bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathways to social evolution: reciprocity, relatedness, and synergy

Jeremy Van Cleve[1] Erol Akçay[2]

Key words: cooperation, non-additive fitness, behavioral responses, genetic assortment, class structure

[1] National Evolutionary Synthesis Center (NESCent)
2024 W. Main Street Suite A200
Durham, NC 27705 USA
e-mail: vancleve@nescent.org

[2] Department of Biology
University of Pennsylvania
433 S. University Avenue
Philadelphia PA, 19104 USA
e-mail: eakcay@sas.upenn.edu

Date modified: 13th November, 2013

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Many organisms live in populations structured by space and by class, exhibit plastic responses to their social partners, and are subject to non-additive ecological and fitness effects. Social evolution theory has long recognized that all of these factors can lead to different selection pressures but has only recently has attempted to synthesize how these factors interact. Using models for both discrete and continuous phenotypes, we show that analyzing these factors in a consistent framework reveals that they interact with one another in ways not previously emphasized. Specifically, behavioral responses (reciprocity), genetic assortment (relatedness), and synergy interact in non-trivial ways that cannot be easily captured by simple summary indices. We demonstrate the importance of these interactions by showing how they have been neglected in previous synthetic models of social behavior both within and between species. These interactions also affect the level of behavioral responses that can evolve in the long run; proximate biological mechanisms are evolutionarily stable when they generate enough responsiveness relative to the level of responsiveness that exactly balances the ecological costs and benefits. Given the richness of social behavior across taxa, these interactions should be a boon for empirical research as they are likely crucial for describing the complex relationship between ecology, demography, and social behavior.

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

Explaining the evolution of social behaviors has been a goal of evolutionary theory going back to Darwin's time. The modern effort to explain the origin of social behaviors both cooperative and noncooperative started with the seminal work of Hamilton (1964a,b), which showed that genetic relatedness has a profound influence on evolutionary dynamics. Later work added other significant factors, starting with conditional and responsive behaviors (Trivers, 1971; Axelrod and Hamilton, 1981) and extending to non-additive fitness interactions (Queller, 1985). Beginning in the 1980's, researchers analyzed many models that apply and extend these mechanisms (e.g., extending reciprocity to include indirect responses, Nowak and Sigmund, 1998). At the same time, empirical researchers tested the role of various mechanisms developed in the theoretical literature (e.g., Queller and Strassmann, 1998; Hughes et al., 2008; Schino et al., 2009).

Whereas early discussions of social evolution usually pitted one mechanism against another, recent efforts are increasingly focused on synthesizing and integrating the various pathways of social evolution (e.g., Sachs et al., 2004; Lehmann and Keller, 2006a; Nowak, 2006; West et al., 2007; Fletcher and Doebeli, 2009; Queller, 2011). This is perhaps the surest sign that social evolution theory has entered a mature phase. These efforts at synthesis are not always without controversy, but they have served to illuminate the connections between various models devised over the decades. In this paper, we aim to contribute to this synthetic effort by focusing specifically on the relation between three important concepts: assortment of genotypes (relatedness), behavioral responsiveness (e.g., reciprocity), and non-additive (e.g., synergistic) effects of social behaviors on fitness. These topics have of course been written about in the past; yet, how these concepts relate to one another is still not well understood.

Two of the most fundamental mechanisms for the evolution of cooperation are responsive (and conditional) behaviors such as direct and indirect reciprocity, and genetic assortment that creates relatedness between individuals (Lehmann and Keller, 2006a). Both factors are clearly in operation in many, if not most social interactions. Cooperative breeding, for example, almost always involves interactions with relatives of varying degree. At the same time, most cooperatively breeding species also exhibit conditional behavior; for example, non-helpers might get evicted from the breeding group (Balshine-Earn et al., 1998). Likewise, reciprocal food-sharing among vampire bats can occur amongst relatives as well as non-relatives (Wilkinson, 1984).

Non-additive interactions between individual phenotypes also play a prominent role in social evolution.

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sometimes, the non-additivity is a direct function of organismal metabolism, as in the case of the plant pathogen *Agrobacterium tumefaciens*, which induces infected hosts to produce opines that can be used as a resource only by those cells that also carry the virulence plasmid (White et al., 2007). In other cases, the ecological context might create non-additive payoffs. For example, in the penduline tits (*Remiz pendulous*), one or both of the parents desert the nest after eggs are laid. Data on breeding success suggests that the reason for this behavior is in part that a single parent has better than even chance of raising the clutch (Szentirmai et al., 2007). Hence, the second parent adds less to breeding success than the first parent, meaning a negative non-additivity. It has been argued that the opposite is true for many bird species with biparental care (Wesołowski, 1994). It is worth noting that in both of these examples, as perhaps with most instances of non-additive interactions between phenotypes, the traits in question are plastic. For example, the *Agrobacterium* pathways inducing opine production are not expressed constitutively and instead depend on quorum-sensing (White et al., 2007). Likewise, the male and female decisions to desert in penduline tits appear context-dependent; the same individual can care or desert according to factors such as brood size (Persson and Öhrström, 1989; Valera et al., 1997).

Yet, despite enormous theoretical work on conditional behaviors, genetic assortment, and non-additive payoffs, the interplay between these three factors has attracted relatively little attention until recently (Queller, 1985; Lehmann and Keller, 2006a; Fletcher and Zwick, 2006; Foster and Wenseleers, 2006; Fletcher and Doebeli, 2009; McGlothlin et al., 2010; Queller, 2011; Akçay and Van Cleve, 2012). Much of this recent work neglects to systematically study the interactions between these factors (with the notable exception of Lehmann and Keller, 2006a; McGlothlin et al., 2010; Akçay and Van Cleve, 2012), which we show can lead to an incomplete or incorrect view of how selection affects social behaviors. Moreover, properly accounting for these interactions allows us to make novel insights into when conditional behaviors should be evolutionarily stable and to empirically assess the level of responsiveness across demographically diverse species.

Our paper begins with two models: first, we use a discrete-action model to clarify the relationship between genetic assortment (e.g., relatedness), behavioral responses (a generic notion of reciprocity), and synergistic or non-additive interactions; second, we generalize the results of the discrete-action model by presenting a population genetic model for the evolution of a continuous trait in a structured population, where the structure can be due to any combination of age, stage, sex, or spatial location. Using these models, we show how previous analyses that do not systematically track the effect of behavioral responses (Fletcher and Doebeli, 2009; Queller, 2011) have neglected a crucial interaction between responsiveness and relatedness

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(McGlothlin et al., 2010; Akçay and Van Cleve, 2012). Rather than combining additively in the equation for the effect of selection on a social behavior (Queller, 2011), the equation has a product of relatedness and responsiveness that implies each of these assortation mechanisms contributes distinct effects (contra Fletcher and Doebeli, 2009). We also show this crucial interaction to be missing in models of interspecific mutualisms (Foster and Wenseleers, 2006), which leads to difficulties in interpreting mechanisms that stabilize mutualisms.

More generally, our discrete model allows us to show how synergistic interactions can affect the evolutionary stability of social behaviors and even change the payoff structure of the game resulting in a different game entirely. This is important because helping is more likely to evolve in Hawk-Dove (Maynard Smith and Price, 1973) or Stag-Hunt (Skyrms, 2001) games than in the Prisoner's Dilemma game. Using our continuous-trait model, we show that proximate physiological or psychological mechanisms evolve in the long run only when they produce "enough" responsiveness relative to the responsiveness level that exactly balances the demographic costs and benefits of the behavior. This evolutionarily stable level of responsiveness is a function of both the absolute level of relatedness and the interaction between relatedness and responsiveness. Finally, we highlight in the discussion how recent work in quantitative genetics that measures relatedness and responsiveness requires a proper accounting of the interaction of these two mechanisms.

2. Discrete action model

Our model beings with a 2x2 game that is played repeatedly between two individuals a large number of times. We take the game to be symmetric for ease of exposition, but this assumption can be relaxed with no consequence for our main argument. Label the two actions $\boldsymbol{C}$ and $\boldsymbol{D}$. Most models of such repeated games focus on particular conditional strategies (e.g., tit-for-tat (Axelrod and Hamilton, 1981) and its numerous relatives). We would like a behavioral model that admits a wider range of propensities to respond to one's partner. Hence, we assume that each individual i has two traits: a_i, which is the action ($\boldsymbol{C}$ or $\boldsymbol{D}$) that the individual i intrinsically "prefers", and ρ_i, the probability that individual i overrides its intrinsic preference and reciprocates its opponent's action in the last round. Hence, in common parlance, a_i corresponds to whether individuals are cooperators or defectors, and ρ_i denotes the degree to which individuals are reciprocators. Both traits are genetically encoded; we take a_i to be binary and ρ_i to be a continuous trait between 0 and 1. As an aside, a straightforward extension of this model could treat a_i as a continuous trait by interpreting it as the intrinsic probability of playing $\boldsymbol{C}$ vs. $\boldsymbol{D}$.

Consider an interaction between an intrinsic cooperator and defector (i.e., $a_1 = \boldsymbol{C}$ and $a_2 = \boldsymbol{D}$). Suppose

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at stage t, player 1 (the C-type) played C and player 2 (the D-type) also played C (because he was reciprocating cooperation in the previous round). We label this situation as (C, C), where the elements correspond to the first and second players' actions, respectively. Let us label the response coefficients of the two individuals as ρ_c and ρ_d in according with intrinsic preferences for C and D, respectively. In that case, the probability that at stage $t+1$ they will keep playing (C, C) is the probability that the D-type individual keeps reciprocating, i.e., ρ_d. Similarly, the probability that the $t+1$ stage will result in (C, D) is $1-\rho_d$, as this is the probability that the D-type will revert to his intrinsic preference. Finally, the probabilities of playing (D, C) and (D, D) at $t+1$ given the actions at t is zero, since regardless of whether she plays her intrinsic preference or reciprocates, the C-type player will always play C. In a similar way, we can work out the probabilities of playing any action pair at $t+1$ conditional on the action pair at t, which is summarized with the transition matrix $\mathbf{M} = [m_{ij}]$,

$$\mathbf{M} = \begin{pmatrix} \rho_d & (1-\rho_d) & 0 & 0 \\ (1-\rho_c)\rho_d & (1-\rho_c)(1-\rho_d) & \rho_c\rho_d & \rho_c(1-\rho_d) \\ 0 & 1 & 0 & 0 \\ 0 & 1-\rho_c & 0 & \rho_c \end{pmatrix}, \tag{1}$$

where m_{ij} gives the probability of being in state j at $t+1$ when the game was in state i at t, and states are ordered as $\{(C, C), (C, D), (D, C), (D, D)\}$. We assume that the game is played for a very large number of periods with no discounting, so that the proportion of time each pair of actions is played is given by the stationary distribution of the Markov chain $\mathbf{M}$. Taking the dominant left-eigenvector of $\mathbf{M}$ and normalizing it, we find the stationary distribution of the game between a C- and D-type individual, $\boldsymbol{\pi}$, to be

$$\boldsymbol{\pi} = \left((1-\rho_c)^2\rho_d\,,\,(1-\rho_c)(1-\rho_d)\,,\,(1-\rho_c)\rho_c(1-\rho_d)\rho_d\,,\,\rho_c(1-\rho_d)^2\right)(1-\rho_c\rho_d)^{-2}\,. \tag{2}$$

The expected payoff of a C-type player from playing D-type player is the fraction of time spent at each outcome (given by the elements of π) multiplied by the payoff to her from that outcome. Suppose that at each period the game has the familiar structure of the Prisoner's Dilemma (in the other-only benefit form (Pepper, 2000)): cooperation yields a benefit of b to the partner, at a net cost of c to the cooperator. Defection carries no cost and yields no benefit. Thus, the expected payoff to a C-type player when playing a D-type, w_{CD}, is

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

given by

$$w_{CD} = \frac{(1-\rho_c)(b\rho_d - c)}{1-\rho_c\rho_d} \,. \tag{3}$$

With a similar calculation, we find the expected payoff to D-type from playing a C-type to be

$$w_{DC} = \frac{(1-\rho_c)(b - c\rho_d)}{1-\rho_c\rho_d} \,. \tag{4}$$

The expected payoff to a C-type from playing another C-type is trivial; since both players start with (C, C) and keep playing these actions, $w_{CC} = b - c$. Similar reasoning yields $w_{DD} = 0$.

