Models of self-organizing inequality

36-313, Fall 2022

29 November 2021 (Lecture 23)

Contents

- Some things we've seen over and over:
  - There is a lot of inequality! Income, wealth, health, lifespan, exposure to crime...
  - There are big differences in all these variables across major social categories
  - There are also big inequalities *within* major social categories
  - The size of between-group and within-group inequalities change over time, but their directions are more stable, and the fact of large inequalities is more stable yet
- What might help explain these larger patterns?
  - Very basic probability models of income and wealth inequality predict a lot of patterns we see
  - Making those models a little more complex has big implications for the persistence of both between- and within- group inequality
  - Other models, based on social learning, show how between-group inequalities can be perpetuated without prejudice or bias, and even how they can arise without prejudice or bias

1 Multiplicative Growth and Its Implications

1.1 Multiplicative Growth Implies Heavy Tails

- Suppose person i in year t earns income $Y_i(t)$.
  - The same ideas will work for wealth
- We can express the change over time as $Y_i(t) = G_i(t)Y_i(t-1)$ where $G_i(t)$ is the factor by which income grew, or shrank, over the year
- So $Y_i(t) = Y_i(0)\prod_{s=1}^{t} G_i(s)$
- It makes sense to treat the Gs as random variables, so we can ask what distribution this leads to
- Distributions of products are ugly, so let's see if taking a log doesn't help:

$$\log Y_i(t) = \log Y_i(0) + \sum_{s=1}^{t} \log G_i(s)$$

- Now *assume* that the Gs are statistically independent over time, so the $\log G$s are too, and that $\mathbb{E}[\log G] = \mu$, $\mathrm{Var}[\log G] = \sigma^2$
- Remember the central limit theorem: if X_i are IID with common mean μ and variance σ^2, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow \mathcal{N}(\mu, \sigma^2/n)$$

  - If you are the kind of purist who wants to make the denominator on the left $\sqrt{n}$ so the right-hand side is $\mathcal{N}(\mu, \sigma^2)$, you can make the correction on your own
  - Multiplying both sides by n, and remembering what doing so does to the mean and variance, we can say that, for large n,

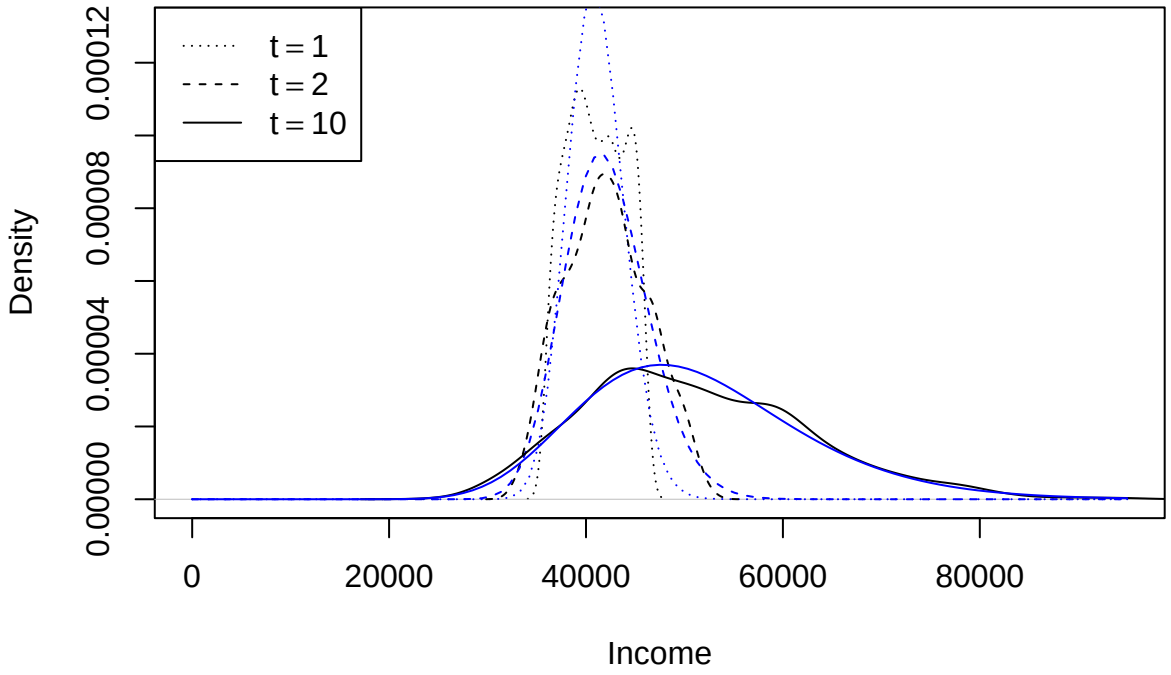
$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

  - So, applied here,

$$\log Y_i(t) \approx \log Y_i(0) + \mathcal{N}(t\mu, t\sigma^2)$$

- This means that $Y_i(t)$ should have an approximately log-normal distribution!
- Illustration: Suppose $G_i(t)$ is independent across i and t, and uniformly distributed on the interval $[0.9, 1.15]$ (to give a slight positive bias). Say $Y_i(0) = 4 \times 10^4$ with probability 1, just to keep things simple. See Figure 1.
  - The black lines show the (approximate) densities when I simulate a population of 1000 individuals whose incomes evolve on this process. The blue lines show the closest-fitting log-normals. You can see that initially the match is very bad, because a uniform distribution and a log-normal are very different, but after 10 steps the match is excellent.

Figure 1



- We saw that the bulk of the distribution of income and wealth is log-normal, across many countries and many decades. This basic argument suggests that this is what we'd expect out of multiplicative growth. But *that* means that right-skewed and heavy-tailed distributions are going to be hard to avoid, *unless* we can somehow make sure income and wealth do not grow multiplicatively[1].
- We also saw that income and wealth distributions have right tails which are too heavy for log-normals; they are, at least very roughly, more like power law or Pareto distributions. How could we explain that?
- One possibility: income growth is *faster* for higher incomes, at least some of the time (Montroll and Shlesinger 1982)

1.2 Refinements

- Seemingly-small differences in average growth rates quickly accumulate into big differences in outcomes
  - This is the power-of-compound-interest lesson again
  - Illustrative example: if $Y(0) = 4 \times 10^4$ and $G(t)$ averages 1.03 for 20 years, $Y(20) = 7.22 \times 10^4$, but if $G(t)$ averages 1.06, $Y(20) = 1.28 \times 10^5$
  - Of course it helps to start out well: if $G(t)$ averages 1.03 but $Y(0) = 5 \times 10^4$ instead of 4×10^4, we're looking at $Y(20) = 9.03 \times 10^4$.
- Some of the transmission of inequality across generations is families trying to equip their children with high $Y(0)$, and some of it is trying to increase the expected value of $G(t)$.
- If belonging to group A rather than group B raises average growth rates, even a modest number of steps can see a big difference in typical outcomes, and *especially* in which group is represented at the top of the income distribution
  - More subtly, if group A and group B have the same average growth rates *and* the same starting values, but group A has higher variance of growth, we'll see more of it in the highest percentiles of the income distribution
    - We'd also see more of group A in the lowest percentiles, but the simple multiplicative growth model will break down there (at least for income)

[1] Taxation is, in principle, one way to do this, though the tax rates would have to increase very rapidly with income levels.

2 Mobility and Transmission of Inequality

- We've seen that rates of economic mobility are not very high
  - E.g., probability that a child born in the bottom 20% of the income distribution around 1980 will be in the top 20% as an adult is about 8% in the US (see homework 7)

2.1 Multiplicative Growth and Social Mobility

Let's consider the relative income (or wealth, etc.) of two individuals:

$$Y_i(t) = Y_i(0) \prod_{s=1}^{t} G_i(s) \tag{1}$$

$$Y_j(t) = Y_j(0) \prod_{s=1}^{t} G_j(s) \tag{2}$$

$$\frac{Y_i(t)}{Y_j(t)} = \frac{Y_i(0)}{Y_j(0)} \prod_{s=1}^{t} \frac{G_i(s)}{G_j(s)} \tag{3}$$

Suppose $Y_i(0) < Y_j(0)$, so that i starts out worse-off than j. For their relative standings to be reversed at some later time, $Y_i(t) > Y_j(t)$, we need $\prod_{s=1}^{t} \frac{G_i(s)}{G_j(s)} > 1$. In words, this means that the initially-poorer individual needs to grow *faster* than the initially-richer one. (How much faster depends on the ratio of initial incomes.) If everyone grows at exactly the same rate, $G_i(t) = G_j(t)$ for all t, then this can't happen and the ratio of incomes is static. For their to be relative mobility, then we need either a *negative* correlation between current income $Y_i(t)$ and growth factors $G_i(t+1)$, *or* we need $G_i(t)$ to vary randomly *across individuals* i, and not just randomly over times t.

To get a concrete sense of this, we'll consider a situation where $Y_i(0) = 2 \times 10^4$ and $Y_j(0) = 6 \times 10^4$. As in the earlier demo, G will be uniform on $[0.9, 1.15]$, independently across people and across time. I'll show *one* trajectory for Y_j, but a whole bunch of trajectories for Y_i.

Here you can see that *some* of the Y_i trajectories overtake the Y_j trajectory, but not many (here, just 2.8% of them do so). If you play with the parameters of the simulation, you can see what happens to this probability; in particular, you can convince yourself that it goes up if you increase $\mathrm{Var}[\log G]$ while leaving $\mathbb{E}[\log G]$ alone.

