Thanks everyone for reading this draft! My hope is for this to grow into a methods paper and possibly become the foundation of my prospectus (hence the long lit review). The paper is long because it contains a lot of graphs from the models. It derails as it goes along, apologies for that. Comments are welcome, in particular I would appreciate feedback on:

1. The goal—is it clear?
2. The model—is the description too much/little? I've presented this to fellow complex systems people, so I'd like to know how it reads to non-modelers.
3. Results/Analysis—Any suggestions on how presenting the model and eventual results might be more intuitive to a PoliSci audience? I'm thinking about trying to do something close to what John Jackson and Ken Kollman (Jackson and Kollman 2010) do in their path dependence paper, but my methods skills aren't there yet.
   1. I currently include a multinomial logit model to try to analyze the results from the model but I fear this does not make much sense. Skim away on this part!
4. Specific examples you have would be **MOST WELCOME**.

Thanks so much,
Jean

**Altered Perspectives: Institutional Path Dependence and Behavioral Outcomes**

**Abstract:** Nations often look to existing institutions for inspiration when designing or altering their own, focusing upon the outcomes they would most like to match. However, there are many influential factors that contribute to the success or failure of such a change beyond regime structure. To understand how behavior may depend upon the relationship between prior and new institutions, I use an agent based model to explore emergent strategies and equilibria induced by institutional structures. I use two games to represent institutional incentives. In the first game, agents play the designated game for a pre-determined number of periods before a change occurs. This change represents some structural change in the institution and the new institution is represented by a second game which is played for some number of periods. Comparing behavioral outcomes to the sequence of games, I find that the initial game can have a lasting effect on the strategies agents employ after the institution changes. Institutions condition agent expectations and strategies as their perspective of available behaviors and meaning is carried into the second game via strategies developed in the first. Understanding the relationship of the second institution to the first is key for effective institutional (re)design.
Current explanations for regime stability in democratic consolidation disagree over the role of influential factors such as institutions and political culture. Acemoglu and Robinson (2012) argue that institutions and critical junctures are responsible for the trajectory of regime success. Others disagree that institutions operate as independently as Acemoglu and Robinson allege, arguing that effects from the initial regime type will linger in the form of institutional preferences, social dynamics, or different survival rates after that initial selection. (Bednar and Page 2007; Inglehart and Welzel 2005; Shugart and Carey 1992) Although these perspectives each offer valuable insight into the process of value formation and influential variables in regime success, many of their components are difficult to measure or lead to incomplete explanations. Given that institutions are designed to influence behavior, how does a change in institutional structure influence already-established behaviors? When are structural changes critical junctures and when are they just changes? As future institutions are introduced to a vibrant political environment, whether a new (or altered) institution fulfills the designers' purpose may be determined before it is even introduced.

My model is simple and offers clear explanations that support existing theories but can be empirically evaluated. I can test existing hypotheses (Geddes 1999) and provide insight to the reasons for influence.¹ Using an agent-based model allows each piece of the puzzle to be carefully examined to understand how they fit together. Some changes may lead to path-dependent outcomes, representing an important moment in the trajectory of institutional development. Preliminary results show that current regimes can influence strategic behavior in ways that persist over many periods. 'Failed' changes may be a function of incompatible institution design while measures of institutional 'success' may vary across regime types over time. The paper proceeds as follows, an introduction to the general literature is followed by an explanation of the model on page 9, and analysis of the model output (using a multinomial logit model) begins on page 18.

¹ Not sure if I'll use this or not. See later discussion.
Institutions

An institution is a “relatively enduring collection of rules and organized practices, embedded in structures of meaning and resources that are relatively invariant in the face of turnover of individuals and relatively resilient to the idiosyncratic preferences and expectations of individuals and changing external circumstances” (March and Olsen 2005). Institutions affect not only how political actors are enabled or constrained but also the governmental capacities of political systems. They guide behavior and provide information so that patterns of behavior might be sustained (McCubbins, Noll, and B. R Weingast 1987; Milgrom, North, and Barry R Weingast 1990; North and Barry R. Weingast 2009; Barry R Weingast 1997). Most interactions are not zero-sum games of conflict: instead, individuals must understand how to coordinate behavior when there are many potential ideal equilibria, which institutions help facilitate. For example, when divergent interests exist, alternate means of providing and gaining the necessary information emerge. Similarly, when multiple equilibria exist, there must exist widespread means to select amongst the equilibria.

Formal institutions, codified rules such as how legislatures are designed, and informal institutions, unwritten rules such as those for sanctioning executives, influence behavior. The development of both types of institutions may have a profound impact on behavioral patterns, perhaps in ways that are initially unanticipated. However, the question remains whether institutions are guiding forces while in force or if their influence persists systematically beyond their tenure.

Formal Institutions

Institutions can shape how individuals orient themselves to and perceive their behavior in political environments. The design of formal institutions, often set forth in a constitution, can influence future institutions implemented in a state. Individuals use institutions and culture as means to communicate information, creating a common knowledge that facilitates interactions over time. Rational behavior motivates individuals to seek common knowledge as group

2 This differs from culture in that culture is context-dependent and not ‘relatively invariant’.
environments necessitate coordination and/or cooperation (Chwe 2002). Often in these interactions, individual behavior is conditioned not upon an individual's knowledge, but what that individual perceives others to know (Chwe 2002). Institutions are one means to standardize information (and generate common knowledge) and can be either formal (written) or informal (unwritten).

