

Portland State University

**PDXScholar**

---

Systems Science Faculty Publications and  
Presentations

Systems Science

---

7-2000

# Simpson's Paradox Can Emerge from the N-Player Prisoner's Dilemma: Implications for the Evolution of Altruistic Behavior

Jeffrey Alan Fletcher

*Portland State University, jeff@pdx.edu*

Martin Zwick

*Portland State University, zwick@pdx.edu*

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/sysc\\_fac](https://pdxscholar.library.pdx.edu/sysc_fac)



Part of the [Other Statistics and Probability Commons](#)

**Let us know how access to this document benefits you.**

---

## Citation Details

Fletcher, Jeffrey A., and Martin Zwick. "Simpson's Paradox can Emerge from the N-Player Prisoner's Dilemma: Implications for the Evolution of Altruistic Behavior." Proceedings of The World Congress of the Systems Sciences and ISSS. 2000.

This Conference Proceeding is brought to you for free and open access. It has been accepted for inclusion in Systems Science Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.edu).

# **Simpson's Paradox Can Emerge from the N-Player Prisoner's Dilemma: Implications for the Evolution of Altruistic Behavior**

Jeffrey A. Fletcher and Martin Zwick

Systems Science Ph.D. Program  
Portland State University  
Portland, Oregon 97207-0751  
jeff@pdx.edu  
zwickm@pdx.edu

Simulations of the n-player Prisoner's Dilemma in multiple populations reveal that Simpson's paradox can emerge in such game-theoretic situations. The relative proportion of cooperators can *decrease* in each separate sub-population, while the proportion of cooperators in the total population can nonetheless *increase*, at least transiently. Factors that determine the longevity of this effect are under investigation. The increase of altruistic behavior exhibited in these simulations is not based on reciprocal altruism, as there are no strategies conditional on other players' past actions, nor does it depend on kin selection via inclusive fitness, as there are no genes. This model is very general in that it can represent both biological and social non-zero sum situations in which utility (fitness) depends upon conditions at different hierarchical levels. The two parameters of the prisoner's dilemma in this model, which determine the gain in individual utility for defection and the dependence of utility on collective cooperation, are respectively analogous to within-group and between-group selective forces in multilevel selection theory.

Keywords Simpson's paradox, prisoner's dilemma, group selection, multilevel selection, tragedy of the commons

## **Introduction**

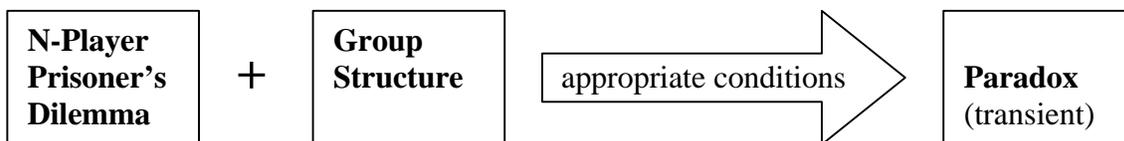
A long-smoldering controversy in evolutionary biology involves the evolution of altruistic behavior and the possibility of multilevel selection. For simplicity, one often considers just two levels where selection of a trait, i.e. altruistic behavior, is the joint result of within-group (individual) and between-group selection (see, e.g., Price 1970). This view, more generally called multilevel selection theory, was initiated by David Sloan Wilson (1975, 1976, 1977) and supported by the empirical experiments and analyses of Michael J. Wade (1977, 1978, 1979). It differs significantly from the idea of species adaptations that ignited the group-selection controversy (e.g., Wynne-Edwards 1962). In a recent book, *Unto Others*, Sober and Wilson (1998) review the history of this controversy and the main alternative theories. These include reciprocal altruism (Trivers

**In *Proceedings of The WorldCongress of the Systems Sciences and ISSS 2000*, Allen, J.K. and Wilby, J.M. eds., Toronto, Canada: International Society for the Systems Sciences.**

## Simpson's Paradox and the N-Player PD

1971) and kin selection via gene-level inclusive fitness (Hamilton 1964, 1975, 1987). Sober and Wilson also highlight how the counterintuitive result of Simpson's paradox, where a trait can decrease within groups, but increase overall, bears on the group selection controversy (1998, pp. 23-25). There is also considerable discussion on the extent to which multilevel selection theory, inclusive fitness, and reciprocal altruism are equivalent explanations (see, e.g., Sober and Wilson 1998 and Reeve 1999). In addition, the gene-level view of evolutionary processes (Dawkins 1976, 1982) generates ongoing controversy and discussion (see, e.g., Sober and Wilson, 1998, pp. 87-92).

Here we hope to avoid these controversies at least partially by capturing the essential nature of multilevel selection within a more fundamental framework that does not depend on genes or reciprocal altruism. The notion that a system (group) does better when it achieves cooperation among its parts (individuals), often against the self-interest of those parts, goes beyond just biological systems undergoing natural selection. It is applicable to hierarchical systems across a variety of fields. The non-zero sum nature of aggregation is general and optimization by subsystems often results in sub-optimization at a higher level (see, e.g., Robertshaw et al, 1978, pp. 207-211). The "prisoners' dilemma" (PD) from game theory is often used to model such non-zero sum situations. Like Simpson's paradox, the PD involves an anomaly of composition. In the PD individually-rational strategies, when aggregated, give a deficient collective outcome. In a recent book, *Non-Zero*, Robert Wright (2000) boldly asserts that non-zero sum interactions are fundamental to the genesis of higher levels of organization in biological and social systems. Here, we more modestly argue that much of what is essential about the tension between hierarchical levels in multilevel selection arises from the existence of non-zero sum interactions. This highlights the similarities between biological and social systems and may allow evolutionary biologists to make use of the extensive work done on competition and cooperation in social systems (see, e.g., Hardin and Baden 1977).



**Figure 1. Schematically illustrates the main finding of this paper: Simpson's paradox can be an emergent of the prisoner's dilemma. "Appropriate conditions" refers to appropriate PD payoff parameter values and initial conditions of group structure.**

As we shall discuss, many other researchers have used game theory via the PD (both 2-player and n-player) to study the evolution of cooperation and altruism. In addition, as noted above, Simpson's paradox (even if not always identified as such) is important in understanding multilevel selection. Here, for the first time as far as we know, a simple relationship is demonstrated between the n-player PD, hierarchical population structure, and Simpson's paradox. Our main finding, illustrated in Figure 1, is that Simpson's Paradox emerges (transiently, but for a wide range of parameters) when a minimal group structure is imposed on an n-player PD. This result is produced by a model which involves an implicit competition between two groups with a simple n-player PD in each.

## Simpson's Paradox and the N-Player PD

The model is based on only two parameters which are easily understood and which correlate with the within-group and between-group selection components in multilevel selection theory.

