Exploring an Evolutionary Paradox: An Analysis of the "Spite Effect" and the "Nearly Neutral Effect" in Synergistic Models of Finite Populations

Abstract:

Forber and Smead (2014) analyze how increasing the fitness benefits associated with prosocial behavior can increase the fitness of spiteful individuals relative to their prosocial counterparts, so that selection favors spite over prosociality. This poses a problem for the evolution of prosocial behavior: as the benefits of prosocial behavior increase, it becomes more likely that spite, not prosocial behavior, will evolve in any given population. In this paper, I develop two game-theoretic models which, taken together, illustrate how synergistic costs and benefits may provide partial solutions to Forber and Smead's paradox.

1. Overview

To date, much of the evolutionary game theory literature has focused on the evolution of prosocial behavior, and relatively less attention has been devoted to the evolution of spite. Nevertheless, in a groundbreaking 2014 paper, Forber and Smead suggest that spite may pose problems for the evolution of prosociality. More specifically, their analysis indicates that – even in a "best-case scenario" where the evolution of prosocial behavior should, in theory, be relatively easy to achieve – spite retains a fitness advantage; and that the greater the benefits of prosocial behavior are, the greater the fitness advantage spiteful individuals have over their prosocial counterparts. As Forber and Smead (2014) point out, the results of their analysis entail a paradox for the evolution of prosocial behavior: as the benefits of prosocial behavior grow larger and larger, it becomes more and more likely that spite – not prosocial behavior – will evolve in any given population. How, then, are we to reconcile the results of their analysis with the modern-day prevalence of prosocial behavior, both throughout the animal kingdom and human society?

In this paper, I attempt to provide a solution to Forber and Smead's paradox. I do this by introducing two new game-theoretic models; both are similar to the model used by Forber and Smead (2014), but involve some additional variables to describe synergistic effects. Specifically, I explore whether it is less difficult for prosocial behavior to evolve if (1) mutual spiteful behavior incurs additional costs or (2) mutual prosocial behavior earns additional benefits. I show that both synergistic models provide partial solutions to Forber and Smead's

paradox, by increasing the fitness of prosocial individuals relative to the fitness of their spiteful counterparts.

2. Introduction

Although evolution typically entails competition for limited resources, the biological world nevertheless abounds with examples of prosocial behavior, or social behavior that benefits other individuals. One form of prosocial behavior, referred to as altruism, involves performing an action that is beneficial to another individual, but detrimental to oneself (Smead and Forber 2012). The study of altruism has attracted significant attention from a variety of disciplines, including philosophy of biology, evolutionary biology, and anthropology.

At first glance, it appears that altruistic behavior is unlikely to evolve. If altruism involves an inherent fitness cost for the actor, how could selection favor altruistic individuals? To answer this question, we can analyze the evolution of altruism using the game commonly known as the prisoner's dilemma. Using a value of one to represent the baseline fitness, b to represent the benefit received from interacting with an altruistic individual, and c to represent the self-inflicted cost of altruism, we obtain the payoffs shown in Table 1:

	altruist	defector
altruist	$1+b-c$	$1-c$
defector	$1+b$	1

Table 1. $b>0$, $c>0$

In this scenario, it is assumed that individuals interact in pairs. Each individual can choose between one of two behavior strategies: perform an altruistic action, or refuse to perform an action at all (defect). Because altruism involves an inherent cost (c), it appears that individuals should always prefer to defect, rather than perform an altruistic action that incurs this cost. In other words, the defection strategy strictly dominates the game (Skyrms 1996; Sober 1992).

However, this interpretation of the prisoner's dilemma assumes that social interactions occur at random in the population; this may not, in fact, be the case (Hamilton 1964). Rather, it is possible for social interactions to occur in a nonrandom, correlated fashion. For example, if altruistic individuals can recognize, and preferentially interact with, other altruistic individuals, there will be correlated interactions between individuals playing the same behavior strategies. In other words, altruistic individuals will be more likely to interact with other altruistic individuals, while defectors will be more likely to interact with other defectors (Smead and Forber 2012).

Generally speaking, the evolution of altruism requires (1) correlated interactions and (2) that altruists receive, with a sufficient degree of likelihood, a benefit that outweighs the cost of their own altruistic behavior. Hamilton expressed these requirements with the inequality $r*b>c$, where r is the degree of correlation, also known as the "coefficient of relatedness." While the coefficient of relatedness is often used to describe correlated interactions among genetic relatives, the same concept may be applied to correlated interactions among specific behavior strategies (Okasha 2002). The larger the coefficient of relatedness, the greater the degree of correlation.

While the evolution of altruistic behavior has been heavily studied, far less attention has been given to the evolution of spite. Spite, which has been referred to as "the dark side of cooperation," can evolve when selfish individuals "exploit the system." Spiteful individuals benefit from interacting with prosocial individuals, but pay a personal fitness cost in order to withhold benefits from those they interact with (Jensen 2010). This definition of spite requires that the spiteful individual incur some cost from its own spiteful behavior. However, it is still possible for spite to be advantageous, in an evolutionary sense, if the benefits accrued from interacting with prosocial individuals outweigh this cost.

As discussed above, correlated interactions are a key factor in the evolution of altruistic behavior. Similarly, anti-correlated interactions are necessary for the evolution of spite. In other words, spite can only evolve when a spiteful individual is more likely to interact with a

prosocial individual, and vice-versa. In finite populations, a slight degree of anti-correlation is always present, and this is because each individual is unable to interact with itself; this effect may be likened to "sampling without replacement." This anti-correlation opens the door to the evolution of spite, leaving populations of prosocial individuals vulnerable to invasion by spiteful individuals (Forber and Smead 2014; Birch 2017).

To summarize, if (1) the cost of performing a spiteful behavior is relatively small, compared to the benefit gained by exploiting prosocial individuals; and (2) there is sufficient anti-correlation among behavior strategies; then (3) it is possible for the fitness of spiteful individuals to exceed the fitness of prosocial individuals, meaning that selection favors spite. Because finite populations always entail some degree of anti-correlation, it is possible for spite to evolve in a finite population, if the benefits exceed the costs.

3. An Evolutionary Paradox

In their 2014 paper, "An Evolutionary Paradox for Prosocial Behavior," Forber and Smead analyze how increasing the benefit from prosocial behavior impacts the fitness of spiteful individuals, relative to the fitness of prosocial individuals. Assuming pairwise interactions in a finite population, Forber and Smead model a game-theoretic scenario similar to the one described in the previous section. However, rather than analyzing a prisoner's-dilemma-style game, Forber and Smead propose a scenario in which mutual prosocial behavior should, theoretically, be less difficult to achieve. In such a game, sometimes referred to as a prisoner's

delight, the greatest payoffs are earned when both individuals play the prosocial strategy [1] (Binmore 2007; Skyrms 2008).

In Forber and Smead's model, individuals are assumed to be one of two types – either prosocial or spiteful – and it is assumed that both behavior strategies are played unconditionally. In this model, interacting with a prosocial individual confers a benefit on the recipient, but prosocial individuals cannot benefit from their own prosocial behavior. Spiteful individuals benefit from interacting with prosocial individuals, while simultaneously incurring a cost due to their own spiteful behavior. Due to the inherent costs associated with spite, one might expect selection to always favor prosocial behavior. However, Forber and Smead's analysis reveals a paradoxical result: increasing the benefit of prosocial behavior can actually increase the fitness of spiteful individuals, so that selection favors spite over prosociality. Forber and Smead describe two effects that drive this relative increase in the fitness of spiteful individuals. The first is the "spite effect," or the tendency of spiteful individuals in a finite population to reap disproportionate rewards from others' prosocial behavior, due to anti-correlated interactions; because prosocial individuals are more likely to interact with spiteful individuals than they are to interact with other prosocial individuals, spiteful individuals are more likely (than their prosocial counterparts) to receive the fitness benefits of prosocial behavior. The second is the "nearly neutral effect," or the tendency for spiteful behavior to

[1] Payoff table given in next section.

evolve by chance when the benefits of prosocial behavior are large and the fitness difference between the two behavior strategies is small; in such a situation, selection is "nearly neutral" with respect to behavior strategies, and the question of which strategy evolves in a given population is largely dependent on drift. As I will explore later in this paper, it is possible to isolate the nearly neutral effect (or, in other words, to remove the spite effect) by removing the anti-correlations from the fitness functions[2] (Forber and Smead 2014).