**Inefficient Institutions**

However, once created, the rate of institutional adaptation to behaviors and the broader political environment is slow. Over time, a gap between formal rules and current behavior and expectations may develop, which means that often institutions are inefficient (March and Olsen 2005). Inefficient institutions may be problematic not just for the functioning of political structures but also because their influence might persist after new institutions are adopted in the future. After institutional change, patterns of behavior and ways of perceiving the institutional environment may persist under new institutional structures if the original institution interacted with cultural traits. The length of time that the pattern has continued may also influence the success of institutional changes. Existing institutional structures, particularly formal structures but also informal rules, can reinforce patterns of behavior. (Grzymala-Busse 2006) The combination of particular formal rules with particular informal environments, such as culture, will be conducive to the emergence of informal institutions.

**Culture**

As institutions persist over time, they continually reinforce existing strategies within the community of individuals/society. This interaction can lead to the formation of different norms and cultural behaviors that can further enforce institutions or aspects of institutions. Authors tend to agree that culture matters but the degree to which it should matter for research is left open (Bednar and Page 2007; DiFranceisco and Gitelman 1984; Gelfand et al. 2011; J. Henrich et al. 2005; Laitin 1986; Putnam, Leonardi, and Nanetti 1993; Richerson, Boyd, and D. 2005; Tabellini 2007). This is due in no small part to the incredible difficulty in understanding what culture is and how its various effects can be measured. Culture is not everything someone is or does but there are particular outgrowths or aspects of culture that may influence perspective. To
understand culture, it must be studied in context, rather than as a thing in itself.

The primary emphasis on culture in this paper is that it is a group-wide phenomenon that potentially shapes how individuals relate to, interact in, and perceive their environment. Particular traits are referenced, but not assumed, and the general properties of a culture are that there is some form of behavioral stickiness, consistency (within and across individuals), and context-dependency. This definition comes primarily from Bednar and Page (2007) and offers a basic framework for the analysis of behavior.

Sharing a culture enables groups to efficiently organize, share information, and interact in environments shaped by institutional structures. Because institutions are often inefficient and slow to change (March & Olsen 2005), cultural traits may interact with incomplete or contradictory institutional structures, promoting solutions to avoid system inefficiencies (Geddes 1999) and coordinate behavior (Tsai 2006, Chwe 2002). Henrich and his coauthors demonstrate the importance of context in behavioral outcomes across societies, and argue for a systematic means to understand the variation observed (J. Henrich et al. 2005).

Given the variations in culture, regimes, and behavior, it can be tempting to try to explain observed differences in institutions and behavior in terms of culture. Studying culture may enable scholars to reach conclusions about institutional environments but not the entire picture, “Culture is unlikely to be of much help in explaining why alternative A was chosen over alternative B—but it may be of great help in understanding why A and B were considered, while no thought was given to C, D or E.” (pg 142, Elkins and Simeon 1979). Often institutions are designed that permit multiple strategies to develop and many outcomes to emerge. However, why these outcomes emerge over others may be a function of the environment itself. The types of research questions asked guides the direction of scholarly work and influence the set of conclusions scholars can reach.

**Institutions and Critical Junctures**

In contrast to many culture-based theories, Acemoglu and Robinson (2012) argue that
institutions are key to nations' success: institutions can support the continued development of successful societies that encourage economic growth. These societies further reinforce the development and maintenance of inclusive institutions, leading to (potentially) continually strong nations. Failed states often have extractive institutions that undermine economic development. Path dependence plays no role in this story. States' trajectories are influenced by critical junctures, moments in time which present definitive opportunities for the creation or destruction of existing institutions and patterns. The decisions made at the junction are often functions of circumstances that exist, other environmental factors, strategies and perceptions of the community, etc.

While institutions may be difficult to change, the critical role for timing, or the idea of a critical juncture is not as definite as some authors imply. (Acemoglu and Robinson 2012; Pierson 2000) These moments may be significant but the change may be reversible in that behavior will quickly change. It depends upon the circumstances, in some instances an institution may be quickly adopted and influence deep-seated behavioral patterns. In other cases, institutional changes may be quickly implemented with a measurable influence on the variable of interest.

When are institutional changes critical junctures? At what point do changes become irreversible? Mahoney and Thelen argue that while exogenous shocks may influence institutional behavior, subtle changes may prove to be more significant for the trajectory of institutional development. As institutions persist over time, changes within the population of agents, their strategies and their strategic environment, continue. (Mahoney and Thelen 2010)

Change

Institutions change over time, either formally (changing a constitution or laws) or informally (new interpretations arise that effectively alter behavior). I do not yet examine causes or reasons for change; the demand or momentum towards change could come from within or without. These could be internal demands, such as those observed and predicted by modernization theory to external influence, such as agencies requiring changes for funding, pressure from the global community. Regardless of the reasons for change, I argue that change occurs. (Note that the
reasons of and for change are likely to be important, and to substantively influence the outcomes, but I do not yet consider them (Dal Bó, Foster, and Putterman 2008).

As institutions are designed to shape behavior, the shift from one institutional structure to another is also context dependent. While the absolute structure of institutions may change, the strategies an institution encourage do not expire in step with the complementary institution. The potential for strategies, behaviors, and conventions to persist over time complicates the story Acemoglu and Robinson (2012) provide as it contradicts their idea of critical junctures and reversibility of paths.

**Future Institutions**

Institutions are changed, by the forces from within and without, but with the underlying belief that the change in structure will have a positive effect in some regard (why change them otherwise?), be it for the rulers, citizens, or some combination thereof. However, we need to consider how the new institutions relate to those which already exist—the new change may be to correct something deficient under the old structure or removing or introducing legislation to a new area. These new institutions are chosen for the same reason, the initial set were selected: they will encourage the development and continued existence of certain