How Irrational Are Subjects in Extensive-Form Games?

(Joint with Drew Fudenberg)

Two views of equilibrium

(1) introspective
axiomatic versions of common knowledge
tracing procedure

(2) learning
common knowledge a conclusion, not an assumption

We ask: to what extent can an equilibrium model drawn from a learning foundation explain experimental data?

Two theoretical ideas:

- Self Confirming Equilibrium
- ε-equilibrium

Two views of experiments

(1) The stakes are too small to matter
(extreme view of ε-equilibrium)

(2) The results do not support the “theory”
(usually means some refinement of Nash equilibrium)

Many proponents of (2) use results to argue against rationality, at least in the narrow sense of maximizing monetary payoff

Self Confirming Equilibrium

$s_i \in S_i$ pure strategies for *i;* $\sigma_i \in \Sigma_i$ mixed
H_i information sets for *i*
$\overline{H}(\sigma)$ reached with positive probability under σ
$\pi_i \in \Pi_i$ behavior strategies
$\hat{\pi}(h_i | \sigma_i)$ map from mixed to behavior strategies
$\hat{\rho}(\pi)$, $\hat{\rho}(\sigma) \equiv \hat{\rho}(\hat{\pi}(\sigma))$ distribution over terminal nodes

μ_i a probability measure on Π_{-i}
$u_i(s_i | \mu_i)$ preferences

$$\Pi_{-i}(\sigma_{-i} | J) \equiv \{\pi_{-i} | \pi_i(h_i) = \hat{\pi}(h_i | \sigma_i), \forall h_i \in H_{-i} \cap J\}$$

Nash equilibrium

a mixed profile σ such that for each $s_i \in \operatorname{supp}(\sigma_i)$ there exist beliefs μ_i such that

- s_i maximizes $u_i(\cdot|\mu_i)$
- $\mu_i(\Pi_{-i}(\sigma_{-i}|H)) = 1$

Unitary Self-Confirming Equilibrium

- $\mu_i(\Pi_{-i}(\sigma_{-i}|\overline{H}(\sigma))) = 1$

(=Nash with two players)

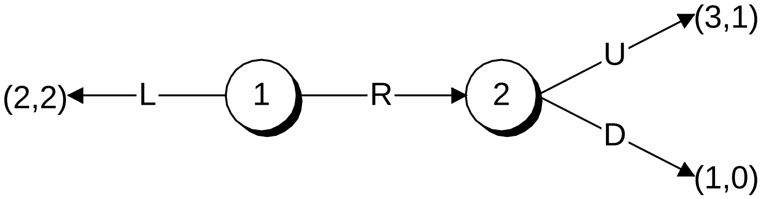
Heterogeneous Self-Confirming equilibrium

- $\mu_i(\Pi_{-i}(\sigma_{-i}|\overline{H}(s_i,\sigma))) = 1$

Can summarize by means of “observation function”

$$J(s_i,\sigma) = H, \overline{H}(\sigma), \overline{H}(s_i,\sigma)$$

Public Randomization



Remark: In games with perfect information, the set of heterogeneous self-confirming equilibrium payoffs (and the probability distributions over outcomes) are convex to go beyond self-confirming in general requires experimentation

might expect self-confirming in the medium run (Roth-Erev simulations; McKelvey-Palfrey estimation), and if enough experimentation Nash in the long-run

another paper “Self-confirming Equilibrium” explores in detail the connection between self-confirming, correlated and Nash

Approximate Equilibrium

- exact: $u_i(s_i|\mu_i) \geq u_i(s_i'|\mu_i)$
  approximate: $u_i(s_i|\mu_i) + \varepsilon \geq u_i(s_i'|\mu_i)$

- Approximate equilibrium can be very different from exact equilibrium

Radner’s work on finite repeated PD
gang of four on reputation

A small portion of the population playing "non-optimally" may significantly change the incentives for other players causing a large shift in equilibrium behavior.

How big is big?

