Detailed appendix for "A Tale of Two Defectors"

In this document, I go through some of mathematical steps in deriving the results for the paper. Unfortunately, many of these steps were omitted from the appendix attached to the paper due to length issues. You should keep in mind that I have not really gone through this thoroughly, so there may be errors. If you find any, please let me know.

In laying out some of the logic of the model, I will show how some of steps that were omitted in the original derivations are calculated. Along the way there may be some things that I discuss that will be helpful in understanding the derivations of and the logic behind other papers that deal with cooperation from and evolutionary game-theoretic perspective. If there are steps that are too basic, I apologize.

In our paper, we are making a model of the social dynamics of reputation. That is, what happens when individuals know how their fellow community members behave with one another and can use that information to behave contingently?

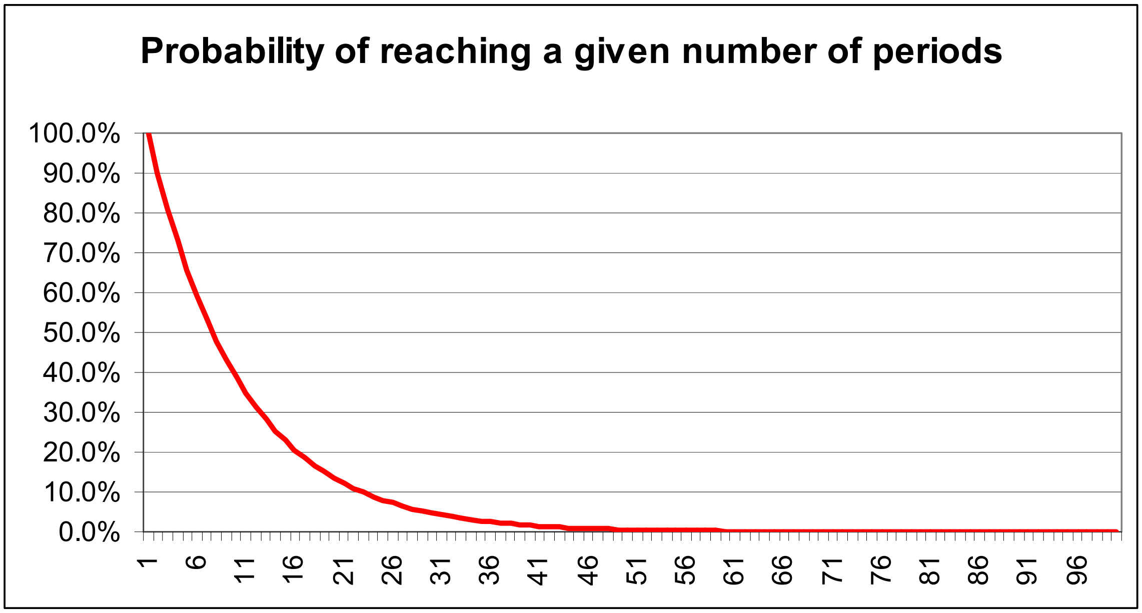
To begin, we define the rules of the game. In our game, individuals interact with a randomly selected partner in each period and play a one-shot prisoners' dilemma. After the first round, individuals know, with some fixed probability, what their current partner did in the last period (i.e., cooperate or defect). Additionally, they know what the standings (or reputation) of both their current partner and his/her previous partner. With these three pieces of information, an individual can come up with an assessment rating (or standing) of their current partner. Obviously, in the real world, this information doesn't come for free. There must some process, such as direct observation or gossip, that disseminates such information. It's okay that we ignore such a process in our model as we are not particularly interested in that question. Instead, building on the pioneering work of Nowak and Sigmund (1998), we want to know what types of reputation-assigning rules can stabilize cooperation. In their model, Nowak and Sigmund consider an *image-scoring* rule where individuals only attend to the behavioral decision of their partners (i.e., cooperate or defect). Such a rule does not take into account the context of that interaction (i.e., was a particular act of defection motivated by greed or an intent to withhold cooperation from a cheater). In our paper, we show that such a strategy is not evolutionarily stable because it ends of punishing individuals who punish cheaters. We show, however, that a *standing* rule is evolutionarily stable. Such a rule takes context into account (i.e., an individual's reputation is not tarnished if she refuses to help a known cheat). Intuitively, such a reputation-assignment rule feels right. However, intuition often fails when processes become complex. That's why we turn to simple model building.

Okay, we know how a particular round of social interaction works. We next set up the rules for deciding how many periods of there will be. This will be important in understanding the evolutionary dynamics of the system.

There is always one bout of indirect reciprocity in our model. Afterwards, with probability *w*, the game goes on for another bout. If we reach a second period, then with the same probability *w* we will reach a third bout. Such a model has no history. Reaching some arbitrary round *n* does not in any way impact the probability of reaching round $n + 1$. It is sort of like radio active decay. This type of model seems implausible and it is. In many models of cooperation, this type of rule is used. We could make things even simpler by assuming a fixed period of *N* rounds. If we do this, however, we must be careful not to introduce strategies which play contingently based on what round in the sequence they are in. Then, we would expect a strategy to always defect on the last round as there are no consequences. Strategies will anticipate this defection and also defect on the second to last round. The process of "backwards induction" goes on until we end up with no cooperation. That is an interesting problem which has vexed economists for a long time. However, that is not what we are interested here. As such, we will not consider strategies which know what round there in or how many rounds there are. All that said, we could have modeled things with a fixed number of rounds. For reasons of convention we did not and so we're stuck with the radio active decay version of time.