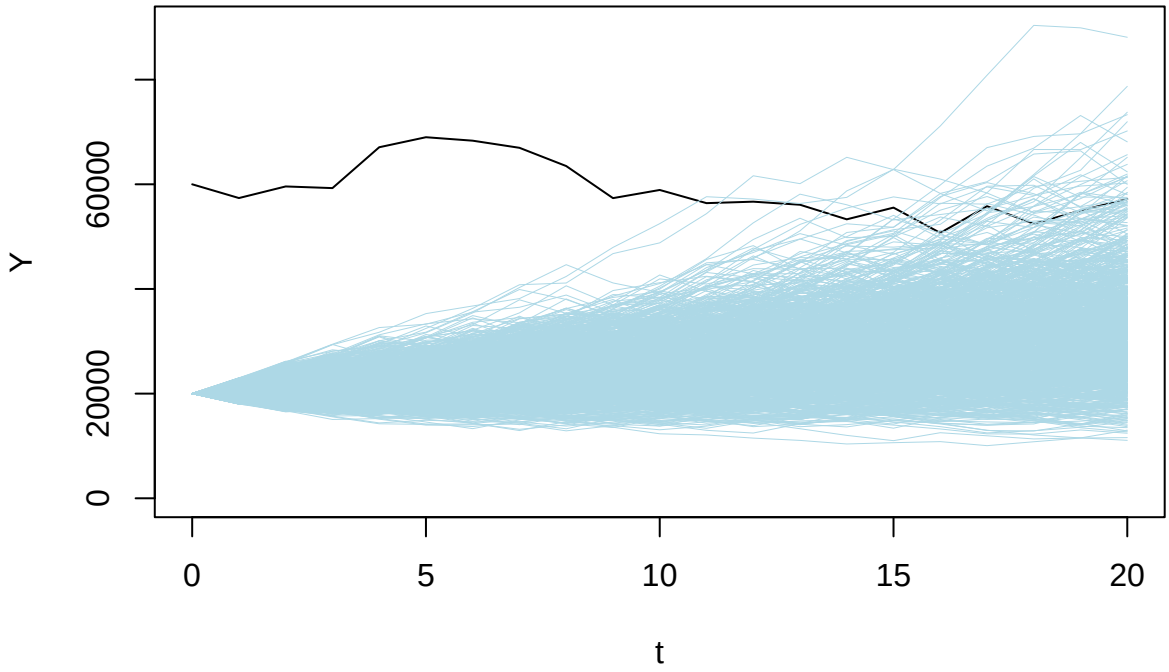


Figure 1: Relative economic mobillity under multiplicative growth. Incomes grow at an average rate of 2.5% per time period, but with a standard deviation of 7.2% per time period. The figure shows one trajectory starting from a high income, and a large number of trajectories starting from a lower income. *Some* of the latter are able to catch up with the former, but it's not what usually happens.

2.2 Transmission of Individual Inequality Implies Persistent Inter-group Inequalities

Suppose (i) in one generation, members of group A tend to be higher in the income distribution than members of group B, (ii) relative inter-generational economic mobility is low, and (iii) (most) children belong to the same group as their parents. Then *even if* inter-generational economic mobility is the *same* for both groups, in the next generation group A will still be better off than group B

- A simple calculation: suppose there are just two economic levels, and the probability of being in the same level as your parents is 80%. Group A starts off as 2/3 better-off, 1/3 worse-off and group B starts off as 1/3 better-off, 2/3 worse-off. Figure 2 shows what happens in this case. You can see that *eventually* the two groups will equalize their economic distributions, but that this takes many generations. Even after 4 generations the disparity between the groups is considerable.
  - You can play around with the precise numbers in the code to convince yourself that the general phenomenon is pretty robust.
  - In more technical terms (take 36-410 or 36-467), the assumption of equal economic mobility in both groups means that economic status follows a Markov chain, and a general theorem about Markov chains with a finite number of states is that they converge to a unique[2] distribution which is "invariant", i.e., doesn't change over time. So under this assumption, the economic distributions of groups A and B will converge to the same limit, and so converge on each other. But the time needed for this depends on how rapidly "mixing" the Markov chain is, and the assumption of low economic mobility implies slow mixing.

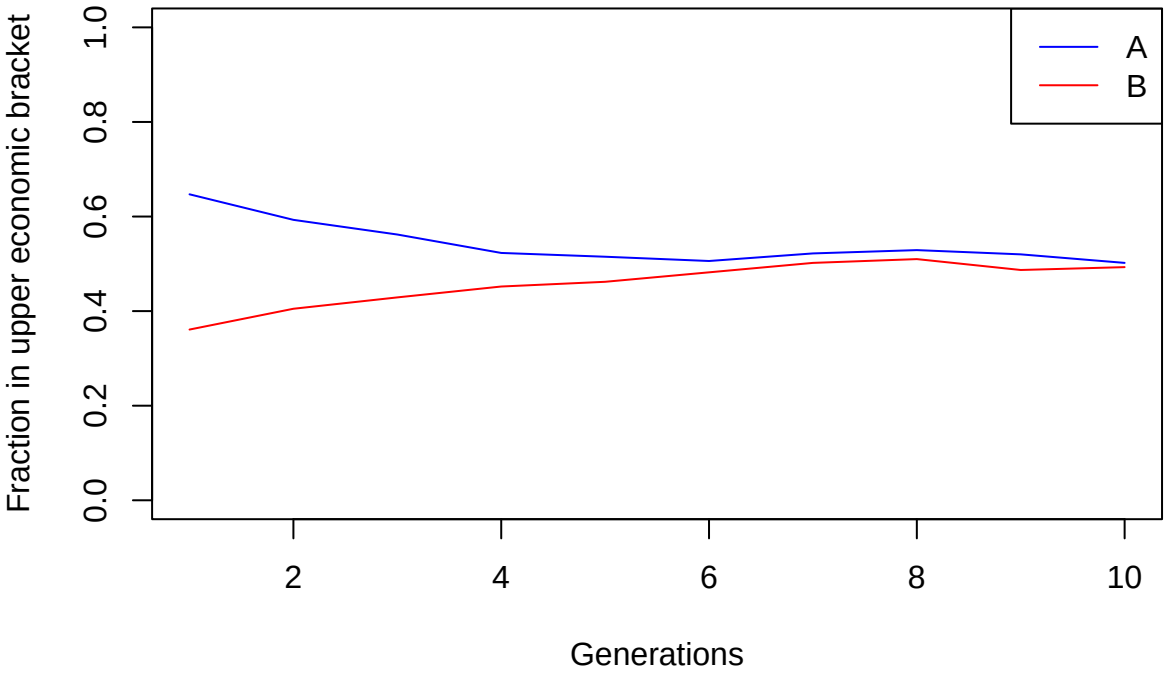


Figure 2: If two sub-populations start with different distributions over economic positions, but thereafter have the same rates of relative economic mobility, they will eventually converge on the same distribution, but large disparities can persist for generations.

Here is the moral of this calculation:

> Once a categorical inequality forms, if individuals' relative mobility is low, and membership in the categories is inherited, then the inequality can persist for a long time, *even if* there is no

[2] If you know enough to quibble here about different ergodic components, you also know enough to see that the state space for this kind of model is necessarily a single irreducible set and that the transition matrix is aperiodic.

discrimination or prejudice

If poverty has other bad consequences (ill health, exposure to crime, being treated badly by the police, etc.), the distribution of those consequences would also be different between groups as well, again, even without discrimination and prejudice.

It's important to the argument that group membership is inherited. So this works better for categories like race, ethnicity or religion than it does for sex.

A not-at-all-random application: the US civil rights acts were passed in the 1960s, i.e., two generations ago. Even if those laws had completely eliminated all forms of racial discrimination, we would still expect to see very large economic disparities between races in the 2020s. Of course, the civil rights acts did not eliminate all forms of racial discrimination and prejudice. Also of course: whether it's acceptable to just (!) eliminate discrimination and let social mobility work to slowly reduce inequalities is an ethical and political value-judgment, not a scientific proposition. But it in making that judgment, it can be *helpful* to know what we'd see in a non-discrimintion-but-nothing-else world, and how close (or far) the actual world is to that baseline.

Remember this is about *relative* mobility, rank in the distribution, not absolute mobility — both groups could be better off in absolute terms over time (or worse off...) If the worse-off group, B, is to catch up to the better-off group, A, in *absolute* terms, it needs *higher* rates of income growth than group A.

I *think* a lot of what our friends in the humanities and some of the social sciences call "systemic" or "structural" inequality is a way of talking about this mechanism, of persistent, inter-generational inequality, or ones very like it. (But I have to admit I am often not sure what exactly, some of them are trying to get at.)

3 Schelling, and the unraveling of integration

- Schelling (1971) introduced a now-classic model which he intended to illustrate a point about how difficult it can be to *maintain* equality, and how the outcomes people produce may be very different from the outcomes they *want*.
  - Two types of people, say A and B
  - Everyone lives on a square lattice (large checker-board or chess-board)
  - Everyone is perfectly happy with neighbors of the opposite type...
  - ... but not comfortable being in a small minority
  - Specifically, if the fraction of your neighbors who are of the same type as you is $< p$, move to a random un-occupied square, otherwise stay put
    - The classic version defines "neighbors" has the 8 squares on the cardinal directions and diagonals
  - If you start with a random scatter of As and Bs, most people are initially fine with the neighborhood...
  - ... but some people are uncomfortable because they're the only B in a block of As (or vice versa)
  - Key point: every time a B moves because there were too many As around, their old neighborhood is now *even more* A-predominant...
    - And their new neighborhood has even more Bs...
  - ... which can induce *others* to move
    - more Bs will leave the first B's old neighborhood, As will leave from the new neighborhood
  - Result: integrated patterns unravel, and we get big homogeneous blobs where almost everyone is of the same type
    - The fraction of neighbors of the same type isn't $\approx p$ but much larger than p
- Schelling's point is that it's hard to maintain integration, unless (i) the initial configuration is *very* carefully engineered, or (ii) you can persuade everyone to be OK with being in a small minority (very low p), or (iii) people are not free to move as they like
  - The government of Singapore has pursued option (iii)
  - It's very hard, legally, to pursue option (iii) in the US; see the discussion of the Starrett City development in New York, which tried, in Ford (2008), ch. 4, pp. 285–294.
- Schelling, not being an idiot or an apologist, was quite clear that the two main reasons why the US was racially segregated in 1971 were a long history of active efforts to *make* it segregated, and the economic inequality which priced blacks out of good neighborhoods. He described his model as one of a "third order" effect, behind these. But the model *does* show that just eliminating active segregationist efforts, and even eliminating racial economic inequality, would not necessarily lead to integration. It also illustrates how the collective outcome of individuals' choices can be very different from what any of the individuals want!

4 Unequal institutions

There are many situations involving the division of labor, gains from cooperation, etc., where everyone benefits if there is *some* recognized rule or convention, rather than having to re-negotiate everything from scratch every time. (Imagine something as simple as roommates dividing up the work of cleaning their apartment.) These conventions or rules are examples of what are called **social institutions**, or just **institutions**. Here are some examples of institutions (in this sense):

- Everyone drives on the right
- The "Pittsburgh left": when a traffic light turns green, cars turning left have right-of-way over those going straight. This is 100% illegal (according to Pennsylvania's traffic code), but it is, or was, a commonly-recognized rule around here
- Greeting acquaintances with hand-shakes, or bows, or kisses on the cheek.
- The grammar and vocabulary of a language.

An institution can often be thought of as a solution to a social problem ("how can we drive multi-ton machines at high speed through cities without killing ourselves or each other?"). That being said, there are usually *many* possible institutions for any particular situation, many solutions to the same problem, which would all be better than sheer disorder. Often, some of these possible institutions are much better for one party than others. (Imagine that one roommate has to take out the trash once a week, but the other roommate has to cook, clean the kitchen, and scrub the bathroom.)