In this paper, I will analyze both the spite effect and the nearly neutral effect by considering two other game-theoretic scenarios. Both scenarios share many characteristics with the model developed by Forber and Smead (2014), while including additional variables to describe the effects of synergism. First, I will analyze a synergistic-cost model, in which spiteful individuals incur an additional cost whenever they interact with each other.[3] Then, I will analyze a synergistic-benefit model, in which mutual prosocial behavior earns an additional

[2] Fitness functions given in next section.

[3] Mutual spite may incur synergistic costs. For example, imagine that spiteful individuals quarrel/fight whenever they interact with each other. Each individual incurs a cost due to its own spiteful behavior (the "inherent" cost of spite), and additionally incurs synergistic costs (e.g., bodily injury, time and energy spent fighting).

benefit.[4] I will show that each of these synergistic models provides a partial solution to Forber and Smead's paradox, by increasing the fitness of prosocial individuals relative to their spiteful counterparts.

A paper by Ventura (2019) likewise analyzed how the addition of synergistic effects to Forber and Smead's model can provide a partial solution to the evolutionary paradox. However, Ventura's definition of synergism differs from mine, with the result that our analyses and proposed solutions to the paradox are quite different. In Ventura's model, the benefit of prosocial behavior is synergistic in the sense that it is a nonlinear function of the number of prosocial individuals, with the benefit decreasing as the number of prosocial individuals approaches the total population size. In my analysis, the benefit of prosocial behavior and the cost of spiteful behavior are not dependent on the number of either type of individual in the population. Rather, in my models, the cost of spiteful behavior is synergistic in the sense that a spiteful individual incurs an additional cost whenever it plays its spiteful strategy against another spiteful individual; similarly, the benefit of prosocial behavior is synergistic in the sense that a prosocial individual reaps an additional benefit when interacting with another

[4] There are several ways in which mutual prosocial behavior may involve synergistic benefits. An example is cooperative hunting, in which prosocial individuals may be able to kill more/larger prey by working in pairs/groups, resulting in more food (benefits) for all the individuals involved.

prosocial individual. The overall result is that mutual prosocial behavior is incentivized, mutual spiteful behavior is disincentivized, and the paradoxical effects noted by Forber and Smead (2014) are dampened to an extent.

4. A Game-Theoretic Model for the Evolution of Spiteful Behavior in Finite Populations

Forber and Smead (2014) investigate the evolution of spite in finite populations. As discussed in the previous section, they assume a game-theoretic scenario referred to as a prisoner's delight, in which the greatest net benefits are gained through mutual prosocial behavior. Their analysis assumes pairwise interactions, with the payoffs shown in Table 2.

	prosocial	spiteful
prosocial	b	1
spiteful	$b - a$	$1 - a$

Table 2.$b>a>0$

Individuals are assumed to be one of two types, prosocial or spiteful. Payoffs are given for the player on the left-hand side. Prosocial behavior confers a benefit b on the recipient, while spite is defined as the withholding of this benefit b, at a cost a to oneself. If two prosocial individuals interact, then both individuals receive the benefit b. If two spiteful individuals interact, they both receive the baseline fitness (1), minus the cost, a, of their own spiteful behavior. If two individuals of different types interact, the prosocial individual receives the baseline fitness (1), while the spiteful individual receives the benefit b (from interacting with

a prosocial individual) minus the cost, *a,* of its own spiteful behavior. In terms of Hamilton's inequality ($r*b>a$), we can say that the evolution of spite entails both a negative benefit (because spiteful individuals withhold benefits from others) and a negative coefficient of relatedness, meaning that the product $r*b$ is a positive quantity (Birch, 2017). In other words, spite can have a fitness advantage if spiteful individuals withhold benefits from prosocial individuals, but do not withhold the same benefits from other spiteful individuals.

Using the payoffs given in Table 2, Forber and Smead (2014) calculate the fitness functions of both types:

$$(1) \quad F(p,N) = \frac{b(x_p - 1) + x_s}{N-1}$$

$$(2) \quad F(s,N) = \frac{(b-a)x_p + (1-a)(x_s - 1)}{N-1}$$

N represents the finite population size, while the numbers of prosocial and spiteful individuals are represented by x_p and x_s, respectively. Prosocial behavior is favored when $F(p,N)>F(s,N)$, while spite is favored when $F(p,N)<F(s,N)$. As discussed in the previous section, when the fitnesses of the two types are approximately equal, selection is relatively neutral with respect to behavior strategies, and drift plays the more significant evolutionary role (Forber and Smead 2014). Because all interactions occur within a finite population, there is a small degree of anti-correlation between behavior strategies; as discussed above, this anti-correlation opens

the door to the evolution of spite, by enabling spiteful individuals to benefit disproportionately from others' prosocial behavior. As Forber and Smead's analysis shows, increasing the benefit (b) of prosocial behavior can actually increase the fitness of spiteful individuals, relative to the fitness of prosocial individuals, so that $F(p,N)<F(s,N)$. Substituting the fitness functions into the two inequalities given above, we see that prosociality is favored in this model when $b < a(N-1)+1$, while spite is favored when $b > a(N-1)+1$. Forber and Smead (2014) conduct simulations to show that, as the benefit (b) increases, spiteful behavior evolves in an increasingly large proportion of populations, and the effect is strongest for small (e.g., $N = 25$) population sizes.

In the final portion of their paper, Forber and Smead (2014) remove the anti-correlations from the fitness functions, in order to analyze the nearly neutral effect in isolation; that is to say, they remove the spite effect from the simulations by getting rid of the slight degree of anti-correlation that results from the finite population size. To facilitate this analysis, Forber and Smead first introduce the concept of a selection coefficient, which is a commonly used method for comparing the fitnesses of two traits.[5] Using s_c to represent the selection coefficient, and the fitness of prosocial individuals as baseline, the relative fitnesses of the two behavior strategies are as follows:

[5] Prosocial behavior is favored when the selection coefficient is greater than zero, and spiteful behavior is favored when the selection coefficient is less than zero.

(3) $F_r(p, N) = 1$

(4) $F_r(s, N) = \dfrac{F(s, N)}{F(p, N)} = 1 - s_c$

The symbol F_r is used to denote relative fitness. Substituting equations (1) and (2) for the fitnesses of the two types, Forber and Smead rearrange equation (4) to yield the following:

(5) $s_c = \dfrac{a(N-1) - b + 1}{b(x_p - 1) + x_s}$

Observe what happens when the benefit (b) of prosocial behavior is increased. Because b is subtracted from the numerator of equation (5), but is a product in the denominator, increasing b decreases the selection coefficient. Decreasing the selection coefficient brings the ratio of the fitnesses [$F(s,N)/F(p,N)$] close to one, meaning that there is only a slight fitness difference between the two behavior strategies. As discussed above, when the fitness difference between the two strategies is small, selection is effectively neutral, and the direction of evolution in any given population is governed primarily by drift (Forber and Smead, 2014).

