Games among relatives revisited

Benjamin Allen[a,b,c,*], Martin A. Nowak[c,d]

[a] Department of Mathematics, Emmanuel College, Boston, MA, 02115
[b] Center for Mathematical Sciences and Applications, Harvard University, Cambridge, MA, 02138
[c] Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, 02138
[d] Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138

Abstract

We present a simple model for the evolution of social behavior in family-structured, finite sized populations. Interactions are represented as evolutionary games describing frequency-dependent selection. Individuals interact more frequently with siblings than with members of the general population, as quantified by an assortment parameter r, which can be interpreted as "relatedness". Other models, mostly of spatially structured populations, have shown that assortment can promote the evolution of cooperation by facilitating interaction between cooperators, but this effect depends on the details of the evolutionary process. For our model, we find that sibling assortment promotes cooperation in stringent social dilemmas such as the Prisoner's Dilemma, but not necessarily in other situations. These results are obtained through straightforward calculations of changes in gene frequency. We also analyze our model using inclusive fitness. We find that the quantity of inclusive fitness does not exist for general games. For special games, where inclusive fitness exists, it provides less information than the

[*] Corresponding author. Email: benjcallen@gmail.com

straightforward analysis.

Keywords: evolutionary game theory; inclusive fitness; family structure; kin selection; fixation probability

1. Introduction

In many biological populations, family members interact frequently with each other. Family structure is an important form of population structure, which can affect evolution in a variety of ways (Nowak *et al.*, 2010*a*). For example, spatial or group structure in a population can promote the evolution of cooperative behaviors by allowing cooperators to cluster togegther and limit exploitation by noncooperators (Nowak & May, 1992; Durrett & Levin, 1994; van Baalen & Rand, 1998; Ohtsuki *et al.*, 2006; Traulsen & Nowak, 2006; Taylor *et al.*, 2007; Allen *et al.*, 2013; Simon *et al.*, 2013; Allen & Nowak, 2014; Débarre *et al.*, 2014). However, this effect is sensitive to the details of the evolutionary process: for some models, spatial or group structure can have no effect (Taylor, 1992; Wilson *et al.*, 1992; Ohtsuki *et al.*, 2006; Nowak *et al.*, 2010*b*) or even a negative effect (Hauert & Doebeli, 2004) on cooperation.

The evolution of cooperation and other social behaviors can be studied mathematically using evolutionary game theory (Maynard Smith & Price, 1973; Maynard Smith, 1982; Hofbauer & Sigmund, 1988, 1998; Weibull, 1997; Nowak & Sigmund, 2004; Nowak, 2006*a*; Broom & Rychtár, 2013). Social behaviors are represented as strategies, and the fitness consequences of an interaction are quantified as payoffs to each participant. First formulated for large, well-mixed populations (Maynard Smith & Price, 1973), evolutionary game theory has since been extended to populations structured in various ways (Nowak *et al.*, 2010*a*), including by finite population size (Nowak *et al.*, 2004; Taylor *et al.*, 2004; Imhof & Nowak, 2006), by space (Nowak & May, 1992; Durrett & Levin, 1994; Killingback & Doebeli, 1996; Ohtsuki *et al.*, 2006; Korolev & Nelson, 2011; Chen, 2013; Allen & Nowak, 2014; Débarre *et al.*, 2014; Rand *et al.*, 2014), by groups (Traulsen & Nowak, 2006; Simon *et al.*, 2013), and by social sets (Tarnita *et al.*, 2009*a*).

Inclusive fitness theory (Hamilton, 1964; Rousset & Billiard, 2000; Wakano *et al.*, 2013; Lehmann & Rousset, 2014) is another approach to studying the evolution of social behavior. In this approach, each individual's fitness (expected number of viable offspring) is expressed as a sum of portions of fitness due to itself and each other indivdiual. An individual's inclusive fitness is then defined as a weighted sum of fitness portions bestowed on self and others, where the weights represent relatedness to the recipient.

Inclusive fitness theory is regarded by its proponents as a general and powerful framework for understanding the evolution of cooperation. Howevever, Nowak *et al.* (2010*b*), building on earlier critiques by Cavalli-Sforza & Feldman (1978), Uyenoyama & Feldman (1982), and Matessi & Karlin (1984), showed that fitness is not generally equal to a sum of portions due to separate individuals, and thus the quantity of inclusive fitness is only well-defined in special cases. Some proponents of inclusive fitness theory responded (Abbot *et al.*, 2011; Gardner *et al.*, 2011) that such portions of fitness can always be identified using linear regression (Hamilton, 1970; Queller, 1992; Frank, 1998; see also Birch, 2014). Yet Allen *et al.* (2013*b*) showed that this regression method relies on invalid use of statistical inference tools and leads to false conclusions.

A different response was given by Bourke (2011), who acknoweldges that calculating inclusive fitness is a technically limited approach to studying social evolution. Bourke argues nonetheless that the more general and powerful methods used in evolutionary game theory and population genetics are still "inclusive fitness approaches", in that they include the effects of interaction between co-bearers of genes affecting social behavior. We agree that all such effects are accounted for in these mathematically exact methods. However, we find it misleading to refer to these methods as "inclusive fitness approaches", since the re-assignment of fitness effects from recipient to actor—central to the concept of inclusive fitness—is generally impossible and always unnecessary in applying them.

Given the controversy surrounding inclusive fitness theory, it is worth asking how the consequences of family structure might be investigated using the tools of evolutionary game theory. An important step was provided by Grafen (1979), who developed a deterministic, infinite-population model of evolutionary game dynamics with a parameter r (sometimes called "relatedness") quantifying assortment between like types. A fraction r of one's interaction partners guaranteed to be of one's same type, while the remainder are drawn from the population at large. We call this model "r-replicator dynamics", because it generalizes replicator dynamics (Taylor & Jonker, 1978; Hofbauer & Sigmund, 1988, 1998) to include assortment. The r-replicator dynamics and variations thereof have been applied to a wide variety of questions in evolutionary dynamics (Eshel & Cavalli-Sforza, 1982; Bergstrom, 2003; Jansen & Van Baalen, 2006; Taylor & Nowak, 2006; van Veelen *et al.*, 2012; Alger & Weibull, 2013; García *et al.*, 2014).

Here we propose a simple model to investigate how family structure affects the evolution of social behavior in a population of finite size. We consider a Wright-Fisher process in which each adult produces a large number of juveniles. Survival of juveniles is determined by their social interactions, which are represented as a game. A fraction r of a juvenile's interaction partners are siblings, and the rest are drawn from the overall juvenile population. Our model extends Grafen's (1979) r-replicator dynamics to populations of finite size.

We derive exact conditions for a strategy to be favored under weak selection. We first obtain results for arbitrary games, and then restrict attention to a subset of games that describe cooperation and defection in social dilemmas. Interestingly, the effect of sibling assortment on the evolution of cooperation depends on the nature of the social dilemma. For the Prisonser's Dilemma and other stringent social dilemmas, cooperation is increasingly favored with r. But for relaxed social dilemmas, sibling assortment can have a negative or even nonmonotonic effect on cooperation.

These results are obtained using straightforward methods based on the probabilities of gene frequency change. In order to connect our results to the literature on inclusive fitness theory, we also attempt to analyze our model using inclusive fitness methods. We find that inclusive fitness is not a well-defined quantity for a general 2×2 payoff matrix, because the contributions that individuals make to each others' fitness cannot be distinguished in a meaningful way. Remarkably, even the linear regression method that is claimed to be "as general as the genetical theory of natural selection itself" (Abbot *et al.*, 2011) fails for this model, because the costs and benefits turn out to be underdetermined. Inclusive fitness is only well-defined for games that satisfy equal gains from switching (Nowak & Sigmund, 1990), but in this case it provides less information than our straightforward analysis based on gene frequencies.

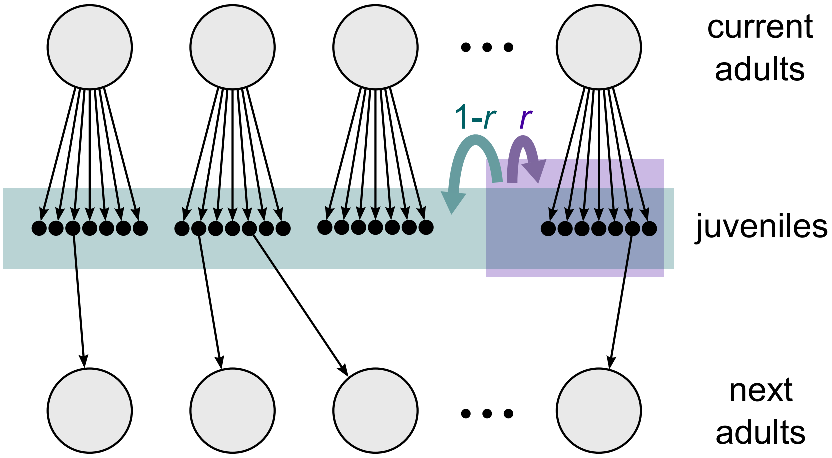


Figure 1: Evolutionary games in a family-structured population. Each generation is founded by N adults. Each of these adults produces a large, equal number of juveniles. The juveniles interact according to a game, with a fraction r of interaction partners chosen from siblings and a fraction $1 - r$ chosen from the population at large. Juveniles survive to adulthood with probability proportional to $1 + w \times$ payoff, where w is a parameter quantifying selection strength.

2. Model

Our model (Figure 1) is a finite-population analogue of the r-replicator dynamics, and can also be described as a Wright-Fisher game process (Imhof & Nowak, 2006) with assortative interactions among siblings. We consider a population of N haploid adults, each having one of two competing genotypes, A and B. Reproduction is asexual. A generation consists of three phases:

1. *Proliferation:* Each adult produces a large number $n \gg 1$ of juveniles, so that each parent contributes a fraction $1/N$ of the juvenile population. Juveniles inherit their parent's genotype.
2. *Interaction:* Each juvenile interacts with a large number of others according to the matrix game
$$\begin{array}{cc} & \begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array} \\ \begin{array}{c} \mathrm{A} \\ \mathrm{B} \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array}. \tag{1}$$
A fraction r of one's interaction partners are drawn from one's siblings, while the remaining fraction $1-r$ are drawn uniformly from the general population (both siblings and nonsiblings). Each juvenile retains the average payoff from these interactions, which we equate to the expected payoff since the number of interactions is large.
3. *Selection:* N juveniles are chosen, independently and with probability proportional to $1 + wf$, where f is the payoff retained from the interaction phase, and $w > 0$ is the strength of selection, i.e. the extent to which game payoff affects reproductive rate. These N juveniles survive to adulthood and form the next generation.

We consider two ways that a new type arises via mutation. First we assume that a new mutation is present in all offspring of a given adult; the mutation arose in the adult (in the germline for multicellular organisms) and was passed to all offspring, but did not affect the phenotype of that adult when it was still a juvenile. In Section 6 we consider the alternative scenario that a mutation first arises in a single juvenile.

3. Analysis

We follow the methods developed by Imhof & Nowak (2006) to study games in a Wright-Fisher process. Since the brood size n is assumed to be large, all of our analysis is performed in the limit $n \to \infty$. For expressions involving n, this limit will be stated explicitly, otherwise it is implied.

Consider a population state with an abitrary number i of adults of type A. The payoffs to A and B juveniles, respectively, can be written as $f_{\mathrm{A}}(i/N)$ and $f_{\mathrm{B}}(i/N)$, where the functions f_{A} and f_{B} are given by

$$\begin{aligned} f_{\mathrm{A}}(x) &= ra + (1-r)[ax + b(1-x)], \\ f_{\mathrm{B}}(x) &= rd + (1-r)[cx + d(1-x)]. \end{aligned} \tag{2}$$

These payoff functions are the same as those used in the r-replicator dynamics (Grafen, 1979), of which our model is a finite-population analogue. We write the rescaled payoffs to A and B as $F_{\mathrm{A}}(i/N)$ and $F_{\mathrm{B}}(i/N)$, respectively, where $F_{\mathrm{A}}(x) = 1 + wf_{\mathrm{A}}(x)$ and $F_{\mathrm{B}}(x) = 1 + wf_{\mathrm{B}}(x)$.