- We could equally well all drive on the left, and many countries do. (This doesn't seem to advantage or disadvantage anyone particularly.)
- *Do* you greet an acquaintance with a hand-shake, or by bowing, or by kissing them on the cheek?
  - Even if you don't care *which* of these, you do want to do the same thing as those around you, to avoid awkwardness and/or looking like a barbarian and/or embarrassing misunderstandings.
- *Which* language are you using? There are very real and obvious advantages to *sharing* a language with those around you. There are also very real and obvious advantages to a firm, school or government office picking *a* language to do business in, which gives an advantage to those who are already good with that language.
  - This is why disagreements about language are often at the core of nationalist disputes, why nationalist movements often begin by creating a standardized literary language that can be used in in education, government and business, etc. (Gellner 1983).
- If you accept a job with a one-year contract, are you free to quit before the year is out? Can your boss fire you before the end of the year? It might seem obvious, to modern Americans, that the answer to both questions is "yes", but this is one of our taken-for-granted institutions. Historically, even in England as late as 1875, workers who quit could be sued or even prosecuted as criminals (Naidu and Yuchtman 2013). While there might be outlandish scenarios where this constraint somehow helped workers, historically they thought it was very much against their interests and struggled, successfully, to get rid of it (and employers, for their part, agreed that benefited them at workers' expense). At the other extreme one can imagine institutions under which an employer who agrees to hire you for a year has committed to paying you for the full year no matter what.

All of this raises three puzzles:

1. How do institutions form?
2. When there are multiple possible institutions, *which* institutions form?
3. When some possible institutions benefit some groups rather than others, who wins?

It turns out that institutions can self-organize even when individual agents follow very simple learning rules. In particular, rules that amount to versions of "do more of what worked for you and less of what didn't" (individual learning), or "copy those who did better than average" (social learning) can do the job, and the precise details often don't matter much.

Simple learning rules like this often lead to situations where (almost) everyone acts the same way, or everyone in the same group acts similarly and this meshes with the actions of other groups. Optimistically, this offers

some hope that society will self-organize *some* kind of solution to coordination problems. Pessimistically, it will not necessarily self-organize to a very *equal* solution. In particular, plausible models of social learning are typically ones where already-advantaged groups have an edge in struggling over which institution to use, further entrenching their advantages.

Once an institutional convention forms, it tends to be very persistent and "sticky", because nobody benefits from being the *only* one pushing against the institution. Institutions *can* change when objective conditions change. Institutions can also change if enough people are stubborn and/or don't care about the (short-term) consequences (collective action), or through random drift. But institutions once formed can be powerfully self-sustaining, even if *everyone* would be better off switching to an alternative.

The last few paragraphs have made a lot of assertions without backing them up; the rest of this section will try to provide some substantiation.

4.1 Game theory: a crash course

Decision theory is about evaluating strategies for taking action in the face of *random* uncertainty. It's an important part of statistics (some argue *the* core idea of statistics). It has an extension to dealing with situations where the decision-maker isn't just facing some unknown, more-or-less random situation, but rather *another* decision-maker. That extension is known as "game theory".

- Two (or more) players or **agents**
- Each agent picks an action (or "move"), and gets a payoff that depends on their action and those of all other players
  - Simple cases: all agents have the same discrete set of actions available
    - Square payoff matrices showing the payoff that each agent gets
  - Continuous actions, distinct action sets for different agents, etc., all possible
  - You can imagine multiple rounds of games where agents respond to each others' previous moves, etc.
    - Reduces to the "one-shot" case by considering *strategies* rather than *moves*
- "Payoff" can be monetary, utility, number of offspring (in evolutionary applications, etc.)
- An **equilibrium**[3] is a situation where no agent benefits from being the *only one* to change their action[4].

4.1.1 The Prisonners' Dilemma

As a classic example, here's the prisoners' dilemma[5]:

[3]The word "equilibrium" is used in a number of distinct-but-related ways in economics. I am about to define a **pure-strategy Nash equilibrium**. "Equilibrium" has even more distinct meanings in other fields, such as physics. Anderson, Arrow, and Pines (1988) offers some classic accounts of social and natural scientists misunderstanding each other because (in part) of the slipperiness of this word.

[4]In symbols, an assignment of actions to agents, $a = (a_1, a_2, \ldots a_n)$, is a (pure-strategy Nash) equilibrium when no agent can increase their payoff by changing *only* their action. Writing $r_i(b_1, b_2, \ldots b_{i-1}, b_i, b_{i+1}, \ldots b_n)$ for the payoff to agent i when agents take actions $b_1, \ldots b_n$, a is a Nash equilibrium when $r_i(a) \geq r_i(a_1, \ldots a_{i-1}, b, a_{i+1}, \ldots a_n)$, for all agents i and for all alternative actions b. (Some people strength this to $>$ rather than $\geq$, others distinguish ordinary from "strict" equilibria.) — A **mixed-strategy** Nash equilibrium, incidentally, is an assignment of *probabilities* of actions to agents where no one can increase their *expected* payoff unilaterally.

[5]The standard parable, which goes back to the 1950s and is the origin of the name, runs as follows. Alice and Bob are political dissidents who have both been brought in by the police for questioning. If both of them keep their mouths shut (C,C), the police have no other evidence against them and have to let them go. (This parable was invented by Americans with very little experience of secret police.) If one of them informs on the other (D,C), the informer (D) gets a reward while the one informed against (C) receives an especially harsh punishment as an example. Finally, if they both turn informer (D,D), they both receive merely ordinary punishment. Here is an alternative parable: Ashoka and Babur are two bandits. If they team up, they can combine their forces to attack caravans and steal valuable loot, which they could split equally (C,C). Or one of them (D) could turn on the other (C) *after* the attack, taking all the loot for himself. Or they could both refuse to cooperate (D,D), and have to content themselves with small-scale mugging in back alleys.

	C	D
C	(2,2)	(0, 3)
D	(3,0)	(1,1)

If both the row and the column player make the move "C", they both get 2. If the row player makes the move C but the column player makes the move D, the row player gets 0 and the column player get 3, and vice versa. If the both play D, they each get 1. Clearly, both players are better off if they can somehow cooperate to both play C, rather than to both play D. But if you know that the other player will play C, you can improve your own pay-off by switching to D. If on the other hand you know the other player will play D, then to protect yourself you must play D as well. The two letters C and D stand for "cooperate" and "defect", and we have just reasoned ourselves into concluding that defection is the equilibrium, and in fact the *only* equilibrium.

The prisoners' dilemma is a very simple game, but it encodes a *lot* of what makes social cooperation difficult to create and sustain. A great deal of ingenuity has gone into trying to figure out how, despite the all-defect equilibrium, cooperation can evolve and persist in the prisoners' dilemma, an effort which has yield a lot of insight into important parts of biology, economics and philosophy. (See further reading.) If this was a different kind of course we could devote months to it. Instead, we're going to focus on a different kind of game.

4.1.2 A coordination game

The prisoners' dilemma has *one* type of agent. We'll now start to worry about situations where we divide up the whole collection of agents into discrete, distinct social groups; I will call these by colors, such as blue, red, etc. But we also have a distribution of types of agents within each group. This makes "the population" ambiguous. To avoid ambiguity, each group will be a "population" of agents, and the complete collection of all agents of all groups will be an "assemblage"[6].

Here's an example pay-off matrix for a game with two types of players, red and blue, and two actions.

	Red makes move A	B
Blue makes move A	(4,1)	(0,0)
B	(0,0)	(2,2)

The first number in each cell shows the pay-off to the row player (blue), the second the pay-off to the column player (red).

This is a **coordination game** — the two players benefit by doing the same thing; the (A,A) and (B,B) outcomes are better for everyone than the (A,B) or (B,A) outcomes. (This is unlike the prisoners' dilemma.) It's also a game with multiple equilibria: both (A,A) and (B,B) are equilibria. (Check this!) But the two types of agents don't benefit equally from the two equilibria. Clearly, the (A,A) equilibrium is much better for blue agents than the (B,B) equilibrium, while the red agents have the opposite preferences[7].

We'll analyze this little coordination game pretty thoroughly, because these features — a shared interest in reaching agreement, multiple equilibria with different agreements, and conflicting preferences over equilibria — make it a clarifying model for struggles over institutions.

[6] I am borrowing "assemblage" from DeLanda (2006), but I am not altogether sure I'm use the term in quite the way he would.

[7] I have made the payoffs for red and blue agents equal at the (B,B) equilibrium, but that's a bit artificial. If I divided all the blue payoffs, but not the red, by a factor of 4, then (A,A) would look egalitarian, with payoffs (1,1), and (B,B) would look biased in favor of reds, (0.5, 2). But you should ask yourself, as we go along, whether there's any way you could *detect* this, without directly observing the payoffs themselves.

4.2 Classical game theory

Classical game theory analyzes the interaction of **strategies**, which are rules saying which move to make as a function of the other players' previous moves, your previous moves, and perhaps some external signals. The usual approach is very "forward looking", where agents are supposed to try to maximize their future payoffs given presently available information, taking into account that the other agents are doing likewise.

This is a mathematically intricate and often beautiful body of theory. I will give just one example of the kind of reasoning it employs, namely "backward induction" in the multi-stage prisoners' dilemma. So suppose the two agents are playing the prisoners' dilemma, with the payoff matrix above, for T rounds; each agent wants to maximize the sum of all the payoffs. If they cooperate each round, each agent will collect $2T$, but let's see if that's really feasible. Start at the end and work backwards: what should the agent do to maximize its payoff on the *last* round, T? Well, that looks just like a one-shot prisoners' dilemma, so it should defect. By symmetry, then, on the last round *both* agents will defect. What about on round $T-1$? You might hope that the agents could sustain cooperation up to that point somehow, but notice that if *one* of them decides to defect at time $T-1$, there's nothing the other agent can do to induce it to change its mind, since they both know they will both be defecting at time T. So both agents will defect at time $T-1$ as well at time T. By mathematical induction, they defect all the way back time 1, and so they only collect a payoff of T each, instead of the $2T$ that could have been theirs if they'd cooperated all the time[8].

Classical game theory is, as I said, a beautiful subject, but it involves a *lot* of this sort of I-know-that-you-know-that-I-know, which becomes very implausible very quickly, at least as an explanation of social phenomena involving lots of ordinary people who are not good at logic puzzles.

4.3 Evolutionary game theory

An alternative way of thinking about games is what's called "evolutionary" game theory[9]. The basic ideas here are that people usually follow habits rather than engaging in complicated forward-looking reasoning, but they're not stupid, they can learn from their own experience about what works and what doesn't, they're willing to try new things at least occasionally, and they can learn from others' experiences, too.

From this point of view, what we really care about is the distribution of strategies within each population, what distribution of payoffs that leads to, and how those distributions change as a result of individual and social learning. I say "strategies", but remember that we're now assuming the agents are mostly just acting by habit, so the strategies will be very simple; in fact let's just reduce them to individual *moves* in the game.