Forber and Smead next apply this concept of a "nearly neutral model" to their original analysis, in order to observe how increasing the benefit (b) of prosocial behavior can cause

the evolutionary dynamics to become more neutral in character (Forber and Smead, 2014). To do this, they isolate the nearly neutral effect by removing the anti-correlations from the original fitness functions. The revised fitness functions are then as follows:

$$(6) \quad F(p,N) = \frac{b(x_p) + x_s}{N}$$

$$(7) \quad F(s,N) = \frac{(b-a)x_p + (1-a)x_s}{N}$$

Forber and Smead then repeat their simulations, using the revised fitness functions; the result is that, when the nearly neutral effect is isolated, spiteful behavior generally evolves less frequently for any given value of *b*. However, spite still becomes more and more likely to evolve in a given population as the benefits of prosocial behavior increase, with the likelihood approaching a limit of 0.5 as the fitness difference between the two strategies approaches zero. Plugging Eqs. 6 and 7 into the inequality *F(s,N)>F(p,N)*, we see that selection will never favor spite as long as the cost of spite is greater than zero ($a > 0$). Nevertheless, increasing the benefit (*b*) of prosocial behavior decreases the fitness difference between the two behavior types, making it more likely that spite will evolve in a given population due to drift alone.

5. The Synergistic-Cost Model

In the following analysis, I will consider two other game-theoretic scenarios. Both retain the general format used by Forber and Smead (2014), while introducing additional variables to describe synergistic effects. The first scenario involves a synergistic cost, *c*, which is incurred by two spiteful individuals whenever they interact with each other (Table 3).

	prosocial	spiteful
prosocial	b	1
spiteful	$b - a$	$1 - a - c$

Table 3.$b>a>0$, $c>0$

The fitness functions for this scenario are as follows:

$$(1) \quad F(p,N) = \frac{b(x_p - 1) + x_s}{N - 1}$$

$$(8) \quad F(s,N) = \frac{(b-a)x_p + (1-a-c)(x_s - 1)}{N - 1}$$

Plugging these formulas into the inequality $F(s,N)>F(p,N)$, we see that selection favors spite when $b > ax_p + a(x_s - 1) + c(x_s - 1)+1$; similarly, selection favors prosocial behavior when

$b < ax_p + a(x_s - 1) + c(x_s - 1) + 1$. Consider a population size of $N = 100$, with $a = 0.1$ and $c = 0.1$ (Fig. 1).[6] Observe that increasing the benefit (b) of prosocial behavior decreases the fitness of prosocial individuals; this is the paradox that Forber and Smead (2014) observed. At relatively low values of b, the selection coefficient remains above the x-axis, meaning that the effect is not strong enough for selection to actually favor spite. However, as shown in Fig. 2, at higher values of b (e.g., $b = 26$), the selection coefficient drops below the x-axis, indicating that selection favors spite.

[6] A value of $a = 0.1$ was also used by Forber and Smead (2014).

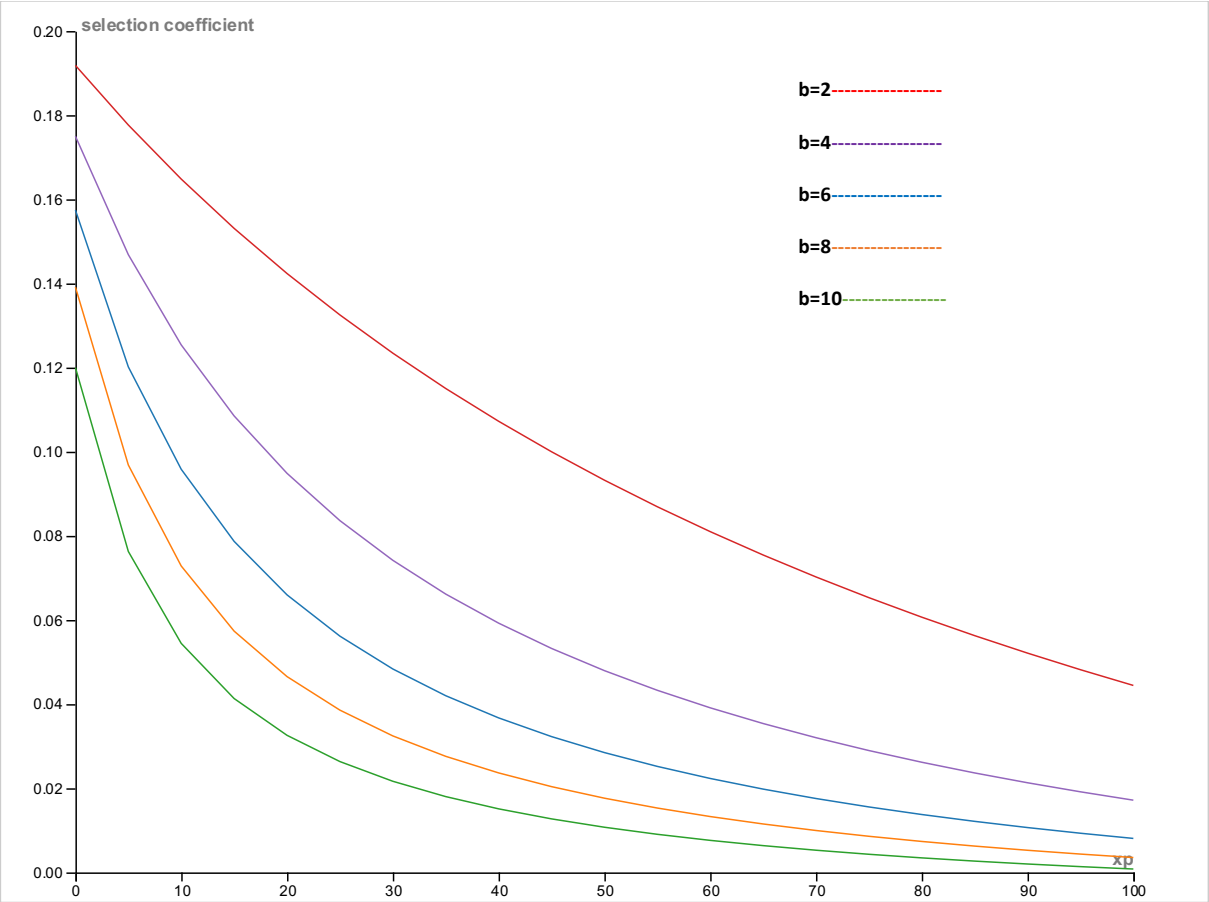

Fig.1. In all figures, a=0.1 and N=100. x_p on x-axis; s_c on y-axis.

