

2001.11.13

## Simpson's paradox, stupidity and the selfish species

John Bigelow and Sam Butchart  
Monash University

Here is a simplified fiction which is based on a real case at a Californian University. The Faculty of Humanities decided to try to increase the number of women on their staff. There were 13 women and 13 men who applied for positions in the Faculty. All the positions were directed towards the study of either time or space, in the departments of History or Geography. There were 13 applicants for the positions in History and 13 applicants for the positions in Geography. Of the women applicants 8 applied for positions in History, and 5 for Geography, whereas 5 of the men applied for positions in History and 8 in Geography.

The History Department appointed 2 of the 8 women applicants (a 25% success rate for the women) and 1 of the 5 men (a 20% success rate for the men), so they congratulated themselves that they had favoured women over men. The Geography Department appointed 4 of the 5 women applicants (an 80% success rate) and 6 of the 8 men (a 75% success rate), so they too congratulated themselves that they had favoured women over men. Yet the Faculty was taken to task by the University administration, because they had had 13 male applicants and 13 female applicants and yet they appointed fewer women than men: 6 females and 7 males, so it appeared that the Faculty had favoured men over women:

|            | Women  |   | Men    |
|------------|--------|---|--------|
| History:   | 2 : 8  | > | 1 : 5  |
| Geography: | 4 : 5  | > | 6 : 8  |
| Faculty:   | 6 : 13 | < | 7 : 13 |

Call this pattern a *Simpson's Reversal* of inequalities. It is called Simpson's *Paradox* because it demonstrates that a certain form of argument can have true premises and a false conclusion even though, to many people in many contexts, arguments of this form appear to be logically valid. These tricky arguments have the same form as the following:

History favoured women applicants;  
Geography favoured women applicants;  
there were no Faculty applicants other than those in History and Geography;

therefore the Faculty favoured women applicants.

Arguments of this form are invalid but it is sometimes hard to see why.

In a nutshell, I submit, the reason why the Faculty appointed more men than women is this:

More of the women were applying for the job which is harder to get.

It was harder to get into History (success rates of 25% and 20%) than to get into Geography (success rates of 80% and 75%).

Generalizing, the pattern which is displayed by the case of the university appointments takes the following form:

|  |                   |     |                 |
|--|-------------------|-----|-----------------|
|  | “women’s team” AC |     | “men’s team” BD |
| “History heats” AB:  | $a : A$           | $>$ | $b : B$         |
| “Geography heats” CD:  | $c : C$           | $>$ | $d : D$         |
| “Faculty tournament” ABCD: $(a + c) : (A + C) < (b + d) : (B + D)$ |                   |     |                 |

Think of the women’s team as having won both the heats, because they had a higher success rate in each of those competitions; yet men won the tournament because they had a higher success rate overall.

There is a theorem to be found which formalizes the idea that men won the tournament because *more of the women were entering the competition in which is harder to win*. This theorem will have the form:

**Theorem:**

There will be a Simpson’s reversal of inequalities just when *this-many* more of the “women applicants” are competing for “jobs” which are *that-much* harder to get.

That is, when a women’s team AC beats a men’s team BD in each of two competitions AB and CD,

$$\begin{array}{rcccl} a & : & A & > & b & : & B \\ c & : & C & > & d & : & D \end{array}$$

the women’s team will nevertheless lose the whole tournament:

$$(a + c) : (A + C) < (b + d) : (B + D)$$

provided that one of the competitions, AB, is “harder” and the other one, CD, is “easier”, and there are *few* enough women, *a*, *winning* the harder competition AB to ensure that:

$$a : C < (b + d) : (B + D)$$

and there are sufficiently *many* women, *A*, *entering* the harder competition AB to ensure that:

$$c : A < (b + d) : (B + D).$$

**Proof.**

Consider the hypothesis that:

$$\begin{array}{rcccl} a & : & C & < & (b + d) : (B + D) \\ c & : & A & < & (b + d) : (B + D). \end{array}$$

This will be true if and only if (by “cross-multiplying”):

$$\begin{array}{rcl} (aB + aD) & < & (bC + Cd) \\ (cB + cD) & < & (Ab + Ad). \end{array}$$

These two entail (by adding together the two smaller left-hand sides, and the two larger right-hand sides) that:

$$(aB + aD) + (cB + cD) < (bC + Cd) + (Ab + Ad).$$

This is true if and only if:

$$(a + c)(B + D) < (b + d)(A + C),$$

which is true if and only if (by the reverse of “cross-multiplying”):

$$(a + c) : (A + C) < (b + d) : (B + D),$$

**which is what was to be demonstrated.**

### **Explanation.**

Here is another way of understanding how it is that a Simpson’s Reversal can occur. Consider the job-seekers example. The success rate for women was higher than for men in History, and in Geography. Nevertheless, the job in History was so much harder to get that the *higher* success rate in History was below the *lower* success rate in Geography. Despite faring *better* in History, those women only had a success rate of  $(2 / 8)$ . Despite faring *worse* in Geography, those men had a success rate of  $(6 / 8)$ . Notice that  $(2 / 8) < (6 / 8)$ , so that the success rate in the *more* successful group in History was higher than that of the *less* successful group in Geography. The women lost the tournament because more of them played in the History competition, and more of the men played in the Geography competition.