In order to determine which type of individual is preferred by natural selection, we calculate the condition when the frequency of intrinsic cooperators with responsiveness ρ_c increases in a population with intrinsic defectors of responsiveness ρ_d. Suppose that some mechanism (e.g., population viscosity, kin-recognition, or greenbeards) causes cooperators to assort themselves so that an intrinsic cooperator interacts with another intrinsic cooperator with probability r_c; likewise, intrinsic defectors assort among themselves with probability r_d. When not assorting specifically with their own types, individuals are paired with a partner randomly chosen from the population (this is analogous to the classic assumption concerning populations with related individuals; see equation 1 in Rousset and Billiard, 2000). A full description of the dynamic of frequency change in the population would require a specification of the mechanism that generates assortment and how it depends on the frequency of genotypes. However, we can leave this unspecified if we only require a qualitative picture of frequency change over the course of a single generation.

Given our assumptions above and that the frequency of the intrinsic cooperative genotype is q, the expected fitnesses of an intrinsic cooperator and an intrinsic defector are, respectively,

$$\begin{aligned} W_C &= (r_c + (1-r_c)q)\, w_{CC} + (1-r_c)(1-q)\, w_{CD} \quad \text{and} \\ W_C &= (1-r_d)q\, w_{DC} + (r_d + (1-r_d)(1-q)\, w_{DD} \,. \end{aligned} \tag{5}$$

The condition for increase of intrinsic cooperators is $W_C - W_D > 0$, which leads to

$$\frac{b}{c} > \frac{1-(1-q)\rho_c - q\rho_d + (1-q)r_c\rho_c\left(1-\rho_d\right) + qr_d\left(1-\rho_c\right)\rho_d}{q\rho_c + (1-q)\rho_d + (1-q)r_c\left(1-\rho_d\right) + qr_d\left(1-\rho_c\right) - \rho_c\rho_d} \,. \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

What this equation entails can be seen better by looking at three special cases. First, suppose individuals do not respond to each other, $\rho_c = \rho_d = 0$, and that the assortment probabilities are equal, $r_c = r_d = r$. In this case, condition (6) becomes $\frac{b}{c} > \frac{1}{r}$, which of course is Hamilton's rule (Hamilton, 1964a). Second, suppose both the intrinsic cooperators and defectors have the same responsiveness, $\rho_c = \rho_d = \rho$, and assortment is zero, $r_c = r_d = 0$. Then, condition (6) evaluates to $\frac{b}{c} > \frac{1}{\rho}$, which is equivalent to other rules for the evolution of responsiveness (Akçay et al., 2009) and reciprocity (Lehmann and Keller, 2006a; André and Day, 2007). Finally, giving intrinsic cooperators and defectors the same assortment probability and responsiveness, $r_c = r_d = r$ and $\rho_c = \rho_d = \rho$, yields

$$\frac{b}{c} > \frac{1 + r\rho}{r + \rho}\,, \tag{7}$$

which is analogous to equation (4) in Lehmann and Keller (2006a). In condition (7), the assortment coefficient r and responsiveness ρ have symmetric effects on the increase condition but are also separate because both the product and the sum of ρ and r appear. Hence, assortment of individuals and conditional behaviors cannot be collapsed to a single index, as Fletcher and Doebeli (2009) have suggested. Not coincidentally, expression (7) is identical in structure to the condition for an evolutionary increase in cooperation we have previously derived for continuous public goods games (Akçay and Van Cleve, 2012) and to other previously derived results (Lehmann and Keller, 2006a; McGlothlin et al., 2010). This further reinforces the point that this basic symmetry and irreducibility is not due to a quirk in this particular model.

3. Population genetic model

While the discrete-action model above highlights the interplay between genetic assortment, and behavioral responsiveness, it is a restrictive model. A more general population genetic framework will allow us to generalize the above results to interactions among an arbitrary number of individuals who may be from different demographic classes (e.g., males and females, queens and workers, juveniles and adults, philopatric and dispersed). A population genetic model will also allow us insight into how the long-term evolutionary stability of social behaviors depends on the proximate production of ecological cost and benefits and the proximate production of behavioral responses through physiological and psychological mechanisms. At the same time, such generality involves a significant increase in the complexity of the model, and hence a different set of simplifying assumptions come into play.

In particular, we are interested in determining when a mutant allele that codes for some phenotype with

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

social effects will, on average, increase in frequency in a population and reach fixation. The proper quantity to calculate in this case is the fixation probability of the mutant allele (Rousset, 2003). For structured populations, the fixation probability may be impossible to obtain analytically, so we follow a standard approach that approximates that probability using a first-order a Taylor-series expansion in δ, which is the difference between the mutant and resident phenotypes (Rousset and Billiard, 2000; Rousset and Ronce, 2004). Population genetics theory shows that the first-order term in the expansion is sufficient to calculate the standard conditions for the evolutionary stability of phenotypes (Rousset and Billiard, 2000; Rousset and Ronce, 2004). Moreover, this first-order term is proportional to a quantity that is much easier to calculate called the "selection gradient", which is the derivative of the expected change in the frequency of the mutant allele with respect to δ or $\mathrm{d}\Delta q/\mathrm{d}\delta$. More intuitively, we can think of $\Delta q \approx (\mathrm{d}\Delta q/\mathrm{d}\delta)\,\delta$ when δ is small. This means that the sign of the selection gradient $\mathrm{d}\Delta q/\mathrm{d}\delta$ tells us when the mutant allele is expected to increase or decrease in frequency or, in terms of fixation probability, when the mutant allele is more or less likely than a neutral allele to fix in the population.

To obtain the selection gradient, $\mathrm{d}\Delta q/\mathrm{d}\delta$, we first write a general expression for Δq in a class structured model using standard theory in population genetics and evolutionary demography (Taylor, 1990; Charlesworth, 1994; Caswell, 2001; Rousset and Ronce, 2004). We do this using an "individually" centered approached that was popularized in population genetics by the Price equation (Price, 1972, 1970). Suppose that there are K classes in the population, where an individual k's class could represent the local population it lives in, its age or developmental stage, its sex, or some other phenotypic feature. Each class contains N_k individuals and the total population size is $N = \sum_k N_k$. In order to track allele frequencies in different classes, we let $\mathbf{q} = (\ldots, q_{jk}, \ldots)$ be a column vector of length N where q_{jk} is the frequency of the allele in individual k in class j. Average allele frequencies in each class in the next time period are calculated by multiplying a transition matrix $\mathbf{W} = [w_{ijk}]$ times $\mathbf{q}$, where w_{ijk} is the probability that an allele in a random individual in class i descended from individual k in class j. The transition matrix is a function of the fitnesses of each class and the transmission processes, such as segregation, mutation, and migration, that affect how alleles move between classes. We assume that the fitness of individual k in class j is function of its phenotype, p_{jk}. In the simplest cases, phenotype depends on genotype in a linear way where $p_{jk} = p + q_{jk}\delta$ and δ is a measure of how much the mutant differs from the resident phenotype p. For non-additive genetic interactions (such as dominance or epistasis), phenotype depends on genotype in a nonlinear way (Gardner et al., 2007).

Calculating Δq requires us to average the class frequencies in such a way that the average frequency

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remains constant in the absence of natural selection. The appropriate weights for this average are the reproductive values of each class (Fisher, 1930; Leslie, 1948; Taylor, 1990), $\boldsymbol{\alpha} = (\ldots, \alpha_k, \ldots)$. Thus, the change in the reproductive value weighted allele frequency is

$$\Delta q = \boldsymbol{\alpha} \cdot \mathbf{W} \cdot \mathbf{q} - \boldsymbol{\alpha} \cdot \mathbf{N} \cdot \mathbf{q} \tag{8}$$

where $\mathbf{N}$ is a $K \times N$ matrix used to create class averages of the allele frequencies $\mathbf{q}$ and whose k-th row has the value $1/N_k$ in the columns $1 + \sum_{l=1}^{k-1} N_l$ to $\sum_{l=1}^{k} N_l$ and zero elsewhere for all k. Using equation (8), the derivative of Δq with respect to δ is given by

$$\frac{\mathrm{d}\Delta q}{\mathrm{d}\delta} = \boldsymbol{\alpha} \cdot \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\delta} \cdot \mathbf{q} = \sum_{i,j,l=1}^{K} \sum_{k=1}^{N_j} \sum_{m=1}^{N_l} \alpha_i \frac{\partial w_{ijk}}{\partial p_{lm}} q_{jk} q_{lm} \tag{9}$$

The components of equation (9) are reproductive value, α_i, the effect of one individual's phenotype on another individual's probability of transmitting an allele from one generation to the next, $\frac{\partial w_{ijk}}{\partial p_{lm}}$, and the probability that those two individuals both have the mutant allele, $q_{jk} q_{lm}$. When this last component, $q_{jk} q_{lm}$, is averaged over all individuals in a class, we obtain a genetic identity probability that can be related to identity by descent probabilities from classic population genetics (Rousset and Billiard, 2000; Rousset and Ronce, 2004; Rousset, 2004) that emerge as more common quantities such as F_{ST} (Wright, 1949) in simple models of population subdivision (Rousset, 2004) such as the island model (Wright, 1931). Moreover, these genetic identity probabilities correspond to the genetic covariance terms found in regression models in social evolution (Frank, 1997), which when normalized by genetic variance correspond to the regression definition of relatedness (Hamilton, 1972; Queller, 1992a).

Although equation (9) allows us to see how evolutionary change should generally rely on genetic relatedness, understanding the interplay between ecological and genetic mechanisms requires unpacking the derivatives of the transition matrix, $\frac{\partial w_{ijk}}{\partial p_{lm}}$. Evaluating these derivatives can be very involved if individuals behave differently in the social interaction as a function of class since we must track the fertility contributions of each class separately. Thus, we will simplify the following model by assuming a relatively simple social interaction that is unaffected by the class structure (though class structure will still affect genetic relatedness). Assume that each individual in each class participates in a social interaction with $n-1$ other individuals from the same class (e.g., members of a patch interact together). Each individual chooses actions in this social

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interaction (e.g., an amount of effort in some helping behavior); we denote the behavioral equilibrium (Akçay et al., 2009; Akçay and Van Cleve, 2012) or time averaged values of the action of individual k in class j by a^*_{jk}. One crucial quantity of interest is the marginal change in the action of one individual in response to another; this is the response coefficient ρ from section 2, which can be written as $\rho_{j,kl} = \frac{\partial a^*_{jl}}{\partial a^*_{jk}}$ for the change in the action of individual l in response to a change in the action of k, both of class j (Akçay et al., 2009; Akçay and Van Cleve, 2012). Since we assume that the social interaction is unaffected by class, we can neglect write the response coefficient as $\rho_{j,kl} = \rho_{kl}$.

