Incomplete Information, Dynamic Stability and the Evolution of Preferences: Two Examples[*]

Jean Paul Rabanal[†]
Economics Department
UC Santa Cruz

Daniel Friedman
Economics Department
UC Santa Cruz

September 13, 2011

Abstract

We illustrate general techniques for assessing dynamic stability in games of incomplete information by re-analyzing two models of preference evolution, the Arce (2007) Principal-Agent game and the Friedman and Singh (2009) Noisy Trust game. The techniques include extensions of replicator and gradient dynamics, and for both models they confirm local stability of the key static equilibria. That is, we obtain convergence in time average for initial conditions sufficiently near equilibrium values.

Keywords: Stability, Perfect Bayesian Equilibrium, evolutionary dynamics.
JEL codes: C62, C73

[*] We thank Dann Arce and Bill Sandholm for their comments on an early draft of this paper.

[†] Corresponding Author. *Email:* jprabana@ucsc.edu *Address:* Engineering 2, 401 Santa Cruz, CA 95064 *Tel:* 1-831-459-5774 *Fax:* 1-831-459-5077

1 Introduction

Standard equilibrium concepts, such as Bayesian Nash Equilibrium and Sequential Equilibrium, offer sophisticated formulations of what one might hope to see in games of incomplete information. These equilibrium concepts, however, beg the crucial dynamical question: would human players ever actually reach such an equilibrium, or even get close?

For games of complete information, such questions of dynamic stability have been addressed in a principled way by evolutionary game theory. That theory shows that certain subsets of Nash equilibrium (e.g., Evolutionary Stable States, ESS) are indeed reachable by players using simple adaptation rules (e.g., replication or imitation or learning); see, for example Weibull (1995), Friedman (1991) and Sandholm (2010) for games in normal form and Cressman (2003) for games in extensive form. For games of incomplete information, however, dynamic stability questions remain largely unresolved.

The present paper begins to address those questions. It does not prove new general results nor offer new models, but it does show how to extend replicator dynamics and gradient dynamics from games of complete information to games of incomplete information. In some cases the extensions continue to yield systems of ordinary differential equations (ODEs) but in other cases they yield partial differential equations (PDEs), and the stability properties of these systems can be investigated analytically or numerically. We illustrate, using two recent models of preference formation due to Arce (2007) and Friedman and Singh (2009).

Both papers use the indirect evolutionary approach to model preference evolution. Players of given preference types are matched pairwise to play a game with known material payoffs. Players learn (on a rapid time scale) to maximize expected utility given their type, but evolve (on a slower time scale) according to realized material payoffs. Previous investigations represented evolution in terms of a static notion such as ESS; e.g. see Güth and Yaari (1992), Ok and Vega-Redondo (2001) and Dekel, Ely and Yilankaya (2007). Here we will instead use standard dynamic specifications — replicator and gradient — to assess the stability of key equilibria of both models.

The current paper is organized as follows. Section 2 presents a Principal-Agent game of incomplete information due to Arce (2007). A population of Agents containing two types (one self-interested and the other autonomy-preferring) is randomly matched pairwise with a population of Principals who can not observe Agents' type. Which preference types and what sort of behavior will survive in the long run? Arce uses static concepts to analyze the dynamic stability of various equilibria, and emphasizes the result that increasing the incentive wage can destabilize an efficient equilibrium for some distributions of preference types. He also finds some cases where the standard agency equilibrium can be destabilized.

Our analysis of the model begins by writing out the expected payoff and expected utilities given all state variables including the population share of each type of Agents, which Arce (2007) takes as exogenous. Then we introduce and analyze a system of four coupled ordinary differential equations (ODEs) that characterizes the evolution of the state variables. That system uses standard replicator dynamics, with fitness given by the realized material payoffs, to model the time path of the population shares. To model adjustment of strategy mixtures, the ODE system focuses on utility and applies gradient dynamics, which in this context looks much like standard replicator dynamics. The parameters include adjustment speed, and we focus on cases where strategy mixture adjustment is faster than the evolution of types.

For some equilibria, eigenvalue techniques allow us analytically to characterize dynamic stability. However, for some key equilibria, these and other tractable analytic techniques are inconclusive. We then rely on numerical solutions of the ODE system, and explain why it is appropriate in such cases to focus on time averages. Our results complement and extend those of Arce (2007), and in particular we find that both kinds of Agents can coexist for a broad set of parameter values.

Section 3 describes the basic trust game and its extension to a game of incomplete information due to Friedman and Singh (2009, henceforth FS09). They proposed a static equilibrium refinement called Evolutionary Perfect Bayesian Equilibrium (EPBE) for population games. At EPBE, each surviving type in each population has the same expected material payoff, and no potential entrant type has higher payoff.

After writing out the expected payoffs and utilities, we derive a system of six coupled differential equations that characterizes the evolution of the state variables. Five of these equations roughly parallel the ODEs for the Arce (2007) model, and the other equation uses gradient dynamics to describe evolution of the preference parameter. In line with FS09 and the previous section, we assume that individual learning enables strategy mixes to adjust more rapidly than population shares (which adjust via entry and exit, or type switching). We assume that preference parameters adjust (via genetic disposition and/or internalized norms) at an even slower rate. Our results support the implicit assumption in FS09 that EPBE is dynamically stable. More specifically, we show for reasonable parameters that the time average state converges to the relevant EPBE from initial conditions near to the equilibrium value. That is, the EPBE of the noisy trust games is locally asymptotically stable in time average.

The last section summarizes our findings and offers suggestions for applying the techniques more broadly. Appendix A (available on request) includes Matlab code for the simulations.

2 The Arce (2007) Principal-Agent Model

After reviewing the model and its known equilibria, we write out the state-contingent payoffs and specify dynamics. Then we assess the dynamic stability of all equilibria.

2.1 Elements of the model

Each Principal (row player) in a large population is randomly matched with one Agent (column player). There are two possible types of Agents: Type 1 or self-interested (comprising a fraction $\varphi \in [0,1]$ of the population) and Type 2 or autonomy-preferring (fraction $1-\varphi$). Either type of Agent decides only to whether to work (W) or to shirk (S); the respective mixture probabilities are denoted q_j and $1-q_j$ for each type $j = 1,2$.[1] Thus Agents' type-contingent strategy space is $[0,1] \times [0,1]$.

Each Principal decides whether to monitor (M) or not (N); the mixture probabilities are p and $1-p$ respectively. Thus Principal's strategy space is $[0,1]$, and the state of the system is a vector $S = (\varphi, p, q_1, q_2) \in [0,1]^4$.

Table 1 shows the payoffs. The Principal receives gross payoff $v > 0$ when the Agent works, offset by the wage $w > 0$ paid to the worker and by the monitoring cost $m > 0$. Agents who work incur an effort cost $e > 0$. In addition to these material payoffs, a Type 2 Agent receives a utility increment $(+\alpha)$ when her Principal does not monitor, and a symmetric decrement $(-\alpha)$ with monitoring. The parametric restrictions $w > e, m$ and $\alpha > e, w-e$ help avoid trivialities. Arce

[1] The *monomorphic* interpretation of a mixture probability q_j is that every type j player adopts exactly the same mixed strategy q_jW + $(1-q_j)$S. The *polymorphic* interpretation is that a fraction q_j of the type j players adopt the pure strategy W and the rest adopt the pure strategy S. The analysis below works for either interpretation, as well as for the more general interpretation that there is a distribution of pure and mixed strategies among the the type j players with overall mean q_j.

(2007) sets wage at the value $w = \sqrt{vm}$ that maximizes the Principal's expected payoff; this implies restrictions on the gross payoff v.

Table 1: **Principal (row) and Agent (column) Payoffs**. The fraction of Type 1 Agents is φ, and mixing probabilities are p for Principal and q_j for Agents of Type j.