Transfer the inequality  $(2 / 8) < (6 / 8)$  in the job-seekers example to the general case described in the Theorem above, and it becomes what we might call a *diagonal* inequality:

$$(a / A) < (d / D).$$

Now draw on the fact that by increasing the *numerator* of a fraction you *increase* its value, while by increasing the *denominator* of a fraction you *decrease* its value. When we add first to the numerator and then to the denominator of a fraction we first increase, and then decrease its value. We could end up with almost any value at all, depending on how much we have increased it and then how much we have decreased it.

You can then visualize a derivation which proceeds from the above inequality

$$(a / A) < (d / D).$$

to the Simpson’s Reversed inequality:

$$(a + c) / (A + C) < (b + d) / (B + D)$$

by the following steps, using positions further to the right to represent larger fractions:

$$\begin{array}{ccc}
 \frac{a}{A} & & \frac{d}{D} \\
 \text{increase} & & \text{increase} \\
 \text{numerator} \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow & & \text{numerator} \rightarrow \\
 & \frac{(a+c)}{A} & \frac{(b+d)}{D} \\
 & \text{increase} & \text{increase} \\
 & \text{denominator} \leftarrow & \text{denominator} \leftarrow \\
 \text{Simpson's Inequality:} & \frac{(a+c)}{(A+C)} & \frac{(b+d)}{(B+D)}
 \end{array}$$

**Varying cases.**

The conditions under which a Simpson's Reversal will occur are in some respects robust and in some respects fragile. In the job-seekers example, if you change any one or more of the numbers the Simpson's Reversal is likely to vanish and this makes the phenomenon seem fragile. Yet the reversal of inequalities can also survive a large number of different *kinds* of variation to the numbers in the example.

There is a natural conjecture that the phenomenon is due to a "skewing" of the sample sizes, so that one sample "swamps" another. However, this is misleading. In the job-seekers example there were exactly as many men as women, and exactly as many applicants in History and in Geography, and yet we still obtained a Reversal. The asymmetries which produced the reversal were these: the success rate in History was low, and the success rate in Geography was high; and more than half the women were applying for the History jobs, and more than half of the men were applying for the Geography job.

The numbers in the initial example were small, but of course they could be converted into large numbers and the pattern could remain:

$$\begin{array}{ccc}
 2000 : 8000 & > & 1000 : 5000 \\
 4000 : 5000 & > & 6000 : 8000 \\
 6000 : 13000 & < & 7000 : 13000
 \end{array}$$

Then this pattern would be robust under a range of deviations around the round numbers given above:

$$\begin{array}{ccc}
 2003 : 8004 & > & 1005 : 5006 \\
 4007 : 5008 & > & 6009 : 8010 \\
 6010 : 13012 & < & 7014 : 13016
 \end{array}$$

The *bounds* of deviation around these round numbers can be marked by the following limiting cases:

$$\begin{array}{lll}
 2 : 8 > 1 : 5 & 2 : 8 > 1 : 5 & 4 : 8 > 2 : 5 \\
 3 : 5 > 4 : 8 & 4 : 5 > 6 : 8 & 4 : 5 > 6 : 8 \\
 \\ 
 5 : 13 = 5 : 13 & 6 : 13 < 7 : 13 & 8 : 13 = 8 : 13
 \end{array}$$

In the left-hand example the success rate in the easier competition has been brought down from the very high 75-80% success rate, to around the much more moderate 50%, at which point the Simpson's Reversal disappears. In the right-hand example the success rate in the harder competition has been brought up from the very low 20-25% to around the much more gratifying 50%, at which point the Simpson's Reversal disappears.

In all these examples the numbers on the two different teams, and the numbers in the two different competitions, are all of the same order of magnitude. It is possible, however, for the women to vastly outnumber the men, or *vice versa*, without disturbing the pattern of the Simpson's Reversal:

$$\begin{array}{ll}
 \text{"Few men":} & \text{"Few women":} \\
 200 : 800 > 1 : 5 & 2 : 8 > 100 : 500 \\
 400 : 500 > 6 : 8 & 4 : 5 > 600 : 800 \\
 \\ 
 600 : 1300 < 7 : 13 & 6 : 13 < 700 : 1300.
 \end{array}$$

This means that the men can increase their frequency not only in the case where they start out in equal numbers to the women, but also when they are vastly outnumbered by the women, and also when women have nearly disappeared from the competition altogether. In the left-hand case men climb from a tenth of the total number of competitors to more than a tenth of the winners; in the right-hand case the women sink from a tenth of the competitors to less than a tenth of the winners.

### Evolutionary and Economic applications

Substitute rats and lemmings for women and men in the story of the job at a Californian university, and for "getting a job" read "reproductive success".

Imagine rats are smart, self-interested and **rational**. Each rat, we may imagine, does whatever boosts the chances of a higher long-term reproductive success rate for itself and its nearest kin, even if this *reduces* the long-term success rate for many of its neighbours.