The juveniles that survive to adulthood are drawn proportionally to rescaled payoff. Technically, this drawing occurs without replacement, since a juvenile cannot grow into multiple adults. However, in the limit of many juveniles ($n \to \infty$), this drawing becomes indistinguishable from a drawing with replacement. Thus, in the limit $n \to \infty$, the number of A's that survive to adulthood has binomial distribution $\mathrm{Binom}(N, p_i)$, where p_i is the probability of choosing an A in one such draw:

$$p_i = \frac{iF_A(i/N)}{iF_A(i/N) + (N-i)F_B(i/N)}. \tag{3}$$

To determine fixation probabilities, we let q_i denote the probability that type A becomes fixed when starting with i individuals. Then the q_i satisfy the recurrence relation

$$q_i = \begin{cases} 0 & i = 0 \\ \sum_{j=0}^{N} \binom{N}{j} p_i^j (1-p_i)^{N-j} q_j & 1 \leq i \leq N-1 \\ 1 & i = N, \end{cases} \tag{4}$$

with p_i given by Eq. (3).

In Appendix A, we solve the recurrence relations (4) under weak selection ($w \to 0$), obtaining

$$\begin{aligned} q_i &= \frac{i}{N} \\ &+ w\frac{i(N-i)}{N}\left[r(a-d) + (1-r)(b-d) + (1-r)\frac{N-1+i}{3N-2}(a-b-c+d)\right] \\ &+ \mathcal{O}(w^2). \end{aligned} \tag{5}$$

Now substituting $i = 1$ yields the fixation probability ρ_{A} starting from a single adult of type A:

$$\begin{aligned} \rho_{\mathrm{A}} &= \frac{1}{N} \\ &+ w(N-1)\left[\frac{r(a-d) + (1-r)(b-d)}{N} + \frac{(1-r)(a-b-c+d)}{3N-2}\right] \\ &+ \mathcal{O}(w^2). \end{aligned} \tag{6}$$

The ratio of fixation probabilities $\rho_{\mathrm{A}}/\rho_{\mathrm{B}}$ represents the relative amount of time the population will consist of all A's versus B's, over long periods of time with low mutation (Fudenberg & Imhof, 2006; Allen & Tarnita, 2014). Using Eq. (5) we calculate this ratio as

$$\frac{\rho_{\mathrm{A}}}{\rho_{\mathrm{B}}} = 1 + w(N-1)\left[2r(a-d) + (1-r)(a+b-c-d)\right] + \mathcal{O}(w^2). \tag{7}$$

4. Conditions for success

How does sibling assortment affect the success of game strategies? To answer this question, we must first clarify what it means for a strategy to succeed in evolution. There are two success criteria we might consider.

First, we can ask whether strategy A, when invading strategy B, has a greater chance of success than a neutral mutant. This means comparing ρ_{A} to $1/N$. If $\rho_{\mathrm{A}} > 1/N$ we say that natural selection *favors the replacement* of B by A.

Second, we can ask whether, over many rounds of invasion and fixation, strategy A will have greater time-averaged frequency than strategy B. For this we must compare ρ_{A} to ρ_{B} (Nowak *et al.*, 2004; Fudenberg & Imhof, 2006; Antal *et al.*, 2009; Allen & Tarnita, 2014). We say A is *favored over* B if $\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$.

We say that these conditions hold *under weak selection* if they hold to first order in w according to Eqs. (6) and (7).

4.1. A replacing B ($\rho_{\mathrm{A}} > 1/N$)

From Eq. (6), we see that A is favored to replace B ($\rho_{\mathrm{A}} > 1/N$), under weak selection, if and only if

$$r(3N-2)(a-d) + (1-r)[Na + 2(N-1)b - Nc - 2(N-1)d] > 0. \tag{8}$$

For large populations, Condition (8) becomes

$$3r(a-d) + (1-r)(a+2b-c-2d) > 0. \tag{9}$$

Condition (9) is an instance of the "one-third rule" (Nowak *et al.*, 2004; Ohtsuki *et al.*, 2007) of evolutionary dynamics: $\rho_{\mathrm{A}} > 1/N$ under weak selection, for all sufficiently large N, if and only if $f_{\mathrm{A}} > f_{\mathrm{B}}$ when the frequency of A is one-third. This can be seen by comparing condition (9) to the formulas for payoff in Eq. (2).

4.2. Time-averaged frequency ($\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$)

From Eq. (7), we see that A is favored over B ($\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$), under weak selection, if and only if

$$(1+r)a + (1-r)b > (1-r)c + (1+r)d. \tag{10}$$

Interestingly, this condition is independent of the population size N. It follows that the structure coefficient (Tarnita *et al.*, 2009*b*; Allen *et al.*, 2013*a*) for this process is $\sigma = (1+r)/(1-r)$.

We observe that $\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$ under weak selection if and only if $f_{\mathrm{A}} > f_{\mathrm{B}}$ when the frequency of A is one-half.

5. The evolution of cooperation

We now return to the question of how family structure affects the evolution of cooperation in social dilemmas. To address this question, we must precisely define the terms "cooperation" and "social dilemma". The effects of population structure on cooperation can depend strongly on the nature of the social dilemma (e.g. Hauert & Doebeli, 2004).

5.1. Social dilemmas

A social dilemma involves a choice to cooperate (C), or defect (D). Payoffs can be represented by the following game matrix:

$$\begin{array}{c|cc} & \text{C} & \text{D} \\ \hline \text{C} & R & S \\ \text{D} & T & P \end{array}. \tag{11}$$

Here R represents the "reward" to two cooperators, S the "sucker" payoff to a cooperator exploited by a defector, T the "temptation" to defect against a cooperator, and P the "punishment" to two defectors.

To call C a cooperative trait, it must provide some benefit to its interaction partners. There are three conditions that might represent such a benefit:

C1. $R > P$ (Mutual cooperation benefits both players)

C2. $R > S$ (Cooperation benefits opposing cooperators)

C3. $T > P$ (Cooperation benefits opposing defectors)

Conditions C2 and C3 represent different forms of help to the other player. C1 specifies that this help is effective, in that it leads to a mutually beneficial outcome.

For a game to be a social dilemma, there must be a reason to play the noncooperative strategy D, i.e., a temptation to defect. Three conditions might constitute such a temptation:

D1. $T > R$ (Temptation to defect against cooperators)

D2. $P > S$ (Temptation to defect against defectors)

D3. $T > S$ (Defector advantage in a C vs. D matchup)

We note that all six pairwise comparisons of payoffs are accounted for in C1–C3 and D1–D3.

We define a *social dilemma* to be a game satisfying C1, at least one of C2–C3, and at least one of D1–D3. That is, cooperation provides help to others, such that it is better for all if everyone cooperates, but there is some temptation to defect. Our definition complements previous definitions of social dilemmas (Dawes, 1980; Kerr *et al.*, 2004; Hauert *et al.*, 2006; Nowak, 2012). If all of C1–C3 and D1–D3 hold, then $T > R > P > S$ and the game is a Prisoners' Dilemma. Thus the Prisoners' Dilemma is the most stringent social dilemma. A social dilemma that does not satisfy all of C1–C3 and D1–D3, such as the Hawk-Dove game (Maynard Smith & Price, 1973), is termed a *relaxed social dilemma.*

5.2. Conditions for sibling assortment to support cooperation

We are now prepared to analyze how sibling assortment affects cooperation in social dilemmas. One might suppose that sibling assortment always supports cooperation in the sense that the fixation probability ρ_{C} increases, and ρ_{D} decreases, as r increases. Interestingly, we find that this is not the case for all social dilemmas. Rather, the effect of sibling assortment depends on the values of $R - S$ and $P - T$:

Theorem 1. (a) $\rho_{\mathrm{C}}/\rho_{\mathrm{D}}$ *increases with* r *if and only if* $R - S > P - T$,

(b) ρ_{C} *increases with* r*, for all sufficiently large* N*, if and only if* $2(R - S) > P - T$,

(c) ρ_{D} *decreases with* r*, for all sufficiently large* N*, if and only if* $R - S > 2(P - T)$.

This theorem follows directly from Eqs. (6) and (7). The set of games satisfying the condition of part (a) is illustrated in Figure 2. The conditions

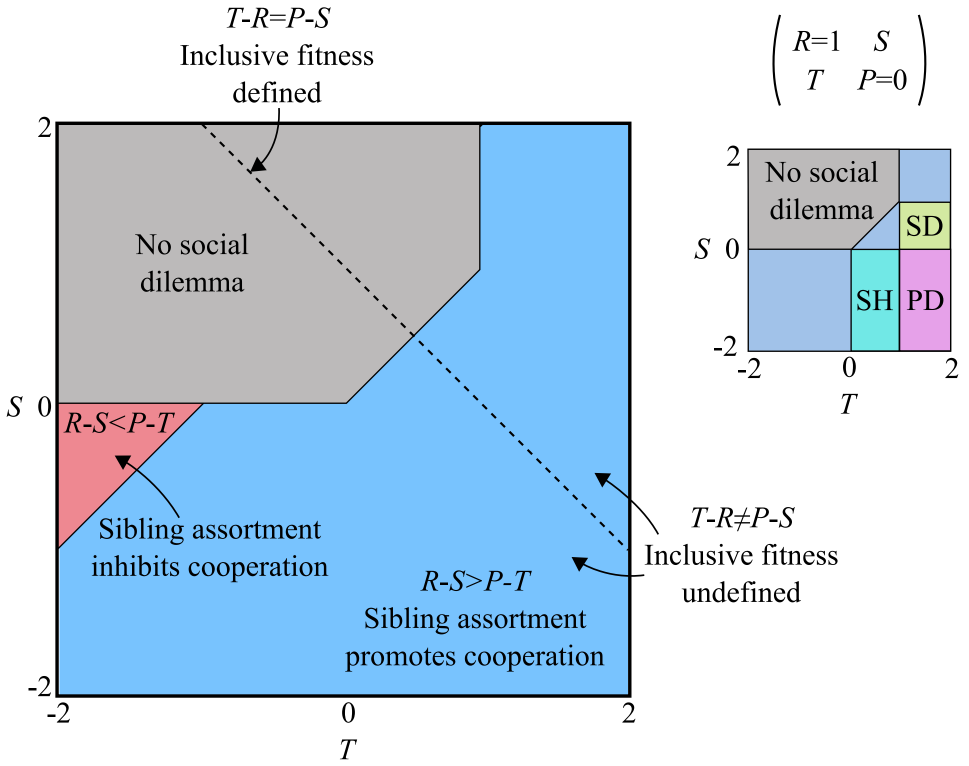


Figure 2: Sibling assortment favors cooperation, in the sense that $\rho_{\mathrm{C}}/\rho_{\mathrm{D}}$ increases with r, if and only if $R - S > P - T$. This condition holds for traditional social dilemmas such as the Prisoners' Dilemma (PD), Snowdrift (SD), and Stag Hunt (SH), but not for all relaxed social dilemmas. This result, illustrated here for the case $R = 1$ and $P = 0$, is proven in Theorem 1 via a straightforward analysis based on gene frequencies. Inclusive fitness, despite being proclaimed as a general theory for the evolution of social behavior, is well-defined only for games that satsify equal gains from switching, $T - R = P - S$ (Theorem 4). Such games are nongeneric: they comprise a subset of measure zero in the space of all possible games.

in Theorem 1 are not satisfied by all relaxed social dilemmas, as we show below. They are, however, satisfied if both C2 and C3 hold, because then $R - S$ is positive and $T - P$ is negative:

Corollary 2. *For any game satisfying C2 and C3, cooperation is increasingly favored with sibling assortment in the sense that*

(a) $\rho_{\mathrm{C}}/\rho_{\mathrm{D}}$ *increases with* r*,*

(b) ρ_{C} *increases with* r *for all sufficiently large* N*,*

(c) ρ_{D} *decreases with* r *for all sufficiently large* N*.*

So, for example, sibling assortment promotes cooperation in Prisoner's Dilemma, Snowdrift, and Hawk-Dove games, all of which satisfy all of C1–C3.

We can also ask what happens if $r = 1$, which means that juveniles interact only with their siblings. In this case, only the payoffs R and P are attained. It therefore follows that any cooperative behavior satisfying C1 is favored:

Theorem 3. *For any game satisfying C1,* $\rho_{\mathrm{C}} > 1/N > \rho_{\mathrm{D}}$ *when* $r = 1$*.*

In other words, for $r = 1$, evolution favors the strategy that gives the largest payoff to the whole population, resolving any social dilemmas. This result also follows directly from Eq. (6).