Okay, so we always play one round and advance to a subsequent round with probability *w*. With this fact, we can show that the probability of reaching round *n* is given by w^{n-1}. To see this, we note that probability of reaching round 1 is given by w^0 which is 1. The probability of reaching the second period is given by w^1 which is *w*. Likewise, the probability of reaching the third round is given by the product of three terms: (1) the probability of reaching the first round, (2) the probability of reaching the second round conditioned on reaching the first round, and (3) the probability of reaching the third round conditioned on reaching the second round. Remember that the probability of reaching any particular round conditioned on reaching the previous round is simply *w*. As such, the probability of reaching round 3 becomes $w^0 * w * w = w^2$. Generally, the probability of reaching round *n* is given by w^{n-1}. If we assume that w=0.9, then the probability of reaching round 6 is given by w^5, which is 0.59.

So, how many rounds will there be? If we set *N* equal to the number of rounds, then we have that $N = 1 + w^1 + w^2 + w^3 + w^4 + w^5 + \ldots$ The series goes on forever. However, there is an interesting thing about such a series. It rapidly reaches equilibrium. We'll first look at this graphically.

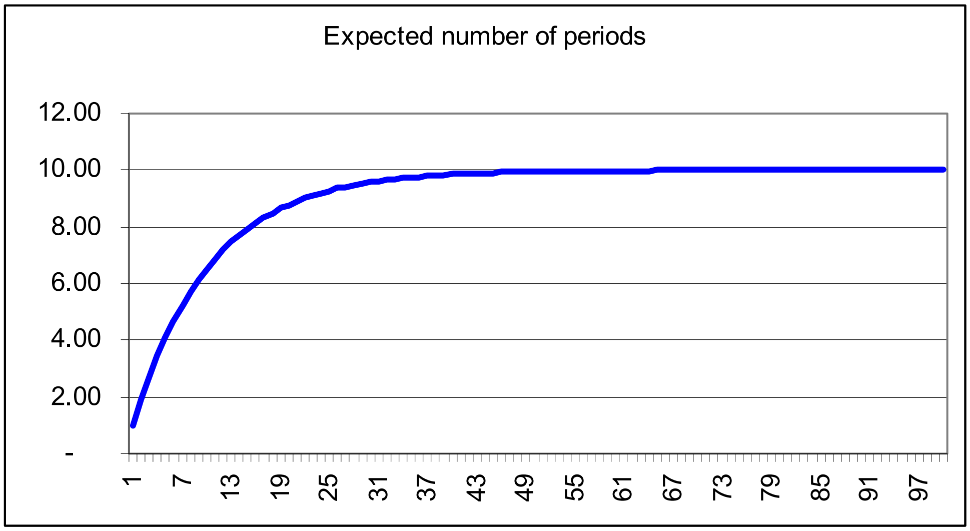


Along the *y*-axis we have the probability of reaching a particular round and along the *x*-axis we have the number of rounds. I have set $w = 0.9$. So, the probability of reaching round 20 is given by 13.5%.

As you notice, the probability quickly decays to zero. This seems counter intuitive because we said that the conditional probability of reaching any particular round, assuming we have reached the previous period, is fixed at *w*. A similarly counter-intuitive problem was laid out by the Greek philosopher Xeno in his famous paradox of the tortoise and the hare.

What concerns us is that while the conditional probability of reaching the next round is fixed at *w*, the probability of reaching any particular round is not fixed; it is given by w^{n-1}. As *n* gets large, the probability rapidly approaches 0. We can see that in the graph above.

Okay, armed with this new knowledge, we can ask our question again: how many rounds will there be? We can't answer the question with certainty because reaching each round is a probability not certainty. However, we can calculate the expected, or average, number of rounds. Let's first look at this graphically.



Along the *y*-axis we have the expected number of periods. Along the *x*-axis, we have the contribution that each additional round makes to that expected number (i.e., w^{n-1}). Again, *w* is set to 0.9. As you can see, for the first few rounds, there is a relatively high probability of reaching the next round and so there is a significant contribution of that marginal round to the expected total. As the probability of reaching a round far off in the future approaches 0, it makes a small contribution to the expected total.

In our case, the expected number of rounds, when $w = 0.9$, is 10. Now, let's see if we can derive this number using *w*. Let us declare *N* as the expected number of periods. We then have

$$N = 1 + w + w^2 + w^3 + w^4 + w^5 \ldots \tag{1}$$

which is shown in the graph above. Let us subtract 1 from both sides of (1).

$$N - 1 = w + w^2 + w^3 + w^4 + w^5 + \ldots \tag{2}$$

Now, we can factor out a *w* from the right-hand side of (2).

$$N - 1 = w(1 + w + w^2 + w^3 + w^4 + \ldots) \tag{3}$$

Notice that the sum inside of the parentheses on the right-hand side of (3) is just the same as the right-hand side of (1). Thus, we can substitute *N* for the sum inside the parentheses of (3).

$$N - 1 = wN \tag{4}$$

Now we solve (4) in terms of *N*.

$$N = \frac{1}{1-w} \qquad (5)$$

We have derived the expected number of periods in (5). From our previous example, where we set w=0.9. If we plug that value into (5), then we get back 10.

After that long digression, we can return to our model of indirect reciprocity. Our indirect reciprocity game lasts for the expected number of periods given in (5). We now want to derive the fitness function for each strategy in our model. In the paper, we consider several strategies. Before getting into specifics, let us see how we can derive a general fitness function for our model. To do this, we simply have to add up the payoff that an individual playing a particular strategy will, on average, receive in each period summed over all periods. How do we do that? Note that we can utilize something like the trick we used to derive the expected sum in (5) to derive our total fitness.

Let us start with Nowak and Sigmund's model from the 1998 *JTB* paper. They have three strategies. We'll call them *ALLC*, *ALLD*, and *DISC*. We'll denote the frequency of each of these strategies with $x1$, $x2$, and $x3$, respectively. Let us denote the payoff of strategy i in round n with $V_n(i)$.[1] Further, let us denote the fraction of type i in good standing in round n with $g_n(i)$.