- we propose to measure how big is, that is to measure the minimal value of ε consistent with players' play

- given the observed distribution over terminal nodes we will "attribute" a loss to each terminal node and report the distribution of losses

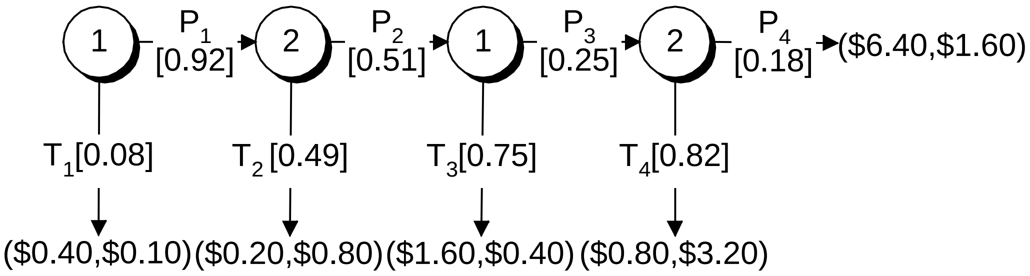
- somewhat involved procedure in general due to the fact that in extensive form games we do not directly observe players' strategies

- while the distribution we report has some arbitrary accounting conventions, such as attributing as much of the loss as possible to the final moves of the game, the mean loss is uniquely defined and independent of the particular accounting convention

distribution over outcome is ρ
loss attributed to z is $\varepsilon_i(z, J(\cdot), \rho)$
mean $\bar{\varepsilon}_i(J(\cdot), \rho)$

$J(\cdot)$ observation function for unitary or heterogeneous

Sample Calculation from Centipede Game



$m_i(a_i)$ “worst subsequent payoff”

for player 1

	P	T
1	$0.20	$0.40
3	$0.80	$1.60

probability distribution over payoffs p_i^y
y where y is a subgame $0, P_1, P_2, P_3$

for player 1
at P_3 (.18 \$6.40, .82 \$0.80)

at $y = P_2$ for $a_i = T_3, P_3$
$$\varepsilon(a_i, \rho) \equiv \max\{0, \max_{a'_i \in g(y)} m_i(a'_i) - \sum_{y'} \sum_u u p_i^{y'}(u) \pi(y' | a_i)\}$$

$$\max_{a'_i \in g(y)} m_i(a'_i) = \$1.60$$
$$\sum_{y'} \sum_u u p_i^{y'}(u) \pi(y' | T_3) = \$1.60$$
$$\sum_{y'} \sum_u u p_i^{y'}(u) \pi(y' | P_3) = .18 \cdot \$6.40 + .82 \cdot \$0.80 = \$1.808$$

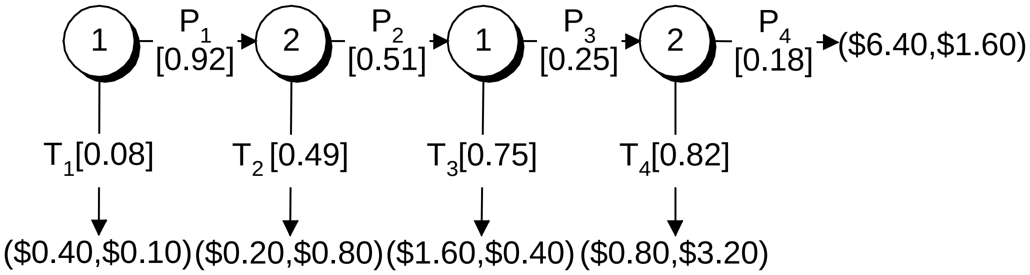
$$\varepsilon(T_3, \rho) = 0, \varepsilon(P_3, \rho) = 0$$

since $\varepsilon = 0$, we assign the actual probabilities to the actual payoffs
$$p_1^{P_3} = (0.75\,\$1,60, 0.25(.18\,\$6.40, .82\,\$0.80))$$

to understand algorithm, if $\varepsilon > 0$ for an action, then the probability of that action is assigned m_i (player knows he could get at least this much)

add up over actions to get terminal node losses

Centipede Game: Palfrey and McKelvey



Numbers in square brackets correspond to the observed conditional probabilities of play corresponding to rounds 6-10, stakes 1x below.