5.3. Sibling assortment can inhibit cooperation in relaxed social dilemmas

For relaxed social dilemmas, sibling assortment does not necessarily facilitate cooperation. For example, consider the coordination game

$$\begin{array}{c c} & \begin{array}{cc} \mathrm{C} & \mathrm{D} \end{array} \\ \begin{array}{c} \mathrm{C} \\ \mathrm{D} \end{array} & \begin{pmatrix} 7 & 5 \\ 1 & 6 \end{pmatrix} \end{array}. \tag{12}$$

This game is a relaxed social dilemma in that it satisfies C1, C2, and D2. Game (12) can describe the situation of a resource that can be utilized two different ways. A cooperative method (C) requires significant effort and utilizes the entire resource, while a noncooperative method (D) makes only partial use of the resource and requires less effort. The best outcome, 7, is achieved via mutual cooperation. A player who performs C alone gains the entire resource (leaving the other with payoff 1) but at significant expense, leaving a net payoff of only 5. It is better not to waste effort on cooperation if one knows that the other player will not cooperate ($6 > 5$). This game resembles a Stag Hunt game in that mutual cooperation and mutual noncooperation are both Nash equilibria, with mutual cooperation providing the better payoff to both players. However, unlike Stag Hunt, the all-D equilbrium is not risk-dominant in this game.

Cooperation is favored in Game (12), in that $\rho_{\mathrm{C}} > 1/N > \rho_{\mathrm{D}}$ under weak selection for all values of r and all population sizes. However, increasing r has a negative effect on cooperation: ρ_{C} decreases and ρ_{D} increases with r under weak selection, for all values of N, as can be seen from Eq. (6).

The negative effect of sibling assortment on cooperation is most prominent in the states in which cooperation is abundant ($i/N \approx 1$). In this case, cooperators have payoff $f_{\mathrm{C}} \approx 7$, while defector payoff is approximately

$$f_{\mathrm{D}} \approx (1-r) \times 1 + r \times 6 = 1 + 5r.$$

Thus the payoff advantage to cooperators, $f_{\mathrm{C}} - f_{\mathrm{D}} \approx 6 - 5r$, decreases steeply with r (i.e. with slope -5) when cooperation is abundant. On the other hand, when cooperation is rare, the payoff advantage to cooperators is $f_{\mathrm{C}} - f_{\mathrm{D}} \approx -1 + 2r$; this payoff increases with r but less steeply (with slope 2) than the decrease when cooperation is abundant. Combining these factors yields an overall negative effect of sibling assortment on cooperation.

To complement our weak selection results, we also obtained fixation probabilities for this game under moderate ($w = 0.1$) and strong ($w = 1$) selection by numerically solving Eq. (4). We find that the fixation probability of defectors ρ_{D} increases monotonically in r while the fixation probability of cooperators ρ_{C} is maximized at intermediate r for these selection strengths (Fig. 3c). Overall the time-averaged frequency of cooperators, as quantified by $\langle x_{\mathrm{C}} \rangle = \rho_{\mathrm{C}}/(\rho_{\mathrm{C}} + \rho_{\mathrm{D}})$ (Fudenberg & Imhof, 2006; Allen & Tarnita, 2014) decreases, while the corresponding quantity for defectors increases (Fig. 3d).

Interestingly, the negative effect of sibling assortment on cooperation is not apparent from equilibrium analysis of the r-replicator dynamics (Fig. 3b). Indeed, the basin of attraction of cooperation expands with r under these dynamics, until for $r \geq 0.5$ full cooperation becomes the only stable equilibrium.

5.4. Simplified Prisoners' Dilemma

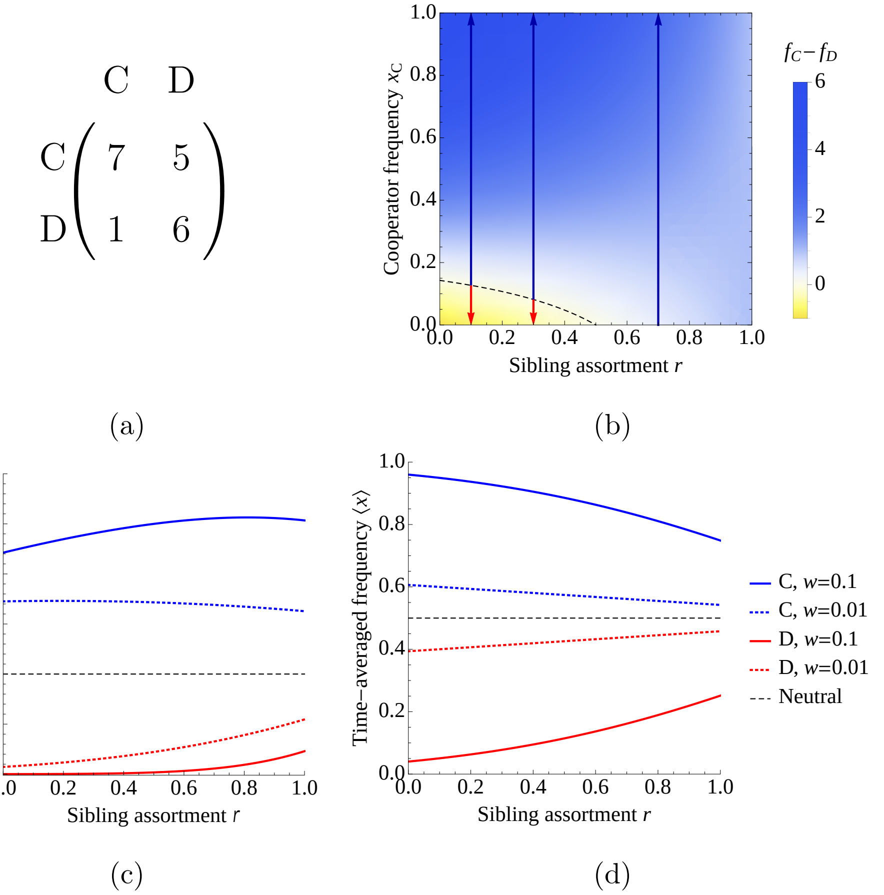
Finally, we consider the simplified Prisoners' Dilemma game (or "donation game"; Nowak & Sigmund, 1998; Nowak, 2006*b*; Sigmund, 2010; Hilbe *et al.*, 2013; Stewart & Plotkin, 2013) in which cooperators pay a cost c to generate a benefit b for the other player:

$$\begin{array}{c c} & \begin{array}{cc} \mathrm{C} & \mathrm{D} \end{array} \\ \begin{array}{c} \mathrm{C} \\ \mathrm{D} \end{array} & \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array}. \tag{13}$$

While only the case $b > c > 0$ describes a Prisoners' Dilemma, our analysis applies to arbitrary b and c.

From Eq. (6) we calculate the fixation probability of cooperators as

$$\rho_{\mathrm{C}} = \frac{1}{N} + w\frac{N-1}{N}(br - c) + \mathcal{O}(w^2). \tag{14}$$



(a) (b)

(c) (d)

Figure 3: Sibling assortment does not necessarily promote cooperation in relaxed social dilemmas. (a) The payoff matrix used in this figure. (b) A bifurcation diagram for the r-replicator dynamics of this game, shaded according to payoff difference $f_{\mathrm{C}} - f_{\mathrm{D}}$. Bistability occurs for $r < 0.5$, with stable equilibria at full defection and full cooperation. For $r \geq 0.5$, full cooperation is the only stable equilibrium. Note that when cooperation is abundant (x_{C} near 1), the rate of increase in cooperation $f_{\mathrm{C}} - f_{\mathrm{D}}$ decreases with r. (c) For moderate ($w = 0.1$) and strong ($w = 1$) selection, cooperator fixation probability ρ_{C} is nonmonotonic (increasing then decreasing) in r, with local maxima at $r \approx 0.18$ for $w = 0.1$ and $r \approx 0.82$ for $w = 1$. The fixation probability ρ_{D} of defectors is increasing in r for both selection strengths. (d) The expected time-averaged frequency of cooperators, calculated as $\langle x_{\mathrm{C}} \rangle = \rho_{\mathrm{C}}/(\rho_{\mathrm{C}} + \rho_D)$ (Allen & Tarnita, 2014) decreases monotonically in r, while $\langle x_{\mathrm{D}} \rangle$ increases, for both $w = 0.1$ and $w = 1$. The adult population size in panels (c) and (d) is $N = 10$.

For this game, the conditions $\rho_{\mathrm{C}} > 1/N$ and $\rho_{\mathrm{C}} > \rho_{\mathrm{D}}$ coincide:

$$\rho_{\mathrm{C}} > \frac{1}{N} > \rho_{\mathrm{D}} \quad \Longleftrightarrow \quad br > c. \tag{15}$$

Thus the condition for success takes the form of Hamilton's (1964) rule in this case. We emphasize, however, that condition (15) applies only to game (13) and only when mutation initially arises in all of an adult's offspring (see Section 6). In other cases, the condition for success does not take the form of Hamilton's rule (Section 4).

6. Fixation from a single juvenile

We now consider the case that a new mutation is initially present in a single juvenile. This situation arises if the occurence of mutation is independent in each offspring.

We return to a general game of the form (1). Consider a single mutant A juvenile in a population otherwise comprised of B's. This mutant has payoff b, while others have payoff d (in the $n \to \infty$ limit). For type A to become fixed, this mutant must first survive to adulthood; we let P_{survive} denote the probability of this event. To obtain a meaningful result for $n \to \infty$, we focus on the product of nP_{survive}. Since N out of Nn juveniles survive into adulthood (with probability proportional to rescaled payoff), we have

$$\begin{aligned} \lim_{n\to\infty} nP_{\text{survive}} &= \lim_{n\to\infty} Nn \frac{1+wb}{(Nn-1)(1+wd)+1+wb} \\ &= 1 + w(b-d) + \mathcal{O}(w^2). \end{aligned} \tag{16}$$

The fixation probability ρ_{A} is the product of the juvenile survival probability P_{survive} and the probability q_1 of fixation from one adult. We focus on the product of ρ_{A} with the total number of juveniles Nn. Combining

Eqs. (6) and (16), we obtain

$$\begin{aligned}\lim_{n\to\infty} Nn\,\rho_{\mathrm{A}} &= (Nq_1)\ \left(\lim_{n\to\infty} nP_{\text{survive}}\right)\\ &= 1 + w\Big[(N-1)r(a-d) + (N-Nr+r)(b-d) \qquad (17)\\ &\qquad + \frac{N(N-1)}{3N-2}(1-r)(a-b-c+d)\Big] + \mathcal{O}(w^2)\end{aligned}$$

The ratio of fixation probabilities (still in the $n \to \infty$ limit) is

$$\frac{\rho_{\mathrm{A}}}{\rho_{\mathrm{B}}} = 1 + w\left[2(N-1)r(a-d) + (N-Nr+r)(a+b-c-d)\right] + \mathcal{O}(w^2). \qquad (18)$$

From Eq. (17), we see that A is favored to replace B, in the sense $\lim_{n\to\infty} Nn\rho_{\mathrm{A}} > 1$, if and only if

$$(N-1)r(a-d) + (N-Nr+r)(b-d) + \frac{N(N-1)}{3N-2}(1-r)(a-b-c+d) > 0.$$

For large adult populations ($N \gg 1$), this condition concides with (9).

A is selected over B in the sense $\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$ if and only if

$$(N+Nr-r)a + (N-Nr+r)b > (N-Nr+r)c + (N+Nr-r)d.$$

Thus the structure coefficient is

$$\sigma = \frac{N+Nr-r}{N-Nr+r}.$$

For large adult populations ($N \gg 1$), this coincides with Condition (10).

For the simplified Prisoners' Dilemma game (13), the condition for the evolution of cooperation changes to

$$\rho_{\mathrm{C}} > \frac{1}{N} > \rho_{\mathrm{D}} \quad \Longleftrightarrow \quad b\frac{N-1}{N}r > c. \qquad (19)$$

We observe that the condition for cooperation is more stringent when starting from a single juvenile (19) than when starting from all offpsring of an adult (15). This is because a single juvenile cooperator must survive an initial generation interacting only with defectors in order for its type to become fixed.

7. Inclusive fitness

Inclusive fitness theory is considered by its proponents to be a powerful approach for analyzing social behavior among relatives. It is therefore worth asking how our model might be analyzed in the inclusive fitness framework.

Inclusive fitness is defined in different ways by different authors. We follow the definition that an individual's inclusive fitness is a sum of the amounts of fitness (expected number of viable offspring) its genotype contributes to each individual, including itself, with each amount weighted according to the relatedness to the recipient. This is the definition proposed by Hamilton (1964), albeit with different relatedness coefficients than are used in modern formulations. It is also the definition used in most contemporary papers on inclusive fitness theory (e.g. Wakano *et al.*, 2013; Lehmann & Rousset, 2014). We note that with this definition, inclusive fitness can only be formulated if each individual's genotype contributes a separate, well-defined amount to each other individual's fitness (Nowak *et al.*, 2010*b*).

