At the end of Class #12, please answer the following:

1. Define invasion and collective stability.
2. Discuss the conditions under which any nice strategy (e.g., TFT) is collectively stable.
3. Use the definition of “invasion in clusters” and perform the algebra (and show your work) necessary to demonstrate the conditions under which TFT can invade ALL D.
4. Identify the four central results in the chronology of cooperation.
5. Describe the rescue fire as an iterated Prisoner’s Dilemma game. Using the algebra associated with invasion by clusters (show your work), explain why cooperation did not flourish on the island.

I. Chronology of Cooperation

A. We have seen that TFT is successful in the computer tournament.
B. We have support for Proposition #1
C. What would happen over time?
   1. Could cooperation get started in a world of defectors?
   2. If so, could it sustain itself against other defectors
D. Analog from evolutionary biology
   1. Single population playing a certain strategy--Native Strategy
   2. A single mutant arises playing a new strategy--Mutant Strategy
   3. The Mutant "invades" Native population if the mutant gets a higher payoff than the typical member of the population gets
   4. A strategy is "collectively stable" if no other strategy can invade it

\[ \text{EV[Mutant|Native]} > \text{EV[Native|Native]} \]

E. Develop a set of hypotheses based upon the Axelrod theory

1. TFT is collectively stable as long as w is sufficiently large
2. ALL D is collective stable regardless of the value of w
3. ALL D can be invaded by clusters of TFT players as long as w is sufficiently large
4. Once in place, TFT cannot be invaded by ALL D as long as w is sufficiently large

II. Question: Under what conditions is TFT (or any nice strategy) collectively stable?

Propositions #2, #3, and #4

TFT is collectively stable if and only if, w is large enough. This critical value of w is a function of the four pay-off parameters T, R, P, and S. Imagine someone playing DC or DD

\[
\begin{array}{cccc}
\text{Native:} & C & D & C & D & C & . & . \\
\text{Mutant:} & D & C & D & C & D & . & . \\
\end{array}
\]

Sufficiently large is the largest of the following:

\[ w > \frac{(T-R)}{(T-P)} \]
\[ w > \frac{(T-R)}{(R-S)} \]

if \( w > \frac{2}{3} \), TFT is collectively stable
if \( \frac{2}{3} > w > \frac{1}{2} \) it pays to defect on every other move
if \( w < 1/2 \), it pays to defect on every move

**Generalize from the results in #2** -- Any strategy that may be the first to cooperate can be collectively stable only when \( w \) is sufficiently large

- Native is TFT and Mutant is ALL D
- Native gets R with other natives
- Native gets S on first move with mutant and P thereafter
- Mutant gets T on first move and P thereafter
- Interaction must last long enough for the gap (T-S) to overcome by future moves

For a nice strategy to be collectively stable, it must be provoked by the first defection of the other player

For TFT to be collectively stable
- \( W \) must be sufficiently large to overcome initial gap (T-S)
- Nice strategy must be provoked by the first defection

### III. Question: Under what conditions is ALL D collectively stable? (Proposition #5)

ALL D is always collectively stable

\[
\begin{align*}
EV[TFT|ALL D] &< EV[ALL D|ALL D] \\
S + wP/(1-w) &< P/(1-w) \\
S(1-w) + wP &< P \\
S(1-w) &< PwP \\
S &< P(1-w) \\
S &< P
\end{align*}
\]

! If other player is certain to defect, there is no point in cooperating
! A world of "meanies" can resist invasion by any other strategy as long as the mutants come one at a time

### IV. Question: Are there strategies that can invade ALL D if they come in clusters or small groups? (Proposition #6)

The strategies that can invade ALL D in a cluster with the smallest value of \( p \) are those which are maximally discriminating, such as TFT; where

- \( p = \) proportion of interactions with someone using your strategy

maximally discriminating = if it will eventually cooperate even if the other has never cooperated yet, and once it cooperates it will never cooperate again with ALL D but will always cooperate with another player using the same strategy it uses.

when mutants arise in clusters, their expected value calculations need to take into account that they will interact with their own kind as well as the natives:
EV[TFT] = p*EV[TFT|TFT] + (1-p)*EV[TFT|ALL D]

Therefore, w/ clusters, the following determines whether a cluster of TFT can invade ALL D

p*EV[TFT|TFT] + (1-p)*EV[TFT|ALL D] > EV[ALL D|ALL D]

if w = .9

EV[TFT|TFT] = R/(1-w) = 30
EV[TFT|ALL D] = S+wP/(1-w) = 9
EV[ALL D|ALL D] = P/(1-w) = 10

Therefore

30p + (1-p)*9 > 10
30p -9p + 9 > 10
21p + 9 > 10
21p > 1
p > 1/21

Conclusion: If TFT has few as 5% of their interactions with other TFT players, they will get a higher return than the ALL D players meeting other ALL D players; A small cluster of TFT players can do well