### The Prisoner's Dilemma

One of the principal ways to model cooperative or altruistic behavior is with game theory (von Neumann and Morgenstern 1947) and the most widely known game of game theory is the PD. In this imaginary situation two prisoners have been arrested for a serious crime and are being interrogated separately by the police. The prisoners cannot communicate and neither knows what the other will do. If they cooperate with each other and keep quiet there is only enough evidence to convict them of a lesser crime and they will both do minimal jail time. If however one of them turns State's evidence and rats on (defects from) the other who remains silent, then the defector does no jail time, but the cooperator receives a very harsh sentence. In the case where both confess (defect), they both receive an intermediate amount of jail time.

### Individual Rationality Leads to Collective Irrationality

The PD is represented by the payoff matrix in Table 1 below where the numbers represent a positive measure of utility. It is simpler to think in terms of positive payoffs rather than the negative payoff of jail time, but the essential features are the same. C represents cooperation and D represents defection. The strategies for player 1 are represented as rows and for player 2 as columns. The payoff values for each pair of strategies that meet are listed as *player 1 payoff/player 2 payoff*.

		<i>Player 2</i>	
		<b>C</b>	<b>D</b>
<i>Player 1</i>	<b>C</b>	3/3	0/5
	<b>D</b>	5/0	1/1

**Table 1. Payoff matrix for a 2-player Prisoner's Dilemma.**

The PD is a non-zero sum game—the sum of the two players' scores varies for different strategy combinations. To understand the paradox at the heart of the PD, imagine player one trying to decide what to do and using the payoff matrix to reason with as follows: Regardless of what player 2 does, I should defect—by defecting I will get 5 instead of 3 if player 2 cooperates, or 1 instead of 0 if player 2 defects. Player 2 is in symmetrical situation and rationally also chooses to defect. So each player follows this *dominant* strategy and gets a payoff of 1, but if they had both cooperated they would have each gotten a payoff of 3. The essential feature of the PD is that the dominant individually-rational strategy for each player leads to a collective sub-optimal or irrational outcome. This outcome (1/1) is non-Pareto optimal because there is another outcome (3/3) to the

## **Simpson's Paradox and the N-Player PD**

game that increases the utility of one player without cost to the other. In fact in this case, both players can do better.

### **Iterated Prisoner's Dilemma, Tit-for-Tat, and Reciprocal Altruism**

The simulation experiments of the political scientist Robert Axelrod (Axelrod and Hamilton 1981, Axelrod 1984) addressed the question of whether individual rationality would still favor defection if instead of playing just once or at random, players were forced to play a series of iterated PD games. This work extended the idea of reciprocal altruism (Trivers 1971), an earlier attempt to explain seemingly altruistic behavior between non-relatives in terms of individual benefit. Axelrod sponsored a tournament in which various strategies implemented in computer programs were played against each other pair-wise in a round robin so that each program played every other program including itself. Each pair-wise iterated game consisted of 200 interactions and the payoff matrix was the one illustrated above in Table 1. Therefore two always-defect (ALLD) strategies playing each other would get a score of 200 each, two always-cooperate (ALLC) strategies playing each other would get a score of 600, and an ALLD playing an ALLC would get respective scores of 1000 and 0. For a summary of subsequent studies of the evolution of cooperation and altruism based on the iterated PD see Dugatkin (1997, p. 25).

Surprisingly the simplest strategy in the Axelrod tournament turned out to also be the best. Proposed by Anatol Rapoport, it is called Tit-for-Tat (TFT). This strategy cooperates in the first interaction and then always plays the strategy its opponent used in the last encounter. Consistent with the theory of reciprocal altruism, TFT players need the capacity to remember previous actions by competitors, but unlike kin selection, no similarity in genes is assumed. The TFT strategy is willing to cooperate, swift to punish a defection, yet forgiving in that it will return to cooperation if its opponent makes the sacrifice of cooperating while TFT is defecting.

Unfortunately, the reason that TFT came out on top has been widely misunderstood as being due to its individual fitness or "unbeatability." As Sober and Wilson (1998, pp. 85-86), and even Rapoport have pointed out, TFT can never beat an opponent in any pair-wise iterated interaction because it never defects unless it has already been on the short end of a defection from its opponent. As Rapoport (1991, pp. 93) put it, "in every paired encounter, Tit-for-Tat must either draw or lose. It can never win a paired encounter." In this sense TFT is altruistic at the individual level because it often gives more points than it gets and it never gains more than its opponent does.

The reason TFT won the tournament hinges on the fact that it often was able to play other TFT or similar altruistic strategies where it could run up its accumulated score. So even though in individual competition TFT is inferior, for example to ALLD which is the most fit unexploitable individual strategy, pairs (groups) of TFT accumulate higher scores than pairs (groups) of ALLD. In the analysis of why TFT was a successful strategy, (Axelrod 1984, p.33) does note that it tended to score especially well (close to 600) when it played

## **Simpson's Paradox and the N-Player PD**

similar strategies, but he does not recognize this as a group effect. Maynard Smith (1982, p. 168) comments that, "the programs were ranked according to the total payoff accumulated (not, it should be noted, according to the number of opponents defeated in the individual matches)." Yet neither he nor Axelrod distinguish the individual and group levels of competition present in this tournament which is obscured by the cumulative method of scoring. Sober and Wilson (1998, pp. 102-116) emphasize the importance of identifying and separating out the selective forces at different hierarchical levels. They argue that much of the controversy surrounding group selection is due to a failure to do so.