Lemmings are stupid and irrational. Each lemming does things which bring about a lower long-term reproductive success rate for itself and its nearest kin. We might then imagine that this often enhances the success rate for many of its neighbours. It is almost as if the lemmings were altruistic, and were refraining from interfering with the reproductive success rates of their neighbours. In extreme cases, when food is scarce we might imagine that many of the lemmings even fling themselves over a cliff in order to bring the population down to a level at which the survivors will have a better chance of long-term survival. Yet we might equally well imagine that

the lemmings have no idea what they are doing. Evolution does not care about their good intentions, only about results.

In Norway we may suppose that at the beginning of Year One there are 8 billion rats and 5 billion lemmings. The survival rate for both rats and lemmings in Norway is low because there are more rats than lemmings and so each rodent has to compete with many ratty neighbours. In Sweden, in contrast, there are only 5 billion rats and 8 billion lemmings. The survival rate for both rats and lemmings in Sweden is higher than it is in Norway because there are fewer rats.

In this situation it is mathematically possible for there to be a replication of the pattern which we found in the appointments at the Californian university. Even though rats do better than lemmings in Norway, and rats do better than lemmings in Sweden, lemmings may nevertheless do better than rats in Scandinavia. It is a separate question whether it is biologically possible for this pattern to occur, but it is worth keeping it in mind that at any rate it is a mathematical possibility.

Lemmings may do better than rats because more of the rats are competing against neighbours who are harder to beat. They are competing against neighbours who are harder to beat because they are living in Norway, where more of the rats are.

The same pattern can be replicated in economic theory. Take a “rat” to be a business which reliably and efficiently acts in such a way as to maximize its own profits and its own chances of long-term survival, even if this reduces the profits and the survival prospects of other businesses in its economic neighbourhood. Take a “lemming” to be a business which regularly acts in ways which *lower* its own profits and its chances of long-term survival, that is, ways which give it lower profits-and-prospects than it would have had if it had done what a business-rat would have done. It is of course possible for a business to forego short-term profits in order to gain *reputation* which will give it higher profits in the longer term. Yet we may imagine that the “lemmings” forego short-term profits in a clumsy way which makes them appear to be ruthless and so gives them a *bad* reputation, resulting in *lower* profits in both the short and the long term. We might imagine also that this behaviour of a business-lemming often permits neighbouring business to draw higher profits and to enjoy better prospects of long-term survival.

The business-lemming may be inhibited from exploiting some business opportunities because it feels bound by moral compunctions which a rattier business would not worry about; or it may be motivated by altruistic concern for the neighbouring businesses. Alternatively it may be acting “emotionally”, say by pursuing a costly vendetta against one particular rival, driven by malice, or by a grievance of some kind; or it may be simply stupid, lazy and inefficient. Whatever the genesis of its shortcomings, we may imagine the upshot to be that the business-lemming acts in such a way as to gain lower profits than it could otherwise have achieved if it had made the choices which the business-rats would have made.

Defining business rats and lemmings in this way, Simpson’s Paradox demonstrates that even if business-rats do better than business-lemmings *locally*, this inequality can reverse *globally*. The inequality can reverse if more of the business-rats are living in neighbourhoods where their neighbours are harder to cheat. More of the rats will be living in neighbourhoods like that just in case they are living in the neighbourhoods which contain more of the rats. That is, a Simpson’s Reversal may occur provided that more of the rats are living in places where there are more rats. Hence Simpson’s Paradox reveals one way in which, under the right conditions, business-lemmings can sustain a significant frequency in a population despite the fact that each of them could have had higher profits and better long-term chances of survival if it had done what a business-rat would have done.

## **Computer simulation**

The mathematical possibility of long-term survival for irrational, emotional, vindictive, stupid, lazy, inefficient, moral and altruistic lemmings can be demonstrated by a computer program. Divide the screen into cells which are aligned in rows and columns, each cell having 8 immediate neighbours. Each cell is then to play a game of Prisoner's Dilemma with each of its neighbours. In a game of Prisoner's Dilemma each player has a choice between two options called "Trust" and "Defect". Each cell is designated as being either a *rat* or a *lemming*; and in each of the games it plays with its neighbours a rat always Defects and a lemming always Trusts.

When a lemming plays a rat it gets a low score; and it would have got a higher score if it had done what a rat would have done, namely Defecting. If a lemming plays against another lemming both of them Trust, and each gets a higher score than it would have got if it had been playing against a rat. However, when a lemming plays against another lemming it gets a lower score than it would have obtained if it had done what a rat would have done, namely Defecting. In either case, whether facing a rat or a lemming, the lemming would have obtained a higher score if it had done what a rat would have done.

Each cell accumulates a total score from the 8 games it plays with its neighbours. Then after a round of these games each cell takes the character of the neighbour with the highest score (where for this purpose we count a square as one of its own neighbours). If the highest score among a lemming's neighbours is a rat, then the lemming's square comes to be occupied by a rat.

The payouts are arranged so that it is true of every lemming cell that it would have obtained a higher total score if it had made the moves which a rat would have made. It is also true of every rat cell that it would have obtained a lower total score if it had made the moves which a lemming would have made.