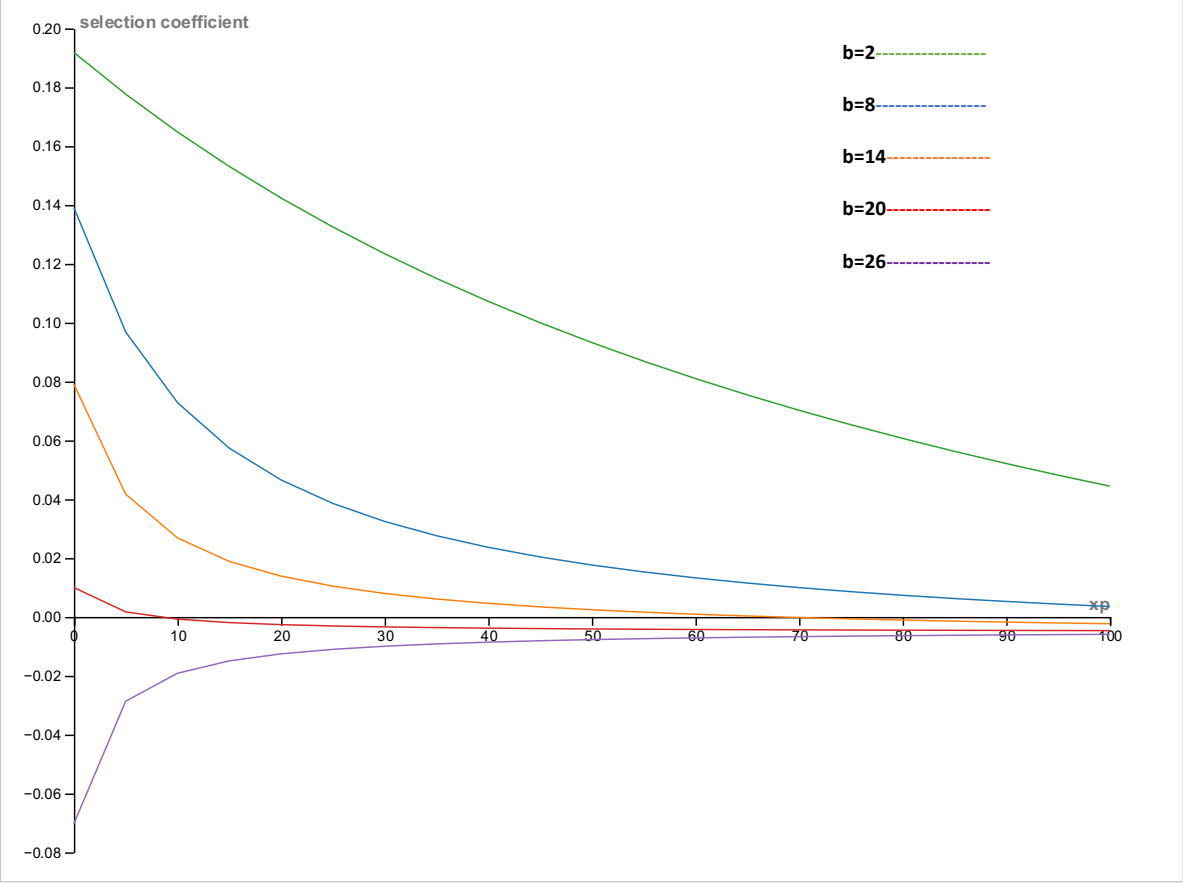

Fig.2

Nevertheless, we see that the introduction of a synergistic cost provides a partial solution to the evolutionary paradox. If, as shown in Fig. 3, we set $c = 0$, with $N = 100$ and $a = 0.1$, and increase the benefit (b) of prosocial behavior, the selection coefficient becomes negative at relatively low values of b (e.g., $b = 14$).[7] However, if we add a large synergistic cost by setting $c = 1$ (Fig. 4), we see that the selection coefficient generally remains positive, unless

[7] When $c = 0$, there is no synergistic cost.

the value of b is high (e.g., $b = 26$) and the number of prosocial individuals is large (e.g., $x_p = 80$). Therefore, if mutual spiteful behavior incurs synergistic costs, it is easier for prosocial behavior to evolve, and more difficult for spite to gain an evolutionary foothold.