### **N-Player Prisoner's Dilemma (The Tragedy of the Commons)**

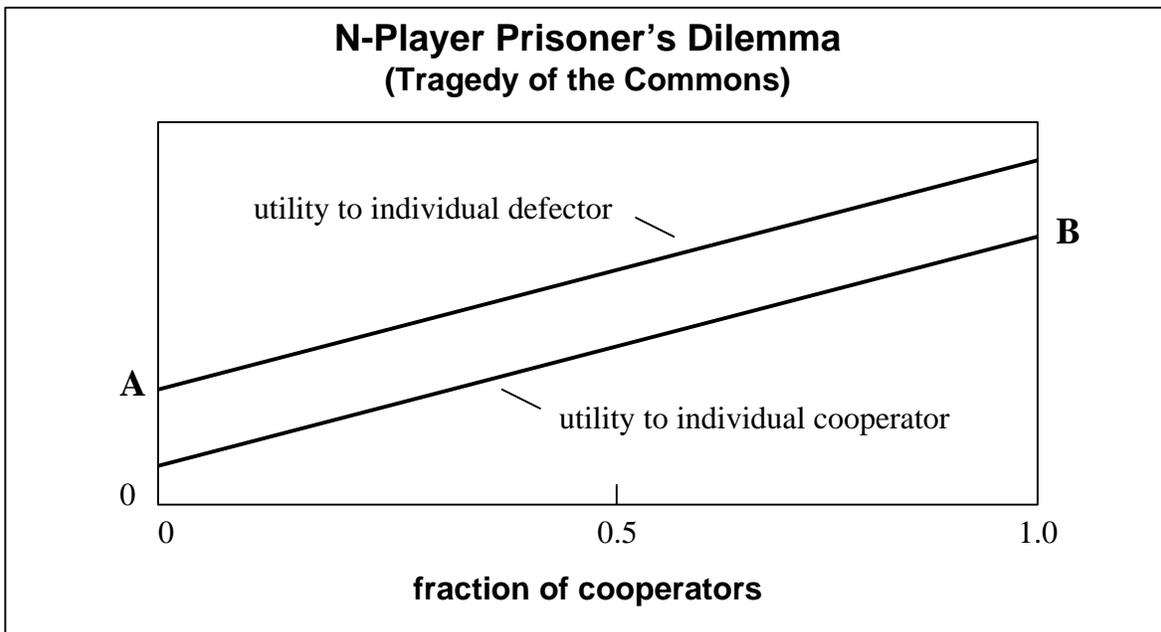
The n-player, as opposed to 2-player, PD offers a straightforward way of thinking about the tension between the individual and group levels. In real-world biological and social systems the effects of cooperation or defection are often distributed diffusely to other members of a group, i.e., they do not necessarily arise via pair-wise interactions. When there is a common and finite resource, each individual benefits by using more than its share of that resource, but when all players apply this individual rationality it can lead to collective irrationality. For example, each country that fishes international waters can increase its utility by taking more of the fish in this common resource, but as more and more countries overfish, the common stock is depleted beyond where it can quickly replenish and so in subsequent years all have less. This leads to decreased utility for both countries that overfish (defectors) and those that don't (cooperators). As another example, imagine policyholders with theft insurance that successfully make false claims. The benefit to an individual cheater is then paid for by an increase in premiums. The more policyholders that cheat in this way, the higher the premiums and therefore the less valuable the insurance is to both those that only make claims for true thefts (cooperators) and cheaters (defectors). Finally, as an evolutionary example, consider social spiders that vary in the sex ratio of their offspring (Aviles 1993). Within a group, spiders with an even sex ratio in offspring (defectors) are more fit than spiders with biased sex ratios (Fisher 1930), but groups that contain individuals biased towards female offspring (cooperators) grow faster and therefore have the potential to do better collectively in competing against groups where the sex ratio is even.

In their simplest form, these n-player examples include no clear role for a TFT strategy, and thus no necessity to remember other players past actions and no reciprocal altruism. However, generalized TFT strategies have been introduced into n-player PD models (see Boyd and Richerson 1988, Joshi 1987, Motro 1991) to study reciprocal altruism. In generalized TFT, a player cooperates at time  $t$  if a certain number of its group members cooperated at time  $t - 1$  or if the player received a certain level of payoff at  $t - 1$ . As with the 2-player PD studies, the distinct within-group and between-group components suggested by multilevel selection theory were not recognized in these studies. There is no distinction made between high individual fitness scores achieved by exploiting others within the same group and high fitness scores due to grouping with other cooperative or

## Simpson's Paradox and the N-Player PD

TFT-type strategies. In this paper, however, using the simplest n-player PD model involving only the pure cooperate or defect strategies, we explicitly capture the individual and group components of selection as two parameters in our model.

An n-player PD involving the exploitation of a common resource (e.g., the fisheries example) is also known as the “tragedy of the commons” (Hardin 1968). A simple payoff scheme for such an n-player PD is illustrated by Figure 2. On the horizontal axis is the fraction of individuals cooperating for the common good. On the vertical axis is the average utility to each individual. For convenience, we assume a linear relationship between utility and percent cooperators. For instance, one can think of insurance premiums going down linearly with the fraction of policyholders that refrain from making false claims. Alternatively, one can think of the growth rate of a spider colony increasing linearly with the number of members who have a female-biased sex ratio in their offspring. The upper line denotes the utility for a defector while the lower line is the utility for a cooperator. The defector's line dominates the cooperator's line, i.e., selfish individual behavior always has a higher utility than cooperating no matter what the fraction of cooperators. The resulting dynamic tends to decrease the number of cooperators within a group.



**Figure 2. Utility lines for defectors and cooperators as a function of the fraction of cooperators for a simple n-player prisoner's dilemma (PD).**

The deficient outcome of the PD here inheres in the fact that the utility to defectors when there is a minimum number of cooperators (point A) is lower than the utility to cooperators when there is a maximum number of cooperators (point B). So even though for a given state of the system an *individual* benefits more by defection than cooperation, still cooperators in a *group* of cooperators get more benefit than defectors in a group of defectors.

## Simpson's Paradox and the N-Player PD

The “tragedy” (and what makes this a PD) is that whatever the current state of the system, individual rationality or individual selection favors defection, which tends to drive the system to a (boundary) equilibrium state less beneficial to all (point A). This state is a non-Pareto, sub-optimal, and irrational collective outcome. To summarize algebraically:  $U_D(x) > U_C(x)$  for all  $x$  ( $U_D$  and  $U_C$  are utility functions of the fraction of cooperators,  $x$ ) causes  $x$  to decrease for all  $x$ , but  $U_C(1.0) > U_D(0.0)$ . The co-parallel lines used here are the simplest of many cooperator and defector utility curves that can satisfy these PD conditions.

Note that it may be tempting to think of cooperation as selfish rather than altruistic because a group of all cooperators gets more utility per individual than a group of all defectors, but this would be incorrect and misses the crux of the PD. In the 2-player PD, the players would be better off if they both cooperate, but defecting is still the rational *individual* strategy because the prisoners have no way to coordinate their actions and enforce any agreements to cooperate. Cooperating is always disadvantageous no matter what the other player does. So in the absence of guarantees of cooperation by other players, cooperating is truly altruistic—it lowers one's individual utility (fitness) while raising the benefit to others. The same situation holds in the n-player game. Given the absence of coordination between players, each player is better off to defect, but benefits others by not doing so in that the system is kept at a state with a higher fraction of cooperators. Of course, this is the dynamic for a single set of players, or for a multi-group system viewed at the intra-group level. As we shall see, at the higher level of organization, i.e. that of the total population which includes all groups, cooperators can thrive, at least for a while, despite their inferior individual utility (fitness).