To see why the distributions matter, let's think about the coordination game. Let's write p_A and p_B for the probability that a blue agent plays A or B, and likewise q_A and q_B for the probabilities for red agents. (This really only gives us two variables, because $p_A + p_B = q_A + q_B = 1$, which is convenient for making plots.) The expected payoff to a blue agent from playing A is $4q_A$. The expected payoff to a blue agent playing B is $2q_B = 2(1-q_A)$. So A has a higher expected payoff than B, for blue, when

$$4q_A > 2(1-q_A) \Rightarrow q_A > 1/3$$

Similarly, the expected payoff of A to a red agent however is just p_A, while red's expected payoff from B is $2p_B = 2(1-p_A)$. Thus, red does better, in expected, from playing A when

$$p_A > 2(1-p_A) \Rightarrow p_A > 2/3$$

[8]Of course, for this sort of backward induction to work, the agents have to *know* that round T is the last one. If the game might continue forever, cooperation *can* be sustained even in classical game theory. It's enough in fact to always have a large-enough probability δ of continuing the game for another round. (If you know that the other player follows a "grim trigger" strategy of cooperating until you defect, and thereafter defecting on you forever, what's the minimum δ which makes it worth your while to always cooperate?) Game theorists call this the "shadow of the future".

[9]The theory actually comes from evolutionary biology (Maynard Smith 1982), but using doesn't commit us to thinking that social change is a process of *genetic* evolution.

Blue agents *really* benefit if everyone plays A rather than B, so much so that they prefer to play A when only a minority (1/3) of reds will go along. But reds have the opposite interests, and so they only prefer to play A when a substantial majority (2/3) of blues are forcing their hand (so to speak).

These calculations matter because we said that agents learn from experience, and prefer actions which work well (i.e., have high payoffs) to those which work poorly (i.e., have low payoffs). Expected values aren't everything, but we should still anticipate that when A is usually better than B for blues, any blues who are playing B are going to tend to switch to A, while those already playing A will tend to stick with what is working for them. (See also the complementary exercises.)

4.4 Replicator dynamics and social learning

This leads us to the important topic of the **replicator dynamic**. Here's how it works with one population, and available strategies i. At time t, a fraction $p_i(t)$ follows that strategy. The expected payoff to that strategy, *given the probabilities of all the other strategies*, is $m_i(t)$, often called the **fitness** of the strategy. The "dynamic" then is to update the population shares according to what's called the "replicator equation",

$$p_i(t+1) = p_i(t)\frac{m_i(t)}{\sum_k p_k(t) m_k(t)}$$

Notice that the denominator on the right-hand side is the average fitness, i.e., the expected payoff to a randomly-selected member of the population. So what this says is that strategies which do better than average will become more common, while strategies which do worse than average will become less common. It's a version of "do more of what worked and less of what didn't".

Now, as for where $m_i(t)$ comes from, it is, implicitly, a function of the distribution of strategies in the population. The payoff matrix is, let us say, $\mathbf{r}$, where $\mathbf{r}_{ij}$ tells us our payoff to playing i when the other agent plays j. Then

$$m_i(t) = \sum_j r_{ij} p_j(t) = (\mathbf{r}p(t))_i$$

abbreviating the vector of all the $p_i(t)$ as $p(t)$.

There are a couple of things to notice about the replicator dynamics (exercises):

- If the $p_i(t)$ are valid probabilities, so $0 \leq p_i(t) \leq 1$, $\sum_i p_i(t) = 1$, and all the $m_i(t) \geq 0$, then the $p_i(t+1)$ are valid probabilities, too.
- If $p_i(t) = 0$ then $p_i(t+h) = 0$ for all $h > 0$; "extinction is forever".
- If one strategy i has a *higher* fitness than all the others, $m_i(t) > m_j(t)$, then that strategy will increase its share of the population, $p_i(t+1) \geq p_i(t)$.
- A **fixed point** of a dynamical system is a state which doesn't change under the dynamics. A fixed point of the replicator dynamic would be a vector of probabilities p where plugging the p_is into the replicator equation on the right-hand side gives us back p_i again on the left. This requires every strategy i where $p_i > 0$ to have the *same* fitness.
- Changing all the payoffs, and so all the fitnesses, by the same factor would leave the replicator dynamic unchanged.

Now, all of this is for one type or population of agents, playing against others from the same population. If we've got an assemblage of multiple types, the changes are pretty straightforward and almost just notational. Let's write $p_i(t)$ for the probability that a member of the "blue" group plays i at time t, and likewise $q_i(t)$ for the probabilities in the red group. The fitnesses will be $m_i(t)$ for the blues and $\ell_i(t)$ for the reds. The

relevant equations become

$$m_i(t) = \sum_j r_{ij}^{(blue)} q_j(t) \quad (4)$$

$$\ell_i(t) = \sum_j r_{ij}^{(red)} p_j(t) \quad (5)$$

$$p_i(t+1) = p_i(t) \frac{m_i(t)}{\sum_k p_k(t) m_k(t)} \quad (6)$$

$$q_i(t+1) = q_i(t) \frac{\ell_i(t)}{\sum_k q_k(t) \ell_k(t)} \quad (7)$$

The important points are that the fitness of members of one population depends on the distribution of strategies in the *other* population, but strategies grow or shrink within a population as they do better or worse than the average for that population.

Why the replicator dynamic? There are several reasons.

- Some evolutionary processes really do consist of copying or "replication". If there are np_i members of the population of type i, and on average each of them creates m_i copies, then there will be $np_i m_i$ new members of type i, and a total population size of $n \sum_k p_k m_k$. The replicator equation then describes the new distribution of types.
- Many processes of individual and social learning which may seem more complicated than simple copying have been shown mathematically to reduce to the replicator dynamic, either exactly or approximately in certain limits (like large population sizes, or gradual learning, etc.). This includes reinforcement learning (Börgers and Sarin 1997), learning by Bayesian updating (Shalizi 2009), and many others. So studying the replicator dynamic gives us at least an approximate handle on what other sorts of learning will do.

4.4.1 Replicator dynamics in the coordination game

To make this all less abstract, and to return to institutions, let's go back to our coordination game. Under the replicator dynamic, the equations of motion are

$$p_A(t+1) = p_A(t) \frac{4q_A(t)}{4p_A(t)q_A(t) + 2(1-p_A(t))(1-q_A(t))} \quad (8)$$

$$q_A(t+1) = q_A(t) \frac{p_A(t)}{q_A(t)p_A(t) + 2(1-q_A(t))(1-p_A(t))} \quad (9)$$

(The denominators here are the expected payoff to a random blue or random red agent, respectively.) The ratios on the right hand sides will be > 1 when A is the preferred action, because that's exactly the time when A has better-than-average payoffs. On the other hand the ratios will be < 1 when A is the dis-favored action. So:

- If at least 1/3 of the reds are playing A, the share of blues playing A will increase
- If less than 1/3 of the reds play A, the share of blues playing A will decrease
- If at least 2/3 of the blues play A, the share of reds playing A will increase
- If less than 2/3 of the blues play A, the share of reds playing A will decrease

What are the fixed points of this, the situations where $p_A(t+1) = p_A(t) = p^*$ and $q_A(t+1) = q_A(t) = q^*$? It would be nice if the equilibria were fixed points, so let's try that. Plugging $p_A(t) = 1$, $q_A(t) = 1$ into the equations gives us back out $p_A(t+1) = 1$, $q_A(t+1) = 1$, so that's one fixed point. Similarly, $q^* = 0, p^* = 0$ is another fixed point. So the two equilibria are indeed fixed points. (See also the complementary exercise.)

Here are some representative trajectories of this system.

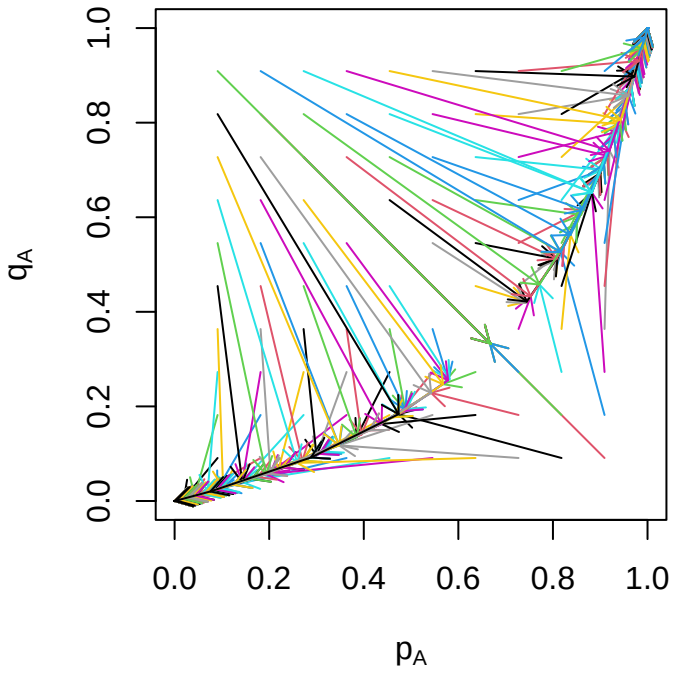


Figure 3: Trajectories for the coordination game, from evenly-spaced initial condtions. (Colored lines connect points on the same trajectory, with arrows showing the direction of motion.) Only the probabilities of each side playing A action are shown. Notice how trajectories move towards the equilibria of the game, which here are $p_A = 1 = q_A = 1$ (referred by the blue agents) and $p_A = 0, q_A = 0$ (preferred by the red agents). Trajectories which start out below the anti-diagonal line will tend to one equilibrium, those above the line to the other.

4.4.2 Some lessons from the example

- Depending on the initial values of p_A and q_A, the assemblage (almost always) ends up at one equilibrium or the other[10]. In social terms, *some* institution forms spontaneously, or **self-organizes**.
  - This is true even though none of the agents has any representation of the institutions or equilibrium, or of the distribution within their population or the other population, or engages in any sort of complicated reasoning. They just do more of what works well right now for other agents like them, and less of what's working poorly.
  - In particular, none of the agents has anything like an *intention* or *plan* to make the population coordinate on their favored equilibrium (i.e., to create institutions that serve their interests).
- If the assemblage starts very near an equilibrium, it approaches that equilibrium (rather than the other one, or doing something weirder).
- Therefore, if the assemblage is at, or close to, an equilibrium, and something perturbs the assemblage, it will tend to return to the old equilibrium. That is, once an institution has been reached, it tends to be *stable*[11].
  - Another way to say this is that the assemblage displays **negative feedback** around each equilibrium: if it's disturbed away from the equilibrium, that sets forces in motion which tend to restore the equilibrium[ˆfeedback]. Negative feedback *looks* an awful lot like purposeful behavior, and a good case can be made that purposeful behavior requires negative feedback (Rosenblueth, Wiener, and Bigelow 1943). But in this case no agent *wants* to maintain the equilibrium, even though some of them benefit from it (and benefit more than others).
  - It's reasonable (but not iron-clad) that the assemblage will spend most of its time close to an equilibrium. ("Stable things persist".)
- A sufficiently large perturbation *could* change the state of the system so much that it would then, if left to its own devices, head to the other equilibrium. It might not be necessary to go all the way to the neighborhood of the other equilibrium[12].