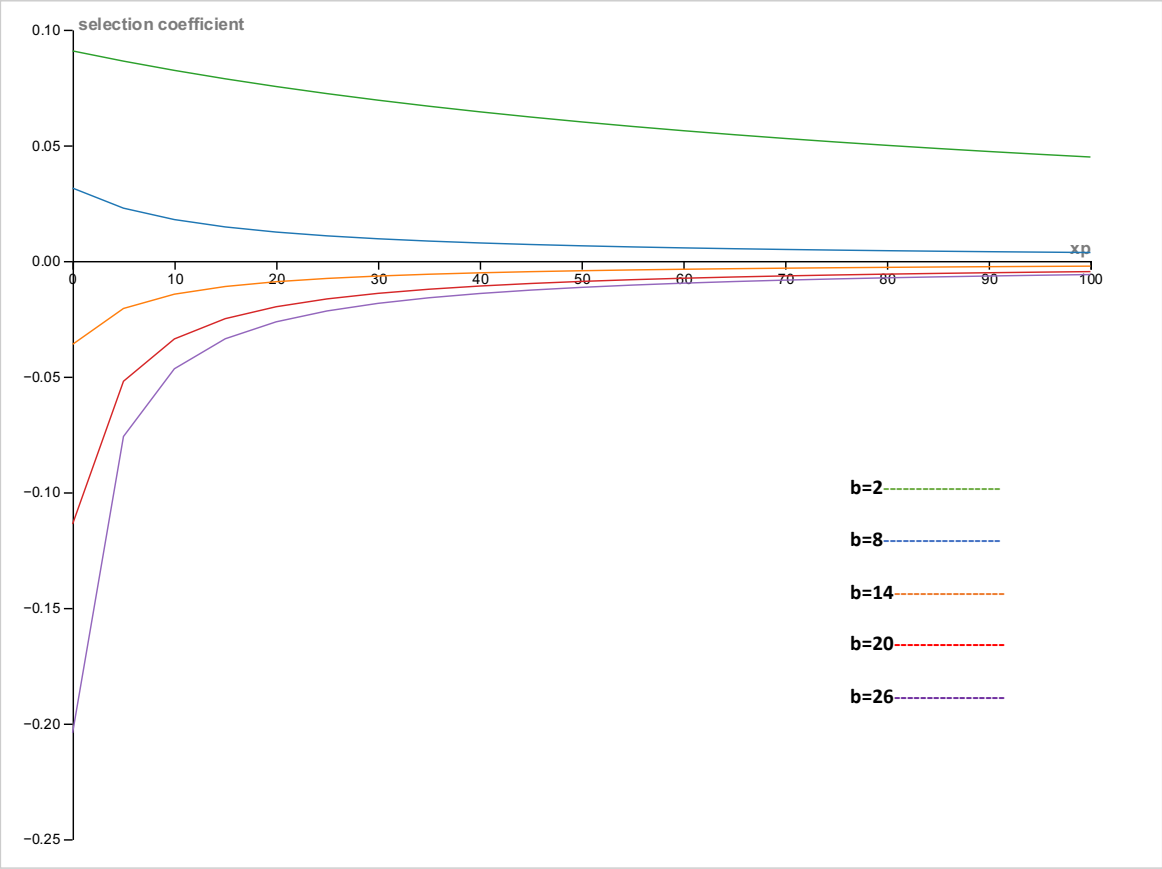

Fig.3

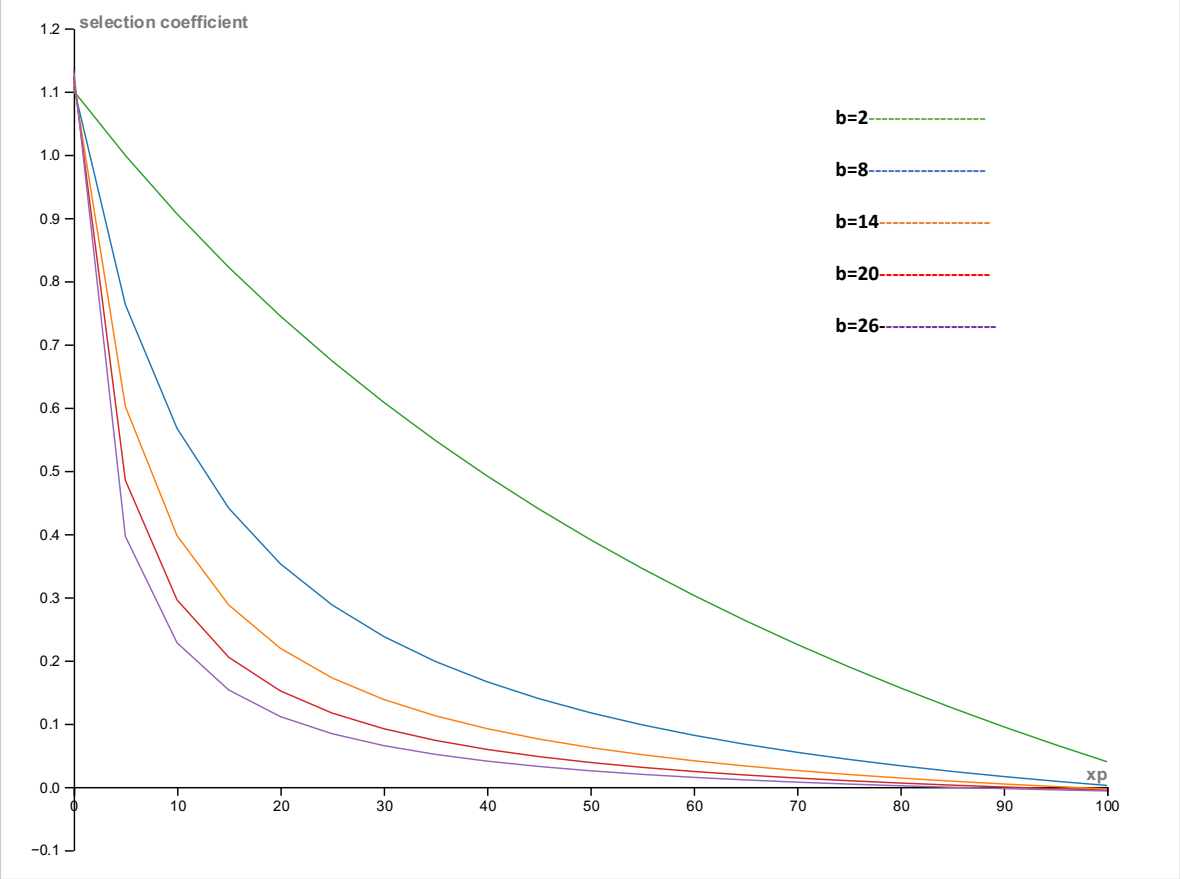

Fig.4

6. The Synergistic-Benefit Model

Having analyzed a synergistic-cost model, let us now turn to a synergistic-benefit model. Suppose that, when two prosocial individuals interact, they mutually confer a synergistic benefit, *d*, on each other. The payoffs for this scenario are given in Table 4.

	prosocial	spiteful
prosocial	$b+d$	1
spiteful	$b-a$	$1-a$

Table 4.$b>a>0$, $d>0$

The fitness functions are now as follows:

$$(9) \quad F(p,N) = \frac{(b+d)(x_p - 1) + x_s}{N-1}$$

$$(2) \quad F(s,N) = \frac{(b-a)x_p + (1-a)(x_s - 1)}{N-1}$$

Now, we see that selection favors spite when $b > d(x_p - 1) + a(N-1)+1$, and that selection favors prosociality when $b < d(x_p - 1) + a(N-1)+1$. Once again, the evolutionary paradox emerges; as the benefit (b) of prosocial behavior increases, the selection coefficient becomes smaller and smaller, and selection can favor spite at very large values of b. Nevertheless, we see that the inclusion of a synergistic benefit, like the inclusion of a synergistic cost, has the potential to significantly diminish these paradoxical effects. Compare Fig. 5 ($d = 0$) to Fig. 6 ($d = 1$). When there is no synergistic benefit (Fig. 5), and we increase the value of b, the paradox is noticeably present; at larger values of b, the selection coefficient falls just below the x-axis, giving spite a slight fitness advantage. However, in the presence of a large

synergistic benefit (Fig. 6, $d = 1$), the selection coefficient generally remains above the x-axis at larger values of b (e.g., $b = 26$); spite is generally selected against, especially if x_p is large.

