

	N	T
N	$\mathbf{P}, \mathbf{P}$	$\mathbf{T}, \mathbf{S}$
T	$\mathbf{S}, \mathbf{T}$	$\mathbf{R}, \mathbf{R}$

Here we can easily see that this fits within equation (1) if and only if $b(f) = \mathbf{R} - \mathbf{S} = \mathbf{T} - \mathbf{P}$ and $c(f) = \mathbf{T} - \mathbf{S}$. (Nowak & Sigmund (1990) introduced the term "equal gains from switching" to indicate a situation where $\mathbf{R} - \mathbf{S} = \mathbf{T} - \mathbf{P}$.)

The division of the population in groups is given by values for f_{NN}, f_{NT} and f_{TT}, which are the frequencies of groups with 0, 1 and 2 carriers of the trait in them, respectively. Naturally, we assume that $f_{NN} + f_{NT} + f_{TT} = 1$. Selection favours the trait if the average payoff of the carriers of the trait is larger than the average payoff of individuals that do not carry the trait:

$$\frac{f_{TT} \cdot \mathbf{R} \cdot 2 + f_{NT} \cdot \mathbf{S} \cdot 1}{2p} > \frac{f_{NT} \cdot \mathbf{T} \cdot 1 + f_{NN} \cdot \mathbf{P} \cdot 2}{2(1-p)} \tag{3}$$

where the groups are weighted by the number of T-players resp. N-players in them, and p is the frequency of the trait in the overall population; $p = \frac{2f_{TT}+f_{NT}}{2}$. Natural definitions of the probabilities for being matched to the different types are $\mathbb{P}(T \mid T) = \frac{f_{TT}}{p}$ and $\mathbb{P}(T \mid N) = \frac{f_{NT}}{2(1-p)}$, with the implication that $\mathbb{P}(N \mid T) = 1 - \mathbb{P}(T \mid T) = 1 - \frac{f_{TT}}{p} = \frac{f_{NT}}{2p}$ and $\mathbb{P}(N \mid N) = 1 - \mathbb{P}(T \mid N) = 1 - \frac{f_{NT}}{2(1-p)} = \frac{f_{NN}}{1-p}$. Then we can rewrite (3) as

(4)

$$\begin{aligned}
\mathbb{P}(T \mid T) \cdot \mathbf{R} + \mathbb{P}(N \mid T) \cdot \mathbf{S} &> \mathbb{P}(T \mid N) \cdot \mathbf{T} + \mathbb{P}(N \mid N) \cdot \mathbf{P} \Leftrightarrow \\
\mathbb{P}(T \mid T) \cdot \mathbf{R} + (1 - \mathbb{P}(T \mid T)) \cdot \mathbf{S} &> \mathbb{P}(T \mid N) \cdot \mathbf{T} + (1 - \mathbb{P}(N \mid N)) \cdot \mathbf{P} \Leftrightarrow \\
\mathbb{P}(T \mid T) \cdot (\mathbf{R} - \mathbf{S}) + \mathbf{S} &> \mathbb{P}(T \mid N) \cdot (\mathbf{T} - \mathbf{P}) + \mathbf{P} \Leftrightarrow \\
\left\{ \begin{array}{c} \mathbb{P}(T \mid T) \cdot (\mathbf{R} - \mathbf{S}) - \mathbb{P}(T \mid N) \cdot (\mathbf{T} - \mathbf{P}) \\ + (\mathbf{S} - \mathbf{P}) \end{array} \right\} &> 0
\end{aligned}$$

If $(\mathbf{R} - \mathbf{S}) = (\mathbf{T} - \mathbf{P})$, then one can replace $(\mathbf{R} - \mathbf{S})$ and $(\mathbf{T} - \mathbf{P})$ with $\mathbf{b}$, replace $(\mathbf{P} - \mathbf{S})$ with $\mathbf{c}$, and $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ with r, which leads to a well-known expression with inclusive fitness:

$$r \cdot \mathbf{b} - \mathbf{c} > 0$$

If however $(\mathbf{R} - \mathbf{S}) \neq (\mathbf{T} - \mathbf{P})$, then it is not possible to rewrite (4) in a way that separates $\mathbb{P}(T \mid T) \cdot (\mathbf{R} - \mathbf{S}) - \mathbb{P}(T \mid N) \cdot (\mathbf{T} - \mathbf{P})$ in a product of a term that only depends on the composition of the population and something that only depends on the fitness function.

For the counterexample, we choose values such that $(\mathbf{R} - \mathbf{S}) \neq (\mathbf{T} - \mathbf{P})$, which implies the game does not satisfy the condition for Theorem 1 to apply; $\mathbf{T} = 3, \mathbf{R} = 2.5, \mathbf{P} = 1$ and $\mathbf{S} = 0$. Figures 5 and 6 in Section 5 depict these payoffs and can be helpful to visualize the game.

With groups of size 2, the composition of the population is uniquely determined by the frequency p of the trait, and a parameter of assortment α, that we will see below equals relatedness.

(5)

$$\begin{aligned}
f_{TT} &= (1 - \alpha) p^2 + \alpha p \\
f_{NT} &= (1 - \alpha) 2p (1 - p) \\
f_{NN} &= (1 - \alpha) (1 - p)^2 + \alpha (1 - p)
\end{aligned}$$

This makes $p = \frac{2f_{TT} + f_{NT}}{2}$ the frequency of the trait in the population and $0 \leq \alpha \leq 1$ a parameter of assortment; $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) = \frac{f_{TT}}{p} - \frac{f_{NT}}{2(1-p)} = \alpha$. From (4) we know that the trait is selected if

(6)

$$\begin{aligned} \mathbb{P}(T \mid T) \cdot (\mathbf{R} - \mathbf{S}) - \mathbb{P}(T \mid N) \cdot (\mathbf{T} - \mathbf{P}) + (\mathbf{S} - \mathbf{P}) &> 0 \\ \frac{f_{TT}}{p} \cdot 2.5 - \frac{f_{NT}}{2(1-p)} \cdot 2 - 1 &> 0 \\ 2.5((1-\alpha)p + \alpha) - 2(1-\alpha)p - 1 &> 0 \\ 0.5(1-\alpha)p + 2.5\alpha - 1 &> 0 \end{aligned}$$

If $p = 0$, then that implies that the trait can invade if $\alpha > \frac{2}{5}$. If on the other end $p = 1$, then the trait is stable if $\alpha > \frac{1}{4}$. Hence, for $\frac{1}{4} < \alpha < \frac{2}{5}$ there is bi-stability (see also Hauert, Michor, Nowak & Doebeli, 2006). The dynamics for different values of assortativity parameter α are given in Figure 4 below.