## Simpson's Paradox

At the University of California at Berkeley in the 1970s, the percentage of women graduate school applicants accepted was significantly lower than the percentage of men accepted (Cartwright 1978). Yet, when the University looked at each department they found none were accepting a smaller percentage of women. The answer to this paradox lies in the fact that different departments varied in their contribution to the whole. Women were applying in greater numbers to departments that accepted a lower percentage of applicants. For example, imagine that 70 women and 30 men apply to department A which has 20 positions. If there is no bias with regard to sex, 14 women and 6 men are accepted. Also, imagine that 30 women and 70 men apply to Department B which has 50 positions. If there is no bias, 15 women and 35 men are accepted. However, if we aggregate these results, 41 of 100 men are accepted, whereas only 29 of 100 women are accepted. This is an example of Simpson's paradox (Simpson 1951).

Like the PD, Simpson's paradox hinges on an anomaly of aggregation. In the PD aggregating individually rational strategies does not lead to a collectively rational outcome. In Simpson's paradox, aggregation yields results qualitatively different from

## **Simpson's Paradox and the N-Player PD**

those apparent at the lower level. As Sober and Wilson (1998, pp. 23-25) note, Simpson's paradox has direct implications for the debate over the evolution of altruism via group selection. If group structure is not taken into account then averaging fitness across groups which vary in their local fitness-relevant conditions may give paradoxical results. For instance, in Axelrod's iterated PD experiments accumulating individual scores in the round robin tournament led TFT to appear the most individually fit even though TFT can never win an iterated match.

### **A Simple Model of Dynamic Change in Altruistic Behavior**

We now demonstrate that the PD and Simpson's paradox are deeply linked by analyzing a very simple (two parameter) model based on the n-player PD which, with the addition of a minimal hierarchical structure, captures the essential tension between group and individual-level selection. As mentioned earlier, the model is more general than evolutionary models and does not involve inclusive fitness or reciprocal altruism.

In the simplest form of the model there are two groups with no migration between them. These groups initially are the same size and vary only in their fraction of cooperators and defectors. There are no other strategies besides ALLC and ALLD. We follow the percentage of cooperators in each group and across the whole population. There are two parameters that influence these dependent variables. The first is the slope of the utility lines (see Figure 2). For simplicity we use linear and parallel utility functions. The slope of both lines affects the disparity in utility for groups of different composition and can be thought of as the magnitude of the group level selective force. At this level groups containing more cooperators have the advantage. The second parameter is the difference in the intercept for the cooperator's and defector's utility lines. Because the lines are parallel, the intercept is the vertical displacement between them at all levels of cooperation. This disparity in utility for defectors vs. cooperators within a group can be thought of as the magnitude of the individual level selective force within each group. At this level defectors have the advantage over cooperators.

This is a dynamic model and at each timestep the following action is implemented: Within each group the number of cooperators is increased in proportion to the cooperators' utility based on the group's composition. Similarly, the number of defectors is increased in proportion to the defectors' utility based on the group's composition.

In addition, the population of each group is proportionally scaled back (preserving the ratio of cooperators and defectors) so that the total population size matches the original total. Scaling is unnecessary and does not do anything substantive, but it helps make the dynamics clear. The changes in the percentage of cooperators and defectors in each group and across the population is unchanged by uniform scaling, although it does have some justification in terms of biological or economic carrying capacities. It is also worth noting that although we have implemented the changes within groups as an increase in population with a subsequent uniform scaling, the resulting changes in group size and composition could also be thought of in terms of an optional behavior for cooperating or

## Simpson's Paradox and the N-Player PD

defecting within a fixed population where members might also be allowed to migrate to other groups.

To calculate the utility (see Figure 2) we use:

- $U_{Ci} = m f_{Ci} + b_C$
- $U_{Di} = m f_{Di} + b_D$

where:

- $U_{Ci}$  is the utility for a cooperator within group  $i$
- $U_{Di}$  is the utility for a defector within a group  $i$
- $m$  is the slope of both the defector and cooperator utility lines
- $f_{Ci}$  is the fraction of cooperators in group  $i$
- $b_C$  is the intercept for the cooperator's utility line (0 in our simulations)
- $b_D$  is the intercept for the defector's utility line ( $\geq 0$  in our simulations)

Note that keeping the cooperator's intercept at 0.0 and the fact that the x-axis (fraction of cooperators) of the utility functions varies from 0.0 to 1.0 means that here the condition for a PD (e.g., point B above point A in Figure 2) is  $m > b_D$ . In all runs reported in this paper this condition is satisfied. The increase of each strategy within a group is directly proportional to that strategy's utility within the group, which equals the number of individuals utilizing the strategy times its utility payoff per individual:

- $N_{Ci}(t+1) = N_{Ci}(t) + N_{Ci}(t) U_{Ci}$
- $N_{Di}(t+1) = N_{Di}(t) + N_{Di}(t) U_{Di}$

where:

- $N_{Ci}$  is the number of cooperators in group  $i$  (before scaling)
- $N_{Di}$  is the number of defectors in group  $i$  (before scaling)

It is also useful to define  $N_1 = N_{C1} + N_{D1}$ ,  $N_C = N_{C1} + N_{C2}$ ,  $N_2 = N_{C2} + N_{D2}$ , and  $N_D = N_{D1} + N_{D2}$ .

An example of one timestep illustrates. For our initial conditions group 1 consists of 90 defectors and 10 cooperators and group 2 is composed of 10 defectors and 90 cooperators. For this example we use a slope of 1.0 and intercepts of 0.0 for cooperators and 0.1 for defectors. Table 2 shows the results of one timestep based on the above equations and description. The calculations of utilities at time 0 are shown below:

- $U_{C1} = m f_{C1} + b_C = (1.0)(0.1) + 0.0 = 0.1$
- $U_{D1} = m f_{D1} + b_D = (1.0)(0.1) + 0.1 = 0.2$
- $U_{C2} = m f_{C2} + b_C = (1.0)(0.9) + 0.0 = 0.9$
- $U_{D2} = m f_{D2} + b_D = (1.0)(0.9) + 0.1 = 1.0$

The population values before uniform scaling depend on the utilities and are calculated below for the first timestep:

- $N_{C1}(1) = N_{C1}(0) + N_{C1}(0) U_{C1} = (10) + (10)(0.1) = 11$
- $N_{D1}(1) = N_{D1}(0) + N_{D1}(0) U_{D1} = (90) + (90)(0.2) = 108$
- $N_{C2}(1) = N_{C2}(0) + N_{C2}(0) U_{C2} = (90) + (90)(0.9) = 171$
- $N_{D2}(1) = N_{D2}(0) + N_{D2}(0) U_{D2} = (10) + (10)(1.0) = 20$

## Simpson's Paradox and the N-Player PD

The shaded cells in Table 2 show that after one timestep the percentage of cooperators in group 1 has dropped from 10.0% to 9.2% and in group 2 from 90.0% to 89.5%, but the overall percentage of cooperators has increased from 50.0% to 58.7%. This is an example of Simpson's paradox, which can be summarized algebraically as:

- $N_C(1) / N(1) > N_C(0) / N(0)$

even though

- $N_{C1}(1) / N_1(1) < N_{C1}(0) / N_1(0)$

- $N_{C2}(1) / N_2(1) < N_{C2}(0) / N_2(0)$

An observer who understood that population changes were due to utility (fitness) differences, but ignored the group structure, would interpret this increase in cooperation as indicating that cooperators were more fit than defectors. Yet it is obvious from Figure 2 that the utility (fitness) for defectors is always above that of cooperators regardless of the fraction of cooperators within their group. Note that in the Berkeley example Simpson's paradox involves an effect seen at a higher (university) level but absent at the lower (department) level, while here effects have opposite directions at the two levels: cooperation is favored at the group level but not favored at the individual level.