I. Question: If a nice strategy cannot be invaded by a single individual, it cannot be invaded by any cluster of individuals either (Proposition #7)

EV[ALL D] = p*EV[ALL D|ALL D] + (1-p)*EV[ALL D|TFT]

therefore, when there are clusters of mutants, the following determines invasion

p*EV[ALL D|ALL D] + (1-p)*EV[ALL D|TFT] > EV[TFT|TFT]

if w = .9

EV[TFT|TFT] = R/(1-w) = 30
EV[ALL D|TFT] = T+wP/(1-w) = 14
EV[ALL D|ALL D] = P/(1-w) = 10

Therefore

10p + (1-p)*14 > 30
10p -14p + 14 > 30
-4p + 14 > 30
-4p > 16
-p > 4
p < -4

II. Summary and Conclusions
A. Four Central Results in the Chronology of Cooperation
   1. TFT is collectively stable if w is sufficiently large
   2. ALL D is collectively stable as long as challengers come one at a time
3. ALL D can be invaded by clusters
4. TFT cannot be invaded by clusters of ALL D as long as w is sufficiently large

B. Cooperation can get started by even a small cluster of individuals
  1. will not work if cooperators are scattered
  2. Cooperation will thrive as long as it is based upon reciprocity
  3. Cooperation will thrive is future casts long shadow
  4. Cooperation can protect itself from invasion
  5. Cooperation can thrive in a variegated environment
  6. Cooperation does not depend upon communication or a central authority

C. Those using a nice strategy can afford to be generous
  1. Nice strategies must be provokable to be stable

VII. Iterated Versions of Other games

A. Stag Hunt/Assurance

B. Chicken

VIII. LORD OF THE FLIES--An Axelrod-Type Interpretation

1. Basic Question
   *It is the structure of the choice situation rather than the nature of the individual that is
customer responsible for the outcomes of the various choice situations.*

2. Characterizing the Boys
   - Rational egoists
   - No central authority
   - Young boys with limited (?) socialization

3. Choice situation
   - Public Goods
   - Prisoner’s Dilemma
   - Payoffs
   - Temptation to Defect \( R > (T+S)/2 \)

4. Probability of Meeting Again/Discount Paramter
   - How high is w?
   - Is w the same for all of the boys?
   - Proposition #1
   - What strategies do they use?
   - Can they calculate expected value?

5. Reasons that EoC may not apply to LoF
   - w is too small
   - p is too small
   - different notion of winning
   - different values of w
• nice strategy that is not maximally discriminating

6. An Explanation--Begin by cooperating indiscriminantly
• Players use ALL C
• Question: can ALL D invade ALL C?

7. Can ALL D invade ALL C?
\[ p \cdot [\text{ALL D|ALL D}] + (1-p) \cdot [\text{ALL D|ALL C}] > [\text{ALL C|ALL C}] \]
\[ p \cdot [T/(1-w)] + (1-p) \cdot [S/(1-w)] > R/(1-w) \]

w=.9, T=5; R=3; P=1; S=0

\[ 10p + 50 - 50p > 30 \]
\[ -40p > -20 \]
\[ -p > -.5 \]
\[ p < .5 \]

8. Conclusions on Invasion of ALL C by ALL D
• ALL D can invade ALL C
• Invasion works only up to a point
• The smaller the number of ALL D players, the easier it is to invade
• Invasion works because there are “prey” for the “predator”
• Once others see how well the ALL D players do, there would be a rush to join in and take a free ride--players would see that there were great gains to be had by playing ALL D
• Breakdown is similar to the “Ecological Collapse” in the Tragedy of the Commons

\[ p \cdot [\text{TFT|TFT}] + (1-p) \cdot [\text{TFT|ALL D}] > [\text{ALL D|ALL D}] \]
\[ p \cdot [R/(1-w)] + (1-p) \cdot [S + wp/(1-w)] > P/(1-w) \]

w=.9, T=5; R=3; P=1; S=0

\[ 30p + 9 > 10 \]
\[ 21p + 9 > 10 \]
\[ p > .047 \]
9. **How do we explain the ending?**

Tragedy of the Commons-type explanation
- rush to exploit the commons
- at some point everyone joined in and there was collapse
- recriminations abounded

Axelrod-type explanation
- unable to enlarge the shadow of the future
- unable to get critical coalition
- did not realize that maximal discrimination was important