Okay, let's us start with round 1. As no social behavior has yet taken place, we assume that everyone is in good standing. Thus, we have,

$$\begin{aligned} g_1(ALLC) &= 1 \\ g_1(ALLD) &= 1 \\ g_1(DISC) &= 1 \end{aligned} \qquad (6)$$

Using the definition of the strategies from Nowak and Sigmund's paper, we have the following payoffs in round 1

$$\begin{aligned} V_1(ALLC) &= b(x_1 + x_3) - c \\ V_1(ALLD) &= b(x_1 + x_3) \\ V_1(DISC) &= b(x_1 + x_3) - c \end{aligned} \qquad (7)$$

This is so, because the *ALLC* and *DISC* will always cooperate, while the *ALLD* will not. After this first period, we need to update the fraction of good standing for each type. Notice that all the *ALLC* and *DISC* individuals will be in good standing, because they cooperated in period 1, while all the *ALLD* individuals are bad because they did not cooperate. Entering period 2, we have

[1] I have changed things slightly from the paper that I published. I'll use V to capture the payoff in any particular round. I'll sum them up to derive the total fitness and use W for this sum.

$$\begin{aligned} g_2(ALLC) &= 1 \\ g_2(ALLD) &= 0 \\ g_2(DISC) &= 1 \end{aligned} \tag{8}$$

Okay, in period 2, the *ALLC* types will cooperate with everyone, the *ALLD* will cooperate with no one and *DISC* will cooperate only with good-standing players, which are the *ALLC* and the *DISC* types in the population. Thus, we have round 2 payoffs

$$\begin{aligned} V_2(ALLC) &= b(x_1 + x_3) - c \\ V_2(ALLD) &= bx_1 \\ V_2(DISC) &= b(x_1 + x_3) - c(x_1 + x_3) \end{aligned} \tag{9}$$

The *ALLC* types get help from other *ALLC*s and from *DISC*s and they help everyone. The *ALLD* types only get help from the *ALLC*s, the *DISC*s will never again offer help to them. The *DISC* types will get help from the *ALLC*s and from other *DISC*s. They will only offer help to *ALLC*s and other *DISC*s, as those are the only types in good standing. Okay, using the image-scoring rules of Nowak and Sigmund's model, any cooperation brings a positive image score and a defection brings a negative one. Based on these rules and their behavior in period 2, we enter round 3 with

$$\begin{aligned} g_3(ALLC) &= 1 \\ g_3(ALLD) &= 0 \\ g_3(DISC) &= 1 - x_2 \end{aligned} \tag{10}$$

Notice that the fraction of the *DISC* type in good standing at the start of period 3 is a function of who they interacted with in period 2. In this case, the *DISC* individuals will maintain their positive image score unless they interacted with an *ALLD* individual. If they did, they would have defected on the bad-standing *ALLD* and thus themselves fallen into bad standing.

To proceed, let us introduce one more variable. Let g_n denote the fraction of the whole population with a positive image score in period *n*. We note that *ALLC* will always have a positive image score as they always cooperate (remember that in the Nowak and Sigmund model, there are no errors). The *ALLD* individuals never cooperate and so will never have a positive image score. The *DISC* individuals will cooperate when they meet someone with a positive image score and defect on those with a negative image score. Thus, the fraction of *DISC* with a positive image score in any particular round is strictly a function of whether or not they meet good or bad people in the previous period. Thus, in round *n* (where *n* > 1) we have

$$\begin{aligned} g_3(ALLC) &= 1 \\ g_3(ALLD) &= 0 \\ g_3(DISC) &= g_{n-1} \end{aligned} \tag{11}$$

From (11) we can calculate what the fraction of the whole population with a positive image score in any particular round.

$$g_n = x_1 + g_{n-1}x_3 \tag{12}$$

Notice that (12) is the same as (1) from Nowak and Sigmund's 1998 *JTB* paper. Okay, before proceeding, let us derive the fitness for *ALLC* and *ALLD*. After the first round, we have

$$\begin{aligned} V_n(ALLC) &= b(x_1 + x_3) - c \\ V_n(ALLD) &= bx_1 \end{aligned} \tag{13}$$

The payoffs in (13) are constant after the first round, so we can calculate the fitness of *ALLC* and *ALLD* by summing up their payoffs in every round. Using the derivation (5) for the expected number of periods, we have

$$\begin{aligned} W(ALLC) &= \frac{1}{1-w}\left[b(x_1 + x_3) - c\right] \\ W(ALLD) &= \frac{1}{1-w}bx_1 + bx_3 \end{aligned} \tag{14}$$

From (14), we see that *ALLC* will get help from other *ALLC*s and the *DISC*s in every period and help everyone in every period. The *ALLD*s get help from *ALLC*s in all rounds and only get help from the *DISC*s in the first period.

Now, let us return to the calculation of the fitness for the *DISC* types. After the first round, the payoff to *DISC* is given by

$$V_n(DISC) = b(x_1 + x_3 g_n(DISC)) - cg_n \tag{15}$$

From (15), we see that an individual *discriminator* always gets help from the *ALLC*s. He will get help from the other *DISCs* when he has a positive image score, which is given by g_n(*DISC*). This *discriminator* will offer help only when he meets someone with a positive image score, which is given by g_n. Using (11), we can substitute g_{n-1} for g_n(*DISC*) and thus rewrite (15) as

$$V_n(DISC) = b(x_1 + x_3 g_{n-1}) - cg_n \tag{16}$$

Now, using (12), we can substitute g_n for $x_1 + x_3 g_{n-1}$ and rewrite (16) as

$$V_n(DISC) = (b - c)g_n \tag{17}$$

To derive the fitness of *DISC* we need to sum up the payoffs in all rounds. We now have

$$W(DISC) = (b-c)\left[g_1 + wg_2 + w^2g_3 + w^3g_4 + ...\right] - bx_2 \qquad (18)$$