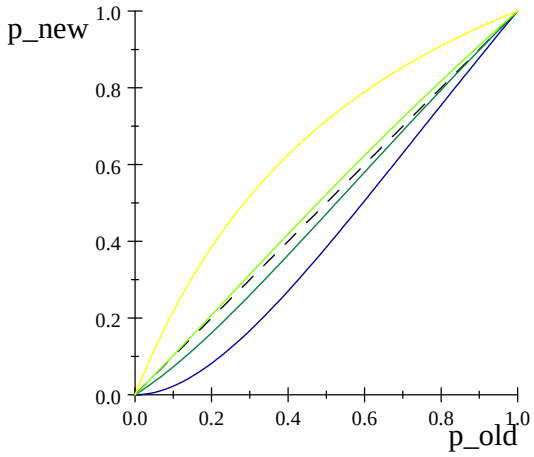

Figure 4: Dynamics for different values of assortment parameter α. The frequency in the next period is plotted as a function of the frequency in the current period for $\alpha = 0$ (blue), $\alpha = \frac{1}{4}$ (dark green), $\alpha = \frac{2}{5}$ (light green) and $\alpha = 1$ (yellow).

If we now take for instance $\alpha = 0.22 < \frac{1}{4}$, then we know that at $p = 1$ the population can be invaded and will be replaced by N-players. Yet, if we would take an inclusive fitness approach and compute the benefit that players confer on their partners (which is $\mathbf{R} - \mathbf{S}$, because all carriers of the trait meet individuals that also carry the trait), the net costs they make (which is $\mathbf{T} - \mathbf{R}$ for the same reason), and relatedness, then we get:

(7)

$$r \cdot \mathbf{b} - \mathbf{c} = 0.22 \cdot 2.5 - 0.5 = 0.05 > 0$$

suggesting, incorrectly, that this behaviour is stable. This indicates that, when fitnesses no longer fit (1), computing inclusive fitnesses no longer correctly indicate the direction of selection correctly. The intuition is provided below.

The values chosen for $(\mathbf{R}-\mathbf{S}), (\mathbf{T}-\mathbf{P})$ and $(\mathbf{S}-\mathbf{P})$ can also be replaced with $w\cdot(\mathbf{R}-\mathbf{S}), w\cdot(\mathbf{T}-\mathbf{P})$ and $w\cdot(\mathbf{S}-\mathbf{P})$, respectively. Letting w be small would then imply w-weak selection (small fitness contribution of the game; see Wild & Traulsen, 2007), but the results above would still hold; the direction of selection in (6) and the sign of the inclusive fitness in (7) remain unchanged. In section 7 we will also discuss δ-weak selection (small distance in phenotype; see again Wild & Traulsen, 2007).

5 Intuition

The fact that linearity of the public goods game is needed in Theorem 1 can at first sight be perceived as counterintuitive. In order to get an intuition why linearity is indeed needed, we can think of how Hamilton (1975) motivated the $r\mathbf{b}$ part of his rule. The idea is that there is a fixed benefit to be gained from an individual that has the trait, and that it is received, in expectation, for a $(1-r)$ share by a random sample from the population, and for an r share by carriers of the trait other than the individual itself. However, when the payoffs of individuals no longer fit within (1), then the benefit that one individual receives from another individual having the trait, depends on the composition of the remainder of the group, including the receiving individual itself. In the simple counterexample given in the appendix, with groups of size 2, that means that what phenotype the receiver is, determines the size of costs and benefits. This is pictured in Figure 5; the fitness transfer if the receiver is a carrier of the trait is not the same as the fitness transfer if the receiver is not a carrier. This implies that one can no longer characterize the fitness effect of the trait as a fitness transfer that is fixed – $\mathbf{b}$ – or that is fixed for a given frequency in the population. Hence one can also no longer break down a fixed benefit into a part that goes to other carriers and a part that goes to non-carriers of the trait, because these two do not receive the same benefits from it.

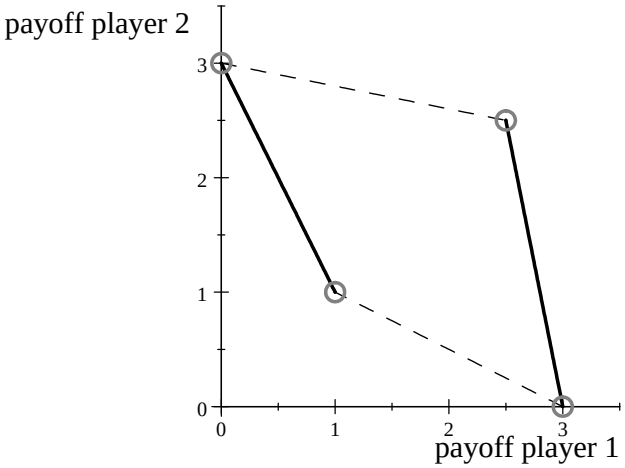

Figure 5: In the counterexample, costs and benefits depend on the phenotype of the receiver. Here we identify player 1 as the donor, or the acting individual, and player 2 as the receiver. The grey circle at $(2.5, 2.5)$ represents the payoffs when both individuals are carriers of the trait, and the grey circle at $(3, 0)$ represents the payoffs when player 2 has the trait and player 1 does not. The line between them therefore represents the fitness transfer by player 1 if player 2 is a carrier. Similarly, the other line, to the left, represents the fitness transfer if player 2 is not a carrier. Costs and benefits of the fitness transfer by player 1 now depend on the phenotype of player 2.