4.5 Perturbations

What would constitute a "perturbation", in this sense, to the replicator dynamic? There are three leading possibilities: shocks, noise, and collective action.

- *Shocks* are some external event or force which (temporarily) compels, persuades or induces a bunch of agents to do something differently for a while.
- *Noise* is any sort of probabilistic or stochastic component in the dynamics. The replicator dynamic can be understood as describing the expected behavior of very large populations, where the law of large numbers makes randomness average itself away, leaving a deterministic limit. But limited populations open the way for randomness to matter, particularly near boundaries between domains of attraction.
- *Collective action* is a group of agents deliberately coordinating their behavior to change the dynamics, or at least the outcome of the dynamics.

I won't say anything more about shocks, but noise and collective action deserve some elaboration.

[10]The set of all states of the assemblage which will tend towards a particular fixed point consistutes that point's **basin of attraction** or **domain of attraction**. Here the two domains of attraction of the equilibria are the two triangles, above and below the anti-diagonal line from $(0, 1)$ to $(1, 0)$. In general the domains do not have to be so nicely-shaped and symmetric.

[11]In dynamical systems theory, we think about dynamical rules like $x_{t+1} = f(x_t)$. We say that u is a **fixed point** of f when $f(u) = u$, so that if the dynamics starts at that state it stays there. A fixed point u is **locally stable** if all trajectories which start sufficiently close to u tend to u. One way to formalize this is to say that there exists a distance $\delta > 0$ such that $0 < \|v - u\| \leq \delta$ implies $\|f(v) - u\| < \|v - u\|$. This would mean that the system showed *negative* feedback, at least locally, because the consequence of moving away from the fixed point would be to head back towards it.

[12]In terms of an earlier footnote, the perturbation just has to move the assemblage into the domain of attraction of the target equilibrium; then the dynamics will do the rest (presuming there are no further perturbations).

4.5.1 Noise

What I've shown above is the pure replicator dynamic. This is a dynamical system for the *distribution* in each population, but the dynamics are *deterministic*: once we say what the initial distributions are, the distributions at all later times are fixed. It is a little unclear, though, how this is supposed to matter to the individuals comprising the members of the population. I waved my hands about how each individual looks at their population and copies the successful, but that was just hand-waving.

Let me now suggest a partially random (or "stochastic") set of rules which will have the replicator dynamic as their *expected* behavior, so what we'll actually see is the replicator dynamic plus noise: - There are n_{blue} blue agents and n_{red} red agents. At every point in time, each agent is either a habitual A player or a habitual B player. - At time t, we randomly pair red and blue agents. (If $n_{blue} \neq n_{red}$ someone has to play more than once.) The paired off agents play their habitual moves and collect their payoffs according to the payoff matrix. - Each red agent selects another red agent with a probability proportional to the latter's payoff. The first one then adopts the habitual action of the selected agent.[13] The blues do likewise. - Repeat.

Under this set-up, the probability that a blue A-player is paired with a red A-player is q_A, in which case that blue agent gets a payoff of 4. The expected total payoff of all the blue agents will be $4n_{blue}p_Aq_A + 2n_{blue}p_Bq_B$. The expected probability that a blue agent will copy *one* of the $n_{blue}p_A$ players is then

$$\frac{4n_{blue}p_Aq_A}{n_{blue}(4p_Aq_A + 2p_Bq_B)} = p_A\frac{4q_A}{4p_Aq_A + 2p_Bq_B}$$

just as it should be for the replicator dynamic. (I will let you verify that this holds for all the other combinations of color and move.) So the on-average behavior of this random process will track the replicator dynamics. Moreover, it should be intuitive that increasing the population sizes, the n_{blue} and n_{red}, will bring the actual random averages increasingly close to the expected values (by the law of large numbers). So very large populations should move especially close to the limiting deterministic behavior.

Here is a demonstration of the trajectories we *can* get from this stochastic replicator dynamic, starting from the same initial conditions for the assemblage as in the previous figure, but with $n_{blue} = n_{red} = 50$.

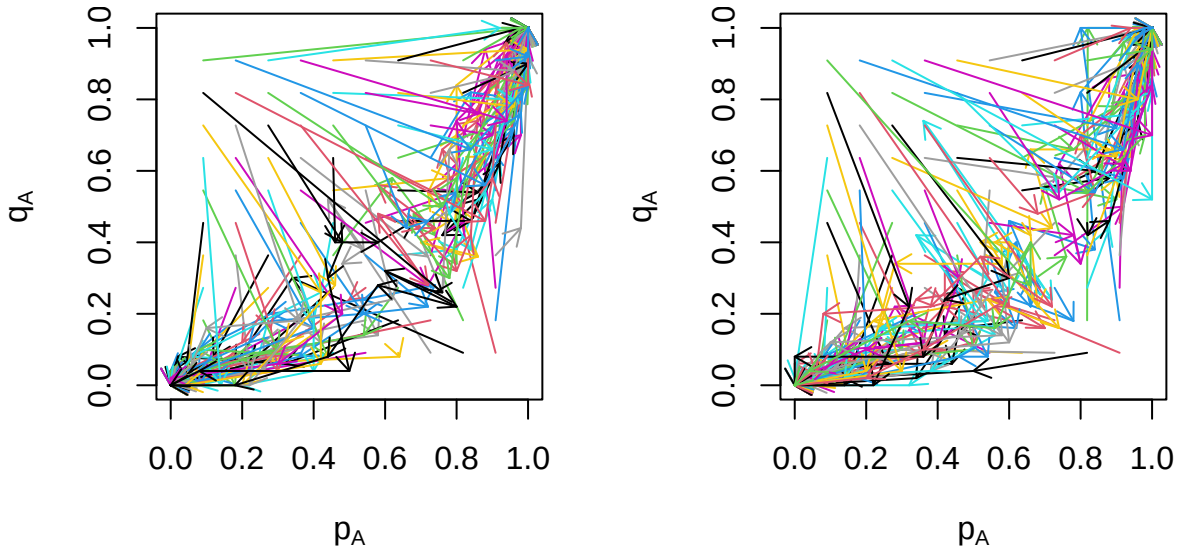


Figure 4: As in the previous figure, but now using finite populations (50 players of each color), randomly paired up in each round of the game. The two panels show two different runs of the stochastic process, from the same initial conditions.

The over-all tendency is clearly still similar to that of the infinite-population deterministic model, but equally

[13] At this point we could even *define* the payoff function this way, through the probability of being an object of emulation.

clearly things are much more erratic. Of course, if I re-run the code, I get a somewhat different set of trajectories (compare the two panels of the figure).

In particular, while the assemblage will still (probably) tend to an equilibrium, *which* equilibrium it reaches depends on where it starts *and* on chance fluctuations along the way. Fluctuations are particularly important near the boundary between two domains of attraction; chance can carry an assemblage which "should" end up at one equilibrium over into the domain of the other. Proverbially: "For want of a nail, a horse was lost", and so on.

If I increase the population size by (say) a factor of 10, the behavior is closer to the deterministic version of the model, because more of the randomness averages away. The long-run outcome is *less* influenced by noise, but not *un*-influenced by noise.

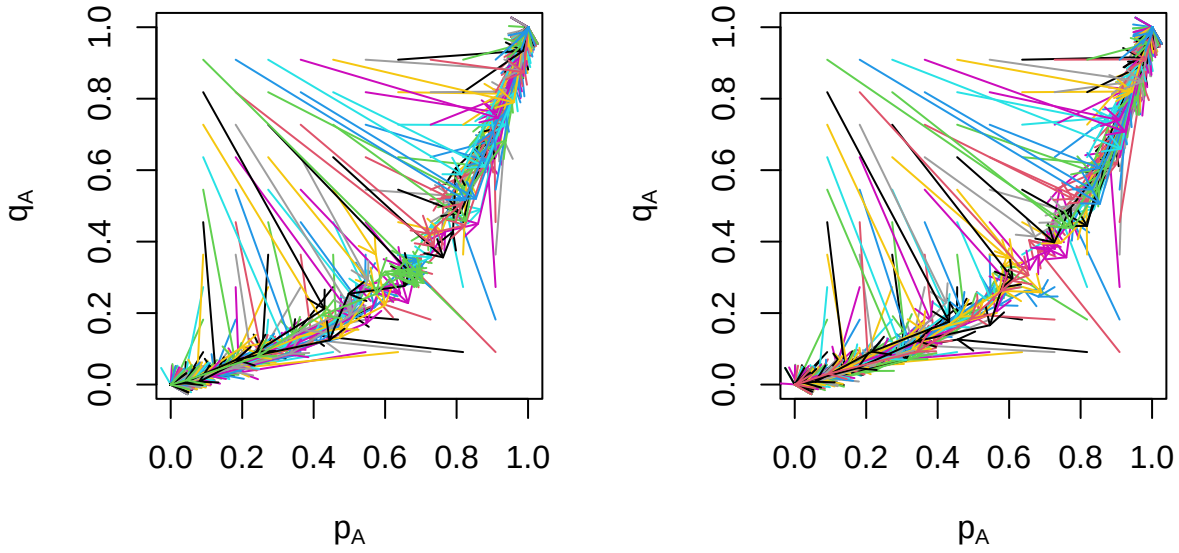


Figure 5: As in the previous figure, but increasing the population size by a factor of 10. Again, the two panels show two independent sets of realizations of the stochastic process, starting from the same initial conditions.

The next figure shows the *distribution* of outcomes that can happen when we start the stochastic replicator dynamic near a boundary between two domains of attraction. You can see behavior that comes out of the deterministic replicator dynamic, which in this case eventually converges to the $p_A = q_A = 0$ equilibrium. That's what we'd anticipate seeing in infinitely large populations. You can also see that with $n_{blue} = n_{red} = 50$, a lot of the stochastic trajectories kind-of-sort-of follow along, but some of them are definitely converging to the other equilibrium.

Deep in a domain of attraction, in particular very close to an equilibrium, we'd need to see very large fluctuations to change the outcome[14]. This is, almost by definition, very unlikely[15], but "very unlikely" is not "impossible". In particular, if we wait long enough, anything which is *merely* very unlikely, but not impossible, *will* happen eventually. There are even techniques for calculating how long it will take for fluctuations to spontaneously move us from the vicinity of one stable fixed point to another, finding the most likely path of this admittedly-unlikely trajectory, and so on.

[14]This could be either a lot of unlikely events at once (e.g, red A players have higher payoffs at time t than red B players, but nonetheless all red agents randomly copy B players), or less individually less unlikely ones that are sustained over time in the same direction (e.g., a slight random excess of copying B's every time step for many time steps).