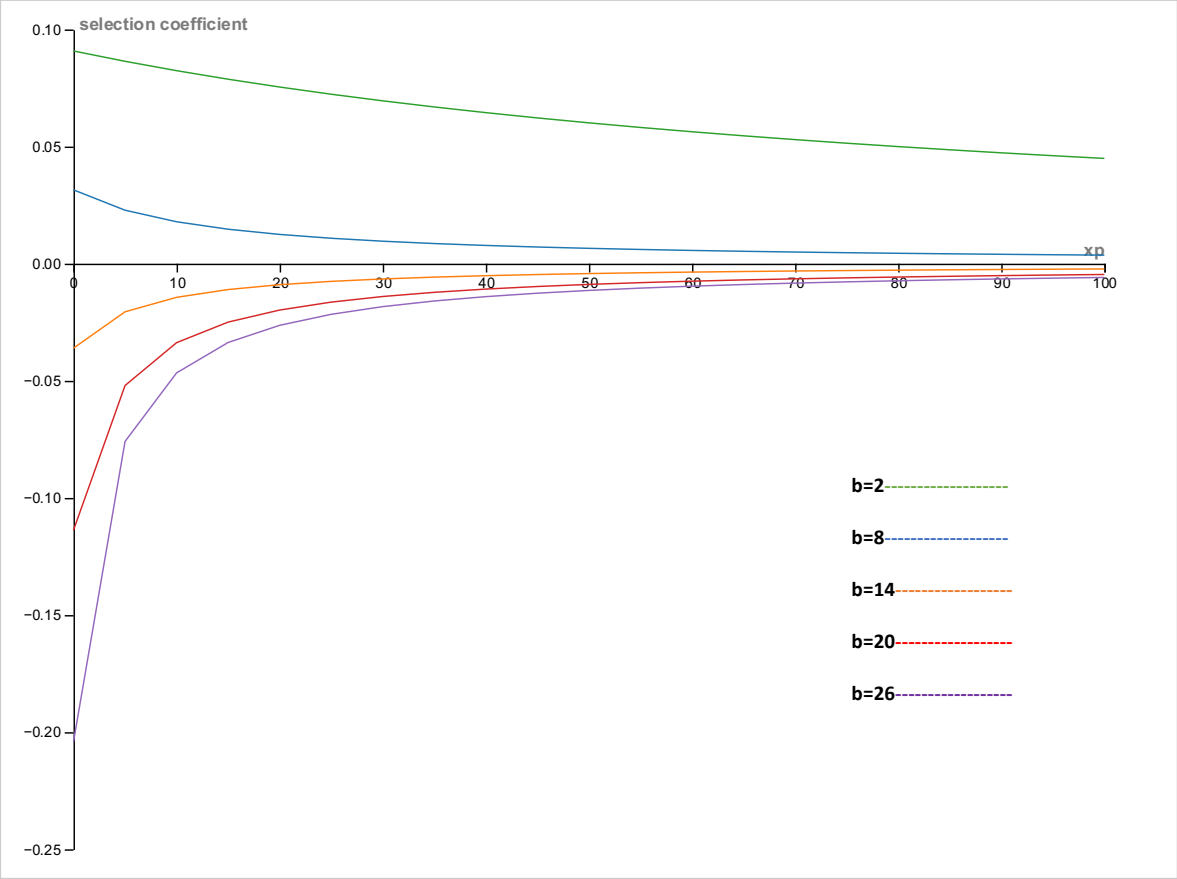

Fig.5

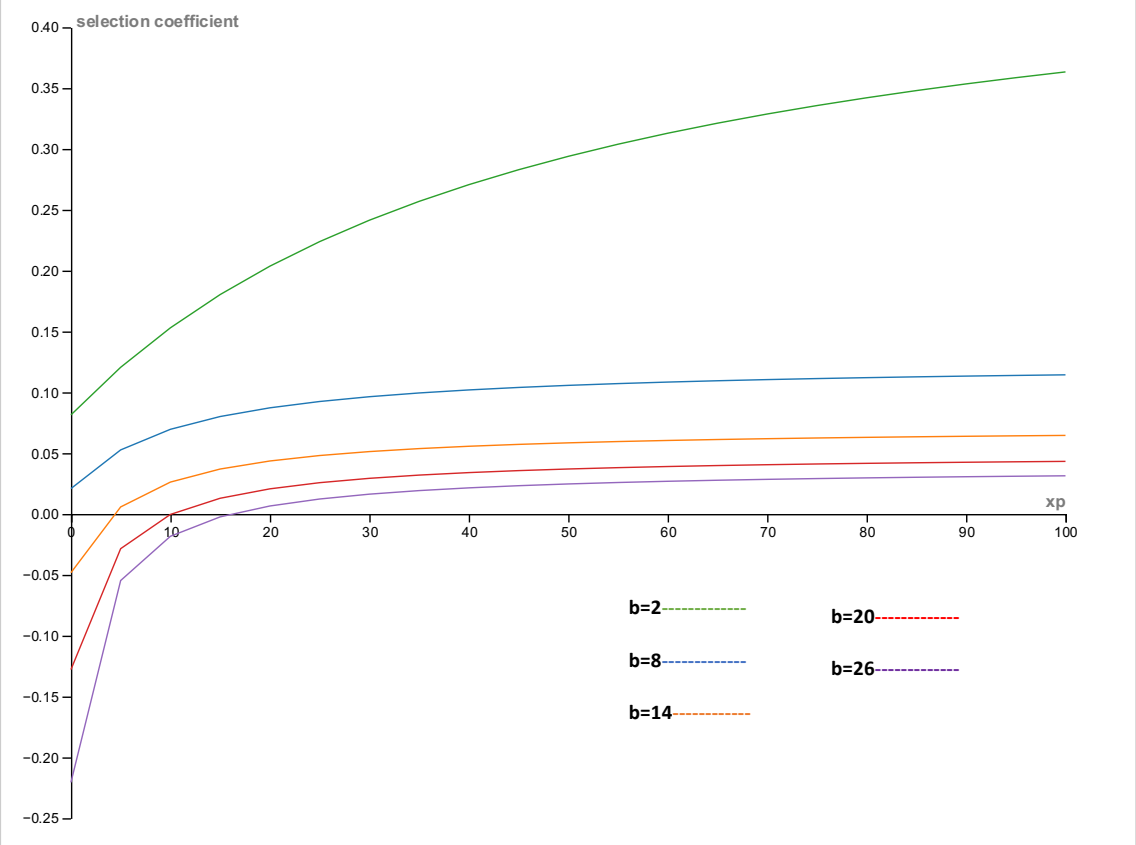

Fig.6

Because only prosocial individuals receive synergistic benefits, the addition of a synergistic benefit generally gives prosocial individuals a fitness advantage over their spiteful counterparts. For the above reasons we see that, like the synergistic-cost model, the synergistic-benefit model provides a partial solution to Forber and Smead's paradox.

7. Isolating the Nearly Neutral Effect

Recall that, in their analysis, Forber and Smead (2014) isolate the nearly neutral effect by removing the anticorrelations from the fitness functions. The result is that, as the benefit (b) of prosocial behavior increases, the fitness difference between the two behavior strategies approaches zero, and selection behaves in a "nearly neutral" manner. That is to say, when the fitness difference between the prosocial and spiteful strategies is very small, selection does not strongly favor one strategy over the other, and the question of which strategy evolves in a given population is largely dependent on drift.

I have already demonstrated that when mutual prosocial behavior earns synergistic benefits, or when mutual spiteful behavior incurs synergistic costs, selection is less likely to favor spite over prosociality, even when the benefit of prosocial behavior is rather large. Next I will show that, when the nearly neutral effect is isolated, selection cannot favor spite if mutual spiteful behavior incurs synergistic costs or mutual prosocial behavior earns synergistic benefits. Still, as we will see, the nearly neutral effect poses a problem for the evolution of prosocial behavior: even if mutual spite entails synergistic costs, or mutual prosociality earns synergistic benefits, the influence of drift becomes stronger as the benefit (b) of prosocial behavior increases. When the benefits of prosocial behavior are very large, and spiteful individuals cannot reap the additional fitness benefits associated with anti-correlated

interactions between behavior strategies,[8] the direction of evolution depends primarily on drift – not selection – and so spite may still evolve in roughly half of all populations.

In order to observe this phenomenon, let us first isolate the nearly neutral effect in the synergistic-cost model. We remove the anticorrelations from the original fitness functions:

$$(6) \quad F(p,N) = \frac{b(x_p) + x_s}{N}$$

$$(10) \quad F(s,N) = \frac{(b-a)x_p + (1-a-c)x_s}{N}$$

Spite is favored when $c < -aN/x_s$ and prosociality is favored when $c > -aN/x_s$. In other words, if (1) the value of a is positive (meaning that spiteful individuals incur a cost from their own spiteful behavior, regardless of whom they interact with) and (2) mutual spiteful behavior incurs positive synergistic costs, selection will always favor prosociality and never favor spite. This result is shown in Figs. 7-8; when $c = 1$, the selection coefficient remains above the x-axis, but when $c = -1$, selection generally favors spite.

[8] This condition may be met if prosocial individuals can, with a sufficient degree of accuracy, identify spiteful individuals and avoid interacting with them.

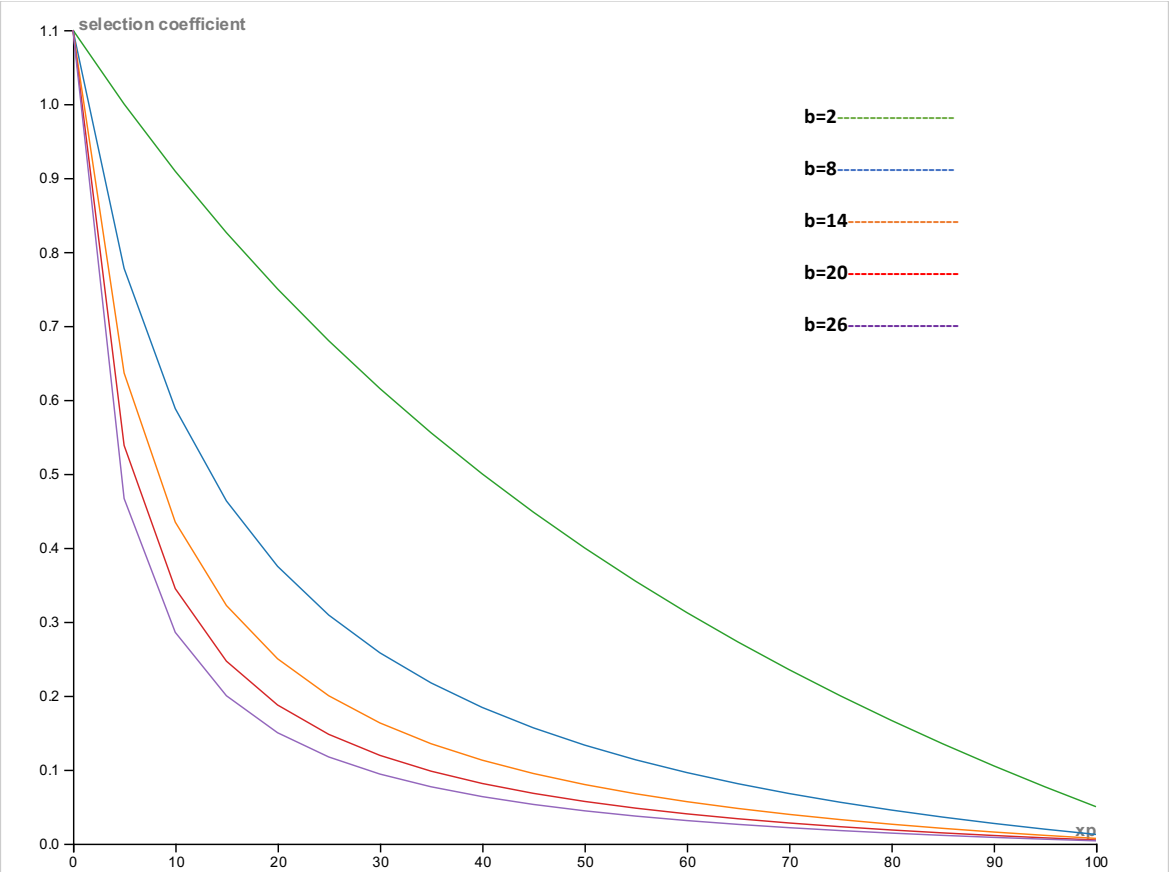
selection coefficient
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
xp
b=2
b=8
b=14
b=20
b=26