Another way of forming an intuition for this result can be to realize that the *marginal* fitness transfers in a situation with all carriers of the trait, measure effects of one-step deviations. These however do not add up to the true combined effect of deviations (in the grey circle in Figure 6 below). With a positive value for $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ inclusive fitness therefore underestimates how well a mutant non-carrier does for this particular payoffs.

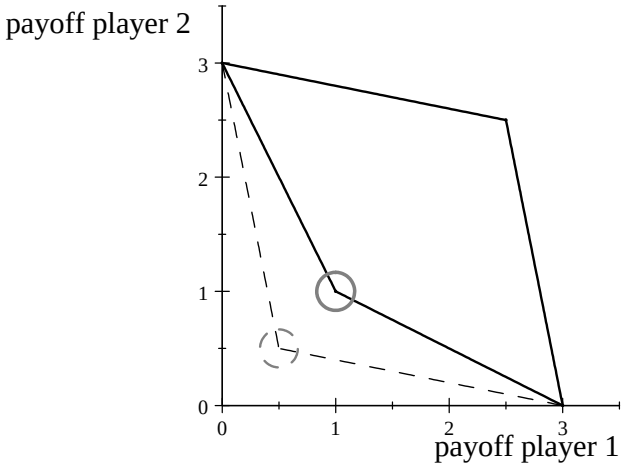

Figure 6: Another way to describe the payoffs of the counterexample is to say that marginal effects of deviations do not add up.

From a close reading of the proof of Theorem 1 and the derivations for the counterexample we have also learned that this cannot be mended by (sophisticated) averaging of cost and benefits, nor by assuming w-weak selection. The sign of inclusive fitness as well as the direction of selection remain what they are when $\pi(T, i, f)$ is scaled down, and hence the divergence between them does not disappear with w-weak selection. If we allow for a continuum of strategies, and assume population states to be monomorphous and moving according to the derivative taken with respect to player's deviations - that is, we examine δ-weak selection - then Hamilton's rule will be restored for this two-player example (see Grafen, 1979, Day & Taylor, 1998, and Wild & Traulsen, 2007. In Grafen (1979) the analysis is done for the Hawk-Dove game; see Appendix C for how this carries over to more general 2 by 2 games). The counterexample in subsection 6.1 shows that for three or more players, this is in general not possible anymore.

6 Relatedness

Above we have defined relatedness as a difference in conditional probabilities: $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$. It should be stressed though that it only summarizes the state of the population, because it is defined as a difference in probabilities in a *hypothetical* chance experiment; *if* we would randomly choose a carrier of the trait from the whole population, with all carriers of the trait having equal probability, and then compute the

probability that a randomly chosen other group member, with all other group members having equal probability, is a carrier too, then that can be written as $\mathbb{P}(T \mid T)$. The second term, $\mathbb{P}(T \mid N)$, is found in an analogous way. These probabilities then still are only a characteristic of a population; they are functions of the distribution of carriers over the groups. This difference is therefore only a measure for the unevenness of the distribution of carriers of the trait over groups (doing the calculations of the bounds on relatedness in Appendix A really helps to form an intuition).

We can think of many evolutionary processes as Markov chains, where states are populations, and transition probabilities between states reflect a combination of population structure and fitnesses (see for instance Chapters 6 to 8 in Nowak, 2006, or Van Veelen & Hopfensitz, 2007, for a Markov chain where states indeed are subdivided populations). For a Markov process, one can first compute $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ for every state in the support of the invariant distribution. Then we can weight these measures by their probability in the invariant distribution. For this Markov chain, we can think of a new chance experiment. Suppose the population follows this model, and we can assume that looking at it today is like taking a draw from the invariant distribution (that is, it has been running for a long time). Then we take a random group, and from that group we pick two random group members, without replacement. The difference in conditional probabilities with which the *second* group member is a carrier - that is, the difference between the conditional probability for carriers and the conditional probability for non-carriers - equals this weighted average.

Luckily, this matches our general idea of what relatedness should be. In Appendix B we show that indeed:

(8)
$$\mathbb{P}(T \mid T) = r + (1 - r)\, p$$

where $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$, and p is the frequency of carriers in the overall population. That is, the probability of someone in my group being a carrier, conditional on me being one, is r plus $(1 - r)$ times the frequency of carriers in the overall population. This matches for instance Grafen's (1985) geometrical view of relatedness. We should be aware though that if we write relatedness as a regression with error terms, then that suggests that we are doing statistics. Statistics is meant to estimate values or test hypotheses concerning the true model. Doing statistics therefore would imply that we do not know the real value, and that we actually carry out this hypothetical chance experiment on an *unknown* Markov chain in order to find out more about the true model.

Our definition of relatedness on the other hand implies that relatedness simply is to follow from model assumptions. We should therefore realise that an assumed model - or the true underlying model - can have many interesting properties other than just $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$. The variance of this $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ across states might for instance differ a lot. But it is also possible, as we will see in the counterexample below, that there are models for which relatedness is the wrong population characteristic to look at.

6.1 Counterexample II

Two players. We begin with the following two player stag-hunt game (following the parable by Rousseau, 1973), which does not satisfy linearity, but where relatedness still helps finding the dynamics:

	N	T
N	$0, 0$	$0, -1$
T	$-1, 0$	$1, 1$

In a picture this looks as follows:

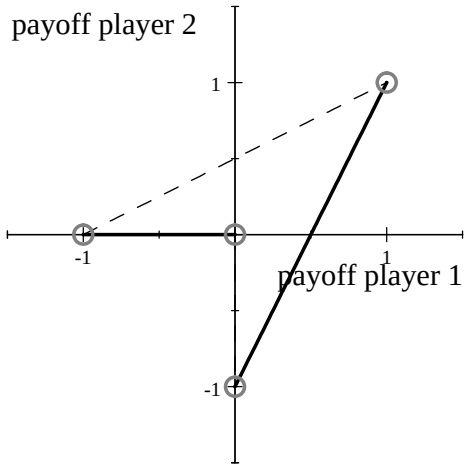

Figure 7: A stag-hunt game with two players. Without players being related, this is a coordination game with two symmetric pure equilibria.