Some authors (e.g. Bourke, 2011) instead use "inclusive fitness" to refer to any concept of fitness that can incorporates effects due to the behavior of others. However, these effects are already included in established definitions of fitness (e.g. Cavalli-Sforza & Feldman, 1978; Metz *et al.*, 1992; Heino *et al.*, 1998; Nowak *et al.*, 2010*b*). Thus, with this expanded definition, there is no meaningful distinction between inclusive fitness and fitness itself—which is already as inclusive as it needs to be. For this reason, we use "inclusive fitness” to refer specifically to the weighted sum described above.

7.1. Fitness

As we have seen in Sections 3–6, analysis of our model does not require the calculation of inclusive fitness or even of fitness itself. Instead, we derived our results directly from the probabilities of gene frequency change. However, to determine whether inclusive fitness applies to this model, we must first calculate fitness. Fitness can be calculated either for juveniles or adults; however, to explore whether inclusive fitness applies, we must restrict our calculations to juveniles since they are the interacting agents. We therefore calculate juvenile fitness, defined here as the expected number of juvenile offspring of a given juvenile in a given population state.

We consider an arbitrary population state in which the juveniles are indexed $j = 1, \ldots, Nn$. We let f_j denote the payoff to juvenile j and $\bar{f} = \frac{1}{Nn}\sum_{j=1}^{Nn} f_j$ denote the average payoff. The fitness of juvenile j, denoted W_j is the expected number of juvenile offspring this juvenile will produce. We can calculate this fitness as the probability of being included in next generation of adults, multiplied by the number n of juveniles per adult. This calculation leads to

$$\begin{aligned} W_j &= \frac{Nn(1 + wf_j)}{\sum_{k=1}^{Nn}(1 + wf_k)} \\ &= 1 + w(f_j - \bar{f}) + \mathcal{O}(w^2). \end{aligned} \tag{20}$$

We note that this fitness already includes effects due to others, through the payoffs f_j (which depend on j’s interaction partners) and through the term $-\bar{f}$, which reflects the constraint of constant adult population size.

7.2. Relatedness

In Appendix B we calculate relatedness coefficients using the method of identity-by-descent (Malécot, 1948; Rousset & Billiard, 2000; Wakano *et al.*, 2013; Allen & Nowak, 2014). When mutations occur in the germ line of an adult prior to reprouction, we derive the following relatedness coefficients for self, siblings, and nonsiblings:

$$\begin{aligned} R_{\text{self}} &= R_{\text{sib}} = 1 \\ R_{\text{nonsib}} &= -\frac{1}{N-1}. \end{aligned} \tag{21}$$

The negative value of R_{nonsib} means that nonsiblings are less related on average than a randomly chosen pair. We note that the parameter r does not appear in these relatedness coefficients, because r quantifies the overall degree of assortment between siblings, not the relatedness of any pair of individuals. However, r does represent the average relatedness of a juvenile to its interaction partners:

$$\lim_{n\to\infty} \left[rR_{\text{sib}} + (1-r)\frac{(n-1)R_{\text{sib}} + (N-1)nR_{\text{nonsib}}}{Nn-1} \right] = r. \tag{22}$$

If instead mutations occur independently in each juvenile, we have

$$\begin{aligned} R_{\text{self}} &= 1 \\ R_{\text{sib}} &= \frac{N-1}{N} \\ R_{\text{nonsib}} &= -\frac{1}{N}. \end{aligned} \tag{23}$$

We observe that, in this case, siblings are less than fully related ($R_{\text{sib}} < 1$), since their genotypes may differ due to mutation. In this scenario, the average relatedness of a juvenile to its interaction partners is $r(N-1)/N$:

$$\lim_{n\to\infty} \left[rR_{\text{sib}} + (1-r)\frac{(n-1)R_{\text{sib}} + (N-1)nR_{\text{nonsib}}}{Nn-1} \right] = \frac{N-1}{N}r. \tag{24}$$

7.3. Failure of inclusive fitness theory for general games

To apply inclusive fitness theory, we must be able to distinguish the contribution that each juvenile makes to each other juvenile's fitness. More precisely, we must be able to partition a juvenile's fitness into additive portions, each due to a single juvenile's genotype. Let us examine whether this is possible for a general game of the form (11).

7.3.1. Nonexistence of separate, well-defined fitness contributions

We start by asking how each juvenile's genotype affects the fitness of each other juvenile in a typical population state. Suppose i out of the N adults are cooperators. The fitness of an individual j is then

$$W_j = 1+w\frac{N-i}{N}\left[r(R-P)+(1-r)\left(\frac{i}{N}(R-T)+\frac{N-i}{N}(S-P)\right)\right]+\mathcal{O}(w^2), \tag{25}$$

if j is a cooperator, and

$$W_j = 1-w\frac{i}{N}\left[r(R-P)+(1-r)\left(\frac{i}{N}(R-T)+\frac{N-i}{N}(S-P)\right)\right]+\mathcal{O}(w^2), \tag{26}$$

if j is a defector.

If the fitnesses W_j were equal to sums of separate, well-defined portions due to each juvenile's genotype, then W_j would depend linearly on the frequency of cooperators. Instead, we see that Eqs. (25) and (26) for W_j are nonlinear in the cooperator frequency i/N, indicating that fitness effects due to others combine in a nonadditive way. This is expected: since the game payoffs in (11) do not separate into portions due to self and others, it would be surprising to find such a partition for fitness.

If we cannot distinguish the contribution that each juvenile makes to each other's fitness, can we at least identify the contribution a juvenile's own genotype makes to its fitness? To address this question we choose a focal juvenile, denoted $\bullet$, and consider a hypothetical change in this juvenile's genotype. If the focal juvenile changes strategy from D to C, its resulting change in payoff is

$$\Delta f_{\bullet}^{\mathrm{D}\to\mathrm{C}} = r(S-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right).$$

The change in average payoff, $\Delta \bar{f}$, becomes zero as $n \to \infty$ because the contribution the focal juvenile makes to average payoff vanishes in this limit. Applying Eq. (20), we find that the fitness of the focal juvenile changes by

$$\Delta W_{\bullet}^{\mathrm{D}\to\mathrm{C}} = w\left[r(S-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right)\right] + \mathcal{O}(w^2). \tag{27}$$

On the other hand, if the focal juvenile changes strategy from C to D, a similar analysis shows that its resulting change in fitness is

$$\Delta W_{\bullet}^{\mathrm{C}\to\mathrm{D}} = w\left[r(T-R) + (1-r)\left(\frac{i}{N}(T-R) + \frac{N-i}{N}(P-S)\right)\right] + \mathcal{O}(w^2). \tag{28}$$

Both of the quantities $\Delta W_{\bullet}^{\mathrm{D}\to\mathrm{C}}$ and $-(\Delta W_{\bullet}^{\mathrm{C}\to\mathrm{D}})$ appear to represent the contribution that a juvenile's own genotype makes to its fitness. However, these quantities do not coincide: $\Delta W_{\bullet}^{\mathrm{D}\to\mathrm{C}} \neq -(\Delta W_{\bullet}^{\mathrm{C}\to\mathrm{D}})$. Additionally, both $\Delta W_{\bullet}^{\mathrm{D}\to\mathrm{C}}$ and $-(\Delta W_{\bullet}^{\mathrm{C}\to\mathrm{D}})$ depend on the frequency i/N of cooperators, and thus these fitness changes cannot be attributed solely to the change in focal juvenile's genotype. Therefore, in this model, there is no well-defined portion of a juvenile's fitness that is due solely to its own genotype.

7.3.2. Failure of the regression method

It is often claimed (Hamilton, 1970; Queller, 1992; Frank, 1998; Gardner *et al.*, 2011) that the partitioning of fitness into portions caused by self and others can be accomplished using linear regression. In this method, the fitnesses and genotypes of each indivdual are treated as statistical data (even if these are exact values derived from a mathematical model). One then performs a multivariate least-squares linear regression of fitness based on the genotypes of oneself and one's interaction partners. The resulting regression coefficients are then interpreted as portions of fitness due to self and others. In other words, this method relies on an equation of the form

$$\begin{aligned}&\text{regression coefficient of } j\text{'s genotype for } i\text{'s fitness}\\&\qquad\qquad\qquad = \text{amount of } i\text{'s fitness caused by } j\text{'s genotype.}\end{aligned} \tag{29}$$

While Eq. (29) is not often stated explicitly, it is implied in the language used to refer to these regression coefficients (e.g. "benefit", "cost", "altruism"; Queller, 1992; Gardner *et al.*, 2011; Marshall, 2011). Proponents of this method assert that the goodness of fit is irrelevant (Gardner *et al.*, 2011), and alternative causal explanations are typically not considered.

We emphasize at the outset that this approach is not a valid procedure for statistical inference. There is no result in mathematics or statistics that would support Eq. (29). On the contrary, it is well-known to be false: correlation does not imply causation. In some cases this method may correctly classify the qualitative nature of a behavior, but in other cases it yields demonstrably false conclusions (Allen *et al.*, 2013*b*; see also Birch & Okasha, 2015). Yet this method is standard in inclusive fitness theory (e.g. Frank, 1998; Gardner *et al.*, 2011; Marshall, 2011; Queller, 2011) and has been trumpeted as "the very foundation of social-evolution theory" (Gardner *et al.*, 2011) and "as general as the genetical theory of natural selection itself" (Abbot *et al.*, 2011).

For the sake of connecting our work with the inclusive fitness literature, we attempt the regession method for our model. To do this, we consider a typical population state in which there are i adults of type C and $N-i$ of type D. We assign numerical values to these genotypes: 1 for C and 0 for D. We represent each juvenile j in this state by a quadruplet $(G_j, \bar{G}_{j,\mathrm{sib}}, \bar{G}_{j,\mathrm{nonsib}}, W_j)$ where

- G_j is the numerical genotype (0 or 1) of juvenile j,
- $\bar{G}_{j,\mathrm{sib}}$ is the average numerical genotype among j's siblings,
- $\bar{G}_{j,\mathrm{nonsib}}$ is the average numerical genotype among j's nonsiblings,
- W_j is the fitness of j.

We then treat these values as if they were statistical data and attempt to fit a linear model of the form

$$W_j = W_0 + \beta_{\mathrm{self}}\, G_j + \beta_{\mathrm{sib}}\, \bar{G}_{j,\mathrm{sib}} + \beta_{\mathrm{nonsib}}\, \bar{G}_{j,\mathrm{nonsib}} + \epsilon_j. \qquad (30)$$

The values W_0, β_{self}, β_{sib}, and β_{nonsib} are defined to be those that minimize the quantity $\sum_{j=1}^{Nn} \epsilon_j^2$. According to the regression method, the regression coefficient W_0 is to be interpreted as baseline fitness, while β_{self}, β_{sib}, and β_{nonsib} are to be interpreted as fitness effects due to ones own genotype, one's siblings' genotypes, and one's nonsiblings' genotypes, respectively.

In a typical state of our model, all juveniles of the same type have the same fitness and interact with the same distribution of genotypes. Thus there are only two distinct quadruplets ("data points"). The point representing cooperator juveniles, which we label $(G_{\mathrm{C}}, \bar{G}_{\mathrm{C,sib}}, \bar{G}_{\mathrm{C,nonsib}}, W_{\mathrm{C}})$, has relative multiplicity i/N and values

$$G_{\mathrm{C}} = 1, \qquad \bar{G}_{\mathrm{C,sib}} = 1, \qquad \bar{G}_{\mathrm{C,nonsib}} = \frac{i-1}{N-1},$$

$$W_{\mathrm{C}} = 1 + w\frac{N-i}{N}\left[r(R-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right)\right] + \mathcal{O}(w^2).$$

The point $(G_{\mathrm{D}}, \bar{G}_{\mathrm{D,sib}}, \bar{G}_{\mathrm{D,nonsib}}, W_{\mathrm{D}})$ representing defector juveniles has relative multiplicity $(N-i)/N$ and values

$$G_{\mathrm{D}} = 0, \qquad \bar{G}_{\mathrm{D,sib}} = 0, \qquad \bar{G}_{\mathrm{D,nonsib}} = \frac{i}{N-1},$$
$$W_{\mathrm{D}} = 1 - w\frac{i}{N}\left[r(R-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right)\right] + \mathcal{O}(w^2).$$

To continue, we should determine the values of W_0, β_{self}, β_{sib}, and β_{nonsib} that minimize $\sum_{j=1}^{Nn} \epsilon_j^2$ in Eq. (30). However, we encounter a problem: the values of W_0, β_{self}, β_{sib}, and β_{nonsib} are mathematically underdetermined. There are infinitely many choices for W_0, β_{self}, β_{sib}, and β_{nonsib} that yield $\sum_{j=1}^{Nn} \epsilon_j^2 = 0$. These regression coefficients are therefore undefined. In statistical terminology, this problem arises because we are trying to fit a model with four undetermined parameters to a "dataset" with only two distinct points (one for cooperators and one for defectors). Geometrically, an infinite number of hyperplanes pass through these two points.