Fig.7

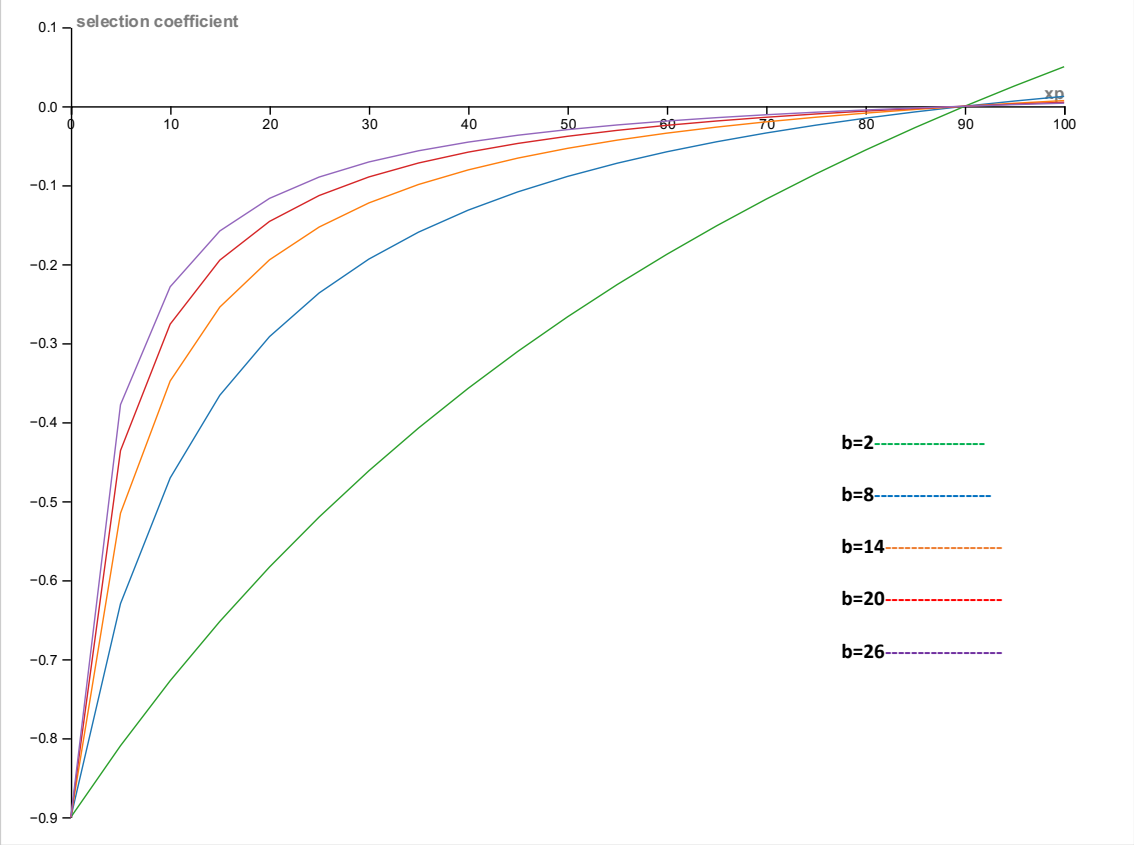

Fig.8

Next, consider the synergistic-benefit model. Again, we isolate the nearly neutral effect by removing the anticorrelations from the original fitness functions:

$$(11)\quad F(p,N) = \frac{(b+d)x_p + x_s}{N}$$

$$(7)\quad F(s,N) = \frac{(b-a)x_p + (1-a)x_s}{N}$$

Spite is favored when $d < -aN/x_p$ and prosociality is favored when $d > -aN/x_p$. In other words, if (1) the value of a is positive and (2) mutual prosocial behavior earns positive synergistic benefits, selection will always favor prosocial behavior and will never favor spite. This result is shown in Figs. 9-10; when $d = 1$, the selection coefficient remains above the x-axis, but when $d = -1$, selection generally favors spite unless x_p is small.