It is not too hard to see that the whole population playing N and the whole population playing T are the two candidates for stability. In order to find their basins of attraction, we compute the (unstable) mixed equilibrium in between, that is, we look for a frequency p of carriers of the trait for which the payoffs of both coincide:

$$\begin{aligned}
\mathbb{P}(T \mid T) \cdot \mathbf{1} + \mathbb{P}(N \mid T) \cdot -\mathbf{1} &= \mathbb{P}(T \mid N) \cdot \mathbf{0} + \mathbb{P}(N \mid N) \cdot \mathbf{0} \quad \Leftrightarrow \\
\mathbb{P}(T \mid T) \cdot \mathbf{1} + \{1 - \mathbb{P}(T \mid T)\} \cdot -\mathbf{1} &= 0 \quad \Leftrightarrow \\
\mathbb{P}(T \mid T) \cdot 2 &= 1 \quad \Leftrightarrow \\
\mathbb{P}(T \mid T) &= \frac{1}{2}
\end{aligned}$$

By (8) we can rewrite that as

$$\begin{aligned}
r + (1 - r)\, p &= \frac{1}{2} \\
p &= \frac{1 - 2r}{2 - 2r}
\end{aligned}$$

For frequencies lower than $\frac{1-2r}{2-2r}$, the dynamics will take the population to $p = 0$, and for frequencies that are higher, the dynamics will take the population to $p = 0$. Please note that, even though the game is not a linear public goods game, and indeed the prediction does not follow Hamilton's rule, the prediction still uses a formula in which r features. The same applies to the first counterexample, where we have shown that the parameter that matters - α - equals relatedness. When we go to games with more than two players, we will see that this no longer holds.

Three players. With three players, we can define a stag-hunt game as pictured below. Again, the candidates for stability are all playing N, with payoffs $(0, 0, 0)$ and all playing T, with payoffs $(1, 1, 1)$

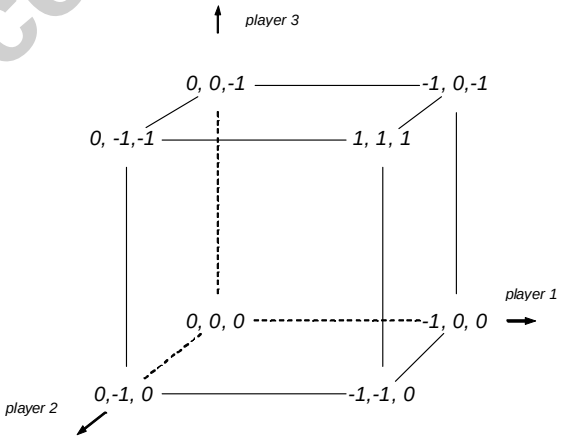

Figure 8: A stag-hunt game with three players. With random matching, this is again a coordination game with two symmetric pure equilibria.

In order to find their basins of attraction, we again compute the (unstable) mixed equilibrium in between, that is, we look for a frequency p of carriers of the trait for which the payoffs of both types coincide. We now denote the probability of facing two more carriers in the group of three, given that an individual is a carrier itself, as $\mathbb{P}(TT \mid T)$. If we realise that the payoff is 0 regardless of the others in the group, if an individual itself is not a carrier, we can write that as follows:

$$\begin{aligned} \mathbb{P}(TT \mid T) \cdot \mathbf{1} + \{1 - \mathbb{P}(TT \mid T)\} \cdot -\mathbf{1} &= 0 \quad \Leftrightarrow \\ \mathbb{P}(TT \mid T) \cdot 2 &= 1 \quad \Leftrightarrow \\ \mathbb{P}(TT \mid T) &= \frac{1}{2} \end{aligned}$$

The last probability can generally not be expressed in terms of p and r. In order to see why, we should realise that for groups of 2, choosing of f_{NN}, f_{NT} and f_{TT} (or, alternatively, f_0, f_1 and f_2) gives us two degrees of freedom; because f_{NN}, f_{NT} and f_{TT} have to add up to one, choosing two of them determines the third. That implies that a distribution of carriers over groups - that is, a choice for f_{NN}, f_{NT} and f_{TT} - is completely determined by a choice of p and r. In other words, any combination of p and r allows for only one combination of group-frequencies such that $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ and $p = \frac{1}{2}(f_{NT} + 2f_{TT})$. With groups of three players, the choice of f_0, f_1, f_2 and f_3 gives us three degrees of freedom. One combination of values of p and r therefore can summarize *different* underlying distributions of carriers over the groups. For our example, that means that the same combination of values for p and r can come with different values for f_3, and hence with different values of $\mathbb{P}(TT \mid T) = \frac{f_3}{p}$.

For this game, relatedness would therefore not be the accurate population characteristic to look at. What matters is not $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ but $\mathbb{P}(TT \mid T)$, and there is no one-to-one mapping between the two of them. We should also realize that $\mathbb{P}(TT \mid T)$ also in no way reflects an alternative definition of relatedness in the literature. For groups larger than 3, the degrees of freedom for choosing $f_0, ..., f_n$ obviously only increase with n.

For 2 by 2 games, we know that Hamilton's rule does make the correct prediction if the players can choose from a continuum of actions, rather than from a finite set (see Grafen, 1979, Day & Taylor, 1998). For the three player stag-hunt game, however, filling up the interval between 0 and 1 and examining δ-weak selection does not do the same. This is shown in Appendix C, but it is not hard to imagine that this indeed is

unavoidable, since also there what matters is how an individual's change in strategy changes $\mathbb{P}(TT \mid T)$, and knowing r is not enough to determine that.