[15]Basically, we say that a fluctuation away from the expected behavior is "large" if it becomes exponentially unlikely as n grows. We then measure the size of large fluctuations by the exponential rate at which they become improbable. The all-red-agents-copy-B-players-at-once event from the previous footnote is going to be exponentially improbable because each red agent has some probability of doing that, say s, but they do so independently so the probability of them *all* doing it will be $s^n = e^{-n \log s}$. Having a slight excess of copying Bs at one time will likewise be exponentially unlikely at any one time, and having that sustained over time will require multiplying together probabilities.

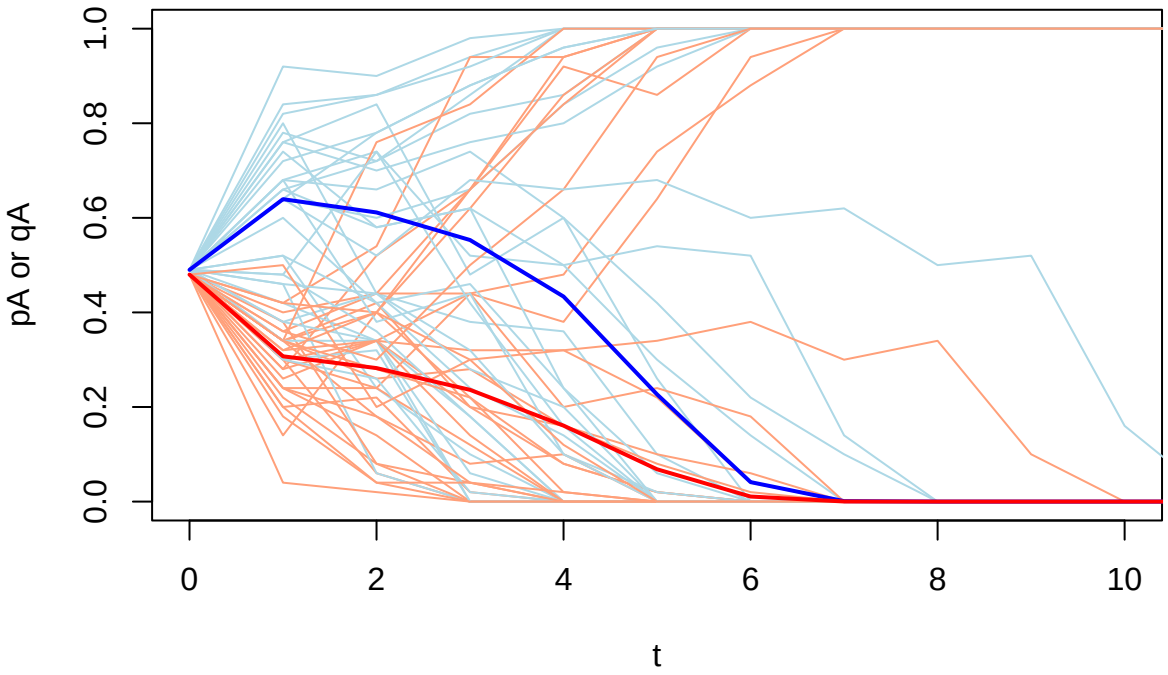


Figure 6: Comparison of the deterministic and stochastic (50 agents per color) dynamics, starting from the same initial conditions. The stochastic trajectories are shown in paler colors, the deterministic one by more saturated colors in thicker lines. While most of the stochastic trajectories sooner or later reach the same equilibrium as the deterministic model, at this small population size we can see that an appreciable number do not.

Now, as I've written out the model above, if the assemblage has *completely* converged on one equilibrium or the other, so $p_A = q_A = 0$ or $p_A = q_A = 1$, it's stuck there, because the *only* way agents change their habits is by copying, so once everyone is doing exactly the same thing they'll keep doing it. An easy fix is to allow everyone *some* (small) probability of spontaneously switching from being an A player to a B player and vice versa. If we think in more human terms about the agents, this switching probability could represent sheer mistakes or mis-understandings, occasional stubbornness, or, frankly, creativity and innovation. The important thing is that it will keep the assemblage from getting completely stuck at an equilibrium.

In the rest of these notes, I am going to "turn the noise back off", and just show the deterministic behavior. I will, however, make some remarks about the effects of noise.

4.5.2 Collective action

As I said, in this sort of model, institutions (tend to) form even though no one is thinking about that, trying to achieve it, or considering alternatives. Of course *people* can think about institutions and alternatives, and in particular about whether *different* institutions would be better *for them*.

In the case of this model, the (B,B) equilibrium is better for the reds than the (A,A) equilibrium is. If the assemblage has coordinated on the (A,A) equilibrium, an *individual* red agent insisting on playing B is not going to achieve anything except a 0 payoff for themselves. But *enough* red agents all agreeing to play B, and to keep playing B, could tip the balance[16]. This would be an example of **collective action** on the part of the red agents.

Of course, agreeing to and carrying out this collective action raises other issues. In the first place, the red agents who insist on playing B when the assemblage has coordinated on (A,A) are, at initially, going to get a payoff of $0 < 1$ for reds who keep playing A. Maybe worse, if the red B-players succeed, and tip the assemblage over to the (B,B) equilibrium, then *all* the red agents will benefit, even those who *didn't* participate in the collective action. This creates an incentive for red agents to "free ride" on the red B-players, by continuing to play A until the new equilibrium is established. But then why would *any* red agent join the collective action?[17] Now, as a matter of fact, collective action to change institutions happens all the time, but social scientists have come to see it as a *puzzle* which requires explanation, and often requires institutional underpinnings of its own[18].

4.6 Persistence in the face of changing conditions

Suppose the payoff matrix used to be the one I gave above, but that conditions have gradually (or suddenly) changed, so that the new payoff matrix is now

	Red makes move A	B
Blue makes move A	(4,1)	(0,0)
B	(0,0)	(5,5)

The two equilibria are still (A,A) and (B,B), but now *both* populations would prefer the (B,B) equilibrium[19]. Nonetheless, if enough members of one populations are playing A, then A is the better action for their partners in the other population, so the (A,A) equilibrium is still stable[20]; see the figure.

One consequence of this stability is that if the assemblage coordinated on the all-A institution in the past, this institution can persist, *even though* the original conditions which helped bring it into being have gone away. It can even persist despite *everyone* preferring the alternative all-B institution. Fundamentally, though, it can persist because inefficient coordination is better (for everyone) than *no* coordination.

[16] Specifically, in this model, at least 2/3 of the reds would have to agree to play B, *and stick to it*. The emphasized phrase is important. If the assemblage has converged on $p_A = q_A = 1$, and then 2/3 of the reds agree to play B *once*, we just move from $(1, 1)$ to $(1, 1/3)$, and a glance at the figure shows that this would return to $(1, 1)$ if the replicator dynamics were left to their own devices. If we're considering one-time perturbations, they need to be big enough to move the assemblage into the domain of attraction of another equilibrium to matter. This will often be easier to achieve if many of the blues can be persuaded to cooperate with the red collective action. (Of course, blue allies help even if the red collective-actors *can* make themselves stick to playing B.)

[17] If this reminds you of the prisoners' dilemma, that's no accident!

[18] Even the little discussion above suggests some possible ways to get collective action to happen: (a) creating trust among the collective actors; (b) rewarding those who join the collective action; (c) punishing those who don't join, or who defect from it; (d) making agents *not care* about their immediate payoffs. Item (d) in turn can take many forms, ranging from material compensation for the lack of payoff (e.g., a fund to support striking workers, car-pooling during a bus boycott, etc.) to instilling the conviction that participating in the action has some invisible or future payoff ("for great is their reward in heaven"), or that participating is is a *duty* that must be fulfilled regardless of theconsequences.

[19] Moving from the (A,A) equilibrium to the (B,B) equilibrium would in fact be a strict Pareto improvement, making everyone better off.

[20] What are the critical values of p_A and q_A above which A is the best action for each population?

On the other hand, raising the (B,B) payoff to (5,5) does make the range of p_A and q_A values which favor the (A,A) equilibrium much smaller than before. (Again, see the figure.) Thus the all-A institution is *less* stable than before. It takes a smaller perturbation — less noise, or less extensive collective action — to move the assemblage to a state where it will converge on the (B,B) equilibrium. (Once more, see the figure.)

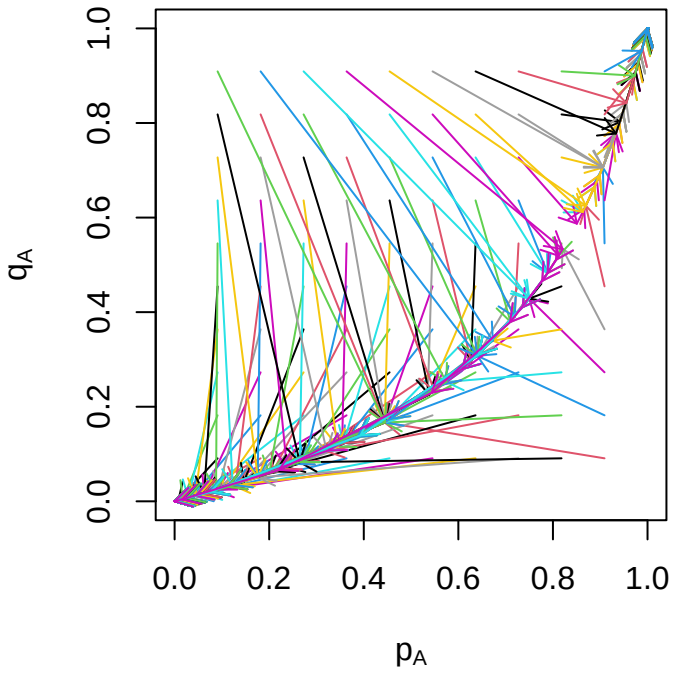


Figure 7: Behavior in the coordination game when the payoff from (B,B) is raised to (5,5), everything else being left unchanged. Notice that there are still trajectories converging to the (A,A) equilbrium where $p_A = q_A = 1$ in the upper-right corner.

4.7 Disadvantage payoffs and bargaining power

So far in these examples, we've made it so that the payoff if the two players fail to coordinate is equal, and is zero for both of them. What happens if we increase this "disagreement payoff" for one side but not the other? Each color would (selfishly) prefer that the other color agrees to play at its favored equilibrium. It can't always get that, because if *enough* agents of the other color play *their* favored move, coordinating with them is better than the disagreement payoff. But if we raise the disagreement payoff for blue but not red agents, we make playing A better for blues, no matter what the reds do. This will lead to more blues playing A. This in turn makes it better for reds to play A, etc. All of this suggests that increasing the disagreement payoff for one side but not the other should make it easier to reach an equilibrium that favors the advantaged side. This is generally correct.