Before I proceed, let me explain why I put the *–bx*$_2$ term at the end. After the first round, the term g_n captures the *ALLC*s and the *DISC*s who have a positive image score. Thus the term g_n accurately reflects the fraction of good people in the population in a round. In the first round, however, we assumed that g_1=1. That is, we assume that, in the absence of information, everyone is good. However, we know that the *ALLD* types will not cooperate in any round by definition. Without making any adjustment, we would have the first-round payoff to DISC to be $V_1(DISC) = (b-c)g_1$. As we have defined g_1=1, we would have $V_1(DISC) = b-c$. This is not accurate, however. The *DISC* will always cooperate (paying the cost –c), but they will only receive help from the *ALLC* and *DISC* types. Thus, we could write $V_1(DISC) = b(x_1 + x_3) - c$. Remember, however, that adding and subtracting the same number does nothing to a sum. Thus, we can write $V_1(DISC) = b(x_1 + x_3) + bx_2 - bx_2 - c$. This can be rewritten as $V_1(DISC) = b(x_1 + x_2 + x_3) - bx_2 - c$, which can then be rewritten as $V_1(DISC) = b(1) - bx_2 - c$ as we know that $x_1 + x_2 + x_3 = 1$. Okay, now knowing that g_1=1, we can write $V_1(DISC) = bg_1 - bx_2 - cg_1$. This can be rewritten as $V_1(DISC) = (b-c)g_1 - bx_2$. When we include this expression in the sum of (18), I just left the term *–bx*$_2$ at the end.

Sorry for that digression. To proceed, let us define a new variable *G* (don't try to hard to think about what *G* 'means', it will just make the make easier)

$$G = g_1 + wg_2 + w^2g_3 + w^3g_4 + ... \qquad (19)$$

Notice that (19) looks a little like our expression for the expected number of rounds in (1). Okay, using factoring a *w* from (19) we have

$$G = g_1 + w\left[g_2 + wg_3 + w^2g_4 + w^3g_5 ...\right] \qquad (20)$$

Using (12), we can rewrite (20) as

$$G = g_1 + w\left[(x_1 + g_1x_3) + w(x_1 + g_2x_3) + w^2(x_1 + g_3x_3) + w^3(x_1 + g_4x_3) + ...\right] \qquad (21)$$

Factoring this in terms of x_1 and x_3, we have

$$G = g_1 + w\left[x_1\left(1 + w + w^2 + w^3 + ...\right) + x_3\left(g_1 + wg_2 + w^2g_3 + w^3g_4 + ...\right)\right] \qquad (22)$$

Alright, now we're getting somewhere. Notice that the sum multiplied by x_1 in (22) is the series from (1). Thus, we can substitute (5) for it. Also, notice that the sum multiplied by

x_3 in (22) is the same as (19). Thus, we can substitute *G* for it. We can now rewrite (22) as

$$G = g_1 + w\left[x_1 \frac{1}{1-w} + x_3 G\right] \tag{23}$$

Remember that g_1=1. Also, let us multiply the expression by (1-w). We now have

$$G(1-w) = (1-w) + w\left[x_1 + x_3 G(1-w)\right] \tag{24}$$

Now, with a little algebra, we can solve (24) in terms of G. First, let rewrite (24) as

$$G(1-w) = (1-w) + wx_1 + wx_3 G(1-w) \tag{25}$$

Let us now bring all the *G* terms to one side.

$$G(1-w-wx_3(1-w)) = 1-w+wx_1 \tag{26}$$

Now we have that

$$G = \frac{1-w+wx_1}{1-w-wx_3(1-w)} \tag{27}$$

Let us rewrite (27) as

$$G = \frac{1-w+wx_1}{1-w(1-wx_3)} \tag{28}$$

Substituting (27) into (18), we can derive the fitness function for *DISC*.

$$W(DISC) = (b-c)\left[\frac{1-w+wx_1}{1-w(1-wx_3)}\right] - bx_2 \tag{28}$$

Okay, we have derived the fitness function of *DISC* from the Nowak and Sigmund model of indirect reciprocity from their 1998 *JTB* paper. It is written slightly differently from the way that Nowak and Sigmund write, but I think it is more clear.

Okay, now let us move on to the derivations that were new in our paper. We will derive the fitness function of the *RDISC* strategy taken from appendix *B1*. Unfortunately, when reputations are modeled as standings as opposed to image-scores, the math becomes much more complicated. As such, a few 'tricks' must be employed. There isn't any funny business going on. The tricks serve to make good approximations while making the equations solvable. Whenever using simplifying tricks, you should always check the results against the true results to make sure that nothing has been lost in the simplification process. A spreadsheet like Excel is a good place to perform such a check.

We will do a simplified version of the model presented in the paper which will make things a little more clear. We will assume that information is complete (i.e., everyone knows exactly what everyone else did in the previous round). You can follow along with the math presented in the paper's appendix if you set $q = 1$.