The actions individuals choose in a social interaction are determined by some proximate behavioral mechanism whose biological components (e.g., neurological structures) are genetically determined (whereas actions themselves are not inherited). For simplicity, we assume that a single phenotype p characterizes the proximate mechanism, and it is the evolution of this phenotype that we track with equation (8). As an example, the phenotype could be a measure of hormone function in some brain region associated with cooperative behavior (see Soares et al., 2010, for a recent review). One way of measuring how the phenotype affects actions is $\frac{\partial a^*_l}{\partial p_k}$, which is the effect of individual k's phenotype on the action of l. The response coefficient can be expressed as a function of these derivatives: $\rho_{kl} = \frac{\partial a^*_l}{\partial a^*_k} = \frac{\partial a^*_l}{\partial p_k} \Big/ \frac{\partial a^*_k}{\partial p_k}$.

The outcome of the social interaction determines the fertility of each individual in the interaction (which is $f(p_{jk})$ for individual k in class j since fertilities are independent of class); for example, an individual's fertility might be determined by the net calories it obtains from participating in group hunting and foraging activities. We can define the "costs" and "benefits" of the social interaction in terms of the affect of actions on fertility where $-c = \frac{\partial f}{\partial a^*_\bullet}$ and "$\bullet$" represents the focal individual and $b = \frac{\partial f}{\partial a^*_\circ}$ for some other individual "$\circ$". Given these assumptions, we show in Appendix B that the selection gradient $\mathrm{d}\Delta q/\mathrm{d}\delta$ is proportional to

$$\mathcal{S} = \underbrace{b\rho(n-1) - c}_{\text{“direct” effect}} + r\ \underbrace{(n-1)\left[b\left(1+\rho(n-2)\right) - \rho c\right]}_{\text{“indirect” effect}}, \tag{10}$$

where ρ is the response coefficient in a resident population monomorphic for the phenotype p and b and c are evaluated at the resident phenotypic value p. This result was previously derived by Akçay and Van Cleve (2012) and is analogous to expressions found in McGlothlin et al. (2010). The first term in (10) is the "direct" effect on the fertility of a focal individual of its own phenotype and actions, whereas the second term is the "indirect" effect on the focal individual of the actions and phenotypes of its social partners. It is important to note that equation (10) is essentially a rearrangement of Hamilton's classic rule (Hamilton, 1964a) where

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the effect of local competition due to kin is moved from the benefit and cost terms into the *scaled* relatedness-coefficient r (Queller, 1994; Rousset and Ronce, 2004; Rousset, 2004; Lehmann and Rousset, 2010). We derive an expression for r in Appendix B.

Assuming that the phenotype p corresponds to investment in a cooperative behavior in a Prisoner's Dilemma type of interaction, the condition for an increase in the level of cooperation derived from equation (10) is $\mathcal{S} > 0$ or

$$\frac{b}{c} > \frac{1 + r\rho(n-1)}{(n-1)(r + \rho + (n-2)r\rho)} \,. \tag{11}$$

Condition (11) is the n-player analog of (7). Even for arbitrarily large interaction groups (n large), the symmetry between relatedness r and responsiveness ρ is maintained. This reinforces the notion that relatedness and responsiveness are distinct causal factors that interact synergistically and cannot be both replaced by a single index of "assortment" (Fletcher and Doebeli, 2009). Moreover, social interactions with more than two individuals ($n > 2$) introduce an additional product of relatedness and responsiveness; this product is scaled by $rb(n-1)(n-2)$ and effectively accounts for the benefit a focal individual obtains when its relatives affect the behavior of a third party that interacts with the focal.

4. Applications

The discrete action and population genetics models developed above show that an interaction between responsiveness and relatedness emerges naturally when considering the effect of selection a plastic social behavior. Below, we illustrate the importance of the interaction between responsiveness and relatedness in four ways: first, we show how neglecting the interaction leads to incorrect predictions of the level of cooperation through the b/c ratio and that past efforts to synthesize mechanisms for cooperation within species and for mutualisms between species have neglected this interaction; second, we show how the interaction between responsiveness and relatedness can be important in interspecific interactions, such as mutualisms; third, we extend the discrete action model to include non-additive or synergistic payoffs and demonstrate how responsiveness, relatedness, and synergy interact; and fourth, we use the population genetic model to derive a stability condition that reveals how proximate mechanisms evolve in the long run only when they generate enough responsiveness relative to the break even point determined by relatedness and the ecological costs and benefits of the behavior (i.e. $\mathcal{S} = 0$).

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.1. The interaction between reciprocity and relatedness

The interaction between responsiveness (i.e., reciprocity) and relatedness in equation (10) naturally breaks into two components: $-c(n-1)r\rho$ represents the cost to a focal individual of increasing its own level of helping in response to the intrinsic increase in helping of its related group mates; and $b(n-1)(n-2)r\rho$ represents the benefit the focal receives from third parties that increase their helping in response to the intrinsic increase in helping of the focal's related group mates. Both of these components are due essentially to the ability of individuals to respond to the actions of the focal's related group mates. Neglecting these responses generates

$$\frac{b}{c} > \frac{1}{r+\rho} \tag{12}$$

as the condition for an increase in cooperation. Condition (12), which lacks the products of relatedness and responsiveness present in condition (11), effectively assumes that related group mates of the focal individual do not respond plastically to other individuals whereas unrelated individuals can respond plastically. This assumption is very restrictive, which suggests that condition (12) is incorrect in most cases.

To visualize how neglecting the interaction between responsiveness and relatedness affects predicted levels of cooperation, we plot in Figure 1 the log of the ratio of the correct b/c expression in (11) to incorrect expression in (12) as a function of responsiveness. For pairwise interactions ($n = 2$), the log ratio is always greater than zero regardless of the value of relatedness, which implies that neglecting the interaction between relatedness and responsiveness leads to an overestimate of the level of cooperation. This is the case because the neglected interaction is only the cost of the focal of responding to its relatives, $-c(n-1)r\rho$. For larger interaction groups ($n > 2$), the interaction effect of third parties comes into play, $b(n-1)(n-2)r\rho$. This leads to an underestimate of the level of cooperation (log ratio less than zero) for small values of responsiveness and relatedness and an overestimate for large values of either factor. Thus, in precisely the situation where cooperation is expected to be low (both low relatedness and low responsiveness), the interaction between responsiveness and relatedness *increases* cooperation.

Although an interaction between responsiveness and relatedness is a generic and important component of selection on social behavior as we have argued above, it can easily be neglected in efforts to synthesize different mechanisms of assortment when those efforts don't systematically track all the effects of plastic responses. Here, we give two examples of such efforts in an attempt to clarify how future syntheses can incorporate all the effects of different assortment mechanisms, whether they derive from behavioral plasticity

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

within a generation or from demographic mechanisms operating across generations. The first example is an "expanded Hamilton's rule" given by Queller (2011, equation 1) as the following inequality

$$-c + \sum_i b_i r_i + \sum_i d_i s_i + \sum_i m_i f_i > 0, \tag{13}$$

where c is the cost of a cooperative act to the focal individual, b_i is the benefit accruing to the i-th relative of the focal individual, r_i the relatedness coefficient between the focal individual and the i-th relative. The third and fourth terms represent an expansion of the classical Hamilton's rule to include multiplicative effects and behavioral feedbacks, which Queller calls "kind" and "kith" selection, respectively. Of these terms, kind selection ($\sum_i d_i s_i$) involves a coefficient quantifying the deviation from additivity, d_i for the i-th interaction partner, which is a "synergistic" fitness effect, and a synergy regression coefficient s_i. Finally, the last term ($\sum_i f_i m_i$) is composed of a coefficient measuring how the i-th social partner's phenotype responds to the focal individual's phenotype, f_i, times the fitness effect of this response m_i. Queller suggests that this expansion of the phenotypic version of Hamilton's rule represents more types of fitness effects in a causal manner than the classical form, and therefore might lead to better disentangling the various pathways through which selection acts on social behaviors. While we agree with the basic premise, we show below that equation (13) lacks the interaction between relatedness and responsiveness and is therefore correct only under restrictive assumptions.

To see this, we simplify equation (13) by dropping the "kind" selection component (i.e. setting $d = 0$) and obtain

$$-c + br + mf > 0 \tag{14}$$

for the case where there is a single class of individuals ($i = 1$ in equation (13)). In condition (14), the kith selection feedback coefficient, f, is analogous to our behavioral responsiveness measure ρ. If we let the kith fitness effect m be equal to the fitness effect for kin, b, then condition (14) immediately becomes condition (12) that we evaluated above that lacks the interaction between responsiveness and relatedness. The lack of this interaction stems from Queller's assumption, without any explicit justification, that the behavioral feedbacks and indirect fitness effects can be decomposed additively. This assumption is correct only under very restrictive circumstances, namely when the only individuals that respond to the focal individual are those with zero relatedness to the focal individual, whereas related individuals do not respond. In Appendix C, we

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

show that adopting this assumption explicitly in Queller's framework leads to condition (12). However, in most interactions among social partners with responsive phenotypes, both related and unrelated individuals will exhibit behavioral feedbacks, which suggests that a generalized Hamilton's rule that includes both kin and kith selection cannot be constructed by simply adding the kin and kith components together.

Our second example comes from Fletcher and Doebeli (2009) who argue for the concept of "interaction environments" as a way to measure assortment genotype and behavior, regardless of the mechanism that generates the assortment. Fletcher and Doebeli argue that the important quantity for determining whether cooperators proliferate more than defectors is the difference between the average interaction environment of cooperators, e_c, and that of defectors, e_d, which using our definitions of b and c yields

$$\frac{b}{c} > \frac{1}{e_c - e_d} \,. \tag{15}$$

The interaction environments depend not only on the genetic assortment probabilities, r_c and r_d, but also how much each genotype ends up playing C and D against other intrinsic cooperators and defectors, respectively. For intrinsic cooperators, $e_c = r_c + (1 - r_c)(q + (1 - q)(\pi_1 + \pi_3))$, where the last term is simply the sum of the fraction of games between a $\boldsymbol{C}$- and $\boldsymbol{D}$-type where the outcome is $(\boldsymbol{C}, \boldsymbol{C})$ and $(\boldsymbol{D}, \boldsymbol{C})$ calculated at the stationary distribution in (2). Similarly, $e_d = (1 - r_d)q(\pi_1 + \pi_2)$. Substituting these expressions for e_c and e_d into condition (15) yields

$$\frac{b}{c} > \frac{1 - \rho_c \rho_d}{r_c(1 - \rho_d)(1 - q) + r_d q(1 - \rho_c) + \rho_d(1 - \rho_c) + q(\rho_c - \rho_d)} \,, \tag{16}$$

which simplifies to

$$\frac{b}{c} > \frac{1 + \rho}{r + \rho} \tag{17}$$

when $\rho_c = \rho_d = \rho$ and $r_c = r_d = r$. Comparing this expression with condition (7), we can see that it is not the correct condition for an increase in cooperation as it lack the interaction between relatedness and responsiveness in the numerator. The reason the interaction environment method does not produce the correct interaction term $r\rho$ is because reciprocity and other types of behavioral responses affect a focal genotype's fitness not only through changing the social environment a genotype experiences, but also through changing the phenotype expressed by the focal genotype in response to this social environment. Therefore, the implication that behavioral responses and genetic relatedness are only two special cases of the same fundamental

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

assortment quantity does not hold in general since different assortment mechanisms can interact.