	Type 1 (φ)		Type 2 ($1-\varphi$)	
	W (q_1)	S (1-q_1)	W (q_2)	S (1-q_2)
M (p)	$v-w-m \quad ,w-e$	$-m,0$	$v-w-m \quad ,w-e-\alpha$	$-m,0$
N (1-p)	$v-w \quad ,w-e$	$-w,w$	$v-w \quad ,w-e+\alpha$	$-w,w$

Source: Arce (2007)

Arce notes that if all Agents are known to be Type 1, then the unique Nash equilibrium is mixed: $p = p^* = e/w$; $q_1 = q_1^* = (w-m)/w$. He also notes that if all Agents are known to be Type 2, there is again a mixed NE in which $p = p^{**} = (\alpha - e)/(2\alpha - w)$, $q_2 = q_2^* = (w-m)/w$ as well as two pure NE: one at (N,W) or $p = 0$, $q_2 = 1$ and the other at (M,S) or $p = 1$, $q_2 = 0$.

2.2 Expected payoffs and utilities

Which equilibria, if any, are dynamically stable? Before introducing evolutionary dynamics to answer that question, we set the stage by writing out expected payoffs and utilities.

The Principal's expected payoff ω^P in equation (1) below arises from receiving v when the Agent works (probability $\varphi q_1 + (1-\varphi)q_2$), minus the monitoring costs incurred (with probably p) and the wages paid to the Agent. Recall from Table 1 that the Principal pays w unless he monitors and the Agent shirks, an event of probability $p(\varphi(1-q_1)+(1-\varphi)(1-q_2))$. Thus

$$\omega^P = (\varphi q_1 + (1-\varphi)q_2)\cdot v - p\cdot m - [1 - p(\varphi(1-q_1) - (1-\varphi)(1-q_2))]\cdot w \qquad (1)$$

Both types of Agent receive material payoff $w-e$ if they work (probability q_i), or w in the event that they do not work $(1-q_i)$ and the Principal does not monitor $(1-p)$. For the self-interested Agent (type 1), expected utility coincides with expected material payoff ω^{A1}, which is therefore

$$\omega^{A1} = q_1\cdot(w-e) + (1-q_1)(1-p)\cdot w. \qquad (2)$$

Similarly, material payoff for Type 2 Agent is

$$\omega^{A2} = q_2 \cdot (w-e) + (1-q_2)(1-p) \cdot w, \tag{3}$$

while her expected utility includes the preference parameter α and is

$$\omega_\alpha^{A2} = q_2 \cdot [p \cdot (w-e-\alpha) + (1-p) \cdot (w-e+\alpha)] + (1-q_2)(1-p) \cdot w. \tag{4}$$

2.3 Dynamic adjustment equations

Recall that the state space is four dimensional, and specifies the fraction of type 1 Agents (φ), Principal's mixing probability (p) and Agents' mixing probabilities (q_j). We therefore specify dynamics as a system of four coupled ordinary differential equations (ODEs), derived from expected payoffs using standard evolutionary principles.

Arce (2007, p. 718) comments, "This then begs the question, what determines the initial distribution of agent types (ρ)?" and cites several exogenous factors. For our purposes it is better to complete the model by endogenizing ρ, or φ in our notation. We invoke the basic principle of evolution that the type with higher material payoff (= fitness) will increase its share of the population. More specifically, we impose standard continuous time replicator dynamics (Taylor and Jonker, 1978; Hofbauer and Sigmund, 1988), which postulate that the growth rate $\dot{\varphi}/\varphi$ of the share of self-interested Agents is proportional (with rate constant β_1) to its payoff ω^{A1} relative to the population average ($\bar{\omega}$). Equation (5) and other equations below use the fact that relative payoff can be written $\omega^{A1} - \bar{\omega} = \omega^{A1} - \varphi\omega^{A1} - (1-\varphi)\omega^{A2} = (1-\varphi)(\omega^{A1} - \omega^{A2})$. Thus $\dot{\varphi}$ is equal to $\varphi(1-\varphi)(\omega^{A1} - \omega^{A2})$ times a positive adjustment speed parameter β_1.

The remaining equations use hybrid gradient-replicator dynamics for the mixture probabilities p, q_1 and q_2. Gradient dynamics are standard for evolution of continuos biological traits (e.g., Wright (1949), Lande (1976) and Kauffman (1993)), and for continuous strategy sets seem more common in economics (e.g., Sonnenschein (1982), Friedman and Ostrov (2010)) than alternative specifications. In these dynamics, the adjustment rate is proportional to its payoff gradient $\frac{\partial \omega}{\partial p}$. To shrink to the range $[0,1]$ of valid mixtures, we include a factor of the form $(1-p)p$, analogous to the binomial variance $(1-\varphi)\varphi$ that keeps φ in the interval $[0,1]$. Of course, the mixing probability q_2 responds to expected utility ω_α^{A2}, rather than to the material payoff that guides the evolution of preference types.

Thus the system of four coupled ODEs is

$$\begin{aligned}
\dot{\varphi} &= \beta_1\varphi(1-\varphi)[\omega^{A1}-\omega^{A2}] && (5)\\
&= \beta_1\varphi(1-\varphi)[(pw-e)(q_1-q_2)]\\
\dot{p} &= \beta_2 p(1-p)\frac{\partial\omega^P}{\partial p} && (6)\\
&= \beta_2 p(1-p)[-m+(\varphi(1-q_1)+(1-\varphi)(1-q_2))\cdot w]\\
\dot{q}_1 &= \beta_3 q_1(1-q_1)\frac{\partial\omega^{A1}}{\partial q_1} && (7)\\
&= \beta_3 q_1(1-q_1)(pw-e)\\
\dot{q}_2 &= \beta_4 q_2(1-q_2)\frac{\partial\omega^{A2}_{\alpha}}{\partial q_2} && (8)\\
&= \beta_4 q_2(1-q_2)[(w-2\alpha)p+\alpha-e]
\end{aligned}$$

where the parameters w, e, m, α, and β_i are exogenous.

We assume that individual learning enables the mixing probabilities to adjust more rapidly than the population fraction φ, which adjusts via entry and exit, or type switching. Thus $0 < \beta_1 < \beta_2 = \beta_3 = \beta_4$. The restrictions noted earlier apply to parameters w, e, m, α (or to v, e, m, α if w is chosen by the formula mentioned earlier). To complete the dynamic specification, take the initial state as given and impose the boundary conditions $0 \leq p \leq 1$, $0 \leq q_j \leq 1$ and $0 \leq \varphi \leq 1$.

It might seem at first that we are dealing with a system of partial differential equations, but the gradients on the right hand side can be expressed in terms of the state variables only, using equations (1-3). For example, using equation (2) in equation (7) we have $\frac{\partial\omega^{A1}}{\partial q_1} = \omega^{A1}|_{[q_1=1]} - \omega^{A1}|_{[q_1=0]} = pw-e$. Indeed, since the functions are linear in the relevant probability, these gradients all can be replaced by fitness difference terms, and the second lines of equations (6 - 8) therefore resemble standard replicator equations.

2.4 Dynamic behavior

Describing the dynamic behavior of a system of 4 ordinary differential equations depending on 8 exogenous parameters sounds like a complicated task. However, for our purposes it suffices to identify the dynamically stable equilibrium (DSE) points — the subset of rest points or steady states that are Lyapunov stable. That is, we seek steady states (states for which the right hand side of the ODE system is zero) such that a solution of the ODE system with initial condition sufficiently close to the steady state will remain close to the steady state forever. The idea is that only neighborhoods of DSE are likely to be empirically relevant; elsewhere behavior is transient and will be hard to identify in field data.

Two technical remarks are in order before proceeding. First, Lyapunov stability does not guarantee asymptotic stability, i.e., does not guarantee that the solution above actually converges to the DSE as $t \to \infty$. For the systems we are studying, asymptotic stability is too much to ask for, because in two or more dimensions Liouville's theorem precludes direct convergence of replicator dynamics to any interior equilibrium from an open neighborhood of initial conditions; see e.g., Fudenberg and Levine, 1998, p.95.

Second, it is well known that Nash equilibria (NE) are a subset of steady states (or dynamic equilibria, DE); see for example Weibull, 1995, Proposition 3.4. It is also well known that DSE are a subset of NE and that, for smooth systems of ODEs like (5 - 8), a necessary condition for DSE is that the Jacobian matrix evaluated at the NE has no eigenvalues with positive real part and a sufficient condition is that all eigenvales have negative real parts; see for example Hirsch and Smale, 1974, Chapter 9. Eigenvalues with zero real part suggest (but do not guarantee) Lyapunov stability, and suggest (again with no guarantee) failure of asymptotic stability.