7 Relation to existing results

As mentioned before, the first counterexample is similar to examples given in Wenseleers (2006) and Gardner, West & Barton (2007). The difference found between models with a discrete strategy space (cooperate or not cooperate) and possibly heterogeneous, mixed populations on the one hand and a continuous strategy space and monomorphic populations on the other hand is documented in Grafen (1979) - who responds to Hines & Maynard Smith (1979) - and also in Day & Taylor (1998) and, slightly differently, in Wild & Traulsen (2007). What is different though, is that those results all concern 2 by 2 games. When we think of group selection, we naturally want to consider groups of any size, and when we think of cooperation, we also want to implicate cooperative efforts done with 3 individuals or more. So while the 2 by 2 counterexample is not new, the formal result that Hamilton's rule does work for groups of any size when the game is a linear public goods game is. What is also new is the 3 by 3 counterexample, and its two implications. When the game is not a linear public goods game, one can, for groups of size two, still make a prediction for the direction of selection in which relatedness features. This allows for an adjusted version of Hamilton's rule, with alternative b and c, for invasion (at $p = 0$). The counterexample shows that with groups of sizes larger than 2, this is not possible. Also, in the 2 by 2 case, changing from a binary to a continuous strategy space guarantees that Hamilton's rule works, even when the game is not a linear public goods game (or, since it is a 2 by 2 game: when it is not a prisoners dilemma with equal gains from switching). This also is no longer true with groups larger than 2.

Another recent article is Traulsen, Shoresh & Nowak (2008), where analytical results are derived for fixation probabilities in a Moran process where fitness is an exponential function of payoff. These can be seen as to imply something similar, in the sense that also there, the ratio of fixation probabilities in a kin selection framework only coincides with the ratio of fixation probabilities in their setting when there is equal gains from switching. Their model, however, and thereby also the model in Traulsen & Nowak (2006), is different from the model here. In their model, pairs of individuals are matched within their group to play a 2 x 2 game, while here the entire group plays an n player game, which, again, is what one can imagine is an appropriate model for

collective action within human groups or for cells within multi-cellular organisms, for instance. Theorem 1, the central result in this paper, can furthermore be understood without reference to the Moran process - which could be a good or a bad thing - and gives a formal proof of a general result that remains relatively close to basic game theory as well as to Hamilton's (1964, 1975) justification for his rule.

The framework is also different than most. Everything is exact here; there are no first order approximations involved. It also does not use the Price equation, the limitations of which I have discussed in Van Veelen (2005). That does not mean there are no great similarities with the existing literature. Section 6 for instance derives that this setting also justifies the geometrical view of relatedness by Grafen (1985). This similarity, I think, is a good thing. But the way relatedness is built up in this framework from a population characteristic, is not superfluous; it gives the geometrical view a proper fundament. Generally, this definition for relatedness gives the *statistical* idea we tend to have about relatedness its proper probabilistic basis.

Furthermore Figure 1 looks similar to Figure 7.2 in Rousset. Again, this similarity is not at all a bad thing. We should realize, however, that the model setup here is so basic and simple, that everything in Figure 1 gets a geometric meaning. Again, this formalism does not use approximations, everything is exact, and the setup allows for formal proofs of exact statements about the validity, and the limits to the validity of inclusive fitness.

8 Discussion

Although Theorem 1 shows that the representation in Figure 1 only applies to models that translate to linear public goods games, one of its main insights nonetheless does carry over to the general model. Also with more general functions $\pi(T, i, f)$ and $\pi(N, i, f)$ there is merit in two ways of looking at the same models. On the one hand there is a whole continuum of models that are similar in the sense that within groups, bearers of the trait do worse, while between groups, groups with many bearers do better. On the other hand these same models can also be distinguished into two fundamentally different types, namely models that do and models that do not need assortative group formation in order to get selection of the group beneficial trait. The reasons why (similar) group selection models can work, can therefore be quite different; they can work because individuals in groups have shared genes or because they have shared interests (or both, see Van Veelen & Hopfensitz, 2007).

The theorem formally shows that deriving the direction of selection in a group selection model by computing inclusive fitness does work if a models translates to a linear public goods game. This linearity is needed; when a model translates to a non-linear public goods game, it is no longer true that inclusive fitness gives the direction of selection. It should also be stressed that the class of models that translate to non-linear public goods games is large and contains realistic models from the existing literature, as well as models that have potential for explaining phenomena ranging from the evolution of multi-cellular life to human sociality. Counterexamples also show that the size of the group makes a qualitative difference. While groups of size 2, or 2 by 2 games, the prediction of the direction of selection can be given with a formula that uses relatedness, this is not possible for groups of 3 or more individuals. Something similar is also true for the alternative model with a continuous action space; Hamilton's rule does hold there for 2 by 2 games, but not for games with more than 2 players, unless, of course, the game is a linear public goods game.

9 Acknowledgements

I would like to thank Leticia Avilés, Maus Sabelis, Martijn Egas, Sébastien Lion and an anonymous referee. This research was done at the Institute for Advanced Studies in Berlin and the University of Amsterdam, and supported by a grant form the Netherlands' Organisation for Scientific Research (NWO).

A Appendix: Bounds on $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$

Suppose there are m groups each consisting of n individuals, adding up to a total of $n \cdot m$ individuals, K of which have the trait. In group i, k_i individuals have the trait. This naturally implies that $0 \leq k_i \leq n$ and $\sum_{i=1}^{m} k_i = K$, which makes K the total number of carriers. Now consider the following expression:

(9)

$$\sum_{i=1}^{m} \frac{k_i}{K} \frac{k_i - 1}{n - 1} - \sum_{i=1}^{m} \frac{n - k_i}{nm - K} \frac{k_i}{n - 1}$$

This expression is a measure for the unevenness of the distribution of carriers the trait over groups. In order to see how, it can help to read this expression, for a given composition of the groups, as a difference between two probabilities that follow from a

hypothetical chance experiment. If we randomly choose a carrier of the trait from the whole population, with all carriers of the trait having equal probability, then the first term in expression (2) is the probability that a randomly chosen other group member, with all other group members having equal probability, is a carrier too, of course conditional on the first one being a carrier. This can be written shortly as $\mathbb{P}(T \mid T)$. The second term can, in an analogous way, be written as $\mathbb{P}(T \mid N)$. This makes expression (9) equal $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$. It should be stressed though that this only summarizes the state of the population, and writing it as a difference in probabilities in a *hypothetical* chance experiment is for the moment only done for practical reasons (see also Van Veelen, 2005).