In Appendix C we consider an alternative formulation of the regression method, in which both sibling and nonsibling interaction partners are represented by a single regressor $\bar{G}_{j,\mathrm{partner}}$ representing the average genotype of all interaction partners. In this case there are three regression coefficients instead of four: W_0, β_{self}, β_{partner}. However, this still exceeds the number of distinct "data points"; thus this formulation of the regression method fails as well. Intuitively, since each cooperator interacts with the same distribution of types, and the same is true for defectors, linear regression cannot identify the extent to which fitness is associated with one's own genotype versus the genotypes of one's partners.

In summary, neither direct methods nor linear regression are able to distinguish well-defined portions of fitness due to self and others. Thus inclusive fitness is not a well-defined quantity for our simple model describing evolutionary games among relatives.

7.4. Conditions for inclusive fitness to be well-defined

There is, however, a special case for which inclusive fitness analysis is possible. If $T - R = P - S$, then Eqs. (25) and (26) become linear in the frequency of cooperators, and the fitness changes in Eqs. (27) and (28) become opposites. This indicates the possibility of a well-defined partitioning of fitness into effects due to the genotypes of different actors in this case. We state this result precisely as

Theorem 4. *The following conditions are equivalent:*

(i) *The fitness W_j of each juvenile varies linearly with the frequencies of C and D,*

(ii) $\Delta W_{\bullet}^{\mathrm{D}\to\mathrm{C}} = -(\Delta W_{\bullet}^{\mathrm{C}\to\mathrm{D}})$

(iii) $T - R = P - S$.

Theorem 4 provides an exact condition for inclusive fitness to be a well-defined quantity in our model. The condition $T - R = P - S$ is known as "equal gains from swtiching" in the evolutionary game theory literature (Nowak & Sigmund, 1990). Games satisfying this condition are nongeneric: they comprise a set of measure zero in the space of all 2×2 payoff matrices (Figure 2). This set excludes snowdrift games, coordination games, and all other games for which the replicator dynamics have an interior fixed point.

Although games satisfying equal gains from switching are nongeneric among 2×2 games, they play a special role when studying competition between phenotypically similar types. If phenotypes can be numerically parameterized so that (a) C and D are numerically close, and (b) game payoff is a differentiable function of the phenotypes of each player, then equal gains from switching holds to first order in the phenotypic difference (Wild & Traulsen, 2007; Allen *et al.*, 2013*a*). Thus equal gains from switching is an appropriate assumption for cooperative traits that evolve via mutations of small effect (Grafen, 1985; van Cleve, 2014). However, if the competing cooperator and defector phenotypes are significantly different, equal gains from switching cannot be assumed (Wild & Traulsen, 2007).

7.5. The "synergism" approach of Queller (1985)

The necessity of equal gains from switching can also be seen using the approach of Queller (1985). Relative to the case that two defectors meet and receive payoff P, we can define a cost $C = -(S - P)$ to switching to cooperation, and a benefit $B = T - P$ received if the other player switches. We also define a quantity $D = R - S - T + P$, which measures the nonadditivity of the game, i.e. the extent to which it deviates from equal gains from switching. Following Queller's (1985) approach (details in Appendix D), the condition $f_{\mathrm{C}}(i/N) > f_{\mathrm{D}}(i/N)$ for expected increase in cooperators from a given number i, under weak selection, can be rewritten as

$$-C + Br + D\left[r + (1-r)\frac{i}{N}\right] > 0. \tag{31}$$

The left-hand side of Eq. (31) is not an inclusive fitness summation in the usual sense, since the final term is the product of D, which is not the fitness given by any individual to any other, with $r + (1 - r)i/N$, which is not the relatedness between any pair of individuals. Thus, although the procedure of Queller (1985) is sometimes spoken of as an inclusive fitness result, Eq. (31) actually highlights how inclusive fitness—as it is customarily defined—fails to capture the dynamics of our model. The left-hand side of

Eq. (31) only becomes an inclusive fitness summation when $D = 0$, which is exactly the case of equal gains from switching. In this case, Eq. (31) reduces to Hamilton's (1964) rule, $Br > C$.

7.6. Inclusive fitness for the additive Prisoners' Dilemma

Theorem 4 implies that the only payoff matrices for which inclusive fitness exists are those that can be written in the form (13), for some real numbers b and c, plus a constant payoff that is unimportant for selection. We therefore calculate fitness for the payoff matrix (13). Although we will use terminology corresponding to the Prisoners' Dilemma (which occurs in the case $b > c > 0$), our analysis applies to arbitrary b and c.

Substituting the payoffs from game (13) into Eqs. (25) and (26), the fitness of a juvenile j reduces to

$$W_j = \begin{cases} 1 + w\frac{N-i}{N}(rb - c) + \mathcal{O}(w^2) & \text{if } j \text{ is type C} \\ 1 - w\frac{i}{N}(rb - c) + \mathcal{O}(w^2) & \text{if } j \text{ is type D.} \end{cases} \tag{32}$$

To partition these fitnesses into effects due to different actors, let us consider again a focal juvenile, denoted $\bullet$. Suppose this focal juvenile changes strategy from D to C. Note that each juvenile has $Nn - 1$ potential interaction partners in total, of which $n - 1$ are siblings. The resulting change in payoff for the focal juvenile, its siblings, and its nonsiblings, are given respectively by

$$\Delta f_\bullet = -c \tag{33}$$

$$\Delta f_{\text{sib}} = \left(\frac{r}{n-1} + \frac{1-r}{Nn-1}\right) b \tag{34}$$

$$\Delta f_{\text{nonsib}} = \frac{1-r}{Nn-1} b. \tag{35}$$

The average payoff to all juveniles changes by

$$\begin{aligned}\Delta \bar{f} &= \frac{\Delta f_\bullet + (n-1)\Delta f_{\text{sib}} + (N-1)n\Delta f_{\text{nonsib}}}{Nn} \\ &= \frac{b-c}{Nn}\end{aligned} \tag{36}$$

Moving now from payoff to fitness by way of Eq. (20), and taking the limit $n \to \infty$, we find that the total fitnesses of the focal individual, its siblings, and its nonsiblings change by

$$\lim_{n\to\infty} \Delta W_\bullet = -wc + \mathcal{O}(w^2) \tag{37}$$

$$\lim_{n\to\infty} [(n-1)\Delta W_{\text{sib}}] = w\frac{(N-1)rb + c}{N} + \mathcal{O}(w^2) \tag{38}$$

$$\lim_{n\to\infty} [(N-1)n\,\Delta W_{\text{nonsib}}] = w\frac{N-1}{N}(-rb + c) + \mathcal{O}(w^2). \tag{39}$$

If instead the focal juvenile changes strategy from C to D, the changes in fitness are the opposite of those in Eqs. (37)–(39). Thus Eqs. (37)–(39) have a consistent interpretation as portions of fitness due to the genotype of the focal individual, allowing for inclusive fitness to be a well-defined quantity in this case. This result applies only to the simplified Prisoners' Dilemma game (13), and is false for general games of the form (11).

The inclusive fitness effect of the focal juvenile's genotype can be calculated as

$$\Delta W_\bullet^{\text{IF}} = \Delta W_\bullet + R_{\text{sib}}\left[(n-1)\Delta W_{\text{sib}}\right] + R_{\text{nonsib}}\left[(N-1)n\Delta W_{\text{nonsib}}\right]. \tag{40}$$

For mutations arising in all offspring of an adult, substituting from

Eqs. (21) and (37)–(39) we obtain:

$$\begin{aligned}\left.\frac{d\Delta W_{\bullet}^{\mathrm{IF}}}{dw}\right|_{w=0} &= -c \quad \text{(self)}\\ &\quad + \frac{(N-1)rb+c}{N} \quad \text{(siblings)}\\ &\quad + \left(-\frac{1}{N-1}\right)\left[\frac{N-1}{N}\left(-rb+c\right)\right] \quad \text{(nonsiblings)}\\ &= rb-c\end{aligned} \tag{41}$$

We find that the inclusive fitness effect is positive ($W_{\bullet}^{\mathrm{IF}} > 0$ to first order in w) if and only if $br > c$. This result coincides with the condition for cooperation (15) that we found via straightforward analysis of gene frequencies. However, it does not lead to the fixation probability that was calculated in Eq. (14) using direct methods.

For mutations arising in a single juvenile, we instead use the relatedness values from Eq. (23), yielding

$$\begin{aligned}\left.\frac{d\Delta W_{\bullet}^{\mathrm{IF}}}{dw}\right|_{w=0} &= -c \quad \text{(self)}\\ &\quad + \frac{N-1}{N}\left[\frac{(N-1)rb+c}{N}\right] \quad \text{(siblings)}\\ &\quad + \left(-\frac{1}{N}\right)\left[\frac{N-1}{N}\left(-rb+c\right)\right] \quad \text{(nonsiblings)}\\ &= \frac{N-1}{N}rb-c\end{aligned} \tag{42}$$

We find that the sign of the inclusive fitness effect is positive if and only if $br(N-1)/N > c$, recovering Condition (19).

8. Discussion

We have introduced a simple model of evolutionary dynamics in family-structured populations. This model allows for the study of social behavior among juveniles, with interactions occurring more frequently among siblings. Our model generalizes the approach of Grafen (1979) to stochastic dynamics and populations of finite size. The "relatedness" parameter r has a natural interpretation in terms of assortment among siblings.

The main conclusions of our model can be summarized as follows:

- *Increasing relatedness does not always support the evolution of cooperation.* It is often thought that—in the absence of local competition for space or resources—assortment of relatives promotes the evolution of cooperation. In our model, this is true for the Prisoners' Dilemma and other games that satisfy conditions C2 and C3. However, for relaxed social dilemmas such as game (12), sibling assortment can have a negative or even nonmonotonic effect on cooperation (Theorem 1; Fig. 3).

  Our finding differs in an important way from previous results showing that, under some models with spatial or group structure, the benefits of cooperator assortment can be negated by local competition among neighbors or group-mates (Taylor, 1992; Wilson *et al.*, 1992; Hauert & Doebeli, 2004; Ohtsuki *et al.*, 2006; Lion & van Baalen, 2008; Nowak *et al.*, 2010*b*). Such local competition does not occur in our model, since the N members of each adult generation are chosen independently from among all juveniles (i.e. there is global competition to survive to adulthood). In our model, the negative effect of sibling assortment on cooperation in game (12) is due instead to the game itself. This game has the property that when cooperation is abundant, assortment helps defectors more than it helps cooperators.

- *Inclusive fitness theory is not needed to study the evolution of coopera-*

tion among relatives. Hamilton (1964) developed the first mathematical model of cooperation, and used it to elucidate the role that kinship can play in its evolution. While this was an important achievement, it does not constrain future researchers to use only the approach that Hamilton (1964) developed. Like any scientific question, cooperation among relatives can be studied using whatever methods are best suited to the problem at hand.

For our model, we found that the most effective method of analysis does not require the calculation of inclusive fitness or even fitness itself. Instead, we obtain fixation probabilities directly from the probabilities of gene frequency change. In this regard, we follow the recommendation of Grafen (1979), who argues that gene frequencies provide the most natural target for analysis, while "arguments using 'fitness' are much more likely to mislead the unwary". For other models, the concept of fitness has proven to be an important analytical tool (e.g. Metz *et al.*, 1992; Dieckmann & Law, 1996; Antal *et al.*, 2009; Tarnita *et al.*, 2009*a*). However, it is important to note that both gene-frequency-based and fitness-based analyses already include all effects of social interaction, without requiring any partition of fitness into portions due to self and others.

- *Inclusive fitness is well-defined only in special cases.* In order for inclusive fitness to be a well-defined quantity (rather than an overarching concept in the sense of Bourke, 2011), each individual's fitness must be equal to a sum of portions due to each individual's genotype. We found that, while each indivdual has a well-defined fitness in all cases, as calculated in Eqs. (25) and (26), these fitnesses depend nonlinearly on gene frequencies and thus are not equal to sums of portions due to each individual's genotype. Even isolating the contribution of a juvenile's own genotype to its fitness is impossible: if this genotype changes, the resulting change in fitness depends on the direction of this change (C to D versus D to C) as well as on the genotypes of others.

Moreover, we found in Theorem 4 an exact condition for inclusive fitness to be well-defined in our model. This condition, $T-R=P-S$ or equal gains from switching, excludes all games with bistability or coexistence. It also excludes relaxed social dilemmas, such as game (12), for which sibling assortment can have a negative or nonmonotonic effect on cooperation. In short, the set of games for which inclusive fitness is well-defined gives an impoverished view of how kin assortment can affect the evolution of social behavior.