We will therefore use the following algorithm to identify DSE:

- find all DE, separately checking corners, edges, faces and interior of the state space;
- identify the subset of DE that are NE, and eliminate the others;
- find the eigenvalues of the Jacobian matrix of (5 - 8) evaluated at each NE, and eliminate any NE which yields an eigenvalue with positive real part; and
- identify as locally stable (and therefore empirically relevant) any NE whose eigenvalues all have negative real parts, and use numerical methods to assess the dynamic stability of any remaining NE that yields an eigenvalue with zero real part.

To begin, recall that by definition a DE for the present model is a solution to

$$\dot{\varphi} = \dot{p} = \dot{q}_1 = \dot{q}_2 = 0. \tag{9}$$

To sort out the many solutions, recall that our state vector $S = (\varphi, p, q_1, q_2) \in [0,1]^4$ is a four dimensional hypercube. Each of the $2^4 = 16$ corners represents a pure strategy profile, and (by virtue of the binomial factors) is a solution to (9). One strategy is mixed along each of the $16 \cdot 4/2 = 32$ edges, two are mixed in each 2-d face ($4 \cdot (4 \cdot 3)/2 = 24$ of them), and three are mixed in each 3-d face ($4 \cdot 2 = 8$ of them), while interior points represent strictly mixed strategy profiles.

The first step in the algorithm, then, gives us 16 corner DE. Checking all edges and faces sounds tedious, but the special structure of the model allows shortcuts. When $\varphi = 0$ (or 1), the value of q_1 (or q_2) is irrelevant, so 8 of the edges and all 16 corners are subsumed in the DE subset $\{(1,0,0,\cdot),(1,0,1,\cdot),(1,1,0,\cdot),(1,1,1,\cdot);(0,0,\cdot,0),(0,1,\cdot,1),(0,1,\cdot,0),(0,0,\cdot,1)\}$. Recall from Section 2.1 that only the last two cases are pure NE in the restricted ($\varphi = 0,1$ face) games, so we can eliminate the other six cases as dynamically unstable. Indeed, the same argument allows us to eliminate all edge DE in either of these faces. So the only remaining edges are of the form (i) $\varphi \in (0,1)$ and (ii) $p, q_1, q_2 \in \{0,1\}$. From equations (9) and (5) we see that (i) entails either $p = w/e$ (which is inconsistent with (ii)) or $q_1 = q_2$, which is "pure pooling" by (ii). But $p = 1$ and $q_1 = 0$ are not mutual best responses, nor are $p = 0$ and $q_1 = 1$. Hence the only remaining edge DE are of the form $(\varphi, 1, 1, 1)$.

Appendix A collects arguments of a similar character that show the full set of NE is

$$\{(0,1,\cdot,0),(0,0,\cdot,1),(x,1,1,1),(1,p^*,\frac{w-m}{w},\cdot),(0,p^{**},\cdot,\frac{w-m}{w}),(\frac{w-m}{w},p^*,1,0)_{[\text{lw}]},$$
$$(\frac{m}{w},p^*,0,1)_{[\text{hw}]},(x,p^*,\frac{-m+wx}{wx},1),(x,p^*,\frac{-m+w}{wx},0)\}. \qquad (10)$$

Here x is in some parameter-dependent subset of [0,1] specified as needed below, while $p^* = e/w$, $p^{**} = \frac{\alpha - e}{2\alpha - w}$, and [hw] (or [lw]) in subscripts indicates that the state is a NE in the high wage region of parameter space $w - 2e > 0$ (or in the low wage region $w - 2e < 0$).

The next step in the algorithm is to write out the Jacobian matrix $((\frac{\partial \text{RHS eq. } i}{\partial \text{state var. } j}))$, evaluate at each NE, and compute the eigenvalues. The Jacobian of ODE system (5 - 8) is

$$J = \begin{pmatrix} \beta_1(1-2\varphi)(pw-e)(q_1-q_2) & \beta_1\varphi(1-\varphi)w(q_1-q_2) & \beta_1\varphi(1-\varphi)(pw-e) & -\beta_1\varphi(1-\varphi)(pw-e) \\ \beta_2 p(1-p)w(q_2-q_1) & \beta_2(1-2p)(w-m-q_2w+\varphi w(q_2-q_1)) & -\beta_2 w(p-p^2)\varphi & -\beta_2 w(p-p^2)(1-\varphi) \\ 0 & \beta_3 q_1(1-q_1)w & \beta_3(1-2q_1)(pw-e) & 0 \\ 0 & \beta_4 q_2(1-q_2)(w-2\alpha) & 0 & \beta_4(1-2q_2)[(w-2\alpha)p+\alpha-e] \end{pmatrix}.$$

As a warmup exercise, we compute the 2x2 Jacobian (sub)matrix for 2-d face where $\varphi = 0$

and $p, q_2 \in (0,1)$. At the pure NE $(0,1,\cdot,0)$ it is

$$J = \begin{pmatrix} -\beta_2(w-m) & 0 \\ 0 & -\beta_4(\alpha-(w-e)) \end{pmatrix}$$

and at the other pure NE $(0,0,\cdot,1)$ it is

$$J = \begin{pmatrix} -\beta_2 m & 0 \\ 0 & -\beta_4(\alpha-e) \end{pmatrix}$$

For these diagonal matrices, the diagonal entries are the eigenvalues, and the parametric restrictions guarantee that all of them are negative. Hence these equilibria are both stable “sinks" with respect to dynamics restricted to the face. To assess overall stability, we have to look at the full Jacobian matrix, and the Appendix shows that these include positive eigenvalues. Hence neither NE is a DSE.

The same Jacobian (sub)matrix evaluated at the mixed NE $(0, p^{**}, \cdot, q_2^*)$, where $q_2^* = \frac{w-m}{w}$, is

$$J = \begin{pmatrix} 0 & -\beta_2 p^{**}(1-p^{**})w \\ -\beta_4 q_2^*(1-q_2^*)(2\alpha-w) & 0 \end{pmatrix},$$

whose eigenvalues are real and have opposite signs, since the parametric restrictions imply that $2\alpha - w > 0$. Therefore the equilibrium is an unstable saddle point even with respect to dynamics restricted to the face, and thus is not a DSE.[2]

The systematic way to assess stability is to evaluate the 4x4 Jacobian matrix at each NE. To illustrate, take the last NE listed, $\left(x, \frac{e}{w}, \frac{w-m}{wx}, 0\right)$, where $x \in [\frac{w-m}{w}, 1]$. The Jacobian evaluated at such states is

$$J = \begin{pmatrix} 0 & \beta_1(w-m)(1-x) & 0 & 0 \\ \frac{\beta_2 e(w-m)(w-e)}{w^2 x} & 0 & -\beta_2 e\left(1-\frac{e}{w}\right)x & -\beta_2 e\left(1-\frac{e}{w}\right)(1-x) \\ 0 & \frac{\beta_3(w-m)(m+w(-1+x))}{wx^2} & 0 & 0 \\ 0 & 0 & 0 & \frac{\beta_4(w-2e)\alpha}{w} \end{pmatrix},$$

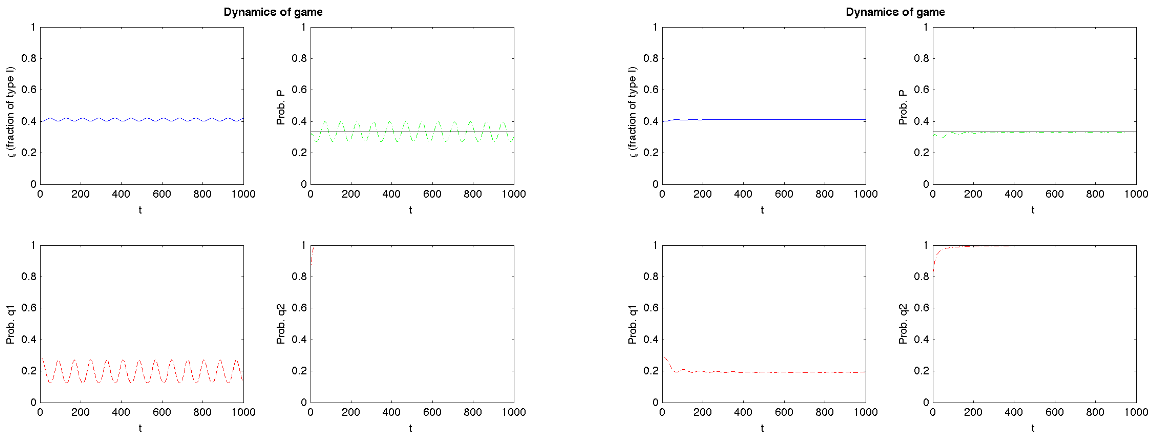
[2] More concretely, one can check that, for small $\varphi > 0$, type I Agent entrants will play a pure strategy that gives them a higher payoff than incumbents' payoff given that Principal plays p^*.