4.2. Interspecific mutualisms

The interaction between responsiveness and relatedness affects not only social behavior between individuals of the same species but it also shapes social behavior between species. We can illustrate this by applying the population genetic model with continuous traits (Section 3 and Appendix B) to an interaction between a population of host individuals that each interact with a group of n symbionts. Label the investment of the host species (H) into the ith symbiont by x_i and the investment of the ith symbiont (S_i) into the host by y_i. The payoffs (i.e., individual fertility as a result of investments) are denoted $u_H(x_1, \cdots, x_n, y_1, \cdots, y_n)$ for the host and $u_{S_i}(x_i, y_i)$ for the ith symbiont (i.e., assume that a symbiont's fertility is only affected by the investment it makes into and it receives from the host).

Suppose further that the investments of both the hosts and symbionts are plastic and can change in response to each other. As a result, all of the investment variables are joint functions of the genotypes of the individual hosts and symbionts. Following the methodology in Section 3 (see also Akçay et al., 2009; Akçay and Van Cleve, 2012), we label the host's and symbionts' genetically determined phenotypes p_H and p_S, respectively. Recall that phenotype in this context refers to a physiological mechanism (neurological, biochemical, etc) that alters investment levels, and that phenotypes are directly genetically determined rather investment levels. To find the increase condition for genotypes that result in higher investments in the mutualism, we apply equation (9) once for each species while treating the partner of the other species as a non-reproductive class.

Calculating the increase condition requires the total derivative of the fertilities (u_H and u_S) with respect to the host and symbiont phenotypes. For example, the change is host fertility with respect to a change in the phenotype of symbiont j is:

$$\begin{aligned} \frac{du_H}{dp_{Sj}} &= \sum_i \frac{\partial u_H}{\partial x_i}\frac{dx_i}{dp_{Sj}} + \sum_i \frac{\partial u_H}{\partial y_i}\frac{dy_i}{dp_{Sj}} \\ &\propto n\left[-c_H + \rho_{HS} b_H\right] , \end{aligned} \tag{18}$$

where $c_H \equiv -\frac{\partial u_H}{\partial x_i}$, $b_H \equiv \frac{\partial u_H}{\partial y_i}$, and $\rho_{HS} \equiv \frac{dy}{dp_{Sj}} / \frac{dx}{dp_{Sj}}$. The latter quantity denotes the response of a symbiont's investment to the change of the host's investment into that symbiont (Foster and Wenseleers, 2006; Akçay and Simms, 2011; Akçay, in review). Calculating the other total derivatives in the same way, we obtain the

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conditions for the increase of the higher investment level phenotypes in hosts and symbionts, respectively, as

$$\frac{b_H}{c_H} > \frac{1 + r_{HS}\left(\rho_{S_\bullet H_\bullet} + (n-1)\rho_{S_\bullet H_\circ}\right)}{\rho_{HS} + r_{HS}\left(1 + (n-1)\rho_{SS}\right)} \tag{19}$$

$$\frac{b_S}{c_S} > \frac{1 + r_{SH}\rho_{HS} + (n-1)r_{SS}\rho_{SS}}{\rho_{S_\bullet H_\bullet} + (n-1)r_{SS}\rho_{S_\bullet H_\circ} + r_{SH}}, \tag{20}$$

where r_{SS} is the relatedness coefficient for the symbionts. The coefficients r_{HS} and r_{SH} measure the genetic covariance between higher investment level phenotypes of the symbionts and the host (relative to host and symbiont genetic variance, respectively) and are analogous to relatedness coefficients (Frank, 1994). Here, the coefficient $\rho_{S_\bullet H_\bullet}$ denotes how much increased investment of a focal symbiont causes the host to increase its investment into the same symbiont, $\rho_{S_\bullet H_\circ}$ measures how the investment in another symbiont by the host responds to a focal symbiont's investment, and ρ_{SS} measures the response of symbionts to each other.

Similar to the intraspecific condition for the increase of cooperative phenotypes (equation 11), conditions (19) and (20) contain both multiplicative and additive interactions between the different relatedness and responsiveness coefficients. Such interactions have not been fully appreciated before (e.g., Foster and Wenseleers, 2006) and have important effects on the level of investments that evolve in the host and symbionts. For example, hosts and symbionts that can respond plastically to changes in the investment levels of one another result in higher levels of investment by both host and symbiont and this increase is mediated by an interaction between relatedness and responsiveness.

4.3. Reciprocity and synergy

In the above analyses, and in many biological models of social interactions, the marginal costs and benefits of cooperation and defection for the focal individual do not depend on the genotype of its partner. However, it is often plausible that cooperators might obtain an extra benefit (or cost) from interacting with other cooperators, and this benefit (or cost) is called positive (or negative) synergy. The importance of synergy, and non-additive interactions more generally, to understanding the evolution of social behavior was demonstrated by Queller in his seminal paper (1985). Since then, much work has emphasized the positive and negative synergies can relax or constrain, respectively, the condition for cooperative behaviors to increase in frequency (Queller, 1992b; Fletcher and Zwick, 2006; Hauert et al., 2006; Lehmann and Keller, 2006a,b; Taylor and Nowak, 2007; Ohtsuki, 2012; Taylor and Maciejewski, 2012). If positive synergies are strong enough, they can also change structure of the evolutionary game (Fletcher and Zwick, 2006; Hauert et al., 2006; Taylor

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and Nowak, 2007), which might turn a Prisoner's Dilemma game into a Stag-Hunt game.

Since repeated interactions, such as reciprocity and behavioral responses, in a Prisoner's Dilemma game can also stabilize cooperation (Axelrod and Hamilton, 1981; Fudenberg and Maskin, 1986), these interactions can also be viewed as changing the structure of the one-shot evolutionary game (Fletcher and Zwick, 2006; Lehmann and Keller, 2006a; André and Day, 2007; Taylor and Nowak, 2007; Akçay et al., 2009). The interaction between behavioral responses and synergy, however, has received little study. Here, we show how behavioral responses and synergy have distinct effects on the dynamics of social traits and map the changes in the structure of the social game induced by both mechanisms simultaneously.

As is customary in models of synergy, we modify the Prisoner's Dilemma game from Section 2 by adding a constant d to the payoffs of both individuals when they both cooperate. When $d > 0$, two cooperators produce more benefit together than twice the benefit of a single cooperator; when $d < 0$, they produce less. These cases relate to cooperative acts by the partners being strategic complements versus substitutes, respectively, as defined in economics (Bulow et al., 1985).

Using the same methodology as in section 2, we calculate the expected payoff to an intrinsic cooperator when paired with an intrinsic defector and an intrinsic defector against an intrinsic cooperator as

$$\begin{aligned} w_{CD} &= \frac{(1-\rho_c)\left((b\rho_d - c)(1-\rho_c\rho_d) + d\rho_d(1-\rho_c)\right)}{(1-\rho_c\rho_d)^2} \quad \text{and} \\ w_{DC} &= \frac{(1-\rho_c)\left((b - \rho_d c)(1-\rho_c\rho_d) + d\rho_d(1-\rho_c)\right)}{(1-\rho_c\rho_d)^2}, \end{aligned} \tag{21}$$

respectively. These expressions indicate that reciprocity and payoff synergism can interact in complex ways. To better illustrate this interaction, suppose that both types of players have the same probability of reciprocation, $\rho_c = \rho_d = \rho$, and the same probability of assortment, $r_c = r_d = r$. Then, as shown in Appendix A, the component that the payoff synergy d adds to the condition for cooperation to increase, $W_C - W_D > 0$, is

$$d\left(p + r(1-p) + \frac{\rho(1-2p)(1-r)}{(1+\rho)^2}\right), \tag{22}$$

where the first component, $p + r(1-p)$ matches previous work on synergy without reciprocity (Wenseleers, 2006; Lehmann and Keller, 2006b). The first thing to note from (22) is that, compared the non-synergistic case in condition (7), synergy adds frequency dependence to the increase condition. This is a well-known result (Lehmann and Keller, 2006b; Gardner et al., 2007; Ohtsuki, 2012; Taylor and Maciejewski, 2012) that

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complicates the analysis of synergy. Essentially, this frequency dependence arises from explicitly accounting for the frequency of different discrete pairings, which is necessary when payoffs are not additive. Another way of thinking about this frequency dependence is that while relatedness or genetic assortment generally accounts for pairwise genetic correlations, accounting for synergy requires calculating triplet correlations, which are generally much more tedious to calculate (Ohtsuki, 2012, for example). Notably, this frequency dependence can be eliminated in certain stochastic models with symmetric mutation and weak selection (e.g., Tarnita et al., 2009; Van Cleve and Lehmann, 2013) even in the case of synergistic payoffs.

Second, the evolutionary game, defined by the expected payoffs from the repeated game, w_{CC}, w_{CD}, w_{DC}, and w_{DD}, will take on different game structures depending on the values of d and ρ. Figure 2 illustrates the different possibilities. With $\rho < c/b$, a high enough synergism can transform the game from a Prisoner's Dilemma into a Stag-Hunt (coordination) game, which will have two evolutionarily stable strategies (ESSs): all cooperators or all defectors. If $\rho > 0$, then still higher d can turn the game into a Mutualism game, where intrinsic cooperators have an unconditional advantage. When $\rho > c/b$, the possibility of generating a coordination game disappears, and increasing d turns the Prisoner's Dilemma into a Hawk-Dove game (anti-coordination or Snowdrift game) and eventually into a Mutualism game as d increases further. To our knowledge, the interaction between reciprocity and synergy demonstrated in Figure 2 has not been previously recognized. This interaction is particularly important in the light of recent results that show that non-additivity by itself does not change the conditions for the evolution of cooperation in homogenous spatially structured populations (Ohtsuki, 2012), unless it changes the structure of the evolutionary game.

4.4. Convergence stability of reciprocity

While the discrete-action model allows us to derive intuitive relationships between relatedness, reciprocity, and synergy, it assumes a specific mechanism of repeated interactions that resembles tit-for-tat or direct reciprocity. One crucial advantage of the population genetic model with continuous traits in section 3 is that we can describe whole families (in the mathematical sense) of behavioral mechanisms with a single function, $\rho(p)$, which tells us how responsive individuals are to their social partners, ρ, as a function of some other phenotype, p. For example, p might measure the strength of an other-regarding preference that causes a focal individual to prefer increases in its partners' payoffs. Each level of other regard will then correspond to a level of behavioral responsiveness through $\rho(p)$. Our final application uses the population genetic model to derive a graphical relationship necessary for a behavioral mechanism given by $\rho(p)$ to be evolutionarily stable

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the long term.

If we are interested in what social behaviors or phenotypes evolve in the long-term, we can set $\Delta q = 0$ in equation (9) and solve for candidate behaviors or phenotypic values that are evolutionarily stable. Setting $\Delta q = 0$ is equivalent to setting the expression in (10) equal to zero (i.e., $\mathcal{S} = 0$). In order to ensure that these candidate phenotypes are actually stable, we need to check whether populations can evolve or converge towards these candidate phenotypes starting from neighboring phenotypes (Eshel and Motro, 1981; Eshel, 1983; Taylor, 1989). This kind of evolutionary stability, called convergence stability, is often more important in the long-term evolution of continuous phenotypes than the traditional notion of strict evolutionary stability (Wakano and Lehmann, 2012).