The inadequacy of inclusive fitness summations to capture the dynamics of our model is highlighted by the synergism approach of Queller (1985). Eq. (31) shows that, if the benefit B and cost C are defined straightforwardly in terms of deviations from the all-defector payoff P, the direction of expected change in cooperator frequency is not given by the sign of $-C+Br$. Rather, this quantity must be augmented by an additional "synergy" term that involves neither the relatedness of any pair nor the fitness given by any indivdiual to any other. This synergy term only vanishes in the case of equal gains from switching.

- *Hamilton's rule, when it holds, is not necessarily an inclusive fitness result.* Hamilton's rule is generally understood to be the statement that an altruistic behavior is favored if its fitness benefit b to the recipient, multiplied by relatedness R to the recipient, exceeds the fitness cost c to the actor: $bR > c$. Unfortunately, the logical status of this rule has been obscured in the literature by misleading regression-based definitions of the benefit b and cost c (e.g. Queller, 1992; Gardner *et al.*, 2011), which can misrepresent the actual fitness costs and benefits of a behavior (Fletcher & Doebeli, 2006; Allen *et al.*, 2013*b*; Birch & Okasha, 2015). We are therefore left with an interesting question: for which evolutionary models does the condition for success take the form $bR > c$, with b and c representing the actual fitness benefits and costs of an altruistic behavior?

This question can be investigated using any valid mathematical method. Although inclusive fitness theory and Hamilton's rule are often seen as inextricably linked, inclusive fitness methods are poorly suited to answer this question, since they are less general than methods based on gene frequency or fitness.

In our model, the condition for success takes the form $bR > c$ only in the case of equal gains from switching, in which case the payoff matrix can be written as the simplified Prisoner's Dilemma game (13) plus a constant. Here R represents a juvenile's average relatedness to its interaction partners, which is $R = r$ for a mutation that arises in all offspring of an adult or $R = r(N-1)/N$ for a mutation that arises in a single juvenile; see Eqs. (22) and (24). No calculation of inclusive fitness is needed to derive these results. For other games, the condition for success cannot be written as $bR > c$ because the fitness costs and benefits of cooperation are not well-defined quantities in general.

For other evolutionary models, with social interactions represented by the simplified Prisoner's Dilemma game (13) and with payoff affecting reproductive rate, the condition for cooperation often takes the form $b \times (\text{something}) > c$ (Nowak, 2006*b*). However, Nowak *et al.* (2010*b*) showed that this "something" is not generally relatedness; indeed, it can differ across models that have exactly the same pattern of genetic assortment. While it is possible to define regression coefficients B and C such that the condition takes the form $BR > C$ (with these B and C not representing actual benefits and costs; Allen *et al.*, 2013*b*; Birch & Okasha, 2015), a more interesting question is, why is this "something" is relatedness in some models but not in others? The answer is provided by Nathanson *et al.* (2009): $bR > c$ is the correct condition for success only for models with global updating, in which individuals compete globally for the chance to survive and reproduce. If instead individuals compete only with local neighbors, the condition for success generally takes a different form.

- *Inclusive fitness theory, when it applies, provides less information than an analysis based on gene frequencies.* Using the probilities of gene frequency change, we were able in Eq. (5) to calculate fixation probability, under weak selection, from any starting frequency. This led to exact conditions for a strategy to succeed under two different criteria: $\rho_{\mathrm{A}} > 1/N$ or $\rho_{\mathrm{A}} > \rho_{\mathrm{B}}$. Applying these result to the simplified Prisoner's Dilemma game (13), we found that $br > c$ is equivalent to both $\rho_{\mathrm{C}} > 1/N$ and $\rho_{\mathrm{C}} > \rho_{\mathrm{D}}$ for all population sizes.

  On the other hand, when we calculated inclusive fitness for the game (13) and "all offspring" mutation, it led only to the condition $br > c$. Without our prior analysis, it would not have been clear what this condition means. Does it imply $\rho_{\mathrm{C}} > 1/N$, $\rho_{\mathrm{C}} > \rho_{\mathrm{D}}$, both, or neither?

For some classes of models, it has been proven that $W_\bullet^{\mathrm{IF}} > 0$ if and only if $\rho_{\mathrm{C}} > 1/N > \rho_{\mathrm{D}}$ (Rousset & Billiard, 2000; Taylor *et al.*, 2007; Wakano *et al.*, 2013; Tarnita & Taylor, 2014; van Cleve, 2014). However, these classes do not include the age-structured model we consider here. Moreover, Tarnita & Taylor (2014) have shown that, for populations with heterogeneous spatial structure, $W_\bullet^{\mathrm{IF}} > 0$ is not equivalent to either $\rho_{\mathrm{C}} > 1/N$ or $\rho_{\mathrm{C}} > \rho_{\mathrm{D}}$ and does not provide the correct condition for an allele to be selected. Thus without the analysis based on gene frequencies, it would be unclear what the condition $W_\bullet^{\mathrm{IF}} > 0$ tells us about the evolutionary process. Furthermore, computing $W_\bullet^{\mathrm{IF}} > 0$ does not tell us the actual fixation probabilities, which we were able to derive in Eq. (14) using straightforward methods.

- *Linear regression does not "save" inclusive fitness theory.* It is clear, and has been shown repeatedly (Cavalli-Sforza & Feldman, 1978; Uyenoyama & Feldman, 1982; Matessi & Karlin, 1984; Traulsen, 2010; Nowak *et al.*, 2010*b*; Simon *et al.*, 2013; van Veelen *et al.*, 2014) that individuals do not generally contribute separate, well-defined amounts to each others' fitness. Yet proponents of inclusive fitness (Hamilton, 1970; Queller, 1992; Frank, 1998; Gardner *et al.*, 2011) argue that such portions of fitness given to self and others can always be identified using linear regression.

  Here we have found that the regression method fails for our model, becuse the regression coefficients are mathematically underdetermined. Such a failure would occur in any situation where the number of distinct profiles of the form (genotype, genotypes of partners, fitness) is

> less or equal to the number of distinct classes of interaction partners. Our result disproves all claims (Abbot *et al.*, 2011; Gardner *et al.*, 2011) that inclusive fitness theory is as general as natural selection itself.
>
> Our model is arguably nongeneric in this regard. One might imagine that most natural populations have sufficient variation in fitness and interaction partners for regression coefficients to be well-defined. However, these regression coefficients, when they exist, are not equal to portions of fitness caused by different genotypes, because correlation does not imply causation. Indeed, one can readily find situations for which these regression coefficients have opposite signs from the true fitness effects (Allen *et al.*, 2013*b*). Thus the regression method does not identify amounts of fitness that individuals give to each other, which are needed for inclusive fitness to be a well-defined quantity.

Our work shows how the consquences of kin assortment can be investigated using the tools of evolutionary game theory. This approach can readily be extended to include effects such as sibling recognition, continuous phenotypes, and diploid genetics. Another interesting variation would be to consider games played among adult siblings rather than juveniles. In this case one must deal with the complication that, depending on the details of the model, the number of siblings of a given adult may be random and possibly zero. One way of resolving this difficulty—used in computer simulations by van Veelen *et al.* (2012) and García *et al.* (2014)—is to consider an adult population divided into pairs, where, with probability r, both members of a pair have the same parent, and otherwise their parents are chosen independently.

We have considered a particularly simple model of social interactions occuing between relatives. This is exactly the biological scenario that inclusive fitness theory was formulated to address. One might think that our model would represent a prime target for inclusive fitness analysis. Instead, we find that the terms that comprise inclusive fitness simply do not exist for a generic payoff matrix, because fitness effects due to self and others cannot be additively separated. Since inclusive fitness analysis fails for our minimal model, it should not be expected to succeed for models with greater biological complexity (e.g. diploid and multilocus genetics, nonlinear and multilateral interactions). Indeed, it was shown decades ago (Cavalli-Sforza & Feldman, 1978; Uyenoyama & Feldman, 1982; Matessi & Karlin, 1984) that inclusive fitness theory does not generally apply to models with such complexity.

Nevertheless, it is often claimed (Abbot *et al.*, 2011; Gardner *et al.*, 2011) that inclusive fitness theory has no limitations at all, and applies to every instance of natural selection. There are two bases for such claims. One is the regression method (Hamilton, 1970; Queller, 1992; Frank, 1998; Gardner *et al.*, 2011). While this method produces a condition that appears to be in the form of Hamilton's rule, its terms do not correspond to their verbal descriptions (Allen *et al.*, 2013*b*; Birch & Okasha, 2015). The other is to argue that inclusive fitness remains valid as a concept whether or not it exists as a quantity (Bourke, 2011). In this line of argument, inclusive fitness refers to the general principle that fitness is affected by the actions of others who may share genes affecting social behavior. But this principle is well-understood, both conceptually and quantitatively, in both modern population genetics and evolutionary game theory. These fields have developed precise mathematical tools for investigating such fitness effects; thus there is no need to invoke a concept (inclusive fitness) that has no quantitative instantiation in most cases.

Finally, we caution against overgeneralizing the conclusions of this and other models in which assortment is captured in a single parameter. In many evolutionary models with spatial structure (Nowak & May, 1992; Durrett & Levin, 1994; Ohtsuki *et al.*, 2006; Allen & Nowak, 2014; Débarre *et al.*, 2014), group structure (Traulsen & Nowak, 2006; Simon *et al.*, 2013), or other forms of social structure (Antal *et al.*, 2009; Tarnita *et al.*, 2009*a*; García *et al.*, 2014), assortment of relatives arises naturally from the population structure. It may be tempting to conjecture that these processes are all equivalent in some sense to some variation of the r-replicator dynamics, with r representing the degree of assortment among relatives. However, this is not the case: there are processes leading to the same degree of kin assortment but different outcomes for the evolution of social behavior (Ohtsuki *et al.*, 2006; Taylor *et al.*, 2007; Nowak *et al.*, 2010*b*; Allen & Nowak, 2014; van Veelen *et al.*, 2014). Overall, population structure can have a variety of effects on evolution (Taylor, 1992; Wilson *et al.*, 1992; Hauert & Doebeli, 2004; Lieberman *et al.*, 2005; Nowak, 2006*b*; Nowak *et al.*, 2010*a*; Adlam & Nowak, 2014; Allen *et al.*, 2015), and these effects are best studied using meaningful mathematical methods tailored to the biological question at hand.

Acknowledgements

We are grateful to Carl Veller for helpful comments on the manuscript.

References

Abbot, P., Abe, J., Alcock, J., Alizon, S., Alpedrinha, J. A., Andersson, M., Andre, J.-B., van Baalen, M., Balloux, F., Balshine, S. *et al.* 2011. Inclusive fitness theory and eusociality. Nature, 471 (7339), E1–E4.

Adlam, B. & Nowak, M. A. 2014. Universality of fixation probabilities in randomly structured populations. Scientific Reports, 4.

Alger, I. & Weibull, J. W. 2013. Homo moralis—preference evolution under incomplete information and assortative matching. Econometrica, 81 (6), 2269–2302.

Allen, B., Gore, J. & Nowak, M. A. 2013. Spatial dilemmas of diffusible public goods. eLife, 2.

Allen, B. & Nowak, M. A. 2014. Games on graphs. EMS Surveys in Mathematical Sciences, 1 (1), 113–151.

Allen, B., Nowak, M. A. & Dieckmann, U. 2013*a*. Adaptive dynamics with interaction structure. The American Naturalist, 181 (6), E139–E163.

Allen, B., Nowak, M. A. & Wilson, E. O. 2013*b*. Limitations of inclusive fitness. Proceedings of the National Academy of Sciences, 110 (50), 20135–20139.

Allen, B., Sample, C., Dementieva, Y., Medeiros, R. C., Paoletti, C. & Nowak, M. A. 2015. The molecular clock of neutral evolution can be accelerated or slowed by asymmetric spatial structure. PLoS Computational Biology, 11 (2), e1004108.

Allen, B. & Tarnita, C. E. 2014. Measures of success in a class of evolutionary models with fixed population size and structure. Journal of Mathematical Biology, 68 (1-2), 109–143.

Antal, T., Ohtsuki, H., Wakeley, J., Taylor, P. D. & Nowak, M. A. 2009. Evolution of cooperation by phenotypic similarity. Proceedings of the National Academy of Sciences of the USA, 106 (21), 8597–8600.

Bergstrom, T. C. 2003. The algebra of assortative encounters and the evolution of cooperation. International Game Theory Review, 5 (03), 211–228.

Birch, J. 2014. Hamilton's rule and its discontents. The British Journal for the Philosophy of Science, 65 (2), 381–411.