whose eigenvalues are $\left\{0, \frac{\beta_4(w-2e)\alpha}{w}, \pm\frac{\sqrt{-e(w-e)(w-m)^2(1-x)\beta_1\beta_2-[m+w(-1+x)]\beta_3}}{w\sqrt{x}}\right\}$. The second eigenvalue is negative in the low-wage region and positive in the high wage region, while the last pair of eigenvalues is pure imaginary since the expression in square brackets is positive for $x \geq \frac{w-m}{w}$. Hence this NE is dynamically unstable in the high wage region but remains a candidate DSE in the low wage region, where we will examine it numerically.

The Appendix examines the other NE using the same techniques. It rules out DSE status for the first five in the list, and confirms that there are no asymptotically stable NE. The only remaining candidate DSE are $\{\left(x, \frac{e}{w}, \frac{w-m}{wx}, 0\right) : x \in [\frac{w-m}{w}, 1]\}$ in the low wage case and $\{(x, \frac{e}{w}, \frac{wx-m}{wx}, 1) : x \in [\frac{m}{w}, 1]\}$ in the high wage case.

2.5 Simulation results

The last step in our algorithm is to investigate convergence behavior of the candidate DSE numerically. Since we do not expect asymptotic stability, we look for convergence in time average of the state variable $S = (\varphi, p, q_1, q_2)$. That is, we shall emphasize numerical approximations of $\lim_{t\to\infty} t^{-1}\int_0^t S(u)du$ more than of $\lim_{t\to\infty} S(t)$.[3]



(a) Levels

(b) Time average

Figure 1: Dynamics in the high wage $w > 2e$ case

[3] Stability in time average is also emphasized in the equilibrium concept Time Average of the Shapley Polygon (TASP) proposed by Benaim, Hofbauer and Hopkins (2006).

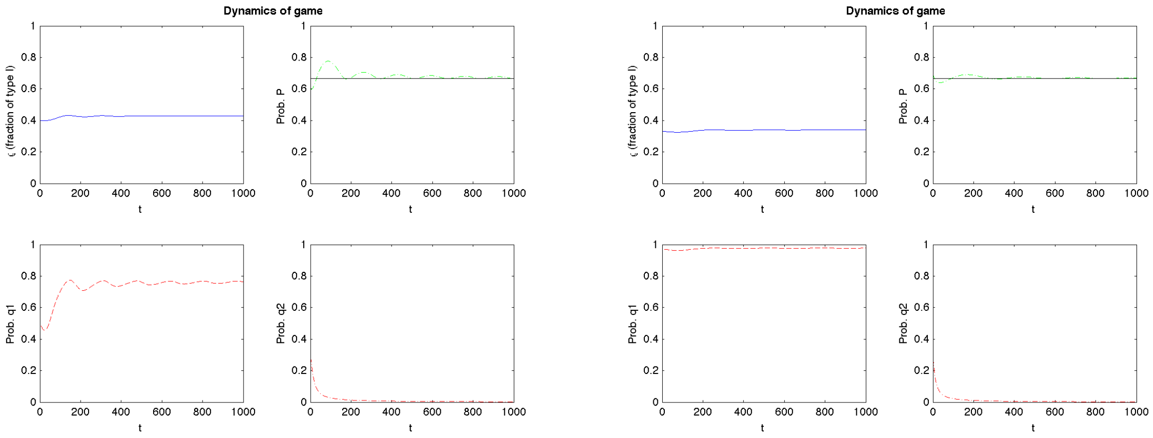
We solve the ODE system numerically using Matlab (codes are available upon request). To illustrate, set speeds of adjustment $\beta_1 = 0.1$ and $\beta_j = 1$ for $j = 2,3,4$ and use baseline parameters $\alpha = 0.5, m = 0.2, e = 0.2, v = w^2 m$, where the high wage is $w = 0.6 > 2e$ and the low wage is $w = 0.3 < 2e$.

For these baseline parameters, the candidate DSE is $(x, 1/3, 1 - \frac{1}{3x}, 1)$, $x \in [\frac{1}{3}, 1]$ for the high wage. Figure 1 shows our simulation results for the high incentive wage. We start from initial values $q_1(0) = p(0) = 0.3$, $q_2(0) = 0.8$ and $\varphi(0) = 0.4$. The left panel shows the simulation results in levels and the right panel in time average. Notice that in the left panel dynamics follow a cycle around the interior solution $\varphi^* = 0.41$ and $q_1^* = 0.19$ with direct convergence to $p^* = 1/3$ and $q_2^* = 1.0$. We achieve convergence in time average to that equilibrium in the right panel.

The dynamics are sensitive to the initial values. The farther the initial values from equilibrium, the wider the cycle around it. The time average convergence is not affected by choices of initial values for p^* and q_2^* sufficiently near to the equilibrium, meanwhile the equilibrium values of q_1^* and φ^* vary accordingly to the model predictions. For initial values far enough from the key equilibrium, we do not achieve convergence. For instance, starting with $q_2(0) < 0.1$, the level of monitoring approaches zero, the fraction of type I agents and the mixing probabilities go to 1. This result is not consistent with the target equilibrium.

The other relevant case is the low wage equilibrium, $(x, 2/3, \frac{1}{3x}, 0), x \in [1/3, 1]$ for baseline parameters. Notice that the level of monitoring is higher compared to the high wage equilibrium and type II Agent shrinks. Figure 2 shows the numerical results. The left panel starts with $\varphi(0) = 0.4$, $p(0) = 0.6$, $q_1(0) = 0.5$ and $q_2(0) = 0.3$. It shows convergence in time average to the equilibrium $(0.43, 0.67, 0.76, 0.00)$. The right panel starts with $\varphi(0) = 0.33$, $p(0) = 0.7$, $q_1(0) = 0.97$ and $q_2(0) = 0.25$. It shows convergence in time average to the equilibrium $(0.34, 0.67, 0.98, 0.00)$.

The dynamics behavior is similar to the high wage case. Once again, starting for initial values far enough from the key equilibrium, we do not achieve convergence. For instance, assuming initial values similar to the left panel but varying $p(0) < 0.2$, the dynamics are explosive and p, q_1 and q_2 go to the corners meanwhile the fraction of types stays in an interior point. This result is not consistent with the target equilibrium.



(a) Low q_1

(b) High q_1

Figure 2: Dynamics in the low wage $w < 2e$ case

We complement Arce's work by endogenously determining the fraction of types and studying the dynamics of the model. As different from his work, we show that both types of Agent coexist independently of the level of the incentive wage for a broad set of parameter values and initial conditions relatively closed to the equilibrium values. *Ceteris paribus*, the high incentive wages leads to lower rates of monitoring from the Principal and foster the working decision of autonomy-preferring Agents; meanwhile, the low incentive wages causes high monitoring levels and not working for the autonomy-preferring Agents. Our numerical solution shows that the dynamics when both Agents are present follow a cycle around the relevant interior equilibrium.

3 The Friedman and Singh (2009) Noisy Trust Game

Analyzing the next model introduces several additional considerations that can be important in games of incomplete information, such as positive tremble rates, evolving preference parameters, and higher dimensional state spaces. We rely more on numerical simulation but are nevertheless able to get a sharp result.