Now partition the cells into the following two sets. The first set comprises all the cells for which at least half of the immediate neighbours are rats; call this the "ratty", or the R-set. The second set comprises all the cells for which less than half of the immediate neighbours are rats; call this the L-set.

Rats in the L-set score better than rats in the R-set. Hence the score-rate for a rat's strategy depends on the *frequency* of rats in the neighbourhood. This echoes a phenomenon in population genetics called *frequency-dependent fitness*. Because of the frequency-dependence of score rates, lemmings in set L score *worse* than rats in L but *better* than rats in R.

In set L, rats score better than lemmings. In set R, rats score better than lemmings. Yet it would be a fallacy, a Simpson's Fallacy, to conclude that rats must have scored better overall.

If more of the rats are in the set where your neighbours are harder to cheat, namely set R, then there can be a Simpson's Reversal of inequalities when we compute the global score rates for rats and lemmings. Hence it is possible for rats to score better than lemmings in each of the sets L and R, and yet to score worse than lemmings in the union of sets L and R. That is why it is possible for lemmings to persist even though each lemming *could* have done better if it had done what a rat would have done in the same situation.

When a program is designed under the description above, the range of patterns which emerge in a simulation will depend on the fine-tuning of the payouts in the games of Prisoner's Dilemma, and on whether scores accumulate over a series of games or whether cells start again from a score of zero in each successive round of games. There are, however, at least some tunings of the variables in which lemmings perform just as well as rats.

Simpson's Paradox can result in the persistence of stupid, inefficient, lazy, altruistic and moral animals or businesses provided that their lemminglike behaviour makes it easier for their neighbours to survive. An individual of this kind often *refrains* from doing what would enhance its own long-term chances of survival. In the biological case it may even happen that a *subset* of

the animals within a species will regularly *fail* to do something which would enhance the long-term reproductive success of *all* the animals in their own reproductive lineage. Provided that this lemminglike abstention enhances the reproductive success of neighbours, it is possible for a Simpson's Reversal to boost the frequency of the lemminglike subset of this species. It can do this even if the neighbours whom the lemmings' behaviour benefits are *genealogically unrelated* to the lemmings whose behaviour benefits them.

Thus Simpson's Paradox reveals a way in which it is at least *mathematically* possible for stupid, inefficient, lazy, emotional, altruistic and moral animals or businesses to persist even under a rigorous imposition of ruthless natural selection, or a heartless free market economy.

### **Metaphysics.**

One way in which a property can have an increasing frequency of instances in a population is by causing its instances to *reproduce* themselves, that is by causing its instances to *cause* there to be other things which will have the same property. Yet sometimes *omissions* can be just as effective as *acts*. If a property causes each of its instances to *fail* to take advantage of a certain range of opportunities to boost its own long-term reproductive chances, this may *boost* the long-term reproductive chances of its neighbours. In neighbourhoods in which there are many individuals who have this property, reproduction is easier.

Hence Simpson's Effect can come into play and can increase the frequency of a property of this kind in a population. One way in which a selfish property can boost its own frequency in a population is by causing its instances to *refrain* from doing things which will depress the reproductive success rates for its neighbours. It is an open empirical question how often this mathematical possibility is realized not only in actual biological or economic situations, but also in *any* case in which things which have a property have a propensity to cause there to be other things which have that same property, whatever ontological category those things might fall under.

### **APPENDIX.**

#### **“Sharks and Suckers I”, variant on John Conway’s “Game of Life”.**

As a short-cut to empirical research on Simpson's Paradox in the social sciences, I propose a computer game, pitting virtual Sharks against virtual Suckers, and seeing which of them win in the long run.

Divide the computer screen into cells, like graph paper. Each cell can be empty, or can be filled with either a Shark or a Sucker. We take turns. I fill a cell with a Shark, then you fill one with a Sucker; we take turns until together we have filled, say, a quarter of the cells on the screen.

Empty = neither 0 nor 1.

Full = 0 or 1.

Sharks = 1; suckers = 0.

After a round of filling each cell with 0 or 1, or leaving it empty, you then set a program to run on the computer, a program which determines the survival and spread of Sharks and Suckers, for generation after generation, until we arrive eventually either at victory for one side, or the other, or

else there is a stalemate, because neither prevails over the other. For each generation, the computer will operate the following rules for determining whether any occupied or unoccupied cell A will, in the next generation, become or remain empty, or will become or remain filled with a Shark, or a Sucker.

**Cell A is empty: “Birth or barrenness”.**

Rule 1: “Too crowded”:

If cell A is empty, and three or more adjoining cells are full, then it stays empty.

Rule 2: “Too lonely”:

If cell A is empty, and fewer than two adjoining cells are full, then it stays empty.

Rule 3: “Sharks rule”:

If cell A is empty, and just two adjoining cells, B and C, are full, then:

- if both B and C are 0s, A becomes a 0;
- if just one of B and C are 1s, A becomes a 1;
- if both of B and C are 1s, A stays empty.

**Cell A is full: “Death or survival”.**

Rule 4: “Too crowded”:

If cell A is full, and three or more adjoining cells are full, then it becomes empty.