A candidate phenotypic value is convergence stable when the derivative of the $\mathrm{d}\Delta q/\mathrm{d}\delta$ with respect to the resident phenotypic value p is negative (Taylor, 1989). This condition amounts to

$$\frac{\mathrm{d}}{\mathrm{d}p}\left[\frac{\mathrm{d}\Delta q}{\mathrm{d}\delta}\right]_{\mathcal{S}=0} \propto \left[\frac{\mathrm{d}\rho}{\mathrm{d}p}\frac{\partial\mathcal{S}}{\partial\rho} + \frac{\mathrm{d}b}{\mathrm{d}p}\frac{\partial\mathcal{S}}{\partial b} + \frac{\mathrm{d}c}{\mathrm{d}p}\frac{\partial\mathcal{S}}{\partial c}\right]_{\mathcal{S}=0} < 0\,, \tag{23}$$

where the derivatives in (23) are *total* derivatives evaluated at $\mathcal{S} = 0$ and measure how ρ, b, and c change due to both direct and indirect changes in the phenotype p. Even in this generic form, condition (23) tells us that the evolutionary stability of a phenotype depends on how a change in the phenotype affects the level of responsiveness, $\frac{\mathrm{d}\rho}{\mathrm{d}p}$, and how responsiveness changes selection for that phenotype, $\frac{\partial\mathcal{S}}{\partial\rho}$. In other words, it is not only the level of responsiveness that matters, but the sensitivity of responsiveness levels to changes in the underlying phenotype and the sensitivity of the selection gradient to changes in responsiveness.

The level of sensitivity of ρ to changes in p, $\frac{\mathrm{d}\rho}{\mathrm{d}p}$, can be calculated for a specific proximate mechanism from the function $\rho(p)$. More specifically, we are interested in this level of sensitivity calculated at candidate evolutionarily stable phenotypes, or $\left[\frac{\mathrm{d}\rho}{\mathrm{d}p}\right]_{\mathcal{S}=0}$. Using the convergence stability condition (23) with $\mathcal{S} = 0$ from equation (10), we find for this level of sensitivity that

$$\left[\frac{\mathrm{d}\rho}{\mathrm{d}p}\right]_{\mathcal{S}=0} < \frac{\mathrm{d}\rho_{\mathcal{S}=0}}{\mathrm{d}p} \tag{24}$$

where $\frac{\mathrm{d}\rho_{\mathcal{S}=0}}{\mathrm{d}p}$ is the change in ρ along the $\mathcal{S} = 0$ curve as a function of the phenotype p. Condition (24) has a simple graphical interpretation given in Figure 3: the curve generated by the behavioral mechanism, $\rho(p)$, must cross the evolutionary stability curve, $\mathcal{S} = 0$, from above. Values of ρ and p on the $\mathcal{S} = 0$ curve can be

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

thought of as phenotypes that generate a level of responsiveness that exactly balances the costs and benefits of the behavior. Thus, a behavioral mechanism that crosses the $\mathcal{S} = 0$ curve from above is one that generate marginally more responsiveness than what is required to exactly balance the costs and benefits, and this is what makes such a behavioral mechanisms stable in the long term. The location of the $\mathcal{S} = 0$ curve will depend on relatedness. For Prisoner's Dilemma type interactions ($b(n-1) > c > 0$), both $\partial\mathcal{S}/\partial r > 0$ and $\partial\mathcal{S}/\partial\rho > 0$, which means that the $\mathcal{S} = 0$ curve shifts down as relatedness increases. In Figure 3, we plot two $\mathcal{S} = 0$ curves for hypothetical values of relatedness, $r_1 < r_2$, and its clear from the figure that increasing relatedness increases the convergence stable level of responsiveness for the proximate mechanism given by $\rho(p)$.

5. Discussion

One of the major goals of evolutionary and behavioral ecology is to elucidate the casual biological pathways that drive the evolution of social traits. However, natural selection rarely operates through a single pathway, and therefore the various causal components of social evolution need to be integrated and their commonalities and interactions explored. Our main goal in this paper is to contribute to this synthesis effort. In so doing, we highlight that genetic assortment, behavioral responses, and non-additive interactions between phenotypes all have both distinct and interacting effects. Therefore, combining these components requires explicit and consistent modeling approaches. Using a model based on discrete phenotypes and a population-genetic model based on continuous phenotypes, we obtain the following results.

(i) Phenotypic assortment is not a univariate quantity: it matters how much assortment comes from both behavioral responses and genetic assortment. Further, behavioral responses and genetic assortment cannot simply added to each other, even when social interactions (payoffs) are additive. Rather, they interact with each other both additively and multiplicatively. Prior results analogous to this can be found in Lehmann and Keller (2006a), McGlothlin et al. (2010) and Akçay and Van Cleve (2012). Yet, despite this interaction being known for some time, its crucial role seems not widely appreciated. Akçay and Van Cleve (2012) is the first paper to our knowledge to highlight the synergistic and symmetric effect of genetic assortment and behavioral responses on the response to selection of social phenotypes in social groups of any size. In that work, we demonstrated how synergism between genetic assortment and behavioral responses can facilitate the transfer of fitness effects from the within-group level to the between-group level, a crucial step in evolutionary transitions in individuality (Maynard Smith and Szathmáry, 1995; Michod, 2005). These previous

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results assumed continuous phenotypes and within-species interactions; here we show that a similar synergism between genetic assortment and behavioral responses is found for discrete phenotypes and is important for species interactions as well.

(ii) Non-additive or synergistic interactions can have important effects on the whether cooperative behaviors increase in frequency in a population and these effect depend behavioral responses and reciprocity. Specifically, synergy and reciprocity can both change the structure of the evolutionary game. When this game starts as a Prisoner's Dilemma where cooperation is unconditionally, increasing synergy or responsiveness levels can stabilize some level of cooperation. Whether all individuals cooperate fully though depends on the combination of synergy and responsiveness that determines whether the game becomes a Stag-Hunt, Mutualism, or Hawk-Dove game.

Even though synergistic interactions are easy to study using a discrete phenotypic model, the population-genetic model with continuous phenotypes shows the importance of non-additive payoffs or fitness in more subtle way. These interactions will affect the evolutionary (convergence) stability of a phenotype and how the stable phenotype changes with underlying parameters. The former effect can be seen in equation (23), which shows that non-additivity in the benefits and costs of a social phenotype (behavior) appear in the convergence stability condition (since b and c are themselves already first-order derivatives of fertility). Non-additivity in benefits and costs relate directly to ecological mechanisms that convert the effort of individuals in the social interaction into payoffs in terms of fertility or survival. For example, individuals who put effort into a group hunt (which might be modeled by a stag-hunt game (Skyrms, 2001)) may receive more return in terms of the amount of prey item per unit of caloric investment when more individuals join the hunt. This kind of ecological synergy, or "complementarity" in economics (Bulow et al., 1985), can have large effects on the level of investment in cooperative interactions both within groups (Akçay et al., 2009) and in structured populations (Akçay and Van Cleve, 2012; Cornforth et al., 2012).

(iii) There is a simple graphical relationship required for a specific proximate behavioral mechanisms (e.g., social preferences, social norms, punishment, etc) to be evolutionarily stable in the long run. When plotting the curve of responsiveness as a function of the evolving phenotype for the mechanism (level of other regard, strength of conformity, probability of punishment, etc), this curve must cross the curve that balances the costs and benefits of the behavior from above. Moreover, since different levels of relatedness result in different curves that balance benefit and cost, what level of responsiveness evolves for a specific proximate

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mechanism can also be determined graphically.

5.1. Empirical issues

The empirical study of the interplay between behavioral responses and genetic assortment and the role of synergistic payoffs is still relatively new. Most of the work on the linkage between behavioral responses and genetic assortment comes from the now rapidly growing application of indirect genetic effect (IGE) models to laboratory, agricultural, and field-based populations (e.g. Moore et al., 2002; Petfield et al., 2005; Mutic and Wolf, 2007; Ellen et al., 2008; Danielson-François et al., 2009; Bleakley and Brodie, 2009; Teplitsky et al., 2010; Frère et al., 2010; Wilson et al., 2011; Hamilton and Ligocki, 2012). IGE models (Moore et al., 1997; Wolf et al., 1999; McGlothlin et al., 2010) extend classical quantitative genetics (Lynch and Walsh, 1998) to include a coefficient, ψ, that measures the effect of a social partner's phenotype on the phenotype of a focal individual. Importantly, the IGE coefficient ψ directly maps to our behavioral responsiveness coefficient ρ (Akçay and Van Cleve, 2012) and is thus also a measure of reciprocity. We know from equation (10) (and analogously equation 18 in McGlothlin et al. 2010) that an interaction between behavioral responses and relatedness is important for selection on social traits, even when no such interaction is assumed to affect individual phenotypes (see equation C.5 in the Appendix or equation 5 in (McGlothlin et al., 2010)). However, most of the empirical work to date spends little time on the interaction of behavioral responses and relatedness with the notable exceptions of Bleakley and Brodie (2009) and Frère et al. (2010), and only Frère et al. (2010) emphasize the crucial role of this interaction. Moreover, some empirical work assumes an explicit interaction term in the regression model for individual phenotypes, and it is not yet clear how empirical estimates of this interaction can be tied to predicted interaction term that arises in measures of selection on the trait. Thus, more work needs to be done understanding the interaction between behavioral responses and relatedness both theoretically and empirically.

Empirical studies explicitly focused on non-additive payoffs are uncommon (Queller et al., 2003; Gore et al., 2009, are notable exceptions). We do not believe that this state of affairs reflects the importance of non-additive interactions between phenotypes in nature. Instead, non-additive and non-linear interactions are likely to be abundant and their effects strong. For example in parental care, many arguments implicitly or explicitly rely on the contributions each parent combining non-additively (Clutton-Brock, 1991; Akçay et al., 2009). We suggest that measuring non-additive interactions at both the payoff (fertility or survival) and fitness levels will prove important for explaining the evolution of social behaviors, regardless of whether the

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenotypes in question vary more or less continuously or are discrete characters with large effects.

To conclude, a careful analysis of behavioral responses, genetic assortment, and non-additive interactions shows a complex interaction between these three pathways. We believe that focusing on this complexity will ultimately deliver a more complete and nuanced understanding of the evolutionary forces shaping social behaviors. The necessary mathematical frameworks for such understanding exists, though as we show in this paper, it needs to be applied with some care. Once crafted, these theories can be tested against the growing abundance of sophisticated datasets that measure both networks of social interactions as well as important demographic and fitness-related variables (e.g. Holekamp et al., 2012; Apicella et al., 2012).

Acknowledgements

J.V. was supported by a Santa Fe Institute Omidyar Fellowship and the National Evolutionary Synthesis Center (NESCent), NSF #EF-0423641. E.A. was supported by National Science Foundation Grant EF-1137894, to S.A. Levin.