This game has a unique self-confirming equilibrium; in it player 1 with probability 1 plays T_1

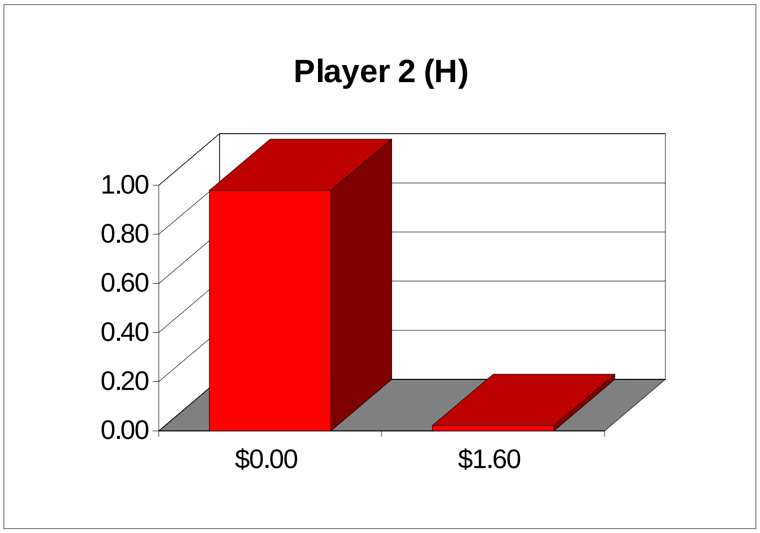
Trials /	Rnds	Stake	Case	Expected Loss			Max	Ratio
Rnd				Pl 1	Pl 2	Both	Gain	
29*	6-10	1x	H	$0.00	$0.03	$0.02	$4.00	0.4%
29*	6-10	1x	U	$0.26	$0.17	$0.22	$4.00	5.4%
	WC	1x	H			$0.80	$4.00	20.0%
29	1-10	1x	H	$0.00	$0.08	$0.04	$4.00	1.0%
10	1-10	4x	H	$0.00	$0.28	$0.14	$16.00	0.9%

Rnds=Rounds, WC=Worst Case, H=Heterogeneous, U=Unitary

*The data on which from which this case is computed is reported above.

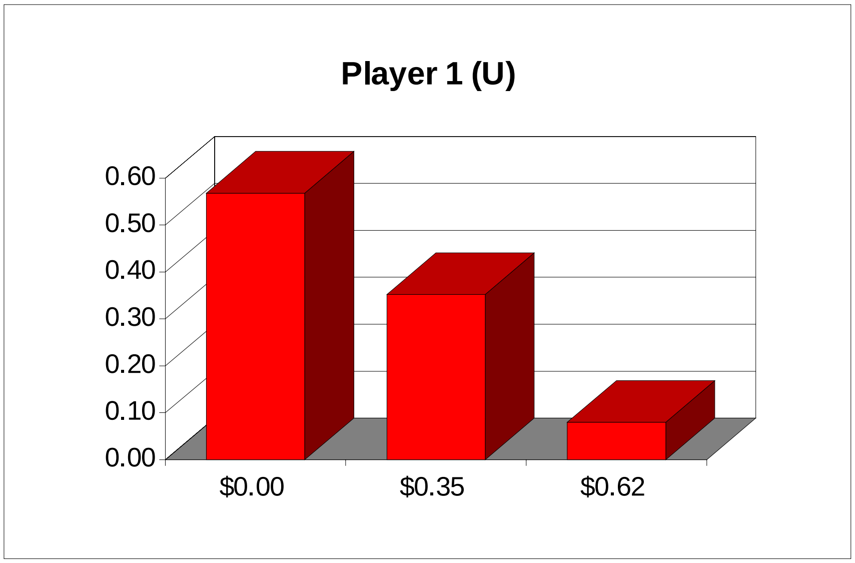
- heterogeneous loss per player is small; because payoffs are doubling in each stage, equilibrium is very sensitive to a small number of player 2’s giving money away at the end of the game.
- unknowing losses far greater than knowing losses
- quadrupling the stakes very nearly causes $\bar{\varepsilon}$ to quadruple
- theory has substantial predictive power: see WC
- losses conditional on reaching the final stage are quite large--inconsistent with subgame perfection. McKelvey and Palfrey estimated an incomplete information model where some “types” of player 2 liked to pass in the final stage. This cannot explain many players dropping out early so their estimated model fits poorly.