Rule 5: “Too lonely”:

If cell A is full, and no adjoining cells are full, then it becomes empty.

Rule 6: “Status quo”:

If cell A is full, and just one adjoining cell is full, then:

- if A is a 0, A stays a 0;
- if A is a 1, A stays a 1.

Rule 7: “Sharks rule”:

If cell A is full and just two adjoining cells B and C are full, then:

- if A is a 0, and both B and C are 0s, then A stays a 0;
- if A is a 0, and either B or C (or both) are 1s, then A becomes empty;
- if A is a 1, and both B and C are 1s, then A becomes empty;
- if A is a 1, and either B or C (or both) are 0s, then A stays a 1.

**Questions:**

How many initial distributions of 0s and 1s would result in a victory for Suckers? And how many would result in a stalemate, under the above rules? And would the answer change dramatically, if you rewrote the above rules?

**“Sharks and Suckers II”, variant on Axelrod’s tournaments of iterated Prisoner’s Dilemmas.**

The screen is divided into *cells*, displayed in a grid of 64 rows and 64 columns, each cell identified by row and column as  $\langle x, y \rangle$ . Rows 0 and 65 enter into calculations but are off-screen; likewise for columns 0 and 65. Cell  $\langle 1, 1 \rangle$  will be at the top left of the screen, and  $\langle 64, 64 \rangle$  will be at the bottom right.

Each player picks a number from 0 to 9, which represents their *strategy* in repeated games of Prisoner’s Dilemma. Each player puts their number in various cells on the screen, in whatever pattern they choose. Each picks any unoccupied cell they choose, when their turn comes, and puts their number in it; and they take turns until there are 10 occupied cells for each player.

Then you run the program and see who wins the tournament. The one who wins is the one whose number is the last to disappear from the screen; or, if several numbers continue to be present on the screen indefinitely, the winner is the one who ends up with the most cells occupied by their number; otherwise, the game results in a stalemate.

The basis of the *program* which determines a tournament winner will be this: each cell plays a series of games of Prisoner’s Dilemma with each of its neighbours. In each game, each participant has to choose either to be “suckered” by their neighbour, or to “shark” them. To be “suckered” is to play “0”; to “shark” your neighbour is to play “1”. The pay-off matrix for the game is the following:

If you play 0 then:

should they play 0, you score minus 1; should they play 1, you score minus 2.

If you play 1 then:

should they play 0, you score plus 1; should they play 1, you score minus 1.

You play a round of eight games, one with each of your eight neighbours. The scores from these eight games are then added onto your overall score at the start of the round, to give a new overall score for the end of that round. This score for any given cell will then determine whether your number will “survive” in that cell, or will “die”. And it will also determine whether this cell “spawns”, and spreads your number into previously unoccupied cells.

Your number, chosen from 0, ... , 9, represents your “strategy” for deciding whether to “shark” or be “suckered” in each of the games of Prisoner’s Dilemma which you have to play with each of your neighbours. Your strategy will be represented by a function which maps any given *sequence of pairs* onto either a 0 or a 1. The strategy tells you whether to be “suckered” by your neighbour, or to “shark” your neighbour; and which you choose to do may depend on what the neighbour, against whom you are playing, has done to you, in similar games in the past. Your record of past games with that neighbour is represented by the sequence of pairs, and a

strategy maps that record of past games onto a decision about what to do in this game. If the function representing your strategy is one which maps the sequence onto a 0, then you let yourself be suckered; if the function maps the sequence onto a 1, then you shark your neighbour.

The sequence of pairs is a record of what you and your neighbour have done in past games. If the first pair in the sequence is  $\langle 0, 0 \rangle$ , that means that in the first game you let yourself be suckered and so did your neighbour. If the second pair in the sequence is  $\langle 1, 0 \rangle$ , that means that in the second game you sharked your neighbour and they let themselves be suckered by you. Given your track record with a given neighbour, your strategy advises you what to next. Here are three strategies:

Strategy0 is the “Sucker Strategy”, which maps any sequence at all onto 0: “Let yourself be suckered, no matter what your neighbour has done before - whatever they have done to you, always turn the other cheek”.

Strategy1 is the “Shark Strategy”, which maps any sequence at all onto 1: “Never give a sucker an even break, and never let yourself be suckered.”

Strategy2 is the “Tit for tat Strategy”; this one maps the empty sequence onto 0 (“If you’ve never played them before then give them the benefit of the doubt”), and any nonempty sequence whose last pair is  $\langle a, b \rangle$  is mapped onto  $b$  (“Do unto them what they have last done unto you”).

Strategy3 is the “Random Strategy”; this one maps every sequence onto whichever number, 0 or 1, is selected by a process of random choice.

Players can choose any one of these strategies “off the rack”, or else they can write one of their own.

Each cell on the screen  $\langle x, y \rangle$  is then assigned an *initial* Strategy: if the players have not assigned to  $\langle x, y \rangle$  any of the strategies 0, ..., 9, then I will say it is unoccupied and it has the “null” strategy. After this initial assignment, there will be a series of Rounds in which each cell is re-assigned a Strategy, along with several other things, according to fixed rules. The initial assignment will be called Round0; succeeding this will be Round1, Round2, and so on.