We can rewrite this expression as follows:

$$
\begin{aligned}
& \left(\sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - 1}{n-1} + \sum_{i=1}^{m} \frac{\frac{K}{m}}{K} \frac{k_i - 1}{n-1}\right) - \left(\sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{k_i}{n-1} + \sum_{i=1}^{m} \frac{\frac{nm-K}{m}}{nm-K} \frac{k_i}{n-1}\right) \\
= & \left(\sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - 1}{n-1} + \sum_{i=1}^{m} \frac{1}{m} \frac{k_i - 1}{n-1}\right) - \left(\sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{k_i}{n-1} + \sum_{i=1}^{m} \frac{1}{m} \frac{k_i}{n-1}\right) \\
= & \sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - 1}{n-1} - \sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{k_i}{n-1} + \sum_{i=1}^{m} \frac{1}{m} \frac{k_i - 1}{n-1} - \sum_{i=1}^{m} \frac{1}{m} \frac{k_i}{n-1} \\
= & \sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - 1}{n-1} - \sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{k_i}{n-1} + \frac{-1}{n-1} \\
= & \sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - 1}{K} \frac{K}{n-1} - \sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{k_i}{nm-K} \frac{nm-K}{n-1} - \frac{1}{n-1}
\end{aligned}
$$

Because $\sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} = 0$ and $\sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} = 0$, we can add any constant to the terms $\frac{k_i - 1}{n-1}$ and $\frac{k_i}{n-1}$, respectively, and hence rewrite this as:

$$
\begin{aligned}
= & \sum_{i=1}^{m} \frac{k_i - \frac{K}{m}}{K} \frac{k_i - \frac{K}{m}}{K} \frac{K}{n-1} + \sum_{i=1}^{m} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{(n-k_i) - \frac{nm-K}{m}}{nm-K} \frac{nm-K}{n-1} - \frac{1}{n-1} \\
= & \frac{K}{n-1} \sum_{i=1}^{m} \left(\frac{k_i - \frac{K}{m}}{K}\right)^2 + \frac{nm-K}{n-1} \sum_{i=1}^{m} \left(\frac{(n-k_i) - \frac{nm-K}{m}}{nm-K}\right)^2 - \frac{1}{n-1} > -\frac{1}{n-1}
\end{aligned}
$$

It is clear that the above calculations imply that this measure for the unevenness of the distribution of carriers the trait over groups should be bounded from below by $-\frac{1}{n-1}$. An obvious upper bound is 1, because $\mathbb{P}(T \mid T)$ cannot be larger than 1 and $\mathbb{P}(T \mid N)$

cannot be smaller that 0. We can however also think of a process of selection as a Markov chain, with transition probabilities between states. These transition probabilities then reflect a combination of population structure and fitnesses. For this Markov process, one can first compute $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$ for every state in the support of the invariant distribution, and then weight these measures by their probability in the invariant distribution. The resulting weighted average can then properly be interpreted as a generalized measure of relatedness for the population- or interaction structure; if one observes a population that indeed follows the model, and we compare the probability with which a random other group member is a carrier between carriers and non-carriers, the weighted average tells us how large the difference is between them. Also $-\frac{1}{n-1}$ would again be the lower bound and 1 would be the upper bound, as they are lower and upper bound, respectively, for all states over which the weighted average is taken.

Please note that $\mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) = [\mathbb{P}(T \mid T) - 1] - [\mathbb{P}(T \mid N) - 1] = \mathbb{P}(N \mid N) - \mathbb{P}(N \mid T)$ which is a symmetry that we would expect.

B Appendix: Relatedness

Before doing the actual computations, it is useful to introduce some notation. The hypothetical chance experiment involves drawing two individuals from the same group, without replacement. The probability with which the second is a T, conditional on that the first one is a N, is written as $\mathbb{P}(T \mid N)$. The probabilities $\mathbb{P}(T \mid T)$, $\mathbb{P}(N \mid N)$ and $\mathbb{P}(N \mid T)$ are defined in an analogous way. Also, obviously, $\mathbb{P}(N \mid T) = 1 - \mathbb{P}(T \mid T)$ and $\mathbb{P}(N \mid N) = 1 - \mathbb{P}(T \mid N)$.

The probability that first an N is drawn, and then a T, is $\mathbb{P}(NT)$. Analogously we define $\mathbb{P}(TN)$, $\mathbb{P}(NN)$ and $\mathbb{P}(TT)$. Obviously, $\mathbb{P}(TT) + \mathbb{P}(TN) = p$ and $\mathbb{P}(NT) + \mathbb{P}(NN) = 1 - p$. By Bayes rule we also know that for instance $\mathbb{P}(T \mid N)$ equals $\frac{\mathbf{P}(NT)}{\mathbf{P}(NT)+\mathbf{P}(NN)}$. Note that drawing without replacement implies that $\mathbb{P}(NT) = \mathbb{P}(TN)$ - or, to put it differently, $\mathbb{P}(N)\,\mathbb{P}(T \mid N) = \mathbb{P}(T)\,\mathbb{P}(N \mid T)$ - as we can see using the formula's from the proof of Theorem 1:

$$\begin{aligned}\mathbb{P}(NT) &= \mathbb{P}(N) \cdot \mathbb{P}(T \mid N) = (1-p)\,\frac{\sum_{i=0}^{n}(n-i)\cdot f_i \cdot \frac{i}{n-1}}{n(1-p)} = \\ &= p \cdot \frac{\sum_{i=0}^{n} i \cdot f_i \cdot \frac{n-i}{n-1}}{np} = \mathbb{P}(T) \cdot \mathbb{P}(N \mid T) = \mathbb{P}(TN)\end{aligned}$$

This, together with $\mathbb{P}(N) = 1 - \mathbb{P}(T)$, implies the following:[2]

$$\begin{aligned} \mathbb{P}(N)\,\mathbb{P}(T \mid N) &= \mathbb{P}(T)\,\mathbb{P}(N \mid T) \qquad \Leftrightarrow \\ (1-\mathbb{P}(T))\,\mathbb{P}(T \mid N) &= \mathbb{P}(T)\,\mathbb{P}(N \mid T) \qquad \Leftrightarrow \\ \mathbb{P}(T \mid N) &= \mathbb{P}(T)\,(\mathbb{P}(N \mid T) + \mathbb{P}(T \mid N)) \end{aligned}$$

which prepares us for the proof of equation (8), which states that $\mathbb{P}(T \mid T) = r + (1-r)\,p$, with $r = \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N)$.