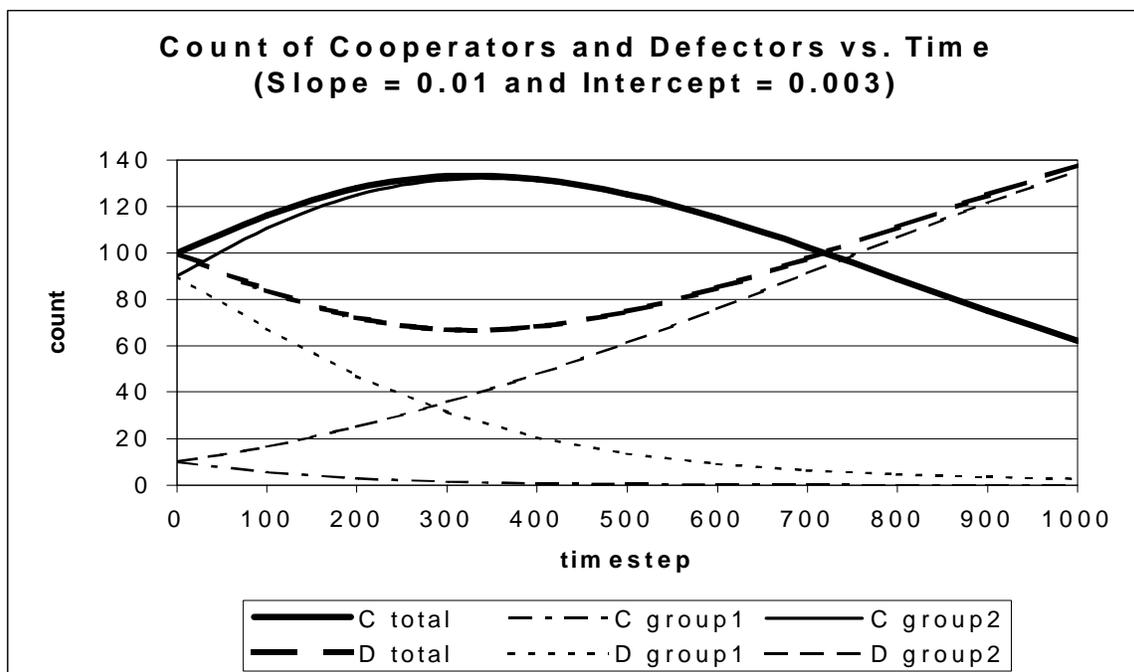
	Group 1			Group 2			Groups 1 and 2		
Time	C	D	%C	C	D	%C	C	D	%C
0	10	90	10	90	10	90	100	100	50
1*	11	108	9.2	171	20	89.5	182	128	58.7
1	7.1	69.7	9.2	110.3	12.9	89.5	117.4	12.9	58.7

**Table 2. The number of cooperators (C), defectors (D), and percent cooperators (%C) in each group at time 0, time 1\* (before scaling), and time 1 (after scaling) for a slope of 1.0 and a defector intercept of 0.1. The shaded cells highlight the changes in %C.**

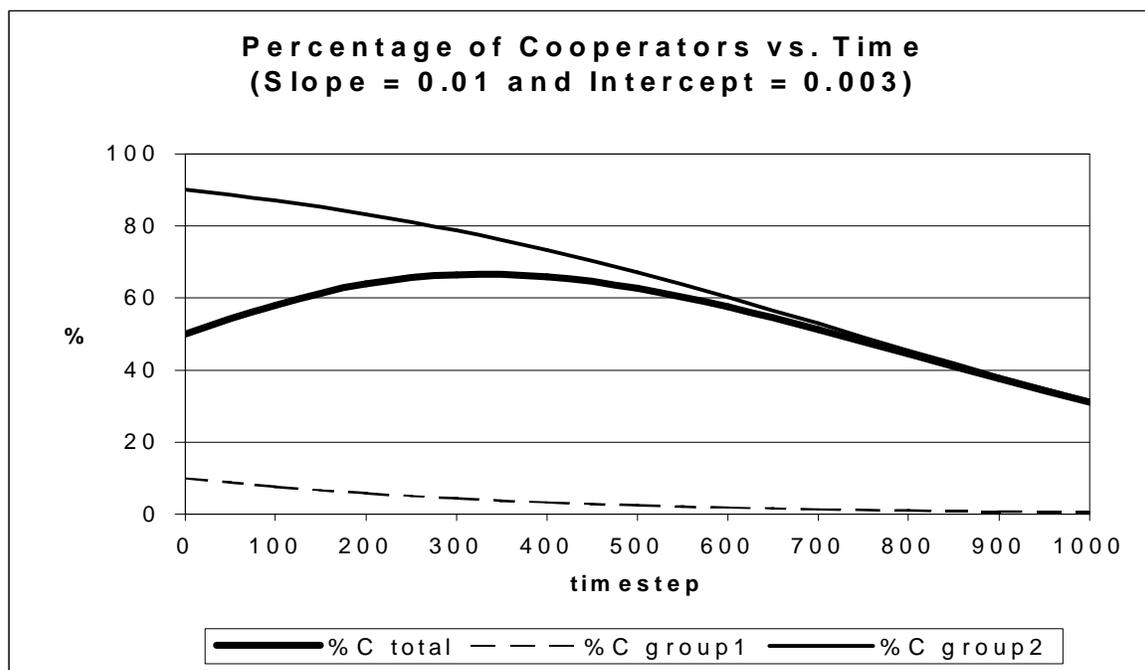
## Experiments and Results

Because the utility for defectors is always higher than that for cooperators, in the long run defectors will dominate both in each group and across the whole population. Yet we have just seen that it is possible, while the percentage of cooperators decreases within each group for the overall percentage of cooperators in the whole population to increase. This effect is transient without mechanisms for reestablishing variation between groups. Several such mechanisms have been proposed elsewhere; they include periodic random founding of groups from a large breeding population (see Williams and Williams 1957, Maynard Smith 1964), non-uniform population densities with viscosity (Mitteldorf and Wilson 2000), groups that remain isolated and fission or bud (Aviles 1993), and various methods of active altruist aggregation (see Sober and Wilson 1998).