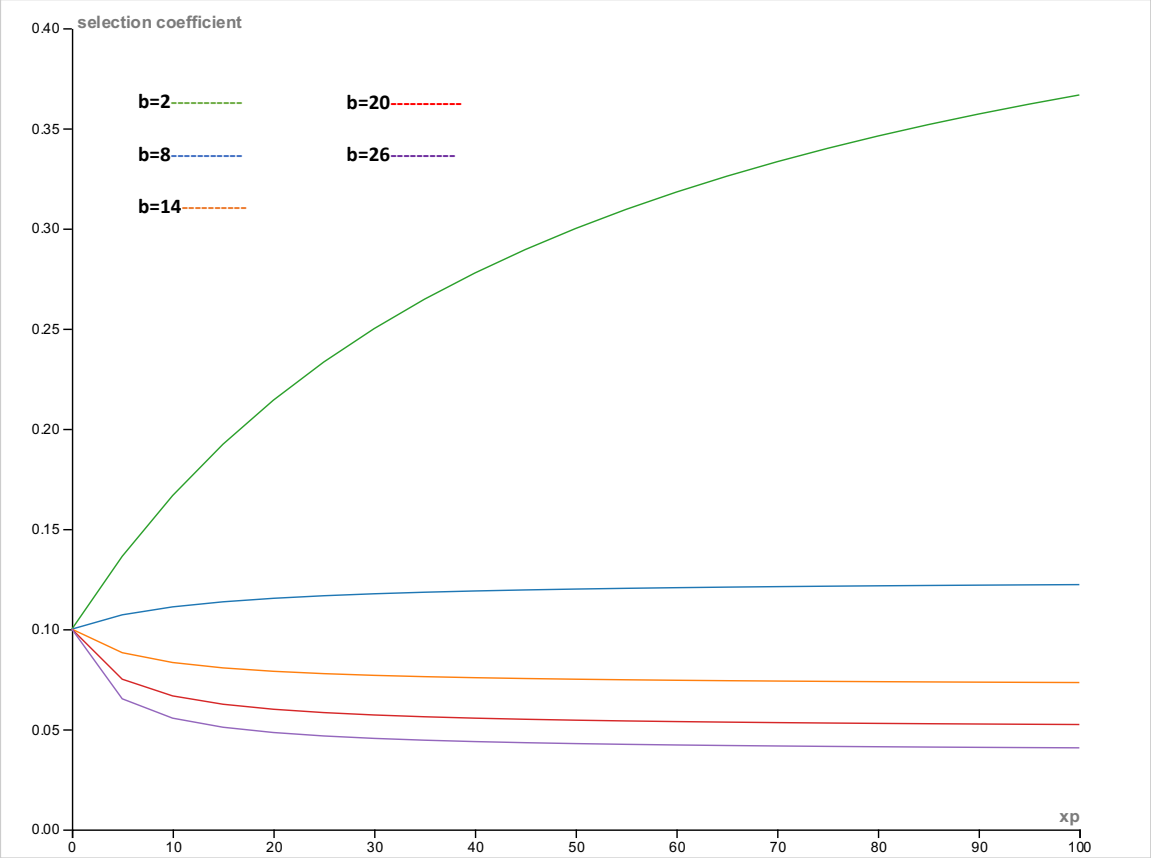

Fig.9

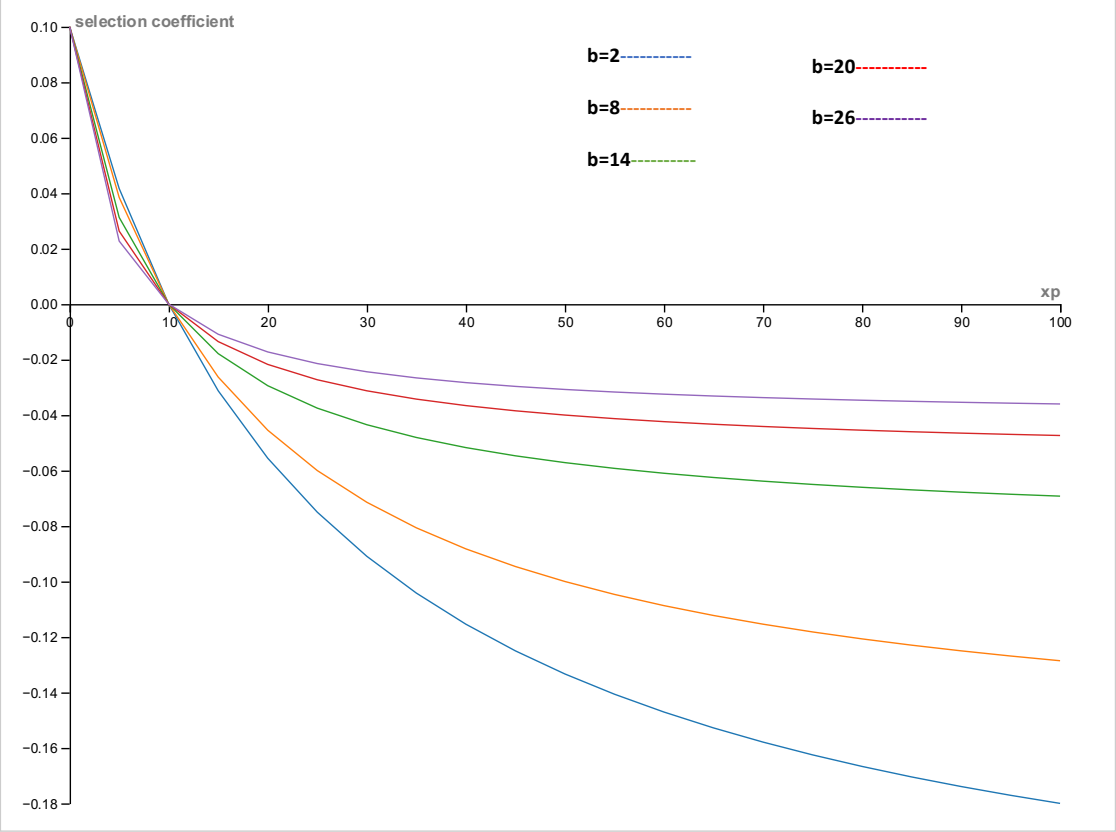

Fig.10

However, as the benefit (*b*) of prosocial behavior increases in both models, the selection coefficient approaches zero. That is to say, as the benefit of prosocial behavior increases, the fitness difference between the two behavior strategies approaches zero; and at very high values of *b*, the question of which behavior strategy evolves in a given population is largely dependent on drift, not selection. It is true that, when the nearly neutral effect is isolated, selection cannot favor spite if mutual spite entails synergistic costs or mutual prosocial behavior earns synergistic benefits. Nevertheless, as the benefit of prosocial behavior

increases in either model, it becomes more and more likely that spite will evolve through the effects of drift alone.

8. Discussion

Forber and Smead (2014) present a paradox for the evolution of prosocial behavior: as the benefits associated with prosocial behavior are increased, the fitness of spiteful individuals also increases. If the fitness of spiteful individuals exceeds the fitness of prosocial individuals, then selection will favor spite; that is to say, increasing the benefits of prosocial behavior may actually decrease the likelihood that prosocial behavior will evolve in a given population. The paradox is salient in models of finite populations, where a slight anti-correlation of behavior strategies typically enables spiteful individuals to reap disproportionate benefits from others' prosocial behavior.

Forber and Smead divide the paradox into two distinct effects: the "spite effect," which they define as the fitness advantage enjoyed by spiteful individuals due to the anti-correlation of interactions in a finite population; and the "nearly neutral effect," or the tendency for spiteful behavior to evolve through drift, when the benefits of prosocial behavior are large and the fitness difference between the two behavior strategies is small. Forber and Smead's analysis shows that, when the nearly neutral effect is isolated – that is to say, when spiteful individuals are prevented from reaping additional fitness benefits associated with anti-correlated interactions between behavior strategies – it becomes relatively more difficult for spite to

evolve, even when the benefits of prosocial behavior are rather large. However, even when the nearly neutral effect is isolated, the paradox still poses a problem for the evolution of prosocial behavior: as the benefits associated with prosocial behavior increase, it becomes more and more likely that spite will evolve through the effects of drift alone.

In an attempt to provide a solution to this problem, I have in this paper developed two game-theoretic models, both involving pairwise interactions in a finite population. Both are similar to the model used by Forber and Smead (2014), but include additional variables to describe synergistic effects. The first involves a synergistic cost, incurred whenever two spiteful individuals interact with each other, and the second involves a synergistic benefit, earned through mutual prosocial behavior. Each model provides a partial solution to the paradox, by increasing the fitness of prosocial individuals relative to their spiteful counterparts; the result is that, when mutual spite incurs synergistic costs, or when mutual prosocial behavior earns synergistic benefits, it becomes more likely that selection will favor prosociality over spite, even when the benefits of prosocial behavior are relatively large.

Nevertheless, isolating the nearly neutral effect reveals that as the benefits of prosocial behavior increase, the fitness difference between the two behavior strategies approaches zero, and it becomes more likely that spite – not prosocial behavior – will evolve in any given population, simply due to drift. So, although synergistic costs and benefits can provide prosocial individuals with a fitness advantage over their spiteful counterparts, Forber and

Smead's paradox remains, albeit in a weaker form; as the benefits of prosocial behavior increase, it becomes more likely that drift will lead spite to evolve in any given population.

References

Binmore, Kenneth. 2007. *Game Theory: A Very Short Introduction*. Oxford: Oxford University Press.

Birch, Jonathan. 2017. *The Philosophy of Social Evolution*. Oxford: Oxford University Press.

Forber, Patrick, and Rory Smead. 2014. "An Evolutionary Paradox for Prosocial Behavior." *The Journal of Philosophy* 111 (3):151-166.

Hamilton, William D. 1964. "The Genetical Evolution of Social Behavior I & II." *Journal of Theoretical Biology* 7:1-32.

Jensen, Keith. 2010. "Punishment and Spite, the Dark Side of Cooperation." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1553):2635-2650.

Okasha, Samir. 2002. "Genetic Relatedness and the Evolution of Altruism." *Philosophy of Science* 69 (1):138-149.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, Brian. 2008. "Evolution and the Social Contract." The Tanner Lectures on Human Values, University of Michigan.

Smead, Rory, and Patrick Forber. 2012. "The Evolutionary Dynamics of Spite in Finite Populations." *Evolution* 67 (3):698-707.

Sober, Elliott. 1992. "The Evolution of Altruism: Correlation, Cost, and Benefit." *Biology and Philosophy* 7:177-187.

Ventura, Rafael. 2019. “The Evolution of Cooperation in Finite Populations with Synergistic Payoffs.” *Biology and Philosophy* 34 (4): 43.