To begin, consider a simple two player game of complete information. The first mover, labelled Self (S), chooses whether to trust (T) or not trust (N). Choice N ends the game with zero payoffs to both players. Choice T gives the move to player Other (O), who can choose either to

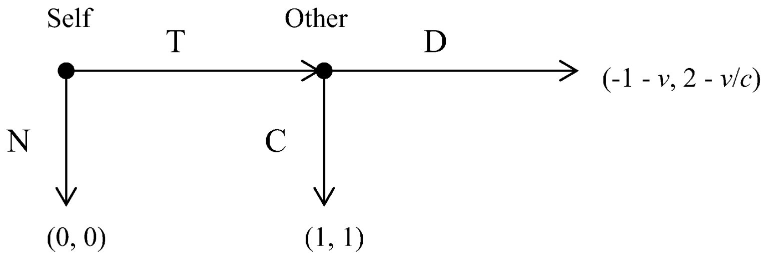


Figure 3: Simple trust game.
Unique subgame perfect NE is (N,D) when $v < c$ and is (T,C) when $v > c$.

to cooperate (C) or defect (D). Choice C gives both players unit payoffs, while choice D yields payoffs 2 to Other and -1 to Self. Following D, a vengeful behavioral type Self ($v = v_H > 0$) will take revenge and, at cost v to himself, will inflict harm v/c on Other, given an exogenous marginal cost parameter $c > 0$. The equilibrium payoffs are inefficient at (0,0) when $v = 0$, but are efficient at (1,1) when $v = v_H > c$.

3.1 Elements of the model

From this simple game, FS09 construct the noisy trust game illustrated in Figure 4. Nature chooses Self's non-vengeful type $v = 0$ with probability $1 - x$, or else chooses a given vengeful type $v = v_H > 0$ with probability x. Nature also independently chooses Other's perception as correct ($s = 0$ for $v = 0$, or $s = 1$ for $v = v_H$) with probability $1 - a$, or incorrect with probability a. The misperception probability is given by:

$$a = A(v_H) = 0.5\exp(-kv_H^2) \tag{11}$$

where $k > 0$ represents a precision parameter.

Let $p_0 = Pr[T|v = 0]$ denote the probability of trusting when S is non-vengeful, and $p_1 = Pr[T|v = v_H]$ the probability of trusting when S is vengeful. These probabilities are constrained by a tremble rate $e \in [0, 1/2)$, so that $e \leq p_0, p_1 \leq 1 - e$. Self's (mixed) strategy space thus is $[e, 1-e] \times [e, 1-e]$. Similarly, let $p_2 = Pr[C|s = 1]$ and $p_3 = Pr[C|s = 0]$ denote the probabilities of cooperating when Other observes a non-vengeful type and a vengeful type, respectively. Other's strategy space is also $[e, 1-e] \times [e, 1-e]$.

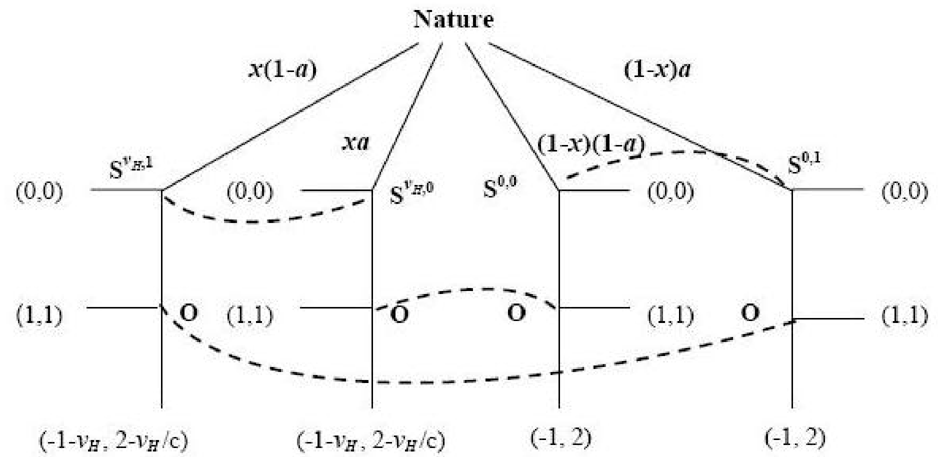


Figure 4: The noisy trust game. O denotes Other, S^{ij} denotes Self with vengeance level i and perception j, as determined by Nature's move. The four branch labels are Nature's move probabilities. Source: FS09

The state of the system is a vector $(v, x, p_0, p_1, p_2, p_3) \in [0, \hat{v}] \times [0,1] \times [e, 1-e]^2 \times [e, 1-e]^2$ that specifies Self's actions (p_0 and p_1), Other's actions (p_2 and p_3), the fraction of the vengeful type (x) and its degree of vengefulness ($v = v_H \leq \hat{v}$). This state space is more complicated than that of the Arce model in several respects. Besides the additional mixing variable, we have here a restricted mixture space (to account for trembles, which are conceptually important according to FS09), and an endogenous preference parameter.[4] Topologically, the state space is the 6-d hypercube $[0,1]^6$ with a specific parametrization.

The equilibrium concept here is perfect Bayesian equilibrium (PBE). Proposition 1 of FS09 identifies seven families of PBE that depend on game parameters x, a, v_H and e. The pure strategy PBE equilibria lie on the 2-d faces $p_i \in \{e, 1-e\}, i = 0,1,2,3$ of state space, and comprise families called "Separating " when $p_0 = p_3 = e$ and $p_1 = p_2 = 1-e$; "Bad Pooling" when $p_0 = p_1 = p_2 = p_3 = e$; and "Good Pooling " when $p_0 = p_1 = p_2 = 1-e$ and $p_3 = e$. The mixed strategy PBE families lie on higher dimensional faces and are called "Bad Mix" when $p_0 = p_3 = e$ and $p_1, p_2 \in (e, 1-e)$; "Bad Hybrid" when $p_0 = p_3 = e$, $p_1 = 1-e$ and $p_2 \in (e, 1-e)$; "Good Mix" when $p_1 = p_2 = 1-e$ and $p_0, p_3 \in (e, 1-e)$; and "Good Hybrid" when $p_0 = p_1 = p_2 = 1-e$ and $p_3 \in (e, 1-e)$.

[4] We could also have endogenized α in Arce's model, but that would not have been useful since material payoffs are flat in α except for a discontinuity at a particular threshold that changes type 2 agents' behavior. We will see that evolving v makes good sense in the FS09 model.

3.2 Expected payoffs and utilities

The expected payoffs w_s^v and w_s of vengeful and non-vengeful types of Self are:

$$w_s^v = (p_1(1-a)p_2 + p_1ap_3) * 1 + (p_1(1-a)(1-p_2) + p_1a(1-p_3)) * (-1-v) \tag{12}$$
$$w_s = (p_o(1-a)p_3 + p_oap_2) * 1 + (p_o(1-a)(1-p_3) + p_oa(1-p_2)) * (-1) \tag{13}$$

And the expected payoffs w_o^s and w_o for Other when he perceives a vengeful or a non-vengeful type are:

$$w_o^s = (x(1-a)p_1p_2 + (1-x)ap_op_2) * (1) + (x(1-a)p_1(1-p_2)) * (2-v/c) + ((1-x)ap_o(1-p_2)) * 2 \tag{14}$$
$$w_o = (xap_1p_3 + (1-x)(1-a)p_op_3) * (1) + (xap_1(1-p_3)) * (2-v/c) + ((1-x)(1-a)p_o(1-p_3)) * 2 \tag{15}$$

Equation (12) is derived as follows. If vengeful Self does not trust (probability $1-p_1$), she receives a zero payoff. On the other hand, if she trusts (probability p_1), she gets payoff 1 or $-1-v$ depending on Other's decision and perception. Her payoff is 1 when Other perceives correctly (probability $(1-a)$) a vengeful type and cooperates (probability p_2), and also when Other misperceives (probability a) and cooperates (probability p_3). She gets $-1-v$ when Other perceives correctly $(1-a)$ and defects $(1-p_2)$, and also when she misperceives (a) and defects $(1-p_3)$. Similar logic yields the expressions for non-vengeful Self's payoff w_s, and the expected payoffs w_o^s and w_o for Other when he perceives a vengeful or a non-vengeful type, respectively.