Birch, J. & Okasha, S. 2015. Kin selection and its critics. BioScience, 65 (1), 22–32.

Bourke, A. F. G. 2011. The validity and value of inclusive fitness theory. Proceedings of the Royal Society B: Biological Sciences, 278 (1723), 3313–3320.

Broom, M. & Rychtár, J. 2013. *Game-theoretical models in biology.* Chapman & Hall/CRC, Boca Raton, FL, USA.

Cavalli-Sforza, L. L. & Feldman, M. W. 1978. Darwinian selection and "altruism". Theoretical Population Biology, 14 (2), 268–280.

Chen, Y.-T. 2013. Sharp benefit-to-cost rules for the evolution of cooperation on regular graphs. The Annals of Applied Probability, 23 (2), 637–664.

Dawes, R. M. 1980. Social dilemmas. Annual Review of Psychology, 31 (1), 169–193.

Débarre, F., Hauert, C. & Doebeli, M. 2014. Social evolution in structured populations. Nature Communications, 5, 4409.

Dieckmann, U. & Law, R. 1996. The dynamical theory of coevolution: a derivation from stochastic ecological processes. Journal of Mathematical Biology, 34 (5), 579–612.

Durrett, R. & Levin, S. 1994. The importance of being discrete (and spatial). Theoretical Population Biology, 46 (3), 363–394.

Eshel, I. & Cavalli-Sforza, L. L. 1982. Assortment of encounters and evolution of cooperativeness. Proceedings of the National Academy of Sciences, 79 (4), 1331–1335.

Fletcher, J. A. & Doebeli, M. 2006. How altruism evolves: assortment and synergy. Journal of Evolutionary Biology, 19 (5), 1389–1393.

Frank, S. A. 1998. *Foundations of Social Evolution.* Princeton University Press.

Fudenberg, D. & Imhof, L. A. 2006. Imitation processes with small mutations. Journal of Economic Theory, 131 (1), 251–262.

García, J., van Veelen, M. & Traulsen, A. 2014. Evil green beards: tag recognition can also be used to withhold cooperation in structured populations. Journal of Theoretical Biology, 360, 181–186.

Gardner, A., West, S. A. & Wild, G. 2011. The genetical theory of kin selection. Journal of Evolutionary Biology, 24 (5), 1020–1043.

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

Grafen, A. 1985. Evolutionary theory: Hamilton's rule OK. Nature, 318, 310–311.

Hamilton, W. D. 1964. The genetical evolution of social behaviour. I. Journal of Theoretical Biology, 7 (1), 1–16.

Hamilton, W. D. 1970. Selfish and spiteful behaviour in an evolutionary model. Nature, 228, 1218–1220.

Hauert, C. & Doebeli, M. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature, 428 (6983), 643–646.

Hauert, C., Michor, F., Nowak, M. A. & Doebeli, M. 2006. Synergy and discounting of cooperation in social dilemmas. Journal of Theoretical Biology, 239 (2), 195–202.

Heino, M., Metz, J. A. J. & Kaitala, V. 1998. The enigma of frequency-dependent selection. Trends in Ecology and Evolution, 13 (9), 367–370.

Hilbe, C., Nowak, M. A. & Sigmund, K. 2013. Evolution of extortion in iterated Prisoner's Dilemma games. Proceedings of the National Academy of Sciences, 110 (17), 6913–6918.

Hofbauer, J. & Sigmund, K. 1988. *The theory of evolution and dynamical systems: Mathematical aspects of selection*. Cambridge University Press Cambridge.

Hofbauer, J. & Sigmund, K. 1998. *Evolutionary Games and Replicator Dynamics*. Cambridge University Press, Cambridge, UK.

Imhof, L. A. & Nowak, M. A. 2006. Evolutionary game dynamics in a wright-fisher process. Journal of Mathematical Biology, 52 (5), 667–681.

Jansen, V. A. & Van Baalen, M. 2006. Altruism through beard chromodynamics. Nature, 440 (7084), 663–666.

Kerr, B., Godfrey-Smith, P. & Feldman, M. W. 2004. What is altruism? Trends in Ecology & Evolution, 19 (3), 135–140.

Killingback, T. & Doebeli, M. 1996. Spatial evolutionary game theory: hawks and doves revisited. Proceedings of the Royal Society B: Biological Sciences, 263 (1374), 1135–1144.

Korolev, K. S. & Nelson, D. R. 2011. Competition and cooperation in one-dimensional stepping-stone models. Physical Review Letters, 107 (8), 088103.

Lehmann, L. & Rousset, F. 2014. The genetical theory of social behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences, 369 (1642), 20130357.

Lieberman, E., Hauert, C. & Nowak, M. 2005. Evolutionary dynamics on graphs. Nature, 433 (7023), 312–316.

Lion, S. & van Baalen, M. 2008. Self-structuring in spatial evolutionary ecology. Ecology Letters, 11 (3), 277–295.

Malécot, G. 1948. *Les Mathématiques de l'Hérédité*. Masson et Cie., Paris.

Marshall, J. A. R. 2011. Group selection and kin selection: formally equivalent approaches. Trends in Ecology & Evolution, 26 (7), 325–332.

Matessi, C. & Karlin, S. 1984. On the evolution of altruism by kin selection. Proceedings of the National Academy of Sciences, 81 (6), 1754–1758.

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

Maynard Smith, J. & Price, G. R. 1973. The logic of animal conflict. Nature, 246 (5427), 15–18.

Metz, J., Nisbet, R. & Geritz, S. 1992. How should we define 'fitness' for general ecological scenarios? Trends in Ecology and Evolution, 7 (6), 198–202.

Nathanson, C. G., Tarnita, C. E. & Nowak, M. A. 2009. Calculating evolutionary dynamics in structured populations. PLoS Computational Biology, 5 (12), e1000615.

Nowak, M. & Sigmund, K. 1990. The evolution of stochastic strategies in the prisoner's dilemma. Acta Applicandae Mathematicae, 20 (3), 247–265.

Nowak, M. A. 2006*a*. *Evolutionary Dynamics.* Harvard University Press, Cambridge, MA, USA.

Nowak, M. A. 2006*b*. Five rules for the evolution of cooperation. Science, 314 (5805), 1560–1563.

Nowak, M. A. 2012. Evolving cooperation. Journal of Theoretical Biology, 299, 1–8.

Nowak, M. A. & May, R. M. 1992. Evolutionary games and spatial chaos. Nature, 359 (6398), 826–829.

Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature, 428 (6983), 646–650.

Nowak, M. A. & Sigmund, K. 1998. Evolution of indirect reciprocity by image scoring. Nature, 393 (6685), 573–577.

Nowak, M. A. & Sigmund, K. 2004. Evolutionary dynamics of biological games. Science, 303 (5659), 793–799.

Nowak, M. A., Tarnita, C. E. & Antal, T. 2010*a*. Evolutionary dynamics in structured populations. Philosophical Transactions of the Royal Society B: Biological Sciences, 365 (1537), 19.

Nowak, M. A., Tarnita, C. E. & Wilson, E. O. 2010*b*. The evolution of eusociality. Nature, 466 (7310), 1057–1062.

Ohtsuki, H., Bordalo, P. & Nowak, M. A. 2007. The one-third law of evolutionary dynamics. Journal of Theoretical Biology, 249 (2), 289–295.

Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. 2006. A simple rule for the evolution of cooperation on graphs and social networks. Nature, 441, 502–505.

Price, G. R. 1970. Selection and covariance. Nature, 227, 520–521.

Queller, D. C. 1985. Kinship, reciprocity and synergism in the evolution of social behaviour. Nature, 318 (6044), 366–367.

Queller, D. C. 1992. A general model for kin selection. Evolution, pp. 376–380.

Queller, D. C. 2011. Expanded social fitness and Hamilton's rule for kin, kith, and kind. Proceedings of the National Academy of Sciences, 108 (Supplement 2), 10792–10799.

Rand, D. G., Nowak, M. A., Fowler, J. H. & Christakis, N. A. 2014. Static network structure can stabilize human cooperation. Proceedings of the National Academy of Sciences, 111 (48), 17093–17098.

Rousset, F. & Billiard, S. 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. Journal of Evolutionary Biology, 13 (5), 814–825.

Sigmund, K. 2010. *The calculus of selfishness.* Princeton University Press.

Simon, B., Fletcher, J. A. & Doebeli, M. 2013. Towards a general theory of group selection. Evolution, 67 (6), 1561–1572.

Stewart, A. J. & Plotkin, J. B. 2013. From extortion to generosity, evolution in the iterated prisoner's dilemma. Proceedings of the National Academy of Sciences, 110 (38), 15348–15353.

Tarnita, C. E., Antal, T., Ohtsuki, H. & Nowak, M. A. 2009*a*. Evolutionary dynamics in set structured populations. Proceedings of the National Academy of Sciences of the USA, 106 (21), 8601–8604.

Tarnita, C. E., Ohtsuki, H., Antal, T., Fu, F. & Nowak, M. A. 2009*b*. Strategy selection in structured populations. Journal of Theoretical Biology, 259 (3), 570 – 581.

Tarnita, C. E. & Taylor, P. D. 2014. Measures of relative fitness of social behaviors in finite structured population models. The American Naturalist, 184 (4), 477–488.

Taylor, C., Fudenberg, D., Sasaki, A. & Nowak, M. 2004. Evolutionary game dynamics in finite populations. Bulletin of Mathematical Biology, 66, 1621–1644. 10.1016/j.bulm.2004.03.004.

Taylor, C. & Nowak, M. A. 2006. Evolutionary game dynamics with non-uniform interaction rates. Theoretical population biology, 69 (3), 243–252.

Taylor, P., Day, T. & Wild, G. 2007. From inclusive fitness to fixation probability in homogeneous structured populations. Journal of Theoretical Biology, 249 (1), 101–110.

Taylor, P. D. 1992. Altruism in viscous populations—an inclusive fitness model. Evolutionary Ecology, 6 (4), 352–356.

Taylor, P. D., Day, T. & Wild, G. 2007. Evolution of cooperation in a finite homogeneous graph. Nature, 447 (7143), 469–472.

Taylor, P. D. & Jonker, L. B. 1978. Evolutionary stable strategies and game dynamics. Mathematical Biosciences, 40 (1-2), 145–156.

Traulsen, A. 2010. Mathematics of kin- and group-selection: formally equivalent? Evolution, 64 (2), 316–323.

Traulsen, A. & Nowak, M. A. 2006. Evolution of cooperation by multilevel selection. Proceedings of the National Academy of Sciences of the USA, 103 (29), 10952–10955.

Uyenoyama, M. K. & Feldman, M. 1982. Population genetic theory of kin selection. ii. the multiplicative model. American Naturalist, 120 (5), 614–627.

van Baalen, M. & Rand, D. A. 1998. The unit of selection in viscous populations and the evolution of altruism. Journal of Theoretical Biology, 193 (4), 631–648.

van Cleve, J. 2014. Social evolution and genetic interactions in the short and long term. bioRxiv, p. 010371.

van Veelen, M. 2005. On the use of the Price equation. Journal of Theoretical Biology, 237 (4), 412–426.

van Veelen, M., García, J., Rand, D. G. & Nowak, M. A. 2012. Direct reciprocity in structured populations. Proceedings of the National Academy of Sciences, 109 (25), 9929–9934.

van Veelen, M., Luo, S. & Simon, B. 2014. A simple model of group selection that cannot be analyzed with inclusive fitness. Journal of Theoretical Biology, 360, 279–289.

Wakano, J. Y., Ohtsuki, H. & Kobayashi, Y. 2013. A mathematical description of the inclusive fitness theory. Theoretical Population Biology, 84, 46–55.

Weibull, J. W. 1997. *Evolutionary game theory.* MIT press, Cambridge, MA, USA.

Wild, G. & Traulsen, A. 2007. The different limits of weak selection and the evolutionary dynamics of finite populations. Journal of Theoretical Biology, 247 (2), 382–390.

Wilson, D. S., Pollock, G. B. & Dugatkin, L. A. 1992. Can altruism evolve in purely viscous populations? Evolutionary Ecology, 6 (4), 331–341.