In each Round, each cell will be assigned the following, in addition to a strategy:

*Files:*

Each cell has, and updates on each Round, eight “files” on its neighbours. These files may be called called the *a-file*, *b-file*, ..., *i-file*. The neighbours of any given cell, call it “e”, are to be visualized as mapped out in the pattern:

```
a b c
d e f
g h i
```

so that if  $\langle x, y \rangle$  is the cell we are interested in, it is in position “e”; its neighbours in the row above are a, b and c; its neighbours to its left and right on its own row are d and f; and its neighbours on the row below are g, h and i.

I will write  $j\langle x, y \rangle$  as a schema for “the neighbour of  $\langle x, y \rangle$  in position  $j$ ”, where “ $j$ ” ranges over “a”, ... , “i”. So for instance  $a\langle x, y \rangle$  will be the neighbour of  $\langle x, y \rangle$  which is in position a - that is,  $a\langle x, y \rangle = \langle x-1, y-1 \rangle$ .

The  $j$ -file for any cell  $\langle x, y \rangle$  will change, round by round. For each round,  $\text{Round}n$ , and for each neighbour  $j$ , we will assign to cell  $\langle x, y \rangle$  a  $j$ -file:  $\text{Round}n.j\text{-file}\langle x, y \rangle$ . This will be the record of all previous rounds of Prisoner’s Dilemma games which have been played between cell  $\langle x, y \rangle$  and its  $j$ -neighbour,  $j\langle x, y \rangle$ .

#### *Plays:*

Each cell, on each Round, plays eight games of Prisoner’s Dilemma, one against each of its neighbours. In these eight games, it makes eight “plays”, one each against each of its neighbours. I call these the  $a\text{-play}$ ,  $b\text{-play}$ , ... ,  $i\text{-play}$ . Each of these plays will be determined by applying cell’s strategy to its file on each neighbour it is playing; and so each of these plays will be either a 0 or a 1:  $\text{Round}n.j\text{-play}\langle x, y \rangle = 0$  or 1.

#### *Takings:*

After making its “play” with each of its neighbours, each cell has to see what “play” its neighbour will make, before it will be clear what its “takings” will be from this game. Given the play  $\langle x, y \rangle$  makes against its  $j$ -neighbour, and the play which that  $j$ -neighbour makes against it, you can determine the  $j$ -takings for cell  $\langle x, y \rangle$ . You determine those  $j$ -takings by looking up the consequences of those two plays, according to the Prisoner’s Dilemma game as defined above. In each round,  $\text{Round}n$ , therefore, each cell  $\langle x, y \rangle$  will be assigned eight “takings”:  $\text{Round}n.a\text{-takings}\langle x, y \rangle$ , ... ,  $\text{Round}n.i\text{-takings}\langle x, y \rangle$ .

#### *Score:*

In each round, you will carry over a score from the previous round, and then you will adjust this score by adding on the “takings” from each of the games of Prisoner’s Dilemma which you have played with your neighbours. Thus, for each cell,  $\langle x, y \rangle$ , and each round,  $\text{Round}n$ , we will have the score for this cell at the end of this round:  $\text{Round}n.\text{score}\langle x, y \rangle$ .

Thus, for each cell  $\langle x, y \rangle$ , and for each round of the tournament, we will have to define the following: a strategy, eight  $j$ -files, eight  $j$ -plays, eight  $j$ -takings, and a score. We define these recursively. First we set each of these for  $\text{Round}0$ ; then we give a rule mapping each of these values in  $\text{Round}(n - 1)$  onto the new value for  $\text{Round}n$ .

### **1. Strategy0:**

Here we set the “initial distribution” of the various contending Strategies in the tournament.

$\text{Round}0.\text{Strategy}\langle x, y \rangle = \text{“null”}$ , and the cell is “unoccupied” or “empty”, unless a player changes it to one of the numbers 0, ... , 9.

### **2. Score0:**

Here we give each occupied cell a “life force” which it can either add to, or squander, by playing games of Prisoner’s Dilemma with its neighbours. All cells are, we will suppose, created equal, with an initial score of (say) 9.

$\text{Round0.Score}\langle x, y \rangle = 9$  if  $\text{Round0.Strategy}\langle x, y \rangle$  is not null, and  $= 0$  otherwise.

### 3. Takings0:

$\text{Round0.j-takings}\langle x, y \rangle = \text{null}$ , for each of the neighbours  $j = a, \dots$ , or  $j = i$ .

### 4. Play0:

$\text{Round0.j-play}\langle x, y \rangle = \text{null}$ , for each of the neighbours  $j = a, \dots$ , or  $j = i$ .

### 5. Files0:

$\text{Round0.j-file}\langle x, y \rangle = \text{null sequence}$ , for each of the neighbours  $j = a, \dots$ , or  $j = i$ .

That completes the “initial settings”. Next, we have five recursive rules for updating each of the above settings as the tournament progresses from  $\text{Round}(n - 1)$  to  $\text{Round}n$ .