In this model we have three strategies, *ALLC*, *ALLD*, and *RDISC*. Let us denote the frequencies of these strategies by x_1, x_2, and x_4, respectively. As in the Nowak and Sigmund model, we have one round with certain probability. A following round occurs with probability w. Thus, condition (5) will yield the expected number of rounds. As in the previous model, the payoff to a strategy in each round will be based on the distribution of types and the probability with which an individual finds himself and his partner in good standing. Remember, that in the Nowak and Sigmund model, they used image-scoring where good people are those that cooperate and bad people are those that defect. In a standing model, people are bad when they defect on a good-standing partner. Any cooperation brings with it good standing. A defection on a bad-standing partner will leave an individual's standing unchanged. Okay, let us first assume that everyone starts out in good standing. Thus we have

$$\begin{aligned} g_1(ALLC) &= 1 \\ g_1(ALLD) &= 1 \\ g_1(RDISC) &= 1 \end{aligned} \tag{29}$$

Afterwards, in round n (where $n > 1$) we have the following recursions. I'll go through the logic of each strategy one by one. First, for *ALLC*, we have

$$g_n(ALLC) = g_{n-1}(ALLC)\left[1 - g_{n-1}\alpha\right] + (1 - g_{n-1}(ALLC))(1-\alpha) \tag{30}$$

This says that the probability an individual *ALLC* will be in good standing in round n is the sum of two components. First, if he was in good standing in the previous round ($g_{n-1}(ALLC)$), he retains this good standing unless he commits an error against a partner who is in good standing ($1 - g_{n-1}\alpha$). Second, if he was in bad standing at the end of the previous round ($1 - g_{n-1}(ALLC)$), he can regain it by cooperation, which he always intends to do. He fails with error sometimes though ($1-\alpha$).

The *ALLD* will never cooperate and so loses its standing and never regains it. Thus, we have,

$$g_n(ALLD) = 0 \tag{31}$$

For the *RDISC* types, we have the following recursion

$$g_n(RDISC) = g_{n-1}(RDISC)\left[1 - g_{n-1}\alpha\right] + (1 - g_{n-1}(RDISC))g_{n-1}(1-\alpha) \qquad (32)$$

Again, this is made up two components. If the *RDISC* was in good standing, he keeps it unless he commits an error against a good-standing partner. If the *RDISC* was in bad standing, he regains it if he cooperates. He only tries to cooperate when he meets a good-standing partner and this must be scaled by the error rate. (Note, that the *CTFT* strategy that we analyze in appendix *B2* and the one that Leimar and Hammerstein (2001) analyze, is slightly different. When it is in bad standing, it will always try and cooperate, regardless of the standing of its partner.)

Let us also use the expression g_n to express the fraction of the population in good standing in round n.

$$g_n = x_1 g_n(ALLC) + x_4 g_n(RDISC) \qquad (33)$$

Okay, let us know write down the payoffs to each strategy in the first round.

$$\begin{aligned} V_1(ALLC) &= (1-\alpha)\left[b(x_1 + x_4) - c\right] \\ V_1(ALLD) &= (1-\alpha)\left[b(x_1 + x_4)\right] \\ V_1(RDISC) &= (1-\alpha)\left[b(x_1 + x_4) - c\right] \end{aligned} \qquad (34)$$

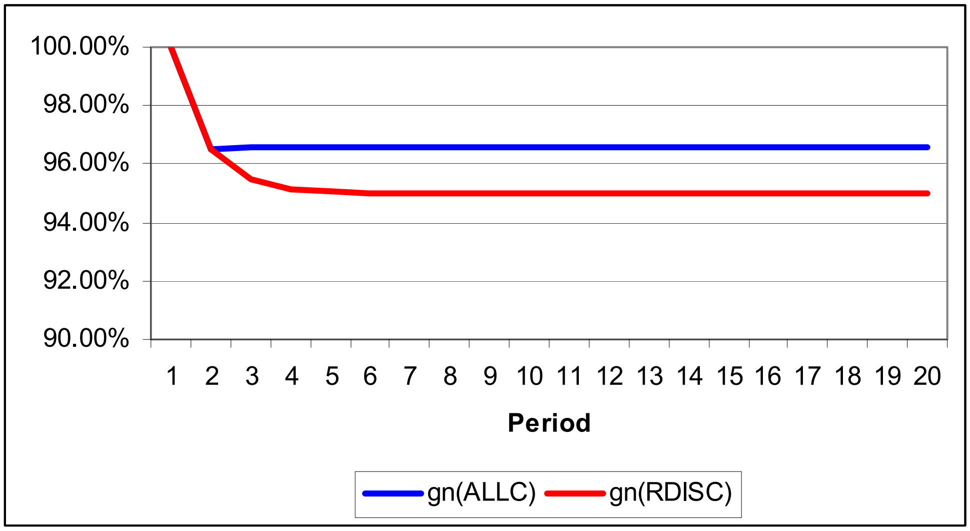
Afterwards, in round *n* (where *n* > 1), we have

$$\begin{aligned} V_n(ALLC) &= (1-\alpha)\left[b(x_1 + x_4 g_n(ALLC)) - c\right] \\ V_n(ALLD) &= (1-\alpha)bx_1 \\ V_n(RDISC) &= (1-\alpha)\left[b(x_1 + x_4 g_n(RDISC)) - cg_n\right] \end{aligned} \qquad (35)$$

To calculate the fitness function for each strategy, we could proceed as did Nowak and Sigmund, which would look something like

$$\begin{aligned} W(ALLC) &= V_1(ALLC) + wV_2(ALLC) + w^2V_3(ALLC) + w^3V_4(ALLC) + ... \\ W(ALLD) &= V_1(ALLD) + wV_2(ALLD) + w^2V_3(ALLD) + w^3V_4(ALLD) + ... \\ W(RDISC) &= V_1(RDISC) + wV_2(RDISC) + w^2V_3(RDISC) + w^3V_4(RDISC) + ... \end{aligned} \qquad (36)$$

We can put in the expressions for the payoffs and try to proceed by calculating something like the *G* expression that we did in the Nowak and Sigmund model. However, the recursions in (30) and (32) are complex compared to the ones from the previous model. Unfortunately, there is no neat way to derive a summation like we did in (28). This means, that we cannot derive accurate expressions for the fitness of each strategy. However, we can note that the recursions in (30) and (32) reach equilibrium and they do it really quickly. Let me show you this with a graph. I have set x_1=0.35, x_4=0.35, and α=0.05.