Heterogeneous Losses

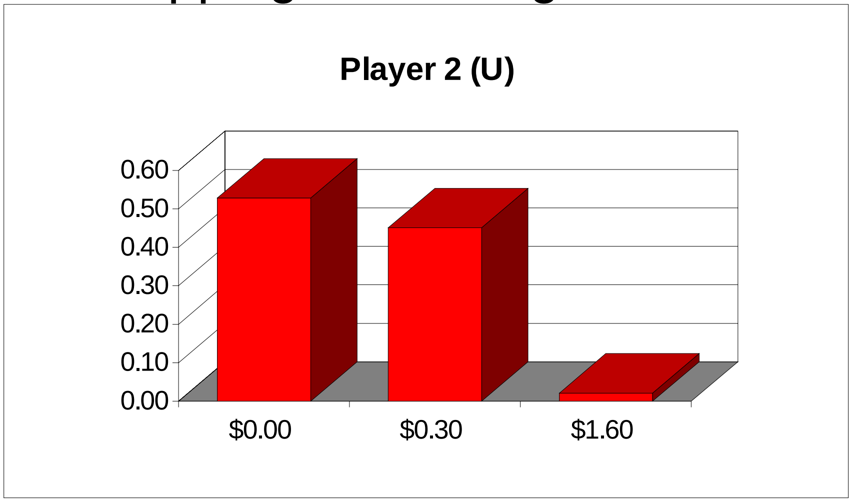


(No player 1 heterogeneous losses)

Unitary Losses



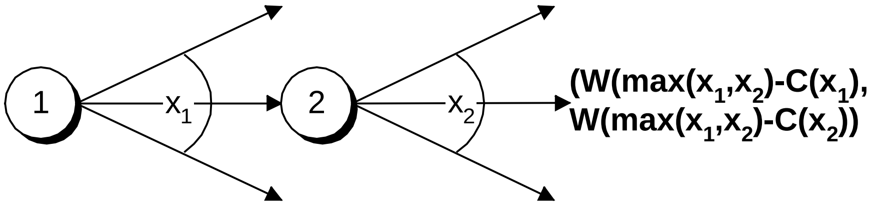
$0.35 for dropping out in stage 3
$0.62 for dropping out in stage 1.



$0.30 for dropping out in stage 2: expected loss of $0.14

$1.60 for giving away money at end: expected loss of $0.03

Best Shot Game: Prasnikar and Roth



x	*W(x)*	*C(x)*
0	$0.00	$0.00
1	$1.00	$0.82
2	$1.95	$1.64
3	$2.85	$2.46
4	$3.70	$3.28
5	$4.50	$4.10
6	$5.25	$4.92
7	$5.95	$5.74
8	$6.60	$6.50

if the other player makes any contribution at all, it is optimal to contribute nothing

unique subgame perfect equilibrium player 1 contributes nothing

another Nash equilibrium player 2 to contributes nothing regardless of player 1’s play

it is not consistent with Nash equilibrium for some player 1’s to play 0 and others 4

any other probability distribution over the two Nash equilibria are heterogeneous self-confirming

Trials	Rnds	Info	Case	Expected Loss			Max Gain	Ratio
				Pl 1	Pl 2	Both		
8	8-10	full	H	$0.00	$0.12	$0.06	$2.06	2.9%
8	8-10	full	U	$0.00	$0.12	$0.06	$2.06	2.9%
10	8-10	part	H	$0.01	$0.15	$0.08	$2.06	3.9%
10	8-10	part	U	$0.39	$0.15	$0.27	$2.06	13.%
	WC		H			$3.41	$2.06	165%