$$\begin{aligned} \mathbb{P}(T \mid T) &= \mathbb{P}(T \mid T) - \mathbb{P}(T \mid N) + \mathbb{P}(T \mid N) \\ &= r + \mathbb{P}(T)\,(\mathbb{P}(N \mid T) + \mathbb{P}(T \mid N)) \\ &= r + \mathbb{P}(T)\,(1 - \mathbb{P}(T \mid T) + \mathbb{P}(T \mid N)) \\ &= r + p\,(1-r) \qquad\qquad Q.E.D. \end{aligned}$$

C Appendix: δ-weak selection with a continuum of phenotypes

Two players. Following Grafen (1979), we first consider the fitness function $F(t,s)$, where t and s are mixed strategies that play T with probability t and s respectively. The definition of relatedness naturally carries over to r being a difference in probabilities concerning what type of player - a t or an s - one is matched with. In slightly abusive, but short notation, that is: $r = \mathbb{P}(t \mid t) - \mathbb{P}(t \mid s)$. The function F now evaluates the fitness of a mutant t in a monomorphic population of incumbent s. Because playing N always gives 0, we have

$$\begin{aligned} F(t,s) &= \mathbb{P}(t \mid t)\left\{t^2 \cdot 1 + t\,(1-t) \cdot -1\right\} + \mathbb{P}(s \mid t)\left\{ts \cdot 1 + t\,(1-s) \cdot -1\right\} \\ &= \mathbb{P}(t \mid t)\left\{2t^2 - t\right\} + \mathbb{P}(s \mid t)\left\{2ts - t\right\} \end{aligned}$$

Examining the success of a mutant t implies that it starts at frequency $p = 0$, by which (8) implies that at invasion $\mathbb{P}(t \mid t) = r$ and $\mathbb{P}(s \mid t) = 1 - r$ (or, rephrased directly; assuming that the frequency is very small implies that $\mathbb{P}(t \mid s) = 0$.) Therefore, taking the derivative with respect to t, we get

$$\frac{dF(t,s)}{dt} = r\,(4t-1) + (1-r)\,(2s-1)$$

[2] This (shorter) version of the proof was suggested by an anonymous referee.

and hence

$$\left.\frac{dF(t,s)}{dt}\right|_{t=s} = (2s-1)+2rs = 2s(1+r)-1$$

so that

$$\left.\frac{dF(t,s)}{dt}\right|_{t=s} = 0 \Leftrightarrow s = \frac{1}{2(1+r)}$$

Looking at the derivative again, we note that $s > \frac{1}{2(1+r)}$ implies that $\left.\frac{dF(t,s)}{dt}\right|_{t=s} > 0$, while $s < \frac{1}{2(1+r)}$ implies that $\left.\frac{dF(t,s)}{dt}\right|_{t=s} < 0$. This $s = \frac{1}{2(1+r)}$ therefore separates the basins of attraction of the two pure equilibria. It is understood that the larger the r, the smaller $\frac{1}{2(1+r)}$, so the larger the basin of attraction of playing T with probability 1.

Three players. Again we consider the fitness function, which now equals

$$\begin{aligned}
F(t,s) &= \mathbb{P}(tt \mid t)\left\{t^3\cdot 1 + t\left(1-t^2\right)\cdot -1\right\} + \\
&\quad +\mathbb{P}(st \mid t)\left\{t^2 s\cdot 1 + t(1-st)\cdot -1\right\} + \\
&\quad +\mathbb{P}(ss \mid t)\left\{ts^2\cdot 1 + t\left(1-s^2\right)\cdot -1\right\} \\
&= \mathbb{P}(tt \mid t)\left\{2t^3 - t\right\} + \mathbb{P}(st \mid t)\left\{2t^2 s - t\right\} + \mathbb{P}(ss \mid t)\left\{2ts^2 - t\right\} \\
&= -t + 2t^3\mathbb{P}(tt \mid t) + 2t^2 s\mathbb{P}(st \mid t) + 2ts^2\mathbb{P}(ss \mid t)
\end{aligned}$$

Taking the derivative with respect to t we get

$$\frac{dF(t,s)}{dt} = -1 + 6t^2\mathbb{P}(tt \mid t) + 4ts\mathbb{P}(st \mid t) + 2s^2\mathbb{P}(ss \mid t)$$

and hence

$$\left.\frac{dF(t,s)}{dt}\right|_{t=s} = -1 + 6s^2\mathbb{P}(tt \mid t) + 4s^2\mathbb{P}(st \mid t) + 2s^2\mathbb{P}(ss \mid t)$$

Note first that $\mathbb{P}(t \mid t) = \frac{f_2+3\cdot f_3}{3p}$ and $\mathbb{P}(t \mid s) = \frac{f_1+f_2}{3(1-p)}$. If p goes to 0, then f_1, f_2 and f_3 must go to 0 too, and so must $\mathbb{P}(t \mid s)$. Relatedness at $\lim p \downarrow 0$ is therefore $r = \lim_{p\downarrow 0}\mathbb{P}(t \mid t) - \mathbb{P}(t \mid s) = \lim_{p\downarrow 0}\frac{f_2+3\cdot f_3}{3p} - 0 = \lim_{p\downarrow 0}\frac{f_2+3\cdot f_3}{3p}$. If we realize that $\mathbb{P}(tt \mid t) = \frac{f_3}{p}$, $\mathbb{P}(ts \mid t) = \frac{2}{3}\frac{f_2}{p}$ and $\mathbb{P}(ss \mid t) = \frac{1}{3}\frac{f_1}{p}$, then it is clear that being able to choose these three variables, being restricted by only two equations - that is, $\left.\frac{dF(t,s)}{dt}\right|_{t=s} = 0$ for finding the value s that separates the basins of attractions and $r = \frac{f_2+3\cdot f_3}{3p}$ for relatedness - allows us to shift the point that separates the basins of

attraction, without affecting the r. This implies that in order to determine whether or not $s = 0$ can be invaded, it is typically not be enough to know r.