3.3 Dynamic adjustment equations

Recall that the state space is six dimensional, and specifies the fraction of vengeful type (x), the degree of vengefulness (v), and four mixing probabilities (p_i). We therefore specify dynamics as a system of six coupled ordinary differential equations (ODEs), derived from expected payoffs using standard evolutionary principles.

For the share x of vengeful types in the Self population, replicator dynamics postulate that the growth rate $\dot{x}/x$ is proportional (with rate constant β_x) to its own payoff w_s^v relative to the population average.

The remaining state variables involve a continuum of alternatives. Here we rely on gradient dynamics. Thus the degree of vengefulness $v = v_H$ changes at a rate proportional to its gradient $\frac{\partial w_s^v}{\partial v}$. As before, we use hybrid gradient-replicator dynamics for each mixing probability p_i. Its adjustment rate is proportional to its payoff gradient $\frac{\partial w_s^{[v]}}{\partial p_i}$. To shrink the range to $[e, 1-e]$, we include factors $(1-e-p_i)(p_i-e)$, analogous to the binomial factors $(1-x)x$ that keep x in the interval $[0,1]$. Thus our system of six ODEs is:

$$\dot{v} = \beta_v \left(\frac{\partial w_s^v}{\partial v} \right) \tag{16}$$

$$\dot{x} = \beta_x (1-x) x (w_s^v - w_s) \tag{17}$$

$$\dot{p}_o = \beta_o (1-e-p_o)(p_o-e) \left(\frac{\partial w_s}{\partial p_o} \right) \tag{18}$$

$$\dot{p}_1 = \beta_1 (1-e-p_1)(p_1-e) \left(\frac{\partial w_s^v}{\partial p_1} \right) \tag{19}$$

$$\dot{p}_2 = \beta_2 (1-e-p_3)(p_2-e) \left(\frac{\partial w_o^s}{\partial p_2} \right) \tag{20}$$

$$\dot{p}_3 = \beta_3 (1-e-p_3)(p_3-e) \left(\frac{\partial w_o}{\partial p_3} \right) \tag{21}$$

We assume that individual learning enables p_i to adjust more rapidly than does x (which adjusts via entry and exit, or type switching), and that v adjusts least rapidly (via genetic disposition and/or internalized norms). Thus $0 < \beta_v < \beta_x < \beta_0 = \beta_1 = \beta_2 = \beta_3$. To complete the dynamic specification, take the initial state as given and impose the boundary conditions $0 \leq x \leq 1, v \geq 0$ and $e \leq p_i \leq 1-e$.

3.4 Dynamic behavior

Which PBE remain when x and v_H can adjust? To answer, FS09 proposes a static refinement called evolutionary perfect Bayesian equilibrium (EPBE). In EPBE, all types in the support of the distribution in each population achieve equal and maximal expected fitness, and no potential entrant (a type outside the support) has higher expected payoff.

Proposition 2 of FS09 shows that only two states survive the EPBE refinement: (i) a unique

“Good Hybrid" EPBE in which Self trusts regardless of her type and Other plays a specific mixed strategy when she perceives a non-vengeful type, and (ii) the extreme “Bad Pooling" EPBE in which (apart from trembles) Self never trusts and Other always defects. The good EPBE has state vector

$$S = (x, v, p_o, p_1, p_2, p_3) = (x^*, v^*, 1-e, 1-e, 1-e, p_3^*),$$

where x^*, v^*, p^* are uniquely determined by the exogenous parameters, and the bad EPBE has state vector $(0, v, e, e, e, e)$, where v is arbitrary (and moot, since the vengeful type has population share zero).

Assuming the baseline parameters $k = 0.2$, $c = 0.8$, $e = 0.1$, $\beta_v = 0.01$, $\beta_x = 0.10$ and $\beta_{1,2,3} = 2$, the “Good Hybrid" is $(0.77, 2.78, 0.90, 0.90, 0.90, 0.68)$. Using standard numerical software, we can compute the eigenvalues of Jacobian matrix evaluated at that point. Of the 6 eigenvalues, 3 are negative, 1 is zero and 2 are pure imaginary. Thus we surmise that the good EPBE typically is neutrally stable. To investigate more carefully, we turn to numerical simulations.

3.5 Simulation results

Panel A figure 5 shows typical numerical solutions for baseline parameters and initial conditions not far from the EPBE. The state indeed cycles around the “good" EPBE with constant amplitude, consistent with Liouville’s theorem. Panel B confirms convergence in time average.

Numerical simulations indicate that the good EPBE is indeed locally stable for all parameters values within the set $c \in (0,1)$, $e \in (0, \hat{e}(k))$and $k \in (0, 0.6)$.[5] In other words, both types of Self survive in the long run and both trust maximally, while Others’ action depends on her perception. If Other perceives a vengeful type she will cooperate with maximal probability $1-e$ and if she perceives a non-vengeful type she will play a specific interior mixed strategy.

There are two caveats. First, we are talking about local stability, so we do not drastically alter the initial state. For baseline parameters we have confirmed convergence from initial states $v^* - 0.04 \leq v(0) \leq v^* + 0.01$, $x^* - 0.2 \leq x(0) \leq x^* + 0.15$, and $0.45 \leq p_i(0) \leq 1-e$. Second, we must restrict the adjustment speeds appropriately: $\beta_v < \beta_x$. This restriction is consistent with the idea from FS09 that slow cultural or genetic adjustment controls v, while exit and entry control x.

[5] See the Appendix “Proof of Proposition 2 and Comparative Statics" in FS09 for the definition of $\hat{e}(k)$

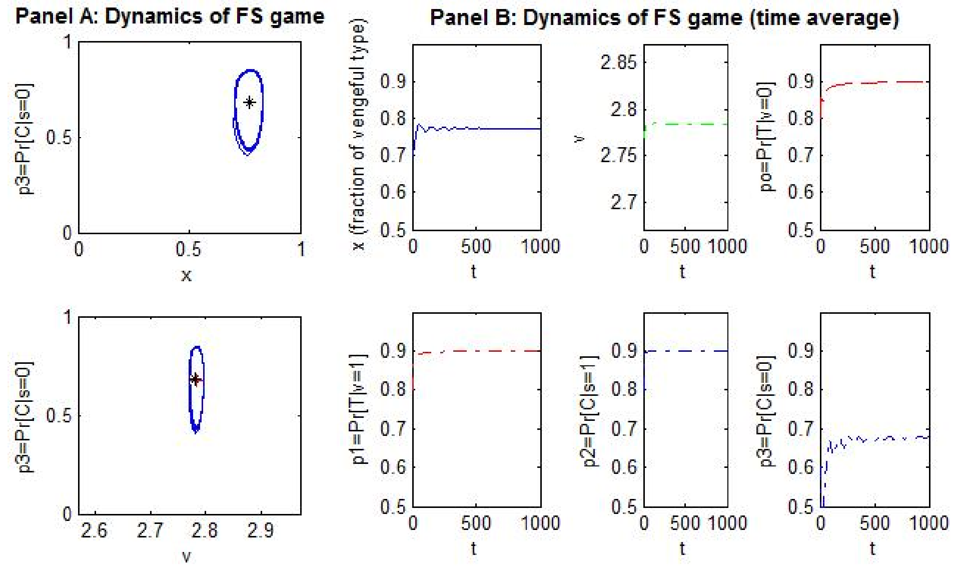


Figure 5: Dynamics (Panel A) and Time-Average dynamics (Panel B) of the FS09 game. In Panel A, the “Good” EPBE is indicated by * and the time average by $+$.

The “Bad" EPBE is at the corner of the state space, where the mixing probabilities are at the lower bound e and the fraction of vengeful type x goes to zero. Liouville’s theorem does not preclude direct convergence to a corner equilibrium. Indeed, we find direct convergence to the “Bad" EPBE given initial values of $x < 0.2$ (not many vengeful types) and $p_2 < 0.2$ (a low probability that Other cooperates).