Rnds=Rounds, WC=Worst Case, H=Heterogeneous, U=Unitary

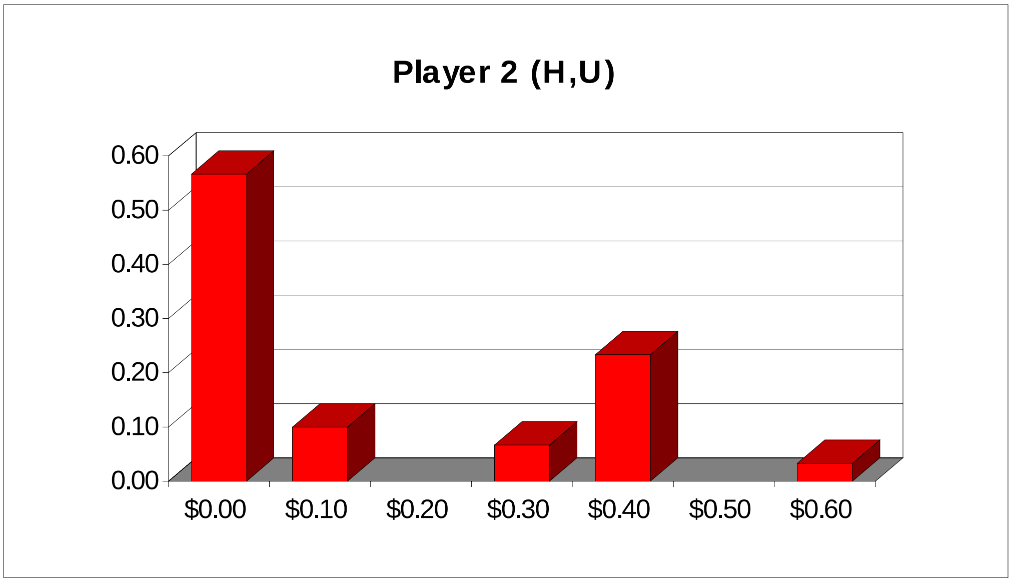
- In the full information case and partial information heterogeneous case player 2 occasionally contributes less than 4 when player 1 has contributed nothing; Note that the player who contributes nothing gets $3.70 against $0.42 for the opponent who contributes 4

- larger losses than centipede game with lower stakes

- full information case heterogeneous losses equal unitary losses-- player 1 never contributed anything, and so never had a loss with either type of information; all losses by player 2 are necessarily knowing losses

- In the partial information case occasionally player 1 contributed 4 and player 2 contributed nothing: looks like public randomization between the two Nash equilibria. This is inconsistent with Nash equilibrium but consistent with self-confirming equilibrium.

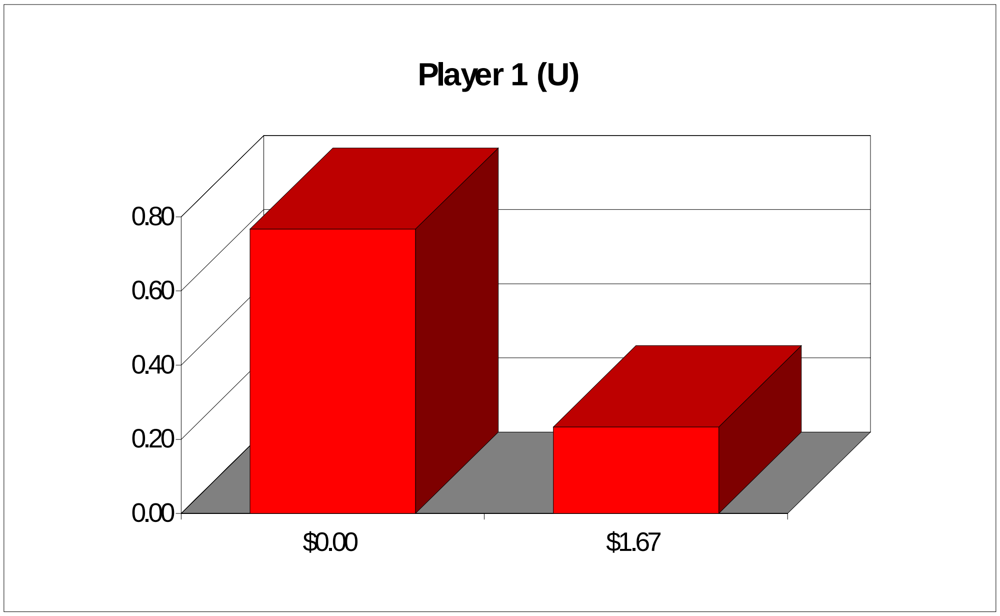
Partial Information Loss Distribution Player 2



losses correspond almost entirely to under contributing when player 1 has failed to contribute