### 6. Strategies, Round $n$ :

The default assumption is “No change in strategy” for any given cell  $\langle x, y \rangle$ ; that is,

$$\text{Round}n.\text{Strategy}\langle x, y \rangle = \text{Round}(n - 1).\text{Strategy}\langle x, y \rangle,$$

except in the following three cases:

- (i) “death from within” occurs when the score has sunk to 0: if  $\langle x, y \rangle$  is occupied and  $\text{Round}(n - 1).\text{Score}\langle x, y \rangle = 0$ , then  $\text{Round}n.\text{Strategy}\langle x, y \rangle = \text{null}$ , and the “strategy number” is erased from the cell;
- (ii) “death from loneliness” or “death from overcrowding” occurs when scores of surrounding neighbours is either too low or too high: if  $\langle x, y \rangle$  is occupied and the sum of scores for all neighbours on  $\text{Round}(n - 1)$  is less than 5 (“loneliness”) or greater than 20 (“overcrowding”), then  $\text{Round}n.\text{Strategy}\langle x, y \rangle = \text{null}$ , and the strategy number is erased from the cell;
- (iii) “birth” occurs in an unoccupied cell, when scores of surrounding neighbours are neither too low nor too high: if the cell is unoccupied, and if the sum of scores for all neighbours on  $\text{Round}(n - 1)$  is both greater than 5 and also less than 20, and if there is a highest score among the neighbours and that highest score is possessed by cell  $\langle j, y \rangle$ , then  $\text{Round}n.\text{Strategy}\langle x, y \rangle = \text{Round}(n - 1).\text{Strategy}\langle j, y \rangle$ .

### 7. Scores, Round $n$ :

The score is obtained by taking the cumulative score for previous rounds, and adding on the “takings” from each of the games of Prisoner’s Dilemma which you play with your neighbours, on this round.

$$\begin{aligned} \text{Round}n.\text{Score}\langle x, y \rangle &= \text{Round}(n - 1).\text{Score}\langle x, y \rangle \\ &+ (\text{Round}n.a\text{-takings}\langle x, y \rangle + \dots + \text{Round}n.i\text{-takings}\langle x, y \rangle). \end{aligned}$$

## 8. Takings, Round $n$ :

Consider first the takings between cell  $\langle x, y \rangle$  and its a-neighbour,  $\langle x - 1, y - 1 \rangle$ . Note that whereas  $\langle x - 1, y - 1 \rangle$  is the a-neighbour for cell  $\langle x, y \rangle$ , we should remember that cell  $\langle x, y \rangle$  is the i-neighbour for  $\langle x - 1, y - 1 \rangle$ . So to find the takings for  $\langle x, y \rangle$  in its game with its a-neighbour  $\langle x - 1, y - 1 \rangle$ , you have to match up the “a-play” which  $\langle x, y \rangle$  makes against its a-neighbour with the “i-play” which that neighbour  $\langle x - 1, y - 1 \rangle$  makes against its i-neighbour  $\langle x, y \rangle$ .

In general, taking  $j$  to be any of the neighbours from a to i, where  $j\langle x, y \rangle$  is the  $j$ -neighbour of  $\langle x, y \rangle$ , let us say that  $\langle x, y \rangle$  is the  $j^*$ -neighbour of  $j\langle x, y \rangle$ . Thus, for instance, the a\*-neighbour of cell  $\langle x - 1, y - 1 \rangle$  is its i-neighbour,  $\langle x, y \rangle$ .

We may then lay down the following as the general rule for determining the “takings” from any given round of Prisoner’s Dilemma between two neighbours  $\langle x, y \rangle$  and  $j\langle x, y \rangle$ :

$$\begin{aligned} \text{Round } n.j\text{-takings}\langle x, y \rangle &= \text{minus } 1, \\ &\quad \text{if } \text{Round } n.j\text{-play}\langle x, y \rangle = 0 \text{ and } \text{Round } n.j^*\text{-play}j\langle x, y \rangle = 0; \\ &= \text{minus } 2, \\ &\quad \text{if } \text{Round } n.j\text{-play}\langle x, y \rangle = 0 \text{ and } \text{Round } n.j^*\text{-play}j\langle x, y \rangle = 1; \\ &= \text{plus } 1, \\ &\quad \text{if } \text{Round } n.j\text{-play}\langle x, y \rangle = 1 \text{ and } \text{Round } n.j^*\text{-play}j\langle x, y \rangle = 0; \\ &= \text{minus } 1, \\ &\quad \text{if } \text{Round } n.j\text{-play}\langle x, y \rangle = 1 \text{ and } \text{Round } n.j^*\text{-play}j\langle x, y \rangle = 1; \\ &= \text{null, otherwise.} \end{aligned}$$

## 9. Plays, Round $n$ :

We derive the “play” which cell  $\langle x, y \rangle$  makes against its  $j$ -neighbour by applying the “strategy” which is assigned to that cell on that round. To apply this “strategy”, we have to look up the “file” which the cell has on its  $j$ -neighbour. The “file” is a sequence of pairs of 0s and 1s, giving the record of past games of Prisoner’s Dilemma. The “strategy” is a *function* which maps such a sequence onto advice about whether to let yourself be suckered, or to shark your neighbour. Take the relevant “file” as an *argument* for the *function* which constitutes the “strategy”, and this function will map that argument onto the required “play” for this game in this round.