See that after 2 or 3 rounds, $g_n(ALLC)$ and $g_n(RDISC)$ have reached their equilibrium. If we assume that they reach equilibrium in round 2, instantly, we lose almost nothing. A spreadsheet check should confirm this. Before proceeding, let me use the variable *G*(*ALLC*) and *G*(*RDISC*) to denote the fraction of *ALLC* and *RDISC*, respectively, in good standing at equilibrium. Also, let *G* be the fraction of the population in good standing at equilibrium. I'm sorry I'm reusing *G*. It means something completely different that the *G* from the Nowak and Sigmund model presented above. If we assume that equilibrium is reached instantaneously in round 2, we can rewrite (35) as

$$\begin{aligned} V_n(ALLC) &= (1-\alpha)\left[b(x_1 + x_4 G(ALLC)) - c\right] \\ V_n(ALLD) &= (1-\alpha) b x_1 \\ V_n(RDISC) &= (1-\alpha)\left[b(x_1 + x_4 G(RDISC)) - cG\right] \end{aligned} \tag{37}$$

This is the same for all rounds after round 1. So, let us know calculate the fitness function for each strategy. I'll do them one by one. First, starting with *ALLC*

$$W(ALLC) = V_1(ALLC) + wV_2(ALLC) + w^2V_3(ALLC) + w^3V_4(ALLC) + \ldots \tag{38}$$

If we substitute in (34) and (37), we have

$$\begin{aligned} W(ALLC) &= (1-\alpha)\left[b(x_1 + x_4) - c\right] + w(1-\alpha)\left[b(x_1 + x_3 G(ALLC)) - c\right] + \\ & w^2(1-\alpha)\left[b(x_1 + x_3 G(ALLC)) - c\right] + w^3(1-\alpha)\left[b(x_1 + x_3 G(ALLC)) - c\right] + \ldots \end{aligned} \tag{39}$$

We can factor out a (1- α) and rewrite (39) as

$$W(ALLC) = (1-\alpha)\left\{ \begin{matrix} \left[b(x_1 + x_4) - c\right] + w\left[b(x_1 + x_3 G(ALLC)) - c\right] + \\ w^2\left[b(x_1 + x_3 G(ALLC)) - c\right] + w^3\left[b(x_1 + x_3 G(ALLC)) - c\right] + \ldots \end{matrix} \right\} \tag{40}$$

Next, let us group up the terms inside of the brackets

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1\left(1+w+w^2+w^3+...\right) \\ +bx_4\left(1+wG(ALLC)+w^2G(ALLC)+w^3G(ALLC)+...\right) \\ -c\left(1+w+w^2+w^3+...\right) \end{array}\right] \tag{41}$$

Now we can substitute result (5) into (41) and get

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1\dfrac{1}{1-w} \\ +bx_4\left(1+wG(ALLC)+w^2G(ALLC)+w^3G(ALLC)+...\right) \\ -c\dfrac{1}{1-w} \end{array}\right] \tag{42}$$

To handle the term in the middle, let us factor out the term *G*(*ALLC*).

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1\dfrac{1}{1-w} \\ +bx_4\left(1+G(ALLC)\left(w+w^2+w^3+...\right)\right) \\ -c\dfrac{1}{1-w} \end{array}\right] \tag{43}$$

Let us now factor out a *w* from the sum to get

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1\dfrac{1}{1-w} \\ +bx_4\left(1+wG(ALLC)\left(1+w+w^2+...\right)\right) \\ -c\dfrac{1}{1-w} \end{array}\right] \tag{44}$$

We can now substitute in (5) to get

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1\dfrac{1}{1-w} \\ +bx_4\left(1+wG(ALLC)\dfrac{1}{1-w}\right) \\ -c\dfrac{1}{1-w} \end{array}\right] \tag{45}$$

Let me rewrite (45) as

$$W(ALLC) = (1-\alpha)\left[\begin{array}{l} bx_1 \dfrac{1}{1-w} \\ + bx_4\left(\dfrac{1-w}{1-w} + wG(ALLC)\dfrac{1}{1-w}\right) \\ -c\dfrac{1}{1-w} \end{array}\right] \tag{46}$$

All that I have done is substitute $\frac{1-w}{1-w}$ for a 1. Now, I will factor out a $\frac{1}{1-w}$ to get

$$W(ALLC) = (1-\alpha)\frac{1}{1-w}\left[bx_1 + bx_4\left(1-w+wG(ALLC)\right) - c\right] \tag{47}$$

Let us next calculate the fitness of ALLD. Using (34), (35) and (36), we have

$$W(ALLD) = (1-\alpha)b(x_1+x_4) + w(1-\alpha)bx_1 + w^2(1-\alpha)bx_1 + \ldots \tag{48}$$

Factoring out a (1- α), we have

$$W(ALLD) = (1-\alpha)\left[b(x_1+x_4) + wbx_1 + w^2bx_1 + w^3bx_1\ldots\right] \tag{49}$$

We can rewrite (49) as

$$W(ALLD) = (1-\alpha)\left[bx_1 + x_4 + wbx_1 + w^2bx_1 + w^3bx_1\ldots\right] \tag{50}$$

Let us group terms, so we have

$$W(ALLD) = (1-\alpha)\left[bx_1(1+w+w^2+w^3+\ldots) + bx_4\right] \tag{51}$$

Substituting in (5), we have the fitness function of ALLD

$$W(ALLD) = (1-\alpha)\left[bx_1\frac{1}{1-w} + bx_4\right] \tag{52}$$

Next, we need to calculate the fitness function of RDISC. Let us use (34), (35) and (36) to have

$$\begin{gathered} W(RDISC) = (1-\alpha)\left[b(x_1+x_4) - c\right] + w(1-\alpha)\left[b(x_1 + x_4G(RDISC)) - cG\right] \\ + w^2(1-\alpha)\left[b(x_1 + x_4G(RDISC)) - cG\right] + w^3(1-\alpha)\left[b(x_1 + x_4G(RDISC)) - cG\right] + \ldots \end{gathered} \tag{53}$$