## Simpson's Paradox and the N-Player PD



**Figure 3. Population of cooperators and defectors in group 1, group 2, and total.**



**Figure 4. Percentage of cooperators in group 1, group 2, and total.**

Here we keep things simple by just noting that (in the absence of the above mentioned necessary mechanisms) the increase in altruistic behavior due to Simpson's paradox is transient. We explore how combinations of our two parameters affect the magnitude and longevity of this effect. All of our experiments included here involve two groups with initial conditions of 90 defectors and 10 cooperators in group 1 and 90 cooperators and

## Simpson's Paradox and the N-Player PD

10 defectors in group 2, but our results can be extended to other initial distributions as will be discussed below. These population sizes can actually represent much larger populations as we have allowed fractional numbers of players. Figure 3 shows the beginning of a typical run with two groups where Simpson's paradox is evident. Here the slope of the utility lines are 0.01 and the intercept of the defectors' line is 0.003.

Notice that in this run both the number of cooperators and defectors increase in group 2, and decrease in group 1 up to about timestep 330. At this point there are 171 players in group 2 (132.1 C, 38.9 D) and only 29 left in group 1 (1.2 C, 27.8 D). Since in the beginning group 2 consists of mostly cooperators, the rapid expansion in group 2 causes the overall number of cooperators to increase from 100 up to 133.3 or from 50% up to 66.6%. After timestep 330 the total population continues to be more and more dominated by group 2, but the defectors in group 2 are steadily increasing. This causes the overall number of cooperators to begin to shrink. By timestep 4,000 (not shown) the overall percentage of cooperators is essentially zero ( $< 0.01\%$ ). Simpson's paradox is easier to see in Figure 4 which tracks the percentages of only the cooperators for the same run. Notice that the percentage of cooperators in both group 1 and group 2 decrease, but the overall percentage of cooperators initially increases.

Run 1 in Table 3 shows the data from the run used in Figures 3 and 4. Run 2 shows the effect of increasing (to 0.05) the slope (group selection) while holding the intercept (individual selection) constant. The magnitude of the Simpson's paradox effect increases as shown by the fact that the total percent cooperator peak increases (from 66.6% to 85.5%) and occurs sooner (121 compared to 328 timesteps). It is interesting to note that the percentages of cooperators and defectors within each group after 100 timesteps are not significantly affected by increasing the slope (7.6% cooperators in both Runs 1 and 2 and 87.0% and 87.1% defectors in Run 1 and Run 2 respectively). The percent cooperators or defectors within a group is an individual selection phenomenon relatively independent of the group forces, although the absolute numbers in each group are affected by the slope change (72.9 in group 1 and 127.1 in group 2 in Run 1 vs. 5.0 in group 1 and 195 in group 2 in Run 2).

Run	Slope	Intercept	Max %C	Time at max	% C group 1	Total # group 1	% C group 2	Total # group 2
initial conditions for all runs:					10.0	100.0	90.0	100.0
					<b>At timestep 100</b>			
1	0.01	0.003	66.6	328	7.6	72.9	87.0	127.1
2	0.05	0.003	85.5	121	7.6	5.0	87.1	195.0
3	0.01	0.001	82.4	520	9.1	65.7	89.1	134.3
4	1.0	0.3	70.2	6	0.0	0.4	0.0	199.6
5	0.0001	0.00003	66.6	32,501	10.0	99.7	90.0	100.3
6	0.01	0.008	50.0	0	4.8	92.2	80.3	107.8

**Table 3. Results of various runs with varied slope and intercept. Shaded area shows percent cooperators and total population of each group after 100 timesteps.**

## Simpson's Paradox and the N-Player PD

Run 3 shows the effect of keeping the slope the same as in Run 1, but decreasing the intercept. Recall that the intercept is a measure of within-group selection for all group compositions because in this model the utility lines for cooperators and defectors are parallel. Decreasing this within-group individual selection also increases the maximum percent of overall cooperators reached (to 82.4%), but unlike the effect of increasing the slope, decreasing the intercept also increases the amount of time to reach the peak (520 compared to 328 timesteps).

These results make sense in terms of our association of group selection and individual selection with slope and intercept, respectively. Increasing the slope gives group 2 a bigger advantage and leads group 2 to dominate the population sooner. Since the rate at which defectors take over group 2 is a within-group phenomena based on the intercept which is unchanged between Run 1 and Run 2, group 2 dominating the total sooner (while it has more cooperators) leads to a higher maximum percentage of cooperators reached overall and this peak is reached in a shorter time. In contrast, in comparing Runs 1 and 3 the intercept is decreased while the slope is held constant. This reduces the advantage defectors have within a group. Here the maximum percentage of total cooperators also increases, but not because group 2 dominates faster. Rather it is because both groups have more cooperators for a longer time due to the decreased individual advantage to defectors. This causes the peak in percent cooperators to be delayed. The effects of the two parameters are most clearly seen at their lowest values. For zero slope, no group will come to dominate. For zero intercept, the composition of cooperators and defectors doesn't change within each group.

Runs 4 and 5 in Table 3 demonstrate that the Simpson's paradox effect is very robust across our two parameter space, but also depends on the balance between the group-level (slope) and the individual-level (intercept) factors. Run 4 shows a similar maximum total cooperators peak to Run 1 (66.6% vs. 70.2%) where the slope and intercept are both increased 100 fold to 1.0 and 0.3 respectively. Run 5 again shows a similar result to Run 1 (both 66.6%) when the slope and intercept are decreased 100 fold compared to Run 1, to 0.0001 and 0.00003. Note again that the time to reach the maximum percent cooperators happens much more quickly for the higher slope values. There are, however, many parameter combinations where, for the initial conditions used in all these runs, the intercept (individual effect) is too high for a given slope (group effect) and the overall percentage of cooperators decreases monotonically. This case is represented by Run 6 in Table 3. In this run there is a PD because  $m > b_D$  ( $0.01 > 0.008$ ), but no Simpson's paradox. In these simulations a PD is a necessary but not sufficient condition to get a Simpson's paradox effect.