(in one case a player 2 wasted money by contributing when player has already contributed--it is hard to find much of a rationale for this, since neither player benefited by 2’s action)

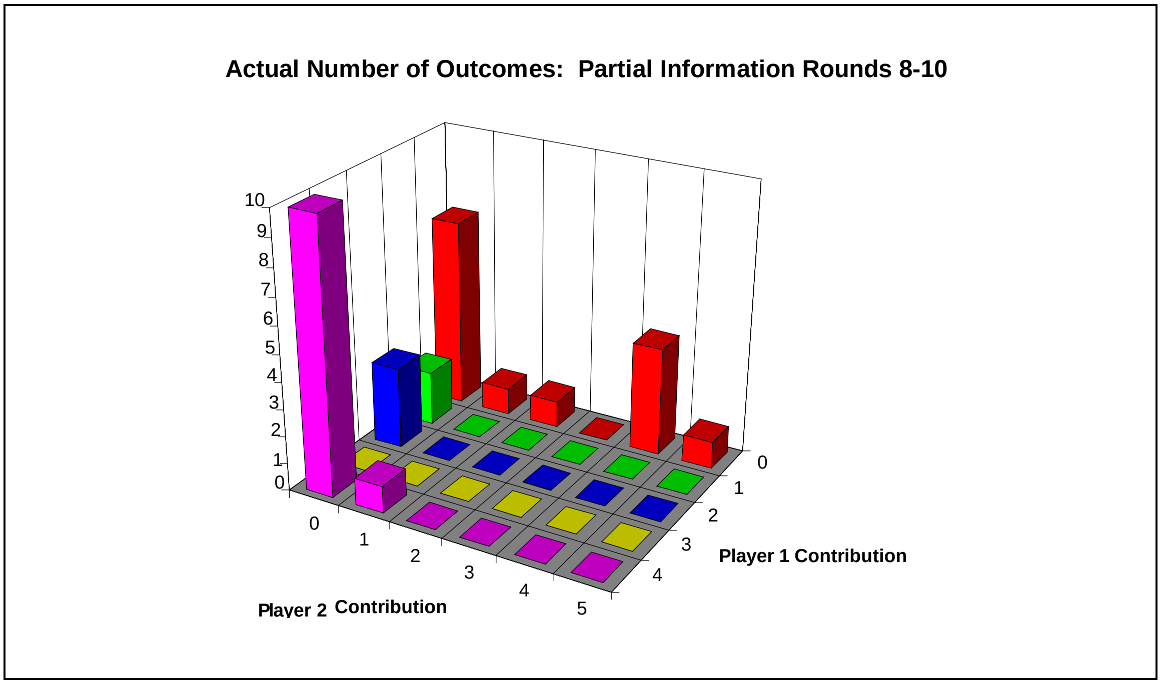
Player 1



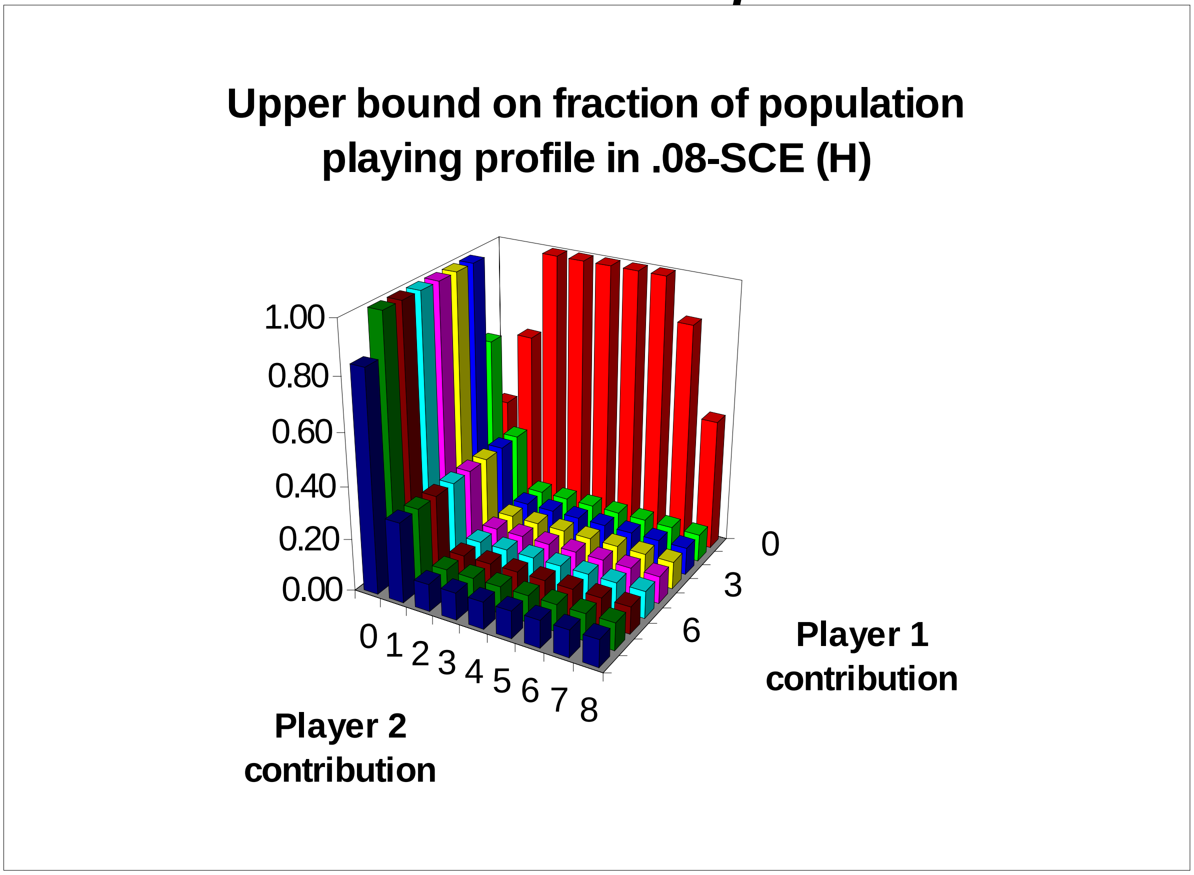
(in the heterogeneous case there was only one game observed in which player 1 failed to play optimally given his information)

unitary losses are from contributing 4, when in fact it is optimal to contribute nothing

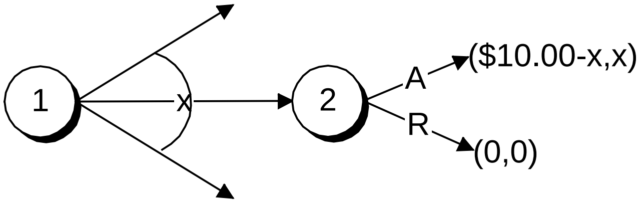
Actual Data



Theoretical Computation



Ultimatum Game:

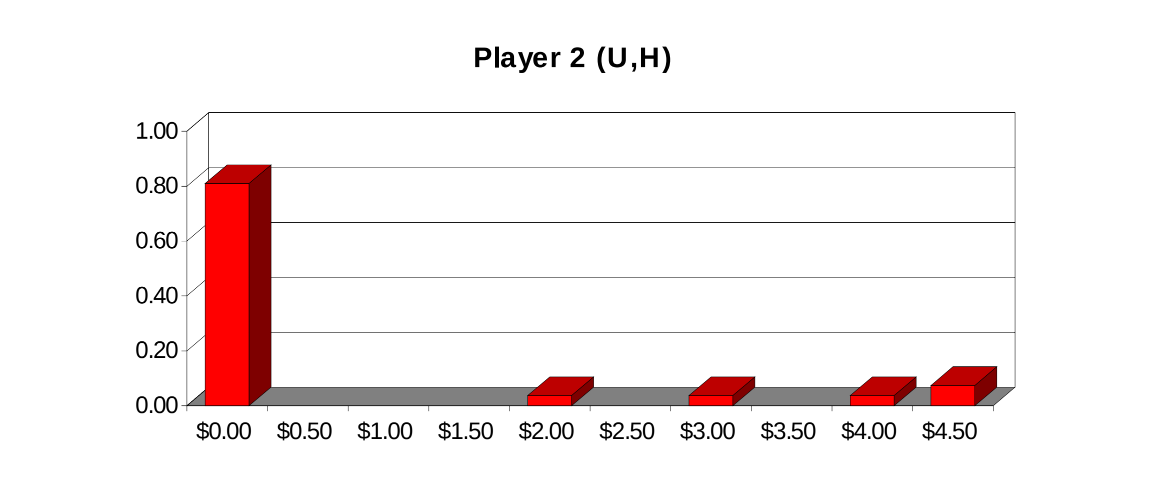


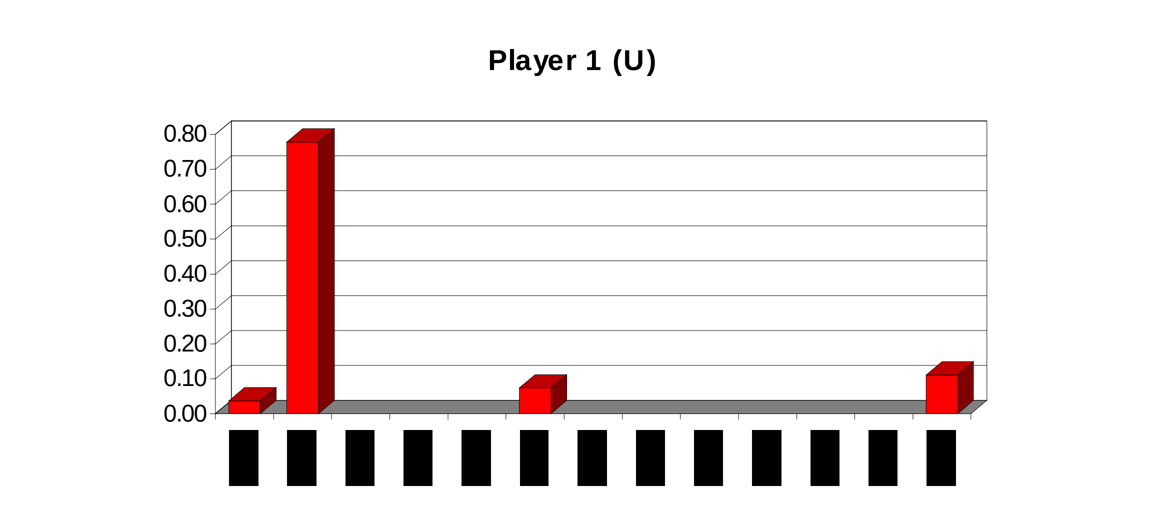
Trials	Rnd	Cntry Stake	Case	Expected Loss			Max Gain	Ratio
				Pl 1	Pl 2	Both		
27	10	US	H	$0.00	$0.67	$0.34	$10.00	3.4%
27	10	US	U	$1.30	$0.67	$0.99	$10.00	9.9%
10	10	USx3	H	$0.00	$1.28	$0.64	$30.00	2.1%
10	10	USx3	U	$6.45	$1.28	$3.86	$30.00	12.9%
30	10	Yugo	H	$0.00	$0.99	$0.50	$10?	5.0%
30	10	Yugo	U	$1.57	$0.99	$1.28	$10?	12.8%
29	10	Jpn	H	$0.00	$0.53	$0.27	$10?	2.7%
29	10	Jpn	U	$1.85	$0.53	$1.19	$10?	11.9%
30	10	Isrl	H	$0.00	$0.38	$0.19	$10?	1.9%
30	10	Isrl	U	$3.16	$0.38	$1.77	$10?	17.7%
	WC		H			$5.00	$10.00	50.0%

Rnds=Rounds, WC=Worst Case, H=Heterogeneous, U=Unitary

- every offer by player 1 is a best response to beliefs that all other offers will be rejected so player 1’s heterogeneous losses are always zero.
- big player 1 losses in the unitary c
- player 2 losses all knowing losses from rejected offers; magnitudes indicate that subgame perfection does quite badly
- as in centipede, tripling the stakes increases the size of losses a bit less than proportionally (losses roughly double).

US Distributions





Raw US Data

x	*Offers*	*Rejection Probability*
$2.00	1	100%
$3.25	2	50%
$4.00	7	14%
$4.25	1	0%
$4.50	2	100%
$4.75	1	0%
$5.00	13	0%
	27	

US $10.00 stake games, round 10