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix A: Two-player game with synergy

We follow the discrete-action model defined in section 2. Individuals are either intrinsic cooperators or intrinsic defectors who can either play their intrinsic strategy or reciprocate their partner's last action. Let ρ_c and ρ_d be the probabilities that an intrinsic cooperator and defector, respectively, reciprocate or copy their partner's last action. We allow for non-randomly interaction among intrinsic cooperators and intrinsic defectors that might results from population structure or other assortation mechanisms and define r_c and r_d to be the probabilities that intrinsic cooperators and defectors, respectively, interact among themselves non-randomly. Two cooperators who play each other each earn payoff $b - c + d$, where b is the benefit, c is the cost of cooperation, and d is a synergy term that account for non-additive payoffs. A cooperator and defector playing together results in a payoff of $-c$ for the cooperator and b for the defector. Two defectors playing each other each earn zero payoff. Given this setup and following the analysis of section 2, the payoffs to an intrinsic cooperator paired with an intrinsic defector and vice-versa are

$$w_{CD} = \frac{(1-\rho_c)\left((b\rho_d - c)(1-\rho_c\rho_d) + d\rho_d(1-\rho_c)\right)}{(1-\rho_c\rho_d)^2} \quad \text{and}$$
$$w_{DC} = \frac{(1-\rho_c)\left((b - \rho_d c)(1-\rho_c\rho_d) + d\rho_d(1-\rho_c)\right)}{(1-\rho_c\rho_d)^2},$$

respectively, which are also given in equation (21) in the main text.

We can plug the above values for w_{CD} and w_{DC}, along with $w_{CC} = b - c + d$ and $w_{DD} = 0$, into the expression for fitness in equation (5) to derive the condition for cooperation to increase, $W_C - W_D > 0$, which simplifies to

$$\begin{aligned} & b\left(\frac{p\rho_c + (1-p)\rho_d + r_c(1-p)(1-\rho_d) + r_d p(1-\rho_c) - \rho_c\rho_d}{1-\rho_c\rho_d}\right) \\ & -c\left(\frac{1-(1-p)\rho_c - p\rho_d + (1-p)r_c\rho_c(1-\rho_d) + r_d p\rho_d(1-\rho_c)}{1-\rho_c\rho_d}\right) \\ & +d\left(\frac{(p + r_c(1-p))(1-\rho_d)\left(1-\rho_c^2\rho_d\right) + (1-p+pr_d)(1-\rho_c)^2\rho_d}{(1-\rho_c\rho_d)^2}\right) > 0\,. \end{aligned} \tag{A.1}$$

This expression simplifies to condition (6) when $d = 0$. Another simplification is to assume that the reciprocation and assortment probabilities are the same for cooperators and defectors, namely $\rho_c = \rho_d = \rho$ and $r_c = r_d = r$. In this case, expression (A.1) simplifies to

$$b\frac{r+\rho}{1+\rho} - c\frac{1+r\rho}{1+\rho} + d\left((p + r(1-p)) + \frac{\rho(1-2p)(1-r)}{(1+\rho)^2}\right) > 0\,, \tag{A.2}$$

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which in turn simplifies to condition (7) when $d = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix B: Allele frequency change in a class structured model

In section 3, we build a general class-structured model of allele frequency change and present a first-order approximation, expression (10), that is valid given the phenotype of one individual can affect the fertility of another individual so long as both individuals are in the same class and the effect of phenotype on fertility is independent of class. Our goal in this section is to show how to arrive at expression (10) beginning with

$$\frac{\mathrm{d}\Delta q}{\mathrm{d}\delta} = \boldsymbol{\alpha} \cdot \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\delta} \cdot \mathbf{q} = \sum_{i,j,l=1}^{K} \sum_{k=1}^{N_j} \sum_{m=1}^{N_l} \alpha_i \frac{\partial w_{ijk}}{\partial p_{lm}} q_{jk} q_{lm} \,. \tag{B.1}$$

Recall that the Taylor series is expanded around $\delta = 0$, so the partial derivatives $\frac{\partial w_{ijk}}{\partial p_{lm}}$ in (B.1) are evaluated at neutrality where natural selection on fertility differences is absent. Additionally, $\delta = 0$ implies that all individuals in a class are equivalent, which means that the partial derivatives above fall into three cases: the first case is the effect of a focal individual's phenotype on its own probability of sending an allele to class j in the next generation from its own class i, $\frac{\partial w_{ij}}{\partial p_{j\bullet}}$; the second is the effect on that probability of another individual's phenotype from the same class as the focal, $\frac{\partial w_{ij}}{\partial p_{j\circ}}$; and the third is the effect on that probability of the phenotype of an individual in a different class than the focal individual, $\frac{\partial w_{ij}}{\partial p_l}$ where $l \neq j$. Substituting these derivates into (B.1), we obtain

$$\boldsymbol{\alpha} \cdot \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\delta} \cdot \mathbf{q} = \sum_{i,j=1}^{K} \alpha_i N_j \left(\frac{\partial w_{ij}}{\partial p_{j\bullet}} \overline{q_{j\bullet}^2} + (N_j - 1) \frac{\partial w_{ij}}{\partial p_{j\circ}} \overline{q_{j\bullet} q_{j\circ}} + \sum_{l \neq j} N_l \frac{\partial w_{ij}}{\partial p_l} \overline{q_j q_l} \right) . \tag{B.2}$$

Further, we assume that the phenotype only affects fertility (though it could affect survival or demographic factors such as carrying capacity or dispersal rates), which means the derivaties in the left-hand side of the equations in (B.3) can be split into two components, the effect of the phenotype on fertility and the effect of fertility on the probability an allele in class i descends from class j. The latter derivatives we represent with τ where $\tau_{ij,j\bullet} = \frac{\partial w_{ij}}{\partial f_{ij\bullet}}$, $\tau_{ij,j\circ} = \frac{\partial w_{ij}}{\partial f_{ij\circ}}$, and $\tau_{ij,l} = \frac{\partial w_{ij}}{\partial f_{il}}$ for $l \neq j$. These derivatives essentially contain transmission probabilities due to dispersal, mutation, allelic segregation, age-dependent mortality, etc, and we assume they are independent of phenotype. Given these two classes of derivatives,

$$\frac{\partial w_{ijk}}{\partial p_{lm}} = \begin{cases} \frac{\partial w_{ij}}{\partial p_{j\bullet}} = \tau_{ij,j\bullet} \frac{\partial f_{ij}}{\partial p_{j\bullet}} + (N_j - 1) \tau_{ij,j\circ} \frac{\partial f_{ij}}{\partial p_{j\circ}} & \text{for } (l,m) = (j,k) \\ \frac{\partial w_{ij}}{\partial p_{j\circ}} = \tau_{ij,j\bullet} \frac{\partial f_{ij}}{\partial p_{j\circ}} + \tau_{ij,j\circ} \left(\frac{\partial f_{ij}}{\partial p_{j\bullet}} + (N_j - 2) \frac{\partial f_{ij}}{\partial p_{j\circ}} \right) & \text{for } l = j \text{ and } m \neq k \\ \frac{\partial w_{ij}}{\partial p_l} = \tau_{ij,l} \left(\frac{\partial f_{il}}{\partial p_{l\bullet}} + (N_l - 1) \frac{\partial f_{il}}{\partial p_{l\circ}} \right) & \text{for } l \neq j \,. \end{cases} \tag{B.3}$$

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we assume that the effect of phenotype on fertility is independent of the source or destination classes of the offspring; that is, we can define $-C = \frac{\partial f_{ij}}{\partial p_{j\bullet}} = \frac{\partial f}{\partial p_\bullet}$ and $B = \frac{\partial f_{ij}}{\partial p_{j\circ}} = \frac{\partial f}{\partial p_\circ}$. Combining C and B with (B.3) into equation (B.2) results in

$$\begin{aligned}\boldsymbol{\alpha} \cdot \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}\delta} \cdot \mathbf{q} = &- C \left[\sum_{i,j=1}^{K} \alpha_i N_j \left(\tau_{ij,j\bullet} \overline{q_{j\bullet}^2} + (N_j - 1)\tau_{ij,j\circ} \overline{q_{j\bullet} q_{j\circ}} + \sum_{l \neq j} N_l \tau_{ij,l} \overline{q_j q_l} \right) \right] \\ & B \left[\sum_{i,j=1}^{K} \alpha_i N_j (N_j - 1) \left(\tau_{ij,j\bullet} \overline{q_{j\bullet}^2} + \left(\tau_{ij,j\bullet} + (N_j - 2)\tau_{ij,j\circ} \right) \overline{q_{j\bullet} q_{j\circ}} \right) + N_j \sum_{l \neq j} N_l (N_l - 1) \tau_{ij,l} \overline{q_j q_l} \right] \\ = & K_C \left(-C + rB \right) ,\end{aligned} \tag{B.4}$$

where K_C and K_B equal the first and second bracket in (B.4), respectively. Scaled relatedness is given $r = K_B/K_C$. The terms in parentheses on the last line of (B.4) are the same as the right-hand side of equation (10), once C and B are evaluated following the assumptions in that section that the phenotype p is a parameter of some physiological/psychological mechanism that affects the underlying actions individuals choose, a^*, and the marginal effect of the action of one individual on the action of another is ρ. A detailed exposition of this analysis is given by Akçay and Van Cleve (2012) where $-C$ corresponds to $\frac{\partial F_i}{\partial p_i}$ in equation (2) and B to $\frac{\partial F_i}{\partial p_j}$ in equation (3) in that paper. In most biological cases of interest, $K_C > 0$, which results in the proportionality in equation (10). The convergence stability condition in (23) can be derived from equation (10) simply by recalling the assumption that the phenotype affects fertility but not demographic quantities contained in r.

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix C: Uni-directional and bi-directional reciprocity in indirect genetic effects models

In this section, we will show that: (i) assuming a uni-directional type of reciprocity where the focal individual can affect its partner's phenotype, but not vice-versa, can lead to the expanded Hamilton's rule presented by Queller (2011, equation 1); and (ii) bi-directional reciprocity leads to condition (7), which is equivalent to results derived by us (Akçay and Van Cleve, 2012) and others (Lehmann and Keller, 2006a; McGlothlin et al., 2010). Notation in the following equations is taken from Queller (2011) and generally follows quantitative genetic approaches to social evolution (e.g., Queller, 1992b; Frank, 1997; Queller, 2011) with a particular focus on the methods in indirect genetic effect (IGE) models (Moore et al., 1997; Wolf et al., 1999; McGlothlin et al., 2010).

We begin with the phenotypic version of Hamilton's rule, derived from the Price equation (Price, 1970), that is equation (7) in Queller (2011), which is

$$\beta_{WP\cdot P'} + \beta_{WP'\cdot P}\frac{\mathrm{Cov}\left[G, P'\right]}{\mathrm{Cov}[G, P]} \,. \tag{C.1}$$

where we use primes to denote variables associated with the social partner. Suppose that the phenotype of the social partner is defined as

$$P' = G' + \beta_{P'P}P + \epsilon' \,, \tag{C.2}$$

which says that the phenotype of the social partner is a linear function of its average breeding value (G'), the effect of the phenotype of the focal ($\beta_{P'P}P$), and a random component with mean zero (ϵ', a standard assumption in quantitative genetics; see Lynch and Walsh, 1998). Suppose also that the phenotype of focal individual is given by

$$P = G + \epsilon \,, \tag{C.3}$$

which says the phenotype of the focal individual is a linear function *only* of its breeding value and a random component; thus, there is no effect of the partner's phenotype on the focal, which embodies the assumption we believe underlies Queller (2011)'s analysis. Plugging P and P' given in equations (C.2) and (C.3) into equation (C.1) yields

$$\beta_{WP\cdot P'} + \beta_{WP'\cdot P}\frac{\mathrm{Cov}\left[G, G' + \beta_{P'P}P + \epsilon'\right]}{\mathrm{Cov}[G, G + \epsilon]} = \beta_{WP\cdot P'} + \beta_{WP'\cdot P}\left(\beta_{P'P} + \beta_{G'G}\right) \,, \tag{C.4}$$

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which can be written as

$$-c + b(\rho + r)$$

in the notation of the main text and is equivalent to the additive combination of kin and kith components in equation (1) of Queller (2011).