Appendix A. Calculation of fixation probability

Here we compute fixation probabilities from the recurrence relation (4). We first note that Eq. (3) admits the following weak-selection expansion for

p_i:

$$p_i = \frac{i}{N} + w\frac{i(N-i)}{N^2}\left[f_A(i/N) - f_B(i/N)\right] + \mathcal{O}(w^2). \tag{A.1}$$

For neutral evolution ($w = 0$), $q_i = i/N$. Thus to analyze weak selection, we write

$$q_i = \frac{i}{N} + wq_i^{(1)} + \mathcal{O}(w^2). \tag{A.2}$$

Substituting Eq. (A.2) into Eq. (4) and using the properties of the binomial distribution, we obtain, for $1 \leq i \leq N-1$,

$$\begin{aligned}\frac{i}{N} + wq_i^{(1)} &= \sum_{j=0}^{N}\binom{N}{j}p_i^j(1-p_i)^{N-j}\left(\frac{j}{N} + wq_j^{(1)}\right) + \mathcal{O}(w^2)\\ &= \frac{1}{N}\sum_{j=0}^{N}\binom{N}{j}p_i^j(1-p_i)^{N-j}j + w\sum_{j=0}^{N}\binom{N}{j}p_i^j(1-p_i)^{N-j}q_j^{(1)} + \mathcal{O}(w^2)\\ &= p_i + w\sum_{j=0}^{N}\binom{N}{j}p_i^j(1-p_i)^{N-j}q_j^{(1)} + \mathcal{O}(w^2).\end{aligned}$$

Now substituting expansion (A.1) for p_i yields

$$\begin{aligned}\frac{i}{N} + wq_i^{(1)} = \frac{i}{N}&\\ + w\left(\frac{i(N-i)}{N^2}\left[f_\mathrm{A}\left(\frac{i}{N}\right) - f_\mathrm{B}\left(\frac{i}{N}\right)\right] + \sum_{j=0}^{N}\binom{N}{j}\frac{i^j(N-i)^{N-j}}{N^N}q_j^{(1)}\right)&\\ + \mathcal{O}(w^2).&\end{aligned}$$

We conclude that the $q_i^{(1)}$ satisfy the recurrence relation

$$q_i^{(1)} = \begin{cases}\frac{i(N-i)}{N^2}\left[f_\mathrm{A}\left(\frac{i}{N}\right) - f_\mathrm{B}\left(\frac{i}{N}\right)\right] + \sum_{j=0}^{N}\binom{N}{j}\frac{i^j(N-i)^{N-j}}{N^N}q_j^{(1)} & 1 \leq i \leq N-1\\ 0 & i = 0, N.\end{cases} \tag{A.3}$$

These recurrence relations can be solved by letting $q_i^{(1)}$ be an arbitrary cubic polynomial in i and solving for the coefficients, yielding the solution in Eq. (5).

Appendix B. Calculation of relatedness

To calculate relatedness coefficients in our model, we consider a related process of neutral drift (see, for example, Rousset & Billiard, 2000; Taylor *et al.*, 2007; Wakano *et al.*, 2013). We introduce a small rate of mutation and consider the stationary probability distribution over population states (Antal *et al.*, 2009; Allen & Tarnita, 2014). A pair of juveniles is identical by descent (IBD) if no mutation separates either of them from their common ancestor. The stationary probability that two distinct sibling juveniles are IBD is denoted q_{sib}, while the probability that two nonsibling juveniles are IBD is denoted q_{nonsib}. We let $\bar{q}$ denote the average IBD probability among all pairs of juveniles. Relatedness is obtained as an expression of the form

$$R = \frac{q - \bar{q}}{1 - \bar{q}}, \tag{B.1}$$

with the appropriate IBD probability substituted for q.

Appendix B.1. Starting from all offspring of a single adult

We begin with the convention that a new mutation initially appears in all offspring of a single adult. In the corresponding neutral drift process, each time an adult reproduces, there is a probabiilty u that all of its offspring are mutant, otherwise they all inheret the genotype of the parent.

It is clear every individual is always IBD to itself; thus $q_{\mathrm{self}} = 1$. Furthermore, since siblings always have identical genotype under this mutation model, we have $q_{\mathrm{sib}} = 1$.

Let us now consider a pair of distinct adults. Since we are considering neutral drift, we can suppose that each adult randomly chooses a parent (with uniform probability) from among the previous generation of adults. With probability $1/N$ they choose the same parent, in which case they are siblings and are guaranteed to be IBD. Otherwise, they choose distinct parents and their IBD probability equals the probablity that their parents are IBD times the probability $(1-u)^2$ that neither is a mutant. This leads to the following equation for the probability q_{adult} that two distinct adults are IBD:

$$q_{\text{adult}} = \frac{1}{N} + \frac{N-1}{N} q_{\text{adult}} (1-u)^2$$

Solving,

$$\begin{aligned} q_{\text{adult}} &= \frac{1}{N-(N-1)(1-u)^2} \\ &= 1 - 2(N-1)u + \mathcal{O}(u^2) \end{aligned}$$

The probability q_{nonsib} that two nonsiblings are IBD equals the probability q_{adult} that their parents were IBD times the probability $(1-u)^2$ that neither was born with a mutation:

$$\begin{aligned} q_{\text{nonsib}} &= (1-u)^2 q_{\text{adult}} \\ &= 1 - 2Nu + \mathcal{O}(u^2). \end{aligned}$$

The average IBD probability $\bar{q}$ among all pairs of individuals is calculated as

$$\begin{aligned} \bar{q} =& \frac{1 + (n-1)1 + (N-1)n q_{\text{nonsib}}}{Nn} \\ =& \frac{1}{N} + \frac{N-1}{N} q_{\text{nonsib}} \\ =& 1 - 2(N-1)u + \mathcal{O}(u^2) \end{aligned}$$

We now calculate relatedness using the definition (B.1), yielding $R_{\text{self}} = R_{\text{sib}} = 1$ and

$$\begin{aligned} R_{\text{nonsib}} &= \lim_{u \to 0} \frac{q_{\text{nonsib}} - \bar{q}}{1 - \bar{q}} \\ &= \frac{-1}{N-1}. \end{aligned}$$

Appendix B.2. Starting from a single juvenile

We now turn to the convention that mutation initially appears in a single juvenile. In the corresponding neutral drift process, each offspring has probability u of being born with a mutation, and these mutation events are independent across juveniles. Thus siblings are not necessarily IBD; instead, the probability that two siblings are IBD can be calculated as

$$q_{\text{sib}} = (1-u)^2.$$

The probability that two (distinct) adults are IBD satisfies the recurrence

$$\begin{aligned} q_{\text{adult}} &= \frac{1}{N} q_{\text{sib}} + \frac{N-1}{N} q_{\text{adult}} (1-u)^2 \\ &= (1-u)^2 \frac{1 + (N-1) q_{\text{adult}}}{N}. \end{aligned}$$

Solving for q_{adult}, we find

$$\begin{aligned} q_{\text{adult}} &= \frac{(1-u)^2}{N - (N-1)(1-u)^2} \\ &= 1 - 2Nu + \mathcal{O}(u^2). \end{aligned}$$

The probability that two nonsiblings are IBD can now be calculated as

$$q_{\text{nonsib}} = (1-u)^2 q_{\text{adult}} = 1 - 2(N+1)u + \mathcal{O}(u^2).$$

The average IBD probability $\bar{q}$ is

$$\bar{q} = \frac{1 + (n-1) q_{\text{sib}} + (N-1) n q_{\text{nonsib}}}{Nn}.$$

In the $n \to \infty$ limit this becomes

$$\bar{q} = 1 - 2Nu + \mathcal{O}(u^2).$$

We can now calculate relatedness using definition (B.1), yielding the values in Eq. (23).

Appendix C. Regression with one category of interaction partner

The regression method implemented in Section 7.3.2 used one regressor for the genotype of one's siblings and another for the average genotype of nonsiblings. Here we consider an alternative formulation in which the genotypes of both sibling and nonsibling interaction partners are combined into single regressor $\bar{G}_{j,\text{partner}}$ representing the average genotype of all interaction partners. In this case, each juvenile j is represented by a triplet $(G_j, \bar{G}_{j,\text{partner}}, W_j)$, and we write

$$W_j = W_0 + \beta_{\text{self}}\, G_j + \beta_{\text{partner}}\, \bar{G}_{j,\text{partner}} + \epsilon_j. \tag{C.1}$$

The values W_0, β_{self}, and β_{partner}, are defined to be those that minimize $\sum_{j=1}^{Nn} \epsilon_j^2$. However, as we saw in Section 7.3.2, since all juveniles of the same type have the same fitness and interact with the same distribution of genotypes, there are only two distinct "data points": one for cooperators (with relative multiplicity i/N) and one for defectors (with relative multiplicity $(N-i)/N$). Each cooperator is represented by the triplet $(G_\text{C}, \bar{G}_{\text{C,partner}}, W_\text{C})$ with

$$\begin{aligned}
&G_\text{C} = 1, \qquad \bar{G}_{\text{C,partner}} = r + (1-r)\frac{i}{N},\\
&W_\text{C} = 1 + w\frac{N-i}{N}\left[r(R-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right)\right] + \mathcal{O}(w^2),
\end{aligned} \tag{C.2}$$

while each defector is represented by the triplet $(G_\text{D}, \bar{G}_{\text{D,partner}}, W_\text{D})$ with

$$\begin{aligned}
&G_\text{D} = 0, \qquad \bar{G}_{\text{D,partner}} = (1-r)\frac{i}{N},\\
&W_\text{D} = 1 - w\frac{i}{N}\left[r(R-P) + (1-r)\left(\frac{i}{N}(R-T) + \frac{N-i}{N}(S-P)\right)\right] + \mathcal{O}(w^2).
\end{aligned} \tag{C.3}$$

Since there are three regression coefficients in Eq. (C.1) but only two distinct "data points", the values of the regression coeffecients are again underdetermined. There are infinitely many choices for W_0, β_{self}, and β_{partner} that yield $\sum_{j=1}^{Nn} \epsilon_j^2 = 0$. Thus this version of the regression method fails as well.

An equivalent formulation of Eq. (C.1), used by Gardner *et al.* (2011), is to write

$$W_j = \bar{W} + \beta_{\text{self}}(G_j - \bar{G}) + \beta_{\text{partner}}(\bar{G}_{j,\text{partner}} - \bar{G}) + \epsilon_j, \tag{C.4}$$

where $\bar{W} = 1$ is the average population fitness and $\bar{G} = i/N$ is the average G-value. At first glance, Eq. (C.4) appears to avoid the problem of underdetermination by eliminating the variable W_0. One might then hope to obtain β_{self} and β_{partner} by setting all $\epsilon_j = 0$ and solving the resulting system of equations:

$$\begin{aligned} W_{\text{C}} &= \bar{W} + \beta_{\text{self}}(G_{\text{C}} - \bar{G}) + \beta_{\text{partner}}(\bar{G}_{\text{C},\text{partner}} - \bar{G}) \\ W_{\text{D}} &= \bar{W} + \beta_{\text{self}}(G_{\text{D}} - \bar{G}) + \beta_{\text{partner}}(\bar{G}_{\text{D},\text{partner}} - \bar{G}). \end{aligned} \tag{C.5}$$

However, upon substituting the values from Eqs. (C.2) and (C.3), it turns out that system (C.5) is singular and thus β_{self} and β_{partner} remain underdetermined.

Appendix D. The "synergism" approach of Queller (1985)

Another way of conceptualizing non-additive fitness effects was developed by Queller (1985). To implement this approach, we consider a particular state in which i out of the N adults are cooperators. We define a baseline fitness W_0, a cost C, a benefit B, and a non-additive effect D by

$$W_0 = 1 + wP - w\bar{f} \tag{D.1}$$

$$B = w(T - P) \tag{D.2}$$

$$C = w(P - S) \tag{D.3}$$

$$D = w(R - S - T + P). \tag{D.4}$$

With these definitions, the fitness W_j of each indivdual j can be written (using the notation of Appendix C) as

$$W_j = W_0 - CG_j + B\bar{G}_{j,\text{partner}} + DG_j\bar{G}_{j,\text{partner}} + \mathcal{O}(w^2). \tag{D.5}$$

The frequency of cooperators is expected to increase if $\text{Cov}[W, G] > 0$, where Cov denotes population covariance (Price, 1970; but see van Veelen, 2005). Substituting the right-hand side of Eq. (D.5) for W_j and dividing through by $\text{Cov}[G, G]$, one obtains that the cooperator frequency is expected to increase under weak selection if and only if

$$-C + B\frac{\text{Cov}[G, \bar{G}_{\text{partner}}]}{\text{Cov}[G, G]} + D\frac{\text{Cov}[G, G\bar{G}_{\text{partner}}]}{\text{Cov}[G, G]} > 0.$$

Computing these covariances using the values in Eqs. (C.2)–(C.3), we obtain the result in Eq. (31) of the main text.

For simplicity, the presentation of this result in the main text omitted the factor of w from B, C, and D. This change does not affect the correctness of Eq. (31), since the left-hand side can be multiplied or divided by w without changing the sign of the inequality.