References

[1] Allee, W. 1951. Cooperation among Animals. Henry Shumann, New York.

[2] Avilés, L. 1999. Cooperation and non-linear dynamics: An ecological perspective on the evolution of sociality. Evol. Ecol. Res. 1, 459-477.

[3] Avilés, L. 2002. Solving the freeloaders paradox: Genetic associations and frequency dependent selection in the evolution of cooperation among relatives. PNAS 99, 14268-14273.

[4] Avilés, L., P. Abbot and A. D. Cutter. 2002. Population ecology, nonlinear dynamics, and social evolution. I. Associations among nonrelatives. Am. Nat. 159, 115-127.

[5] Avilés, L., J. Fletcher and A. D. Cutter. 2004. The kin composition of social groups: Trading group size for degree of altruism. Am. Nat. 164, 132-144

[6] Bowles, S., J-K. Choi and A. Hopfensitz. 2003. The co-evolution of individual behaviors and social institutions, J. Theor. Biol. 223, 135-147.

[7] Day, T., and P. D. Taylor. 1998. Unifying genetic and game theoretic models of kin selection for continuous traits. J. Theor. Biol. 194, 391-407.

[8] Dawkins, R. 1976. The Selfish Gene. Oxford University Press, New York.

[9] Gardner, A., S. A. West and N. H. Barton. 2007. The relation between multilocus population genetics and social evolution theory. Am. Nat. 169, 207-226.

[10] Grafen, A. 1979. The hawk-dove game played between relatives. Anim. Beh. 27, 905-907.

[11] Grafen, A. 1985. A geometric view of relatedness. Oxford Surveys in evolutionary Biology 2, 28-90.

[12] Grafen, A. 2007. Detecting kin selection at work using inclusive fitness. Proc. R. Soc. B 274, 713-719.

[13] Hamilton, W. D., 1964. The genetical theory of social behaviour (I and II), J. Theor. Biol. 7, 1-16, 17-32.

[14] Hamilton, W. D., 1975. Innate social aptitudes of man: Approach from evolutionary genetics. Pages 133-155 *in* Fox, R., ed. Biosocial Anthropology. Wiley, New York.

[15] Hauert, C., F. Michor, M. A. Nowak, and M. Doebeli. 2006. Synergy and discounting of cooperation in social dilemma's. J. Theor. Biol. 239, 195-202.

[16] Hines, W. G. S., and J. Maynard Smith.1979. Games between relatives. J. Theor. Biol. 79, 19-30.

[17] Killingback, T, J. Bieri, and T. Flatt. 2006. Evolution in group-structured populations can resolve the tragedy of the commons. Proc. R. Soc. B 273, 1477-1481.

[18] Lehmann, L., L. Keller, S. West and D. Roze. 2007. Group selection and kin selection: two concepts but one process. PNAS 104, 6736-6739.

[19] Nowak, M. A. Evolutionary Dynamics. Cambridge, MA: Harvard University Press.

[20] Nowak, M. A., and K. Sigmund. 1990. The evolution of stochastic strategies in the prisoner's dilemma. Acta Applicandae Mathematicae 20, 247-265.

[21] Nunney, L., 1985. Group selection, altruism, and structured-deme models. Am. Nat. 126, 212-230

[22] Pepper, J. H. 2000. Relatedness in Trait Group Models of Social Evolution. J. Theor. Biol. 206, 355-368.

[23] Rousseau, J.-J., 1973. A Discourse on the Origin of Inequality. *In*: The Social Contract and Discourses (translated by G. D. H. Cole), pp. 27-l 13. London: Dent

[24] Rousset, F. 2004. Genetic structure and selection in subdivided populations. Princeton University Press, Princeton.

[25] Sober, E., and D. S. Wilson. 1998. Unto Others; the evolution and psychology of unselfish behavior. Harvard University Press, Cambridge, MA.

[26] Traulsen, A., N. Shoresh, and M. A. Nowak. 2008. Analytical results for individual and group selection of any intensity. Bull. Math. Biol. 70, 1410–1424.

[27] Traulsen, A., and M. Nowak. 2006. Evolution of cooperation by multilevel selection. PNAS 103, 10952–10955.

[28] Van Veelen, M., 2005. On the use of the Price equation. J. Theor. Biol. 237, 412-426

[29] Van Veelen, M., and A. Hopfensitz. 2007. In love and war: Altruism, norm formation, and two different types of group selection. J. Theor. Biol. 249, 667-680.

[30] West, S. A., A. S. Griffin and A. Gardner. 2007. Evolutionary explanations for cooperation. Current Biology 17, R661-R672.

[31] West, S. A., A. S. Griffin and A. Gardner. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. 20, 415-432

[32] West, S. A., A. S. Griffin and A. Gardner. 2008. Social semantics: how useful has group selection been? J. Evol. Biol. 21, 374-385.

[33] Wild, G., and A. Traulsen. 2007. The different limits of weak selection and the evolutionary dynamics of finite populations. J. Theor. Biol. 247, 382–390.

[34] Williams, G. C. 1966. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton Univ. Press, Princeton.

[35] Wilson, D. S. 1975. A theory of group selection. PNAS 72, 143-146.

[36] Wilson, D.S. 1979. Structured demes and train-group variation. Am. Nat. 113, 606-610.

[37] Wilson, D.S. 1990. Weak altruism, strong group selection. Oikos 59, 135-140.

[38] Wilson, D. S., and L. A. Dugatkin. 1997. Group selection and assortative interactions. Am. Nat. 149, 336-351.

[39] Wilson, D.S., and E. O. Wilson. 2007. Rethinking the theoretical foundations of socio-biology. Q. Rev. Biol. 82, 327-348.

[40] Wynne-Edwards, V. C. 1962. Animal Dispersion in Relation to Social Behavior. Oliver and Boyd, Edinburgh.