4 Discussion

We have analyzed the dynamic stability of two games of incomplete information in the context of preferences evolution. We complement Arce (2007) results by endogenously determining the fraction of Agent types and studying the dynamics of the state variables. We show that both types of Agents coexist independently of the level of incentive wage. The second example is a noisy trust game due to FS09. In this game, we add dynamics to their static EPBE concept and the numerical results show convergence in time average to the key equilibrium (in which Self trusts regardless of her type and Other cooperates if she perceives a vengeful type and plays a specific mixed strategy if she observes a non-vengeful type).

Perhaps the main contribution of the present paper is to illustrate a toolbox for investigating the dynamic stability of equilibrium in a wide class of games of incomplete information. The toolbox first asks the researcher to write down the expected payoffs and expected utilities at all feasible states. Then it applies standard evolutionary concepts to describe the evolution of types (preference parameters in our examples), their population shares, and the action mixtures. It uses gradient dynamics for a continuous space of types, and uses hybrid gradient-replicator dynamics for the rest of the state vector, the population shares and mixture probabilities. This approach will yield systems of ordinary differential equations (ODEs) in applications like those just analyzed, but it can yield partial differential equations (PDEs) when a continuum of active types is possible.

In view of the fact that asymptotic stability can not be expected in key equilibria of games of incomplete information, the toolbox emphasizes convergence in time average and numerical methods. One can check robustness by sampling the economically feasible parameter space. Our toolbox also calls for appropriate restrictions on the adjustment speed parameters. For instance, in the FS09 game, slow cultural or genetic adjustment controls the type variable (the preference parameter v), while the exit and entry allow population shares to adjust at a moderate rate and individual learning allows very rapid adjustment of action mixtures.

We close with two more philosophical remarks. The toolbox presented here did not include some of the more advanced techniques from dynamical systems theory, such as center manifold techniques or bifurcation techniques. In our experience so far, the return on applied researchers' investment seems much higher for the simple numerical methods we emphasize.

Second, in specifying a game of incomplete information, the set of active types and the population shares of those types is exogenously given in most mainstream analyses. That seems to us to push off stage the most interesting part of the story. Hence our toolbox features methods for describing the evolution of these state variables and for characterizing their long-run behavior.

5 Appendix: Mathematical Details

5.1 Finding DE and NE in the Arce (2007) Model

Recall that section 2.4 already identified all corner and edge DE and the subset that are NE.

On the 2-d faces that lie inside the 3-d faces $\varphi \in \{0,1\}$, section 2.1 noted that the only additional NE are the mixes $(\varphi, p, q_1, q_2) = (1, \frac{e}{w}, \frac{w-m}{w}, \cdot)$ and $(0, \frac{\alpha - e}{2\alpha - w}, \cdot, \frac{w-m}{w})$. The remaining 2-d faces involve $\varphi \in (0,1)$ and a strict mix of only one of the state variables p, q_1, q_2. The last two cases entail one of the q_j pure and the other strictly mixed, but (5) then implies that p is strictly mixed, contradicting the definition of this 2-d face. The remaining 2-d possibility involves $\varphi, p \in (0,1)$, which by (6) implies that $\varphi^* = \frac{m/w + q_2 - 1}{q_2 - q_1}$. Ruling out[6] $q_2 - q_1 = 0$, we see from (5) that $p^* = e/w$. Consequently the only new candidate equilibria are $(\varphi^*, p^*, 1, 0)$ and $(\varphi^*, p^*, 0, 1)$. The dynamics of q_2 depends on the sign of $\frac{\alpha(w-2e)}{w}$ after plugging p^* in (8). The case $w - 2e > 0$ is called high incentive wages, and yields the $q_2^* = 1$ equilibrium, while low incentive wages, the case $w - 2e < 0$, yields the equilibrium above with $q_2^* = 0$.

We have already found all NE in the 3-d faces $\varphi \in \{0,1\}$. The 3-d faces $p \in \{0,1\}$ have no NE, since q_j is strictly mixing for states in such faces, and therefore $p = p^*$ by (8), contradicting $p \in \{0,1\}$. Similarly, the faces $q_1 \in \{0,1\}$ contain no new NE since a strictly mixed strategy for q_2 implies $p = p^{**} = (\alpha - e)/(2\alpha - w)$ which contradicts the solution of p^* in (5). On the faces $q_2 \in \{0,1\}$ we pick up two new NE, $(\varphi^*, \frac{e}{w}, \frac{-m + w\varphi^*}{w\varphi^*}, 1)$ and $(\varphi^*, \frac{e}{w}, \frac{-m+w}{w\varphi^*}, 0)$; the argument parallels that for the 2-d face where $\varphi, p \in (0,1)$. Keeping the third component $q_1 \in [0,1]$ implies the restriction $\varphi \in [\frac{m}{w}, 1]$ for the first new NE and $\varphi \in [\frac{w-m}{w}, 1]$ for the second.

Finally, the interior points are unstable since we already know that the dynamics of q_2 depends on the sign of $\frac{\alpha(w-2e)}{w}$ which forces it to 1 (or zero) in the case of high (or low) wage.

5.2 Evaluating the Jacobian matrix at the NE

The text analyzed stability for the first three NE and the last NE listed in (10). In this section, we find Jacobian matrices and eigenvalues for the remaining NE.

[6] It is hard to keep φ^* finite in that case, and by (7-8), it also entails non-generic parameters ($w = 2e$).

The Jacobian matrix for (5 - 8) evaluated at the equilibrium $(\varphi, p, q_1, q_2) = (0, 1, \cdot, 0)$ is

$$J = \begin{pmatrix} \beta_1 q_1 (w-e) & 0 & 0 & 0 \\ 0 & -\beta_2(w-m) & 0 & 0 \\ 0 & \beta_3(1-q_1)q_1 w & \beta_3(1-2q_1)(w-e) & 0 \\ 0 & 0 & 0 & -\beta_4(\alpha-(w-e)) \end{pmatrix},$$

whose eigenvalues are $\{-\beta_4(\alpha-(w-e)), -\beta_2(w-m), \beta_1 q_1(w-e), \beta_3(1-2q_1)(w-e)\}$. As noted in the text, the first two are always negative in our parameter space. The third is positive except when $q_1 = 0$, in which case the last eigenvalue is positive. Hence this NE is definitely not a DSE.

The Jacobian matrix evaluated at $(0, 0, \cdot, 1)$ is

$$J = \begin{pmatrix} \beta_1 e(1-q_1) & 0 & 0 & 0 \\ 0 & -\beta_2 m & 0 & 0 \\ 0 & \beta_3(1-q_1)qw & \beta_3 e(-1+2q_1) & 0 \\ 0 & 0 & 0 & -\beta_4(\alpha-e) \end{pmatrix},$$

whose eigenvalues are $\{-\beta_4(\alpha-e), -\beta_2 m, \beta_1 e(1-q_1), \beta_3 e(-1+2q_1)\}$. The third is positive except when $q_1 = 1$, in which case the last eigenvalue is positive. Hence this NE also is definitely not a DSE.

The Jacobian at $(x, 1, 1, 1)$ is

$$J = \begin{pmatrix} 0 & 0 & \beta_1(w-e)(1-\varphi)\varphi & \beta_1(w-e)(1-\varphi)\varphi \\ 0 & \beta_2 m & 0 & 0 \\ 0 & 0 & -\beta_3(w-e) & 0 \\ 0 & 0 & 0 & -\beta_4(\alpha-(w-e)) \end{pmatrix},$$

whose eigenvalues are $\{0, -\beta_4(\alpha-(w-e)), \beta_2 m, -\beta_3(w-e)\}$. The second and third are positive, so this equilibrium is not a DSE.