To include bi-directional reciprocity, we follow the analysis presented by McGlothlin et al. (2010); Akçay and Van Cleve (2012) obtain the same results and present a comparison of the relative merits of the current approach and the IGE approach. We start by setting

$$P = G + \beta_{PP'} P' + \epsilon\,, \tag{C.5}$$

which expresses the fact that P' also has an effect on P. Before we can plug this definition of P or the definition of P' into equation (7) of Queller (2011), we must insert our definition of P' into (C.5) and our definition of P into (C.2) and solve the resulting equations; the solutions are

$$\begin{aligned} P &= \frac{G + \epsilon + \beta_{PP'}\left(G' + \epsilon'\right)}{1 - \beta_{PP'}\beta_{P'P}} \\ P' &= \frac{G' + \epsilon' + \beta_{P'P}(G + \epsilon)}{1 - \beta_{PP'}\beta_{P'P}}\,. \end{aligned} \tag{C.6}$$

These solutions can be compared to the analogous solution in McGlothlin et al. (2010) (equation 6). Plugging equations (C.6) into equation (7) of Queller (2011) yields

$$\begin{aligned} \beta_{WP.P'} + \beta_{WP'.P}\frac{\mathrm{Cov}\left[G, \frac{G'+\epsilon'+\beta_{P'P}(G+\epsilon)}{1-\beta_{PP'}\beta_{P'P}}\right]}{\mathrm{Cov}\left[G, \frac{G+\epsilon+\beta_{PP'}(G'+\epsilon')}{1-\beta_{PP'}\beta_{P'P}}\right]} &= \beta_{WP.P'} + \beta_{WP'.P}\frac{\mathrm{Cov}\left[G, G' + \epsilon' + \beta_{P'P}(G+\epsilon)\right]}{\mathrm{Cov}\left[G, G + \epsilon + \beta_{PP'}\left(G' + \epsilon'\right)\right]} \\ &= \beta_{WP.P'} + \beta_{WP'.P}\frac{\mathrm{Cov}\left[G, G'\right] + \beta_{P'P}\,\mathrm{Var}[G]}{\mathrm{Var}[G] + \beta_{PP'}\,\mathrm{Cov}\left[G, G'\right]} \\ &= \beta_{WP.P'} + \beta_{WP'.P}\frac{\beta_{G'G} + \beta_{P'P}}{1 + \beta_{PP'}\beta_{G'G}} \end{aligned}$$

Thus, the Hamilton's rule one gets from equation (7) of Queller (2011) should be

$$\beta_{WP.P'} + \beta_{WP'.P}\frac{\beta_{G'G} + \beta_{P'P}}{1 + \beta_{PP'}\beta_{G'G}} > 0$$

bioRxiv preprint doi: https://doi.org/10.1101/000521; this version posted November 16, 2013. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or

$$\left(1 + \beta_{PP'}\beta_{G'G}\right)\beta_{WP\cdot P'} + \beta_{WP'\cdot P}\left(\beta_{G'G} + \beta_{P'P}\right) > 0\,. \tag{C.7}$$

Assuming that $\beta_{P'P} = \beta_{PP'}$, equation (C.7) becomes in our notation

$$-c(1 + \rho r) + b(r + \rho) > 0\,,$$

which is exactly condition (7) from our model. This is also exactly equivalent to equation (18) given by McGlothlin et al. (2010) for pairwise interactions once the appropriate mapping between ρ and their IGE coefficient ψ is made (see Appendix A9 in ref Akçay and Van Cleve, 2012).

References

Akçay, E. in review. Evolutionary models of mutualism. J. L. Bronstein, ed., Mutualisms. Oxford University Press.

Akçay, E., and J. Van Cleve. 2012. Behavioral Responses in Structured Populations Pave the Way to Group Optimality. Am. Nat. 179:257–269.

Akçay, E., J. Van Cleve, M. W. Feldman, and J. Roughgarden. 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc. Natl. Acad. Sci. U. S. A. 106:19061–19066.

Akçay, E., and E. L. Simms. 2011. Negotiation, sanctions, and context dependency in the legume-Rhizobium mutualism. The American naturalist 178:1–14. URL `http://www.ncbi.nlm.nih.gov/pubmed/21670573`.

André, J.-B., and T. Day. 2007. Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. J. Theor. Biol. 247:11–22.

Apicella, C. L., F. W. Marlowe, J. H. Fowler, and N. A. Christakis. 2012. Social networks and cooperation in hunter-gatherers. Nature 481:497–501.

Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. Science 211:1390–1396.

Balshine-Earn, S., F. C. Neat, H. Reid, and M. Taborsky. 1998. Paying to stay or paying to breed? Field evidence for direct benefits of helping behavior in a cooperatively breeding fish. Behav. Ecol. 9:432–438.

Bleakley, B. H., and E. D. Brodie, III. 2009. Indirect genetic effects influence antipredator behavior in guppies: estimates of the coefficient of interaction psi and the inheritance of reciprocity. Evolution 63:1796–806.

Bulow, J. I., J. D. Geanakoplos, and P. D. Klemperer. 1985. Multimarket Oligopoly: Strategic Substitutes and Complements. J. Polit. Economy 93:pp. 488–511.

Caswell, H. 2001. Matrix Population Models. Sinauer Associates Sunderland, Sunderland, Mass.

Charlesworth, B. 1994. Evolution in Age-Structured Populations. Cambridge University Press, Cambridge, 2nd ed ed.

Clutton-Brock, T. H. 1991. The Evolution of Parental Care. Princeton University Press, Princeton, NJ.

Cornforth, D. M., D. J. T. Sumpter, S. P. Brown, and Å. Brännström. 2012. Synergy and Group Size in Microbial Cooperation. Am. Nat. 180:pp. 296–305.

Danielson-François, A. M., Y. Zhou, and M. D. Greenfield. 2009. Indirect genetic effects and the lek paradox: inter-genotypic competition may strengthen genotype x environment interactions and conserve genetic variance. Genetica 136:27–36.

Ellen, E. D., J. Visscher, J. A. M. v. Arendonk, and P. Bijma. 2008. Survival of Laying Hens: Genetic Parameters for Direct and Associative Effects in Three Purebred Layer Lines. Poult. Sci. 87:233–239.

Eshel, I. 1983. Evolutionary and continuous stability. J. Theor. Biol. 103:99 – 112.

Eshel, I., and U. Motro. 1981. Kin selection and strong evolutionary stability of mutual help. Theor. Popul. Biol. 19:420–433.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. The Clarendon Press, Oxford.

Fletcher, J. A., and M. Doebeli. 2009. A simple and general explanation for the evolution of altruism. Proc. R. Soc. B 276:13–9.

Fletcher, J. A., and M. Zwick. 2006. Unifying the theories of inclusive fitness and reciprocal altruism. Am. Nat. 168:252–62.

Foster, K. R., and T. Wenseleers. 2006. A general model for the evolution of mutualisms. J. Evol. Biol. 19:1283–1293.

Frank, S. A. 1994. Genetics of mutualism: the evolution of altruism between species. J Theor Biol 170:393–400. URL http://dx.doi.org/10.1006/jtbi.1994.1200.

Frank, S. A. 1997. The Price Equation, Fisher's fundamental theorem, kin selection, and causal analysis. Evolution 51:1712 – 1729.

Frère, C. H., M. Krützen, J. Mann, R. C. Connor, L. Bejder, and W. B. Sherwin. 2010. Social and genetic interactions drive fitness variation in a free-living dolphin population. Proc. Natl. Acad. Sci. U. S. A. 107:19949–19954.

Fudenberg, D., and E. Maskin. 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. Econometrica 54:533–554.

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

Gore, J., H. Youk, and A. van Oudenaarden. 2009. Snowdrift game dynamics and facultative cheating in yeast. Nature 459:253–6.

Hamilton, I. M., and I. Y. Ligocki. 2012. The extended personality: indirect effects of behavioural syndromes on the behaviour of others in a group-living cichlid. Anim. Behav. 84:659–664.

Hamilton, W. 1972. Altruism and Related Phenomena, Mainly in Social Insects. Annu. Rev. Ecol. Syst. 3:193–232.

Hamilton, W. D. 1964a. The genetical evolution of social behaviour. I. J. Theor. Biol. 7:1–16.

Hamilton, W. D. 1964b. The genetical evolution of social behaviour. II. J. Theor. Biol. 7:17–52.

Hauert, C., F. Michor, M. A. Nowak, and M. Doebeli. 2006. Synergy and discounting of cooperation in social dilemmas. J. Theor. Biol. 239:195–202.

Holekamp, K. E., J. E. Smith, C. C. Strelioff, R. C. Van Horn, and H. E. Watts. 2012. Society, demography and genetic structure in the spotted hyena. Mol. Ecol. 21:613–632.

Hughes, W. O. H., B. P. Oldroyd, M. Beekman, and F. L. W. Ratnieks. 2008. Ancestral monogamy shows kin selection is key to the evolution of eusociality. Science 320:1213–6.

Lehmann, L., and L. Keller. 2006a. The evolution of cooperation and altruism–a general framework and a classification of models. J. Evol. Biol. 19:1365–1376.

Lehmann, L., and L. Keller. 2006b. Synergy, partner choice and frequency dependence: their integration into inclusive fitness theory and their interpretation in terms of direct and indirect fitness effects. J. Evol. Biol. 19:1426–1436.

Lehmann, L., and F. Rousset. 2010. How life history and demography promote or inhibit the evolution of helping behaviours. Philos. Trans. R. Soc. B 365:2599–2617.

Leslie, P. H. 1948. Some further notes on the use of matrices in population mathematics. Biometrika 35:213–245.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, Mass.

Maynard Smith, J., and G. R. Price. 1973. The logic of animal conflict. Nature 246:15–18.

Maynard Smith, J., and E. Szathmáry. 1995. The major transitions in evolution. W.H. Freeman, Oxford.

McGlothlin, J. W., A. J. Moore, J. B. Wolf, and E. D. Brodie III. 2010. Interacting phenotypes and the evolutionary process. III. Social evolution. Evolution 64:2558–2574.

Michod, R. 2005. On the transfer of fitness from the cell to the multicellular organism. Biol. Philos. 20:967–987. 10.1007/s10539-005-9018-2.

Moore, A. J., E. D. Brodie, III, and J. B. Wolf. 1997. Interacting Phenotypes and the Evolutionary Process: I. Direct and Indirect Genetic Effects of Social Interactions. Evolution 51:1352–1362.

Moore, A. J., K. F. Haynes, R. F. Preziosi, and P. J. Moore. 2002. The evolution of interacting phenotypes: genetics and evolution of social dominance. Am. Nat. 160 Suppl 6:S186–97.

Mutic, J. J., and J. B. Wolf. 2007. Indirect genetic effects from ecological interactions in *Arabidopsis thaliana*. Mol. Ecol. 16:2371–2381.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. Science 314:1560–3.

Nowak, M. A., and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. Nature 393:573–577.

Ohtsuki, H. 2012. Does synergy rescue the evolution of cooperation? An analysis for homogeneous populations with non-overlapping generations. J. Theor. Biol. 307:20 – 28.