To explore the effect of varying slope and intercept on the magnitude of Simpson's paradox, we did 651 runs systematically varying the parameter for slope from 0.0 to 0.2 in 0.01 increments and for intercept from 0.0 to 0.15 in 0.005 increments. Figure 5 shows the peak values of total percent cooperators plotted on the z-axis. The results match our expectations and again support the use of slope and intercept as representing the between-group and within-group selective forces, respectively. The plane at the 50% mark represents a region of parameter space where there is no Simpson's paradox seen, where

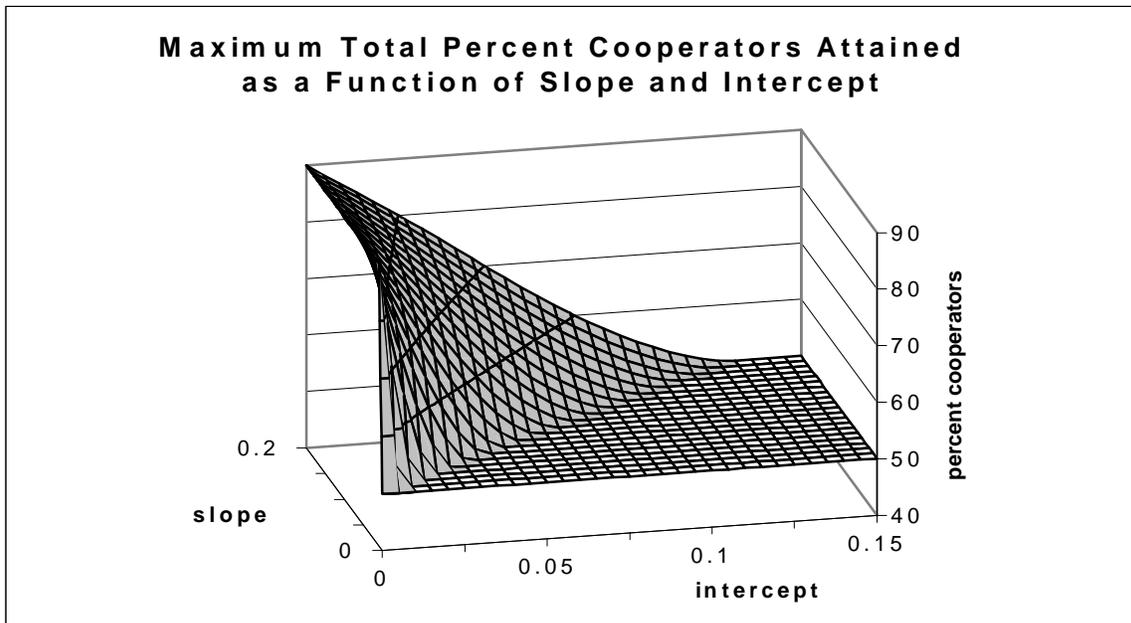
## Simpson's Paradox and the N-Player PD

the individual selection (intercept) is high in relation to group selection (slope). In our model, where for simplicity  $N_1 = N_2 = N_C = N_D$ , the condition for Simpson's paradox expressed in terms of our two parameters and initial conditions is:

- $m / b_D > N_i^2 / (N_{Ci} - N_{Di})^2$  where  $i = 1$  or  $2$  (see Appendix A)

Any initial difference between groups in their cooperator and defector composition can, with appropriate intercept and slope values, generate a Simpson's paradox. For example, initial conditions of 55 defectors and 45 cooperators in group 1, the opposite ratio in group 2, a slope of 1.1, and an intercept of 0.01 produces Simpson's paradox because the condition  $1.1 / 0.01 > 100^2 / 10^2$  is met. Our 10-90:90-10 initial distribution was just chosen to make results easier to visualize. Note however, that the above condition is derived specifically for our symmetrical initial distributions. We have not yet derived a more general condition for any initial distribution.

It is clear from Figure 5 that increasing the slope for a given intercept leads to an increase in the maximum attained overall percent cooperation (which cannot exceed the initial percent cooperation in the cooperator-dominated group, here 90%). For a given level of group selection (slope), decreasing the individual selection (intercept) also leads to a higher level of cooperation reached. Although not shown here, as discussed above, decreasing the intercept also slows down the time for defectors to take over within a group. This extends the temporal length of the Simpson's paradox effect as well as increasing its magnitude.



**Figure 5.** Shows the results of 651 runs where slope and intercept were varied systematically. The z-axis shows the maximum total percent cooperators reached where the initial value was always 50%.

# Simpson's Paradox and the N-Player PD

## Conclusions

Like the 2-player PD, the n-player PD illustrates how the aggregation of actions based on individual rationality can lead to a sub-optimal or irrational collective outcome. This model can be mapped onto many real-world situations in which the optimization of subsystems does not lead to an optimal result at the system level. Here we have demonstrated that by adding a minimal group structure onto the n-player PD, Simpson's paradox emerges such that the *total* percentage of cooperators can increase even though within groups cooperation always gains less utility and steadily decreases relative to defection. The maximum percentage of cooperators attained is a measure of the magnitude of this effect. This increase in total percent cooperators is transient as defectors eventually take over both groups and therefore the total population. We have not yet added mechanisms which might sustain the increase in cooperators.

In its simplest form, this model has only two parameters: the slope of the utility lines and the difference in their intercepts. For two groups that vary only in their initial composition of cooperators and defectors, these parameters are analogous to the between-group selection (slope) and within-group individual selection (intercept difference) in the evolution of altruistic behavior via multilevel selection. Observing Simpson's paradox depends also on initial conditions in a predictable way. For any initial condition where the groups differ (so far we have only examined symmetrical groups), a Simpson's paradox can be generated by a strong enough PD, i.e., by a slope sufficiently greater than the intercept.

The model described in this paper is simpler than models of reciprocal altruism based on the iterated 2-player and n-player PD in that here there are no actions conditioned on past behaviors of other players. It is also more abstract than inclusive fitness models in that there are no genes. An increase of altruism (in the PD) merely requires a suitable higher level of organization. This is consistent with multilevel selection theory. Increasing the slope (group selection parameter) increases the disparity between group size such that the cooperator-dominated group increases and this accounts for the overall increase in cooperators. Decreasing the intercept (individual selection parameter) causes cooperators within each group to be sustained longer and therefore also contributes to an increase in overall cooperators. This simple model lets us tease out the within-group and between-group components of utility or fitness and is applicable to both biological and social systems in which there is competition at multiple levels.

## Further Research

We are continuing to explore the applicability and further elaboration of this model. This includes:

- Exploring mechanisms that might give rise to and/or maintain high levels of cooperation indefinitely, e.g., random re-assortment of groups, variable population

## Simpson's Paradox and the N-Player PD

density with viscosity, group fission or budding, and other methods of altruist aggregation.