First, let us factor out the error term

$$W(RDISC) = (1-\alpha)\left[\begin{array}{l} \left[b(x_1 + x_4) - c\right] + w\left[b(x_1 + x_4 G(RDISC)) - cG\right] \\ + w^2\left[b(x_1 + x_4 G(RDISC)) - cG\right] \\ + w^3\left[b(x_1 + x_4 G(RDISC)) - cG\right] + \ldots \end{array}\right] \tag{54}$$

We group up terms to get

$$W(RDISC) = (1-\alpha)\left[\begin{array}{l} bx_1\left(1 + w + w^2 + w^3 + \ldots\right) \\ + bx_4\left(1 + wG(RDISC) + w^2 G(RDISC) + w^3 G(RDISC) + \ldots\right) \\ - c\left(1 + wG + w^2 G + w^3 G + w^4 G + \ldots\right) \end{array}\right] \tag{55}$$

We can substitute in (5) to get

$$W(RDISC) = (1-\alpha)\left[\begin{array}{l} bx_1 \dfrac{1}{1-w} \\ + bx_4\left(1 + wG(RDISC) + w^2 G(RDISC) + w^3 G(RDISC) + \ldots\right) \\ - c\left(1 + wG + w^2 G + w^3 G + w^4 G + \ldots\right) \end{array}\right] \tag{56}$$

We next factor the middle and last sum to get

$$W(RDISC) = (1-\alpha)\left[\begin{array}{l} bx_1 \dfrac{1}{1-w} \\ + bx_4\left(1 + wG(RDISC)\left(1 + w + w^2 + w^3 + \ldots\right)\right) \\ - c\left(1 + wG\left(1 + w + w^2 + w^3 + \ldots\right)\right) \end{array}\right] \tag{57}$$

Again, substituting in (5) we have

$$W(RDISC) = (1-\alpha)\left[\begin{array}{l} bx_1 \dfrac{1}{1-w} \\ + bx_4\left(1 + wG(RDISC)\dfrac{1}{1-w}\right) \\ - c\left(1 + wG\dfrac{1}{1-w}\right) \end{array}\right] \tag{58}$$

Substituting in a $\frac{1-w}{1-w}$ for a 1, we now have

$$W(RDISC) = (1-\alpha)\begin{bmatrix} bx_1 \frac{1}{1-w} \\ +bx_4\left(\frac{1-w}{1-w} + wG(RDISC)\frac{1}{1-w}\right) \\ -c\left(\frac{1-w}{1-w} + wG\frac{1}{1-w}\right) \end{bmatrix} \tag{59}$$

Factoring out a $\frac{1}{1-w}$, we have

$$W(RDISC) = (1-\alpha)\frac{1}{1-w}\left[bx_1 + bx_4\left(1-w+wG(RDISC)\right) - c\left(1-w+wG\right)\right] \tag{60}$$

We now have the fitness function for each strategy. Let me just rewrite (47), (52), and (60).

$$\begin{aligned} W(ALLC) &= (1-\alpha)\frac{1}{1-w}\left[bx_1 + bx_4\left(1-w+wG(ALLC)\right) - c\right] \\ W(ALLD) &= (1-\alpha)\left[bx_1\frac{1}{1-w} + bx_4\right] \\ W(RDISC) &= (1-\alpha)\frac{1}{1-w}\left[bx_1 + bx_4\left(1-w+wG(RDISC)\right) - c\left(1-w+wG\right)\right] \end{aligned} \tag{61}$$

To complete these fitness functions, we must figure out what *G*(*ALLC*), *G*(*RDISC*) and *G* are. Remember that I said that the recursions (30) and (32) reach equilibrium. Let's calculate that equilibrium. First we'll start with *RDISC*. Rewriting (32)

$$g_n(RDISC) = g_{n-1}(RDISC)\left[1 - g_{n-1}\alpha\right] + (1 - g_{n-1}(RDISC))g_{n-1}(1-\alpha) \tag{62}$$

At equilibrium, we can substitute in G(RDISC) and G to get

$$G(RDISC) = G(RDISC)\left[1 - G\alpha\right] + (1 - G(RDISC))G(1-\alpha) \tag{63}$$

Expanding (63) we have

$$G(RDISC) = G(RDISC) - G(RDISC)G\alpha + G(1-\alpha) - G(RDISC)G(1-\alpha) \tag{64}$$

The two G(RDISC) terms cancel each other out so we have

$$0 = -G(RDISC)G\alpha + G(1-\alpha) - G(RDISC))G(1-\alpha) \tag{65}$$

Next, we can divide through by *G*

$$0 = -G(RDISC)\alpha + 1 - \alpha - G(RDISC)(1-\alpha) \tag{66}$$

Expanding the last term, we have

$$0 = -G(RDISC)\alpha + 1 - \alpha - G(RDISC) + G(RDISC)\alpha \tag{67}$$

The first and last terms cancel each other out, so we have

$$0 = 1 - \alpha - G(RDISC) \tag{68}$$

Thus, we have

$$G(RDISC) = 1 - \alpha \tag{69}$$

Next, we'll move onto ALLC. We have (30)

$$g_n(ALLC) = g_{n-1}(ALLC)\left[1 - g_{n-1}\alpha\right] + (1 - g_{n-1}(ALLC))(1-\alpha) \tag{70}$$