The Jacobian at $(1, e/w, (w-m)/w, \cdot)$ is

$$J = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{\beta_2 e(w-e)(m+(-1+q_2)w)}{w^2} & 0 & -\beta_2 e\left(1-\frac{e}{w}\right) & 0 \\ 0 & \frac{\beta_3 m(w-m)}{w} & 0 & 0 \\ 0 & \beta_4(1-q_2)q_2(w-2\alpha) & 0 & \frac{\beta_4(-1+2q_2)(2e-w)\alpha}{w} \end{pmatrix},$$

with eigenvalues $\left\{0, \pm\frac{\sqrt{-\beta_2\beta_3 em(w-m)(w-e)}}{w}, \frac{\beta_4(-1+2q_2)(2e-w)\alpha}{w}\right\}$. The last of these is positive in the low wage case when $q_2 > 0.5$ and in the high wage case when $q_2 < 0.5$; in either case it can be destabilized by an invasion of type 2 agents and so is not a DSE.[7]

The Jacobian at $(0, (\alpha-e)/(2\alpha-w), \cdot, (w-m)/w)$ is

$$J = \begin{pmatrix} -\frac{\beta_1(w-2e)(m-(1-q_1)w)\alpha}{w(w-2\alpha)} & 0 & 0 & 0 \\ \frac{-\beta_2(m-(1-q_1)w)(\alpha-(w-e))(\alpha-e)}{(w-2\alpha)^2} & 0 & 0 & \frac{-\beta_2 w(\alpha-(w-e))(\alpha-e)}{(w-2\alpha)^2} \\ 0 & \beta_3(1-q_1)q_1 w & \frac{-\beta_3(-1+2q)(w-2e)\alpha}{2\alpha-w} & 0 \\ 0 & -\frac{\beta_4 m(w-m)(2\alpha-w)}{w^2} & 0 & 0 \end{pmatrix},$$

with eigenvalues $\left\{\pm\sqrt{\frac{\beta_4\beta_2 m(w-m)(\alpha-e)(\alpha-(w-e))}{w(2\alpha-w)}}, \frac{\beta_3(1-2q_1)(w-2e)\alpha}{2\alpha-w}, \frac{-\beta_1(w-2e)(m-(1-q_1)w)\alpha}{w(2\alpha-w)}\right\}$. The first pair is real with opposite signs, so this NE is not a DSE.

The Jacobian at $(\frac{w-m}{w}, \frac{e}{w}, 1, 0)$ is

$$J = \begin{pmatrix} 0 & \frac{\beta_1 m(w-m)}{w} & 0 & 0 \\ -\beta_2 e\left(1-\frac{e}{w}\right) & 0 & \frac{-\beta_2 e(w-e)(w-m)}{w^2} & \frac{-\beta_2 em(w-e)}{w^2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\beta_4(w-2e)\alpha}{w} \end{pmatrix},$$

with eigenvalues $\left\{0, \frac{\beta_4(w-2e)\alpha}{w}, \pm\sqrt{\frac{-\beta_1\beta_2 em(w-m)(w-e)}{w}}\right\}$. The second is negative in the relevant case of low wages, $w-2e < 0$, and the last pair is pure imaginary. Hence this NE remains a candidate DSE, requiring further investigation. It can be seen to be an extreme case of the NE family listed last in (10) and already analyzed in the text.

[7] Indeed, performing simulations, q_2^* goes to 1 when $w-2e > 0$ and q_2^* goes to 0 when $w-2e < 0$ for initial values $q_2(0) \in (0,1)$, and $\varphi(0)$, $p(0)$ and $q_1(0)$ relatively closed to the key equilibrium. See section 2.5 for the simulation results of the equilibria that survive in the long run.

At $(\varphi^*, e/w, 0, 1)$, we have $\varphi^* = m/w$ and the Jacobian is

$$J = \begin{pmatrix} 0 & -\beta_1 m\left(1-\frac{m}{w}\right) & 0 & 0 \\ \beta_2 e\left(1-\frac{e}{w}\right) & 0 & \frac{-\beta_2 em(w-e)}{w^2} & \frac{\beta_2 e(w-e)(w-m)}{w^2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{-\beta_4(w-2e)\alpha}{w} \end{pmatrix},$$

with eigenvalues $\left\{0, \frac{-\beta_4(w-2e)\alpha}{w}, \pm\sqrt{\frac{-\beta_1\beta_2 em(w-m)(w-e)}{w}}\right\}$. The second is negative in the relevant case of high wages, $w-2e>0$, so this NE also remains a candidate DSE. It is an extreme case of the next NE family.

The Jacobian at $\left(\varphi^*, \frac{e}{w}, \frac{-m+w\varphi^*}{w\varphi^*}, 1\right)$ is

$$J = \begin{pmatrix} 0 & -\beta_1 m(1-\varphi^*) & 0 & 0 \\ \frac{\beta_2 em(w-e)}{w^2\varphi^*} & 0 & -\beta_2 e\left(1-\frac{e}{w}\right)\varphi^* & -\beta_2 e\left(1-\frac{e}{w}\right)(1-\varphi^*) \\ 0 & \frac{\beta_3 m(-m+w\varphi^*)}{w\varphi^{*2}} & 0 & 0 \\ 0 & 0 & 0 & \frac{-\beta_4(w-2e)\alpha}{w} \end{pmatrix},$$

with eigenvalues $\left\{0, \frac{-\beta_4(w-2e)\alpha}{w}, \pm\frac{\sqrt{\beta_2 em(w-e)(m(-1+\varphi^*)\beta_1+(m-w\varphi^*)\beta_3)}}{w\sqrt{\varphi^*}}\right\}$. The second is negative in the high wage case, and the last pair is pure imaginary since $\frac{-m+w\varphi^*}{w\varphi^*} \geq 0$, so the entire family with $\varphi^* \in [\frac{m}{w}, 1]$ is a candidate DSE in the high wage case.

References

[1] **Arce, D.**, 2007, "Is agency theory self-acting?," *Economic Inquiry*, 45 (4), pp. 708–720.

[2] **Benaim, M., J. Hofbauer and E. Hopkins**, 2009, "Learning in games with unstable equilibria," *Journal of Economic Theory*, 144(4), pp. 1694–1709.

[3] **Cressman, R.**, 2003, *Evolutionary dynamics and extensive form games*, MIT Press.

[4] **Dekel, E., J.C. Ely and O. Yilankaya**, 2007, "Evolution of Preferences," *Review of Economic Studies*, 74(3), pp. 685–704.

[5] **Friedman, D.** , 1991, “Evolutionary games in economics." *Econometrica*, 69, pp. 637-666.

[6] **Friedman, D. and D. Ostrov**, 2010, ”Gradient Dynamics in Population Games: Some Basic Results,” *Journal of Mathematical Economics*, 46(5), pp. 691–70 .

[7] **Friedman, D. and N. Singh**, 2009, “Equilibrium vengeance.” *Games and Economic Behavior*, 66, pp. 813-829.

[8] **Fudenberg, D. and D. Levine**, 1998, *The Theory of Learning in Games*, MIT Press.

[9] **Güth, W. and M. Yaari**, 1992, “An evolutionary approach to explaining reciprocal behavior.” In: Witt, U. (Ed.), *Explaining Process and Change — Approaches to Evolutionary Economics.* The University of Michigan Press, Ann Arbor.

[10] **Hofbauer, J. and K. Sigmund**, 1988, *The Theory of Evolution and Dynamical Systems*, Cambridge University Press.

[11] **Hirsch, M. and S. Smale**, 1974, *Differential Equations, dynamical system and linear algebra*, Academic Press.

[12] **Kaufman, S.**, 1993, *The Origins of Order: Self-Organization and Selection in Evolution*, NY: Oxford U Press.

[13] **Lande, R.**, 1976, Natural Selection and Random Genetic Drift in Pheontypic Evolution, Evolution, 30 no. 2, 314–334.

[14] **Ok, R. and F. Vega-Redondo**, 2001, “On the evolution of individualistic preferences: An incomplete information scenario", *Journal of Economic Theory*, 97, pp. 231–254.

[15] **Sonnenschein, H.**, 1982, “Price dynamics based on the adjustment of firms. ” *American Economic Review*, 72 (5), pp. 1088-1096.

[16] **Sandholm, W.**, 2010, *Population games and evolutionary dynamics*, MIT Press.

[17] **Taylor, P.D. and L.B. Jonker**, 1978, “Evolutionary stable strategies and game dynamics.” *Mathematical Biosciences*, 40, pp. 145-156.

[18] **Weibull, W.**, 1997, *Evolutionary game theory*, MIT Press.

[19] **Wright, S.**, 1949, “Adaption and Selection,” in L. Jepsen, G.G. Simpson, and E. Mayr eds., *Genetics, Paleontology, and Evolution.* Princeton, N.J.: Princeton University Press.