$$\text{Round } n.j\text{-play}\langle x, y \rangle = \text{Round } n.\text{Strategy}\langle x, y \rangle(\text{Round } n.j\text{-file}\langle x, y \rangle).$$

## 10. Files, Round $n$ :

To up-date the file on a neighbour, you just add to the pre-existing file a new entry, registering what you and that neighbour “played” against each other, in your previous round of Prisoner’s Dilemma. Your pre-existing file was a sequence of pairs; to update it, you just have to add a new pair to the sequence. Thus, we have:

$$\begin{aligned} \text{Round } n.j\text{-file}\langle x, y \rangle &= \langle \text{Round}(n - 1).j\text{-file}\langle x, y \rangle, \langle u, v \rangle \rangle, \text{ where} \\ u &= \text{Round}(n - 1).j\text{-play}\langle x, y \rangle, \text{ and} \\ v &= \text{Round}(n - 1).j^*\text{-play}j\langle x, y \rangle. \end{aligned}$$

**Questions:**

Will Suckers ever prosper? Or do Sharks rule okay? Does the Simpson's effect result in a higher frequency of those with a lower score?

And if you change the pay-offs for the Prisoner's Dilemma games which feature in the tournament - either giving Sharks more, or else less, advantage relative to Suckers - will that significantly change the success-rates for different Strategies? And what would be the effects of different fine-tunings of various of the other variables in the program?

In initial pilot trials by Brian Weatherson at Monash, it appears as though Suckers can sometimes prosper. There are ways of tweaking variables so that, most of the time, no strategies at all will survive long-term, and for most initial distributions when you run the program the screen very soon becomes empty. When everyone is, as it were, on the verge of extinction, then it sometimes happens that there will be a small cluster of Suckers who are barely clinging to life, while everywhere else on the screen Sharks are initially wiping Suckers out, but then in the end the Sharks, too, fall over the brink into extinction - leaving only the tenuously surviving cluster of Suckers behind. This is suggestive: genuine self-sacrifice is to be expected less often in conditions of affluence than under conditions of hardship.

Further program development by Sam Butchart confirms the hypothesis that, thanks to Simpson's Reversals, Suckers can prosper.

**References.**

- Axelrod, Robert M. (1984)  
*The Evolution of Co-operation*, Basic Books, New York NY.
- Bigelow, Robert S. (1969)  
*The Dawn Warriors: Man's Evolution Toward Peace*, Little, Brown, Boston MA.
- Darwin, Charles (1871-4)  
*The Descent of Man and Selection in Relation to Sex*, 2nd edn., John Murray, London UK, 1906; Part I and the concluding chapter of Part III reprinted in the *Thinker's Library*, Watts & Co., London UK, 1930.
- Dawkins, Richard (1976)  
*The Selfish Gene*, Oxford University Press, Oxford UK.
- Dobzhansky, Theodosius G. (1962)  
*Mankind Evolving: Gthe Evolution of the Human Species*, Yale University Press, New Haven CT.
- Gould, Stephen Jay and Eldridge, Niles (1977)  
"Punctuated equilibria: the tempo and mode of evolution reconsidered", *Paleobiology* 3, pp.115-151.
- Gould, Stephen Jay and Lewontin, R.C. (1979)  
"The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme", *Proceedings of the Royal Society, London*, B205, pp.581-598.
- Gould, Stephen Jay (1980)  
*The Panda's Thumb: More Reflections in Natural History*, Penguin, London UK.
- Hamilton, W.D. (1971)  
"Geometry for the selfish herd", *Journal of Theoretical Biology* 31, pp.295-311.
- Maynard Smith, John and Parker, G.A. (1976)

- “The logic of asymmetric contests”, *Animal Behaviour* 24, pp.159-175.
- Maynard Smith, John and Price, G.R. (1973)  
“The logic of animal conflicts”, *Nature* 246, pp.15-18.
- Maynard Smith, John (1976)  
“Evolution and the theory of game”, *American Scientist* 64, pp.41-45.
- Mittal, Y.,  
*Journal of the American Statistical Association* [Theory and Methods Section] 86  
(1991), pp.167-172.
- Ridley, Matt (1996)  
*The Origins of Virtue*, Viking, Penguin Books, Harmondsworth, England.
- Sober, Elliott (1988)  
*Nature of Selection: Evolutionary Theory in Philosophical Focus*,
- Wilson, Edward O. (1971)  
*The Insect Societies*, Harvard University Press, Cambridge MA.
- Wilson, Edward O. (1975)  
*Sociobiology*, Harvard University Press, Cambridge MA.
- Wright, Sewall (1917)  
“The average correlation within subgroups of a population”, *Journal of the Washington Academy of Science* 7, pp.523-535.
- Wright, Sewall (1931)  
“Evolution in Mendelian populations”, *Genetics* 16, pp.97-159.
- Wright, Sewall (1968, 1969, 1977, 1978)  
*Evolution and the Genetics of Populations: A Treatise in Four Volumes*, University of Chicago Press, Chicago IL and London UK.