- Examining the dependence of the magnitude and longevity of the Simpson's paradox effect on initial conditions and model parameters.
- Exploring models with cooperator and defector utility curves other than straight lines.
- Examining the relationship between non-zero sum games other than the PD on the evolution of altruism.
- Studying if and how this model might explicitly encompass reciprocal altruism and inclusive fitness.

### Acknowledgements

We would like to thank Andreas Rechtsteiner for a careful reading of an earlier draft of this paper and fruitful discussions.

### References

- Aviles, L. (1993). "Interdemic Selection and the Sex Ratio: A Social Spider Perspective," *The American Naturalist*. 142(2):320-345.
- Axelrod, R. (1984). *The Evolution of Cooperation*. New York:Basic Books, Inc.
- Axelrod, R. and Hamilton, W.D. (1981). "The Evolution of Cooperation," *Science*. 211:1390-96.
- Boyd, R. and Richerson, P.J. (1988). "The Evolution of Reciprocity in Sizeable Groups," *Journal of Theoretical Biology*. 132:337-356.
- Cartwright, N. (1978). "Causal Laws and Effective Strategies," *Nous*. 13:419-437.
- Dawkins, R. (1976). *The Selfish Gene*. New York:Oxford University Press.
- Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford:W.H. Freeman and Co.
- Dugatkin, L.A. (1997). *Cooperation Among Animals, An Evolutionary Perspective*. New York:Oxford University Press.
- Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. New York: Dover.
- Hamilton, W.D. (1964). "The Genetical Evolution of Social Behavior I and II," *Journal of Theoretical Biology*. 7:1-52.
- Hamilton, W.D. (1975). "Innate Social Aptitudes of Man: An Approach from Evolutionary Genetics," in *Biosocial Anthropology*, (R. Fox, ed.) New York:John Wiley and Sons.
- Hamilton, W.D. (1987). "Discriminating Nepotism: Expectable, Common, Overlooked," in *Kin Recognition in Animals*, (D.C. Fletcher, C.D. Michener, eds.) New York:John Wiley and Sons.
- Hardin, G. (1968). "The Tragedy of the Commons," *Science*. 162:1243-48.
- Hardin, G. and Baden, J. (1977) *Managing the Commons*. San Francisco:W.H. Freeman and Co.

To appear in *Journal of Theoretical Biology*.

- Motro, U. (1991). "Cooperation and Defection: Playing the Field and the ESS," *Journal of Theoretical Biology*. 151:145-154.
- Price, G.R. (1970). "Selection and Covariance," *Nature*. 277:520-521.
- Rapoport, A. (1991). "Ideological Commitments and Evolutionary Theory," *Journal of Social Issues* 47:83-99.
- Reeve, H.K. (1999) "Multi-Level Selection and Human Cooperation," *Evolution and Human Behavior*. 21:65-72.
- Robersshaw, J.E., Mecca, S.J., and Rerick, M.N. (1978) *Problem Solving: A Systems Approach*. New York:Petrocelli Books, Inc.
- Simpson, E.H. (1951). "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society B*. 13:238-241.
- Sober, E. and Wilson, D.S. (1998). *Unto Others, The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA:Harvard University Press.
- Trivers, R.L. (1971). "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*. 46:35-57.
- von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton:Princeton University Press.
- Wade, M.J. (1977) "An Experimental Study of Group Selection," *Evolution*. 31:134-153.
- Wade, M.J. (1978) "A Critical Review of the Models of Group Selection," *The Quarterly Review of Biology*. 53(2):101-114.
- Wade, M.J. (1979) "The Primary Characteristics of *Tribolium* Populations Group selected for Increased and Decreased population Size," *Evolution*. 33(2):749-764.
- Williams, G.C. and Williams, D.C. (1957). "Natural Selection of Individually Harmful Social Adaptations Among Sibs with Special Reference to Social Insects," *Evolution* 11:32-39.
- Wilson, D.S. (1975) "A Theory of Group Selection," *Proceedings of the National Academy of Science USA*. 72(1):143-146.
- Wilson, D.S. (1976) "Evolution on the Level of Communities," *Science*. 192:1358-1360.
- Wilson, D.S. (1977) "Structured Demes and the Evolution of Group-Advantageous Traits," *The American Naturalist*. 111(977):157-185.
- Wright, R. (2000) *Non-Zero, The Logic of Human Destiny*. New York:Pantheon Books.
- Wynne-Edwards, V.C. (1962). *Animal Dispersion in Relation to Social Behavior*. Edinburgh: Oliver and Boyd.

## Simpson's Paradox and the N-Player PD

### Appendix A

Here we derive an expression for conditions necessary to produce Simpson's paradox in our model in terms of the initial conditions and the n-player PD parameters of slope and intercept. We assume that the initial number of players in groups 1 and 2 are equal and the initial distribution of cooperators and defectors is opposite and symmetrical in the two groups.

First note that because the utility line for defectors dominates the utility line for cooperators for all distributions of cooperators and defectors, the percentage of cooperators within a group always decreases. So the condition for Simpson's paradox reduces to an initial increase in total cooperators that exceeds any increase in total defectors or:

- $\Delta N_C > \Delta N_D$  (at  $t = 0$ )

The changes in cooperators and defectors in each group at  $t = 0$  are given by the following expressions:

- $\Delta N_{C1} = N_{C1} - m N_{C1} / N_1$
- $\Delta N_{C2} = N_{C2} - m N_{C2} / N_2$
- $\Delta N_{D1} = N_{D1} (m N_{C1} / N_1 + b_D)$
- $\Delta N_{D2} = N_{D2} (m N_{C2} / N_2 + b_D)$

Substituting into the above inequality and noting that  $\Delta N_C = \Delta N_{C1} + \Delta N_{C2}$  and  $\Delta N_D = \Delta N_{D1} + \Delta N_{D2}$ , the condition for Simpson's paradox is:

- $m(N_{C1}^2 / N_1 + N_{C2}^2 / N_2) > m(N_{D1} N_{C1} / N_1 + N_{D2} N_{C2} / N_2) + b_D(N_{D1} + N_{D2})$

Note that for our simulations  $N_C = N_D = N_1 = N_2$  and  $N_{D1} N_{C1} = N_{D2} N_{C2}$  so we can express this inequality in terms of the initial conditions in group  $i$  ( $i = 1$  or  $2$ ):

- $(m / N_i) (N_{Ci}^2 + N_{Di}^2) > (m / N_i) (2 N_{Ci} N_{Di}) + b_D N_i$
- $m (N_{Ci}^2 - 2 N_{Ci} N_{Di} + N_{Di}^2) > b_D N_i^2$
- $m (N_{Ci} - N_{Di})^2 > b_D N_i^2$
- $m / b_D > N_i^2 / (N_{Ci} - N_{Di})^2$

This is the condition for PD parameters and initial distributions in our simulation for Simpson's paradox to emerge.