For instance, suppose we make the disagreement payoff for blues $h > 0$, but leave everything else unchanged. A blue agent's expected payoff to playing A is now $4q_A + hq_B = h + (4-h)q_A$, while their expected payoff from playing B is $hq_A + 2q_B = 2 + (h-2)q_A$ (since $q_B = 1 - q_A$). Thus blues prefer to play A whenever

$$h + (4-h)q_A > 2 + (h-2)q_A \Rightarrow q_A > \frac{2-h}{6-2h}$$

Before, when $h = 0$, at least 1/3 of the red agents had to agree to play A for blue agents to favor it. Increasing the blue agents' disagreement payoff lowers this threshold. For instance if we set $h = 1.5$ the threshold value of q_A drops to 1/6 (see figure). The larger we make h, the less blues will care about the difference between their disagreement payoff and what they'd get at the (B,B) equilibrium. If we raised h all the way to 2, the q_A threshold would fall to zero — blues would always be at least as well off playing A as B. In *that* extreme, no matter what the initial values of q_A and p_A, p_A would always increase, heading towards 1, and so q_A would always eventually increase too[21].

So: one-sided improvement to the disagreement payoffs, say favoring the blues, increases the range of states of the assemblage which will eventually coordinate on the blues' preferred equilibrium. If the assemblage hasn't reached an equilibrium yet, this makes it more likely to reach the blues' preferred equilibrium. If the assemblage *has* reached that equilibrium, the equilibrium is more stable — it'd require a larger perturbation to upset it. If the assemblage has coordinated on some other equilibrium, a *smaller* perturbation is now needed to upset *that* equilibrium and tip us over into the domain of attraction of the blues' preferred equilibrium. In particular, fewer blues would need to go against their immediate interest to *force* a change in the institutions on the whole assemblage.

In this situation, when a blue agent plays A and meets a red playing A, the blue's payoff, 1.5, is almost as good as what it would get if it had played B and accepted an unfavorable agreement. On the other hand if the blue played A and met a red that played A, the blue would get 4. Because the disagreement payoff is so close to the unfavorable agreement payoff, and both are so much lower than the favorable agreement payoff, reds playing A must be rather uncommon to make it worthwhile for a blue to *not* also play A. (Imagine what would happen if blue agents were indifferent between an unfavorable agreement and disagreement.)

4.7.1 Disagreement payoffs and social advantage

All of this raises the question of which populations / social positions will have the more favorable disagreement payoffs. To *some* extent, this will vary depending on the exact situation we're interested in, but we can, nonetheless, make some generalizations. Basically: the group which is *already* better off will usually need *this* agreement less, and/or have other resources they can bring to bear to improve their disagreement payoff.

Thus an employer who has an on-going business and some savings does need to hire workers, but they don't need any particular worker all *that* much, compared to the intensity with which workers need jobs[22]. In many

[21] Specifically, because we haven't change the payoffs for red, q_A would start increasing once $p_A > 2/3$.

[22] Institutions like unemployment insurance can thus be seen as improving the prospective employee's disagreement payoff, improving their bargaining positon. Similarly, if all the employees are in a union, the employer *and the union* have much more similar disagreement payoffs than do the employer and an isolated worker. (Whether these institutions are ethically desirable, or achieve desirable aims at acceptable costs, are separate questions.)

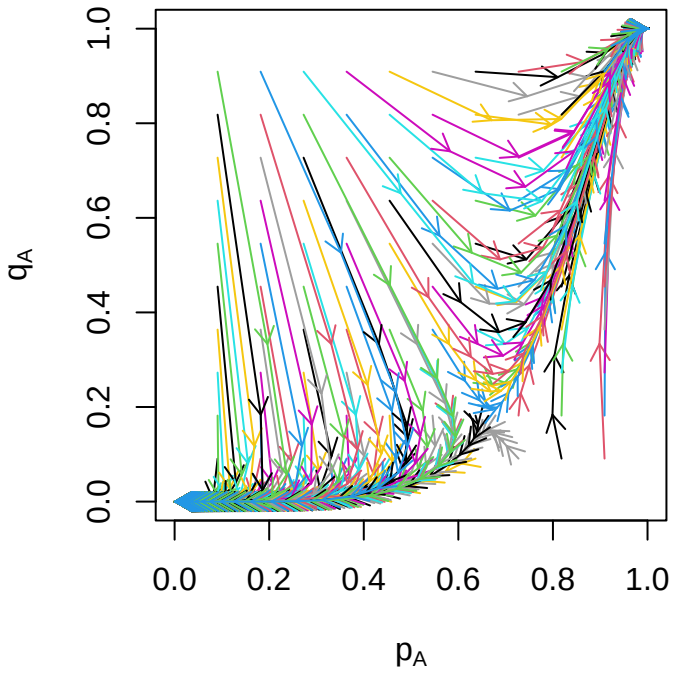


Figure 8: Replicator dynamics for the coordination game, with an improved disagreement payoff for blue agents (and only for blue agents).

times and places, the negotiations between landlords and tenant farmers have been more than a little shaped by the fact that the land-*lords* were highly trained and equipped professional providers of armed violence, and the tenants very much were not. Other examples are easy to multiply; see the further reading.

5 Some common themes

This lecture has covered a bunch of rather different ideas, but there are some themes in common across them.

0. **Generative modeling**: All of the large-scale patterns we see in these models — whether in terms of income distribution, or residential segregation, or the formation of institutions — are logical consequences of the assumptions of the model. They may not be *obvious* consequences, we may need the computer to simulate many steps of the model, but they do follow from those assumptions. (Otherwise, the simulation wouldn't produce them!) So we know that the assumptions we put in, about individual agents and their behavior and their interactions, lead to these conclusions[23]. If we don't want to see those conclusions come to pass, we need to make sure the world doesn't fit those assumptions.
1. **Self-organization**: If we *start* these models in random or dis-organized initial configurations, their dynamics, the "laws of motion" which they follow, tend to produce organized large-scale patterns, if not necessarily ones we like. The Schelling model produces segregated neighborhoods; social learning in the coordination game converges on an equilibrium institution; multiplicative growth leads to a heavy-tailed distribution. Again, these are *consequences* of the ways the individuals interact. They don't require any sort of hidden organizer, intelligent designer, or diabolical conspiracy.
2. **Randomness and path dependency**: *Which* institution the assemblage converges on is at least somewhat random; similarly the ultimate borders between neighborhoods in the Schelling model are also random. That there will be *some* institution, or some set of neighborhoods, is overwhelmingly probable and predictable, but exactly how that turns out is much less so. Random fluctuations early on can have large consequences, persistent consequences for shaping the social structure. This is an example of **path dependence** (Arthur 1994; Page 2006).
3. **Emergence and abstraction**: We can describe the behavior of the models entirely in terms of the individual agents, their actions, their interactions, etc. But it can be at least as illuminating to talk about things like the distribution of incomes across a population, or the institution the assemblage has converged on. This involves aggregating information across lots of agents — dealing with abstractions that ignore many details about individuals. (We say "95% of red agents play A", not "Alice plays A, Babur plays A, Chandra plays B, Dieter plays A..."). These variables aren't *properties* of any individual agent, but they are functions of the states of all the agents. We say that things like income distributions or institutions **emerge from** the agents' behavior, that they are **emergent**. The advantage of such abstraction is that it's a lot simpler, because it hides lots of messy detail. Sometimes we can describe the dynamics in these models entirely in terms of these higher-level, emergent abstractions, so we can talk about (say) the evolution of institutions, without ever having to talk about the behavior of individual agents, even though "the institutions" is nothing but a way of describing the collective behavior of the agents. (The disadvantage of abstractions is that sometimes those details turn out to matter a great deal.)
4. **Objectivity and alienation**: From the viewpoint of any individual agent, those self-organized patterns and emergent phenomena are just facts about its world. "The assemblage has converged on the (A,A) equilibrium" is just as much an objective fact, from Alice's point of view, as "Babur always plays A", or "red's payoff from (A,A) is half of red's payoff from (B,B)". (And similarly for neighborhoods in the Schelling model, or income distributions, etc.) It's true that if everyone acted differently, those facts could change, but they *are* objective facts (within the model). The agents have to deal with the massed consequences of everyone's behavior as though that was something outside them, something *other than* their own choice. The Latin for "other's" being *alienus*, some philosophers describe this as the agents being **alienated** from phenomena they themselves produce, or just **alienation**.
5. **Feedback and cycles**: We have seen a lot of feedback loops. Sometimes these feedbacks stabilize an pattern — small departures from the pattern create responses which tend to restore the pattern. (Think of the stable, equilibrium institutions in coordination games.) Other times, feedbacks *de*-stabilize a pattern, as when an integrated configuration in the Schelling model unravels. (Every red that moves out of a mostly-blue neighborhood makes that neighborhood even more blue, causing more reds to move out.) These are often distinguished as "negative" and "positive" feedback, respectively. Feedback

[23] Of course, other assumptions might *also* lead to those same large-scale patterns. (The mathematicians would say that the "inverse problem" here is "ill-posed".)

loops *can* be a bit difficult to reason about, and to model statistically in the ways we're used to. For instance, if we look at snapshots of the distributions of strategies in a lot of assemblages playing the same coordination game, what we'll see is a bunch of small, random fluctuations around the equilibria, and no particular pattern to which assemblage is near which equilibrium. If we have access to data *over time*, however, we can "unroll the cycles": $p_A(t)$ and $q_A(t)$ depend on $p_A(t-1)$ and $q_A(t-1)$ and so on back in to the past.

6 Further reading

On models of income and wealth leading to heavy tails, see Arnold (2015) for an exposition of many of the classic models (with common notation!), comparisons, references to the original and subsequent papers, etc.

That the outcome of individuals' social interactions is often very different from what any of them sought to achieve is the theme of Schelling's great book Schelling (1978). That book includes an extensive discussion of his segregation model, among many other fascinating topics. Of course the idea of unintended consequences is much older than Schelling (Boudon 1982), but he did bring unusual rigor to showing *how* it can happen.

The literature on institutions is immense, and literally centuries old[24]. Modern studies of institutions in economics and political science (and to a lesser extent, sociology) owes a lot to the works of Douglass North (e.g., North (1990)); Eggertsson (1990) is a good introduction to this literature if you are already familiar with microeconomics.