Substituting in G(ALLC) and G, we have

$$G(ALLC) = G(ALLC)\left[1 - G\alpha\right] + (1 - G(ALLC))(1-\alpha) \tag{71}$$

Expanding this, we have

$$G(ALLC) = G(ALLC) - G(ALLC)G\alpha + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \tag{72}$$

The two G(ALLC) terms cancel out, leaving us with

$$0 = -G(ALLC)G\alpha + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \tag{73}$$

Now, we have to use (33) to put in something for *G*. We now have

$$0 = -\alpha G(ALLC)\left[x_1 G(ALLC) + x_4 G(RDISC)\right] + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \tag{74}$$

Substituting (69), we have

$$0 = -\alpha G(ALLC)\left[x_1 G(ALLC) + x_4(1-\alpha)\right] + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \tag{75}$$

Expanding, we have

$$0 = -x_1\alpha\big(G(ALLC)\big)^2 - x_4\alpha G(ALLC)(1-\alpha) + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \qquad (76)$$

As this expression is getting complicated, I need to employ a few more tricks. First, let me expand (76) one more time

$$\begin{aligned} 0 = &-x_1\alpha\big(G(ALLC)\big)^2 - x_4\alpha G(ALLC) + x_4\alpha^2 G(ALLC) \\ &+1 - \alpha - G(ALLC) + G(ALLC)\alpha \end{aligned} \qquad (77)$$

If we assume that errors rates (i.e., α) are low, then terms like α^2 will be really small. For example, if $\alpha = 0.05$, then $\alpha^2 = 0.0025$. Not much will be lost from the calculations if we assume that $\alpha^2 \approx 0$. We can thus drop one term from (77) and have

$$0 = -x_1\alpha\big(G(ALLC)\big)^2 - x_4\alpha G(ALLC) + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \qquad (78)$$

Next, if we assume that G(ALLC) is close to 1 (an excel check will prove this to be true), we can make one more simplification. It should not be surprising that G(ALLC) is close to 1 because ALLC always tries to cooperate and so at equilibrium, most ALLC should be in good standing. If we assume that G(ALLC) is close to 1, then we can make the following estimation

$$\big(G(ALLC)\big)^2 \approx 2G(ALLC) - 1 \qquad (79)$$

This estimation is taking advantage of the fact that near the value of 1, you can make a linear approximation of a quadratic function with (79). You should convince yourself of this approximation using values of G(ALLC) near 1 to verify (79). Now, substituting (79) into (78), we have

$$0 = -x_1\alpha\big(2G(ALLC) - 1\big) - x_4\alpha G(ALLC) + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \qquad (80)$$

Expanding (80), we have

$$0 = -2x_1\alpha G(ALLC) + x_1\alpha - x_4\alpha G(ALLC) + 1 - \alpha - G(ALLC) + G(ALLC)\alpha \qquad (81)$$

Let us break up the first term into two pieces, so we have

$$\begin{aligned} 0 = &-x_1\alpha G(ALLC) - x_1\alpha G(ALLC) + x_1\alpha - x_4\alpha G(ALLC) \\ &+1 - \alpha - G(ALLC) + G(ALLC)\alpha \end{aligned} \qquad (82)$$

Now we can group terms to get

$$0 = G(ALLC)(-1 + \alpha - x_1\alpha - x_4\alpha - x_1\alpha) + 1 - \alpha + x_1\alpha \qquad (83)$$

If we factor terms, we get

$$0 = -G(ALLC)(1-\alpha(1-x_1-x_4-x_1))+1-\alpha(1-x_1) \tag{84}$$

We can move the first term to the left to get

$$G(ALLC)(1-\alpha(1-x_1-x_4-x_1)) = 1-\alpha(1-x_1) \tag{85}$$

Next, we get

$$G(ALLC) = \frac{1-\alpha(1-x_1)}{1-\alpha(1-x_1-x_4-x_1)} \tag{86}$$

Let's rewrite this as

$$G(ALLC) = 1-\alpha(1-x_1)\frac{1}{1-\alpha(1-x_1-x_4-x_1)} \tag{87}$$

Okay, here's another trick. Let's us introduce δ as an arbitrary variable. The following approximation will be good so long as δ is small.

$$\frac{1}{1-\delta} \approx 1+\delta \tag{88}$$

If we use the trick in (88), we can rewrite (87) as

$$G(ALLC) = (1-\alpha(1-x_1))(1+\alpha(1-x_1-x_4-x_1)) \tag{89}$$

As we have three strategies which add up to 1, we can say that x_2=1-x_1-x_4. We can thus rewrite (89) as

$$G(ALLC) = (1-\alpha(1-x_1))(1+\alpha(x_2-x_1)) \tag{90}$$

We now expand (90) to get

$$G(ALLC) = 1-\alpha(1-x_1)+\alpha(x_2-x_1)-\alpha^2(1-x_1)(x_2-x_1) \tag{91}$$

Remember that we can assume that $\alpha^2 = 0$. We thus have

$$G(ALLC) = 1-\alpha(1-x_1)+\alpha(x_2-x_1) \tag{92}$$

Let us expand (92) to get

$$G(ALLC) = 1-\alpha+\alpha x_1+\alpha x_2-\alpha x_1 \tag{93}$$

Canceling terms, we get

$$G(ALLC) = 1 - \alpha + \alpha x_2 \tag{94}$$

Which can be factored to get

$$G(ALLC) = 1 - \alpha(1 - x_2) \tag{95}$$

We can substitute (95) and (69) into (61) to then derive the fitness functions of the three strategies. I'll leave that to you, I'm tired!

References Cited:

Nowak, M. A. and K. Sigmund (1998). "The dynamics of indirect reciprocity." Journal of Theoretical Biology **194**(4): 561-574.