Pepper, J. W. 2000. Relatedness in trait group models of social evolution. J. Theor. Biol. 206:355–68.

Persson, O., and P. Öhrström. 1989. A New Avian Mating System: Ambisexual Polygamy in the Penduline Tit *Remiz pendulinus*. Ornis Scandinavica 20:105–111.

Petfield, D., S. F. Chenoweth, H. D. Rundle, and M. W. Blows. 2005. Genetic variance in female condition predicts indirect genetic variance in male sexual display traits. Proc. Natl. Acad. Sci. U. S. A. 102:6045–50.

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

Price, G. R. 1972. Extension of covariance selection mathematics. Ann. Hum. Genet. 35:485–490.

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

Queller, D. C. 1992a. A general model for kin selection. Evolution 46:376 – 380.

Queller, D. C. 1992b. Quantitative Genetics, Inclusive Fitness, and Group Selection. Am. Nat. 139:540–558.

Queller, D. C. 1994. Genetic relatedness in viscous populations. Evol. Ecol. 8:70–73.

Queller, D. C. 2011. Expanded social fitness and Hamilton's rule for kin, kith, and kind. Proc. Natl. Acad. Sci. U. S. A. 108:10792–10799.

Queller, D. C., E. Ponte, S. Bozzaro, and J. E. Strassmann. 2003. Single-gene greenbeard effects in the social amoeba *Dictyostelium discoideum*. Science 299:105–106.

Queller, D. C., and J. E. Strassmann. 1998. Kin selection and social insects. Bioscience 48:165–175.

Rousset, F. 2003. A minimal derivation of convergence stability measures. J. Theor. Biol. 221:665–668.

Rousset, F. 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton, New Jersey.

Rousset, F., and S. Billiard. 2000. A theoretical basis for measures of kin selection in subdivided populations: Finite populations and localized dispersal. J. Evol. Biol. 13:814 – 825.

Rousset, F., and O. Ronce. 2004. Inclusive fitness for traits affecting metapopulation demography. Theor. Popul. Biol. 65:127–141.

Sachs, J. L., U. G. Mueller, T. P. Wilcox, and J. J. Bull. 2004. The evolution of cooperation. Q. Rev. Biol. 79:135–160.

Schino, G., F. Di Giuseppe, and E. Visalberghi. 2009. Grooming, rank, and agonistic support in tufted capuchin monkeys. Am. J. Primatol. 71:101–5.

Skyrms, B. 2001. The Stag Hunt. Proc. Adresses Am. Philos. Assoc. 75:pp. 31–41.

Soares, M. C., R. Bshary, L. Fusani, W. Goymann, M. Hau, K. Hirschenhauser, and R. F. Oliveira. 2010. Hormonal mechanisms of cooperative behaviour. Philos. Trans. R. Soc. B 365:2737–2750.

Szentirmai, I., T. Székely, and J. Komdeur. 2007. Sexual conflict over care: antagonistic effects of clutch desertion on reproductive success of male and female penduline tits. J. Evol. Biol. 20:1739–1744.

Tarnita, C. E., H. Ohtsuki, T. Antal, F. Fu, and M. A. Nowak. 2009. Strategy selection in structured populations. J. Theor. Biol. 259:570–81.

Taylor, C., and M. A. Nowak. 2007. Transforming the dilemma. Evolution 61:2281–2292.

Taylor, P. 1989. Evolutionary stability in one-parameter models under weak selection. Theor. Popul. Biol. 36:125–143.

Taylor, P., and W. Maciejewski. 2012. An inclusive fitness analysis of synergistic interactions in structured populations. Proc. R. Soc. B 279:4596–4603.

Taylor, P. D. 1990. Allele-frequency change in a class-structured population. Am. Nat. 135:95 – 106.

Teplitsky, C., J. A. Mills, J. W. Yarrall, and J. Merilä. 2010. Indirect genetic effects in a sex-limited trait: the case of breeding time in red-billed gulls. J. Evol. Biol. 23:935–944.

Trivers, R. L. 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46:35–57.

Valera, F., H. Hoi, and B. Schleicher. 1997. Egg burial in penduline tits, *Remiz pendulinus*: its role in mate desertion and female polyandry. Behav. Ecol. 8:20–27.

Van Cleve, J., and L. Lehmann. 2013. Stochastic stability and the evolution of coordination in spatially structured populations. Submitted .

Wakano, J. Y., and L. Lehmann. 2012. Evolutionary and convergence stability for continuous phenotypes in finite populations derived from two-allele models. J. Theor. Biol. 310:206 – 215.

Wenseleers, T. 2006. Modelling social evolution: the relative merits and limitations of a Hamilton's rule-based approach. J. Evol. Biol. 19:1419–1422.

Wesołowski, T. 1994. On the origin of parental care and the early evolution of male and female parental roles in birds. Am. Nat. 143:39–58.

West, S. A., A. Griffin, and A. Gardner. 2007. Evolutionary explanations for cooperation. Curr. Biol. 17:R661–R672.

White, C., S. Winans, C. White, and S. Winans. 2007. Cell–cell communication in the plant pathogen *Agrobacterium tumefaciens*. Philos. Trans. R. Soc. B 362:1135–1148.

Wilkinson, G. S. 1984. Reciprocal food sharing in the vampire bat. Nature 308:181–184.

Wilson, A. J., M. B. Morrissey, M. J. Adams, C. A. Walling, F. E. Guinness, J. M. Pemberton, T. H. Clutton-Brock, and L. E. B. Kruuk. 2011. Indirect genetics effects and evolutionary constraint: an analysis of social dominance in red deer, *Cervus elaphus*. J. Evol. Biol. 24:772–83.

Wolf, J. B., E. D. Brodie, III, and A. J. Moore. 1999. Interacting phenotypes and the evolutionary process. II. Selection resulting from social interactions. Am. Nat. 153:254–266.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97–159.

Wright, S. 1949. The genetical structure of populations. Ann. Eugen. 15:323–354.

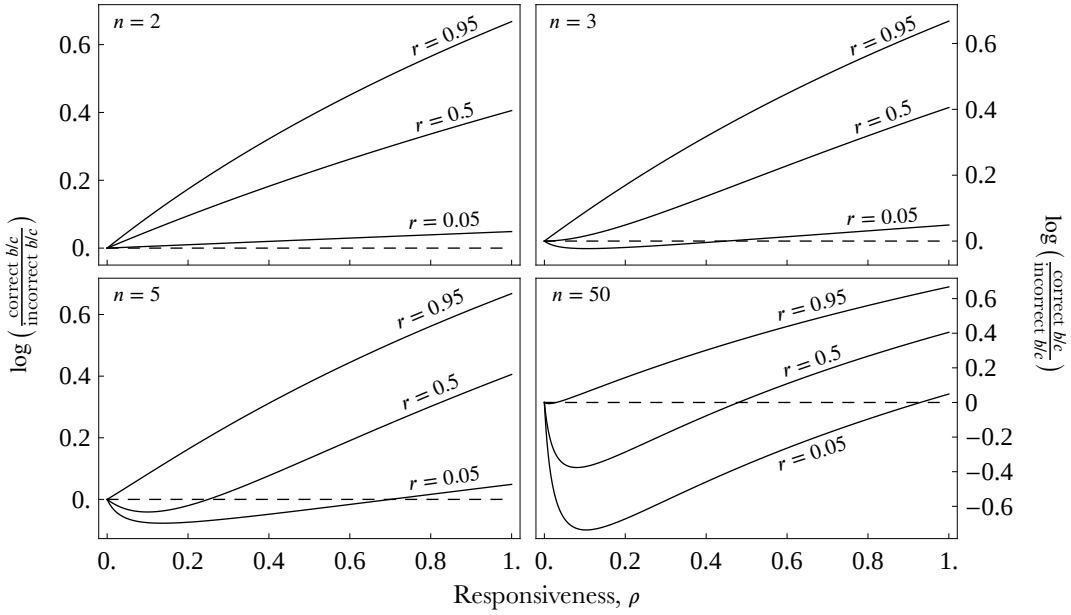

Figure 1: These panels plot the log ratio of the correct b/c expression given in (10) to an incorrect one given in (12) that lacks products of relatedness and responsiveness. Group sizes n are given in each panel. Each curve represents a different value of relatedness r where the lower curve has $r = 0.05$, the middle curve $r = 0.5$, and the upper curve $r = 0.95$.

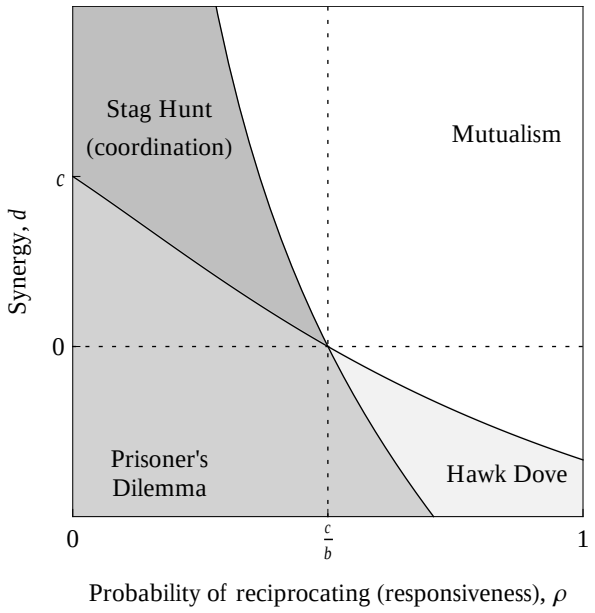

Figure 2: The different possibilities for how a Prisoner's Dilemma game with reciprocity and non-additive payoffs can get transformed, depending on the probabilities of reciprocation ρ and the non-additive payoff or synergy component d. Pure cooperation is unstable in the Prisoner's Dilemma, and pure defection is unstable in the Mutualism game. Both pure cooperation and defection are stable in the Stag-Hunt (or coordination) game. Pure cooperation and defection are unstable in the Hawk-Dove game leaving an intermediate level of both stable (stable mixed strategy or polymorphism).

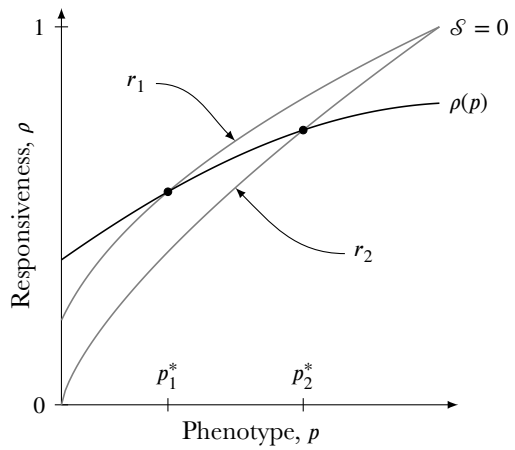

Figure 3: The graphical relationship between candidate evolutionarily stable response coefficients (ρ) as a function of phenotype (p) and the response coefficients generated by a hypothetical proximate mechanism that we label $\rho(p)$. The candidate evolutionary stable values of ρ and p solve $\mathcal{S} = 0$ (see equation 10) and are plotted as gray curves for two hypothetical values of relatedness, $r_1 < r_2$. The proximate mechanism above generates convergence stable phenotypes, p_1^* and p_2^*, since its response coefficient curve crosses the $\mathcal{S} = 0$ curves from above (see equation 24).