Game theory originates with Neumann and Morgenstern (1944); Poundstone (1992) is at once a biography of von Neumann, a conceptual introduction to game theory, and a horror story about the Cold War. Evolutionary game theory starts with the work of Maynard Smith and colleagues around the 1970s, with the deservedly-classic reference being Maynard Smith (1982). Sigmund (1996) is a very readable, but intellectually serious, popular account of evolutionary game theory. Much of evolutionary game theory has focused on the problem of the evolution of cooperation, using the prisoners' dilemma as a test-bed or prototype [Axelrod (1984); Sigmund (2010);]. Gintis (2000) is a good introduction to both classical and evolutionary game theory. Hofbauer and Sigmund (1988);Hofbauer and Sigmund (1998) are thorough treatments of the mathematics of the deterministic replicator dynamics and related systems. Sandholm (2010) is a mathematically advanced textbook aimed at economists, giving a lot of attention to the effects of noise on the dynamics, using the tools of "large deviations theory". This is an important branch of probability theory which has a lot to say not only about stochastic dynamics (Freidlin and Wentzell 1998), specifically how long it takes to leave the vicinity of a fixed point (Kautz 1987, 1988). Large deviations theory is also important for foundational issues in statistical inference (Bahadur 1971). Unfortunately I don't know of any truly elementary introduction aimed at statisticians, but see Hollander (2000).

The idea of combining models of learning with game theory to study the formation and development of institutions goes back to the 1990s; Young (1998) was an important milestone. That book, like Foster and Young (2003), used more elaborate models of individual learning, and less purely social learning like replicator dynamics, but derives very similar results to what we've gone over. Bowles (2004) is a broad-ranging synthesis of institutional economics and evolutionary game theory.

Knight (1992) was one of the first works to connect the new ideas about institutional evolution to social conflict, and specifically to conflict over *which* institution should be selected; many later results in the literature were first announced here, without the benefit (or burden) of advanced mathematics. Axtell, Epstein, and Young (2001) showed explicitly how evolutionary games could evolve unequal and inefficient institutions which could nevertheless persist for very long periods of time[25]. Similar results on the persistence of unequal institutions were later found by Bowles and Naidu (2008), who also studied the connection to the transmission of inequality. More recently, O'Connor (2019) applies similar ideas to inequality between the sexes and to gendered divisions of labor. All of these authors (among others) note how advantaging one group, such as by increasing its disagreement payoff, makes it easier for the whole assemblage to converge on an institution which benefits that group.

Allen, Farrell, and Shalizi (n.d.) tries to pull together ideas about institutions, evolutionary game theory, inequality, large deviations and social networks, if we ever finish writing it.

[24] Machiavelli's *Discoures on Livy*, ibn Khaldun's *Muqaddimah*, and Plato's *Republic* are all precious parts of our common cultural heritage.

[25] In the examples above, I *assumed* that there are red and blue agents, that reds and blues are paired up with each other, and that blue agents only learn from blues and reds only learn from reds. The more ambitious model of Axtell, Epstein, and Young (2001) actually shows how such social divisions (they say "classes") can emerge spontaneously, as agents come to condition their behavior on "tags", arbitrary characteristics assigned to agents which have no intrinsic significance (Holland 1995; Epstein and Axtell 1996).

7 Complementary exercises

Not for turning in.

0. Suppose that a population playing prisoners' dilemma consists of a proportion p of cooperators and $1-p$ defectors. Find the expected payoffs of playing C and D. Verify that playing D always has a higher expected payoff. What will happen to p under the replicator dynamic?
1. *Cheerful facts about the replicator dynamic* Assume we're only dealing with one population.
   a. Show that if the $p_i(t)$s are valid probabilities, so $p_i(t) \geq 0$ and $\sum_i p_i(t) = 1$, and if the fitnesses are all non-negative, $m_i(t) \geq 0$, then the $p_i(t+1)$s will also be valid probabilities.
   b. Show that if $p_i(t) = 0$ then $p_i(t+h) = 0$ for all $h > 0$.
   c. Show that if $p_i(t) = 1$ then $p_i(t+h) = 1$ for all $h > 0$.
   d. Show that if $m_i(t) > m_j(t)$ for all $j \neq i$, then $p_i(t+1) \geq p_i(t)$. If we also assume that $0 < p_i(t) < 1$, can we conclude that $p_i(t+1) > p_i(t)$?
   e. Show that if $m_i(t) = m_j(t)$ and $p_j(t) \neq 0$, then $\frac{p_i(t+1)}{p_j(t+1)} = \frac{p_i(t)}{p_j(t)}$.
   f. Explain why, at a fixed point, all the strategies with positive probability have equal fitness, i.e., either $p_i = 0$ or $m_i = m > 0$ for some common m.
2. The two-player, two-move coordination game used as a running example has *three* fixed points under the replicator dynamic. Two are the equilibria where $p_A = q_A = 0$ and $p_A = q_A = 1$, but there is also a third one. Find it and describe it in words. *Hint 1:* Figure 3. *Hint 2:* At a fixed point of the replicator dynamic, every action which has positive probability must have the *same* payoff. (See previous exercise.)
3. Take the running-example coordination game, and expand it to have *three* actions, as in the following pay-off matrix.
   a. Show that (A,A), (B,B) and (C,C) are all equilibria, and all fixed points under the replicator dynamics.
   b. Are there any other fixed points?
   c. Are the equilibria stable?

	Red makes move A	B	C
Blue makes move A	(4,1)	(0,0)	(0,0)
B	(0,0)	(2,2)	(0,0)
C	(0,0)	(0,0)	(1,4)

References

Allen, Danielle, Henry Farrell, and Cosma Rohilla Shalizi. n.d. "Evolutionary Theory and Endogenous Institutional Change." Manuscript in preparation.

Anderson, Philip W., Kenneth J. Arrow, and David Pines, eds. 1988. *The Economy as an Evolving Complex System*. Reading, Massachusetts: Addison-Wesley.

Arnold, Barry C. 2015. *Pareto Distributions*. Boca Raton, Florida: CRC Press.

Arthur, W. Brian. 1994. *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axtell, Robert L., Joshua M. Epstein, and H. Peyton Young. 2001. "The Emergence of Classes in a Multi-Agent Bargaining Model." In *Social Dynamics*, edited by Steven M. Durlauf and H. Peyton Young, 191–211. Cambridge, Massachusetts: MIT Press. https://doi.org/10.7551/mitpress/6294.003.0009.

Bahadur, R. R. 1971. *Some Limit Theorems in Statistics*. Philadelphia: SIAM Press.

Boudon, Raymond. 1982. *The Unintended Consequences of Social Action*. London: Macmillan.

Bowles, Samuel. 2004. *Microeconomics: Behavior, Institutions, and Evolution.* New York: Princeton University Press. https://doi.org/10.2307/j.ctvcm4gc3.

Bowles, Samuel, and Suresh Naidu. 2008. "Persistent Institutions." 08-04-015. Santa Fe Institute. http://www.santafe.edu/~bowles/PersistentInst.pdf.

Börgers, Tilman, and Rajiv Sarin. 1997. "Learning Through Reinforcement and Replicator Dynamics." *Journal of Economic Theory* 77:1–14. https://doi.org/10.1006/jeth.1997.2319.

DeLanda, Manuel. 2006. *A New Philosophy of Society: Assemblage Theory and Social Complexity.* London: Continuum.

Eggertsson, Thráinn. 1990. *Economic Behavior and Institutions.* Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511609404.

Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom up.* Cambridge, Massachusetts: MIT Press. https://doi.org/10.7551/mitpress/3374.001.0001.

Ford, Richard Thompson. 2008. *The Race Card: How Bluffing About Bias Makes Race Relations Worse.* New York: Farrar, Straus; Giroux.

Foster, Dean P., and H. Peyton Young. 2003. "Learning, Hypothesis Testing and Nash Equilibrium." *Games and Economic Behavior* 45:73–96. https://doi.org/10.1016/S0899-8256(03)00025-3.

Freidlin, M. I., and A. D. Wentzell. 1998. *Random Perturbations of Dynamical Systems.* Second. Berlin: Springer-Verlag.

Gellner, Ernest. 1983. *Nations and Nationalism.* Ithaca, New York: Cornell University Press.

Gintis, Herbert. 2000. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction.* Second. Princeton: Princeton University Press. https://doi.org/10.2307/j.ctvcm4gjh.

Hofbauer, Josef, and Karl Sigmund. 1988. *The Theory of Evolution and Dynamical Systems: Mathematical Aspects of Selection.* Cambridge, England: Cambridge University Press.

———. 1998. *Evolutionary Games and Population Dynamics.* Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9781139173179.

Holland, John H. 1995. *Hidden Order: How Adaptation Builds Complexity.* Reading, Massachusetts: Addison-Wesley.

Hollander, Frank den. 2000. *Large Deviations.* Providence, Rhode Island: American Mathematical Society.

Kautz, R. L. 1987. "Activation Energy for Thermally Induced Escape from a Basin of Attraction." *Physics Letters A* 125:315–19. https://doi.org/10.1016/0375-9601(87)90151-4.

———. 1988. "Thermally Induced Escape: The Principle of Minimum Available Noise Energy." *Physical Review A* 38:2066–80. https://doi.org/10.1103/PhysRevA.38.2066.

Knight, Jack. 1992. *Institutions and Social Conflict.* Cambridge, England: Cambridge University Press.

Maynard Smith, John. 1982. *Evolution and the Theory of Games.* Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511806292.

Montroll, Elliott W., and Michael F. Shlesinger. 1982. "On $1/f$ Noise and Other Distributions with Long Tails." *Proceedings of the National Academy of Sciences (USA)* 79:3380–3. https://doi.org/10.1073/pnas.79.10.3380.

Naidu, Suresh, and Noam Yuchtman. 2013. "Coercive Contract Enforcement: Law and the Labor Market in 19th Century Industrial Britain." *American Economic Review* 103:107–44. https://doi.org/10.1257/aer.103.1.107.

Neumann, John von, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princeton: Princeton University Press. https://www.jstor.org/stable/j.ctt1r2gkx.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511808678.

O'Connor, Cailin. 2019. *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198789970.001.0001.

Page, Scott E. 2006. "Path Dependence." *Quarterly Journal of Political Science* 1:87–115. https://doi.org/10.1561/100.00000006.

Poundstone, William. 1992. *Prisoner's Dilemma*. New York: Doubleday.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. 1943. "Behavior, Purpose and Teleology." *Philosophy of Science* 10:18–24. http://www.jstor.org/stable/184878.

Sandholm, William H. 2010. *Population Games and Evolutionary Dynamics*. Cambridge, Massachusetts: MIT Press.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology*. https://doi.org/10.1080/0022250X.1971.9989794.

———. 1978. *Micromotives and Macrobehavior*. New York: W. W. Norton.

Shalizi, Cosma Rohilla. 2009. "Dynamics of Bayesian Updating with Dependent Data and Misspecified Models." *Electronic Journal of Statistics* 3:1039–74. https://doi.org/10.1214/09-EJS485.

Sigmund, Karl. 1996. *Games of Life: Explorations in Ecology, Evolution and Behavior*. London: Penguin.

———. 2010. *The Calculus of Selfishness*. Princeton, New Jersey: Princeton University Press. https://www.jstor.org/stable/j.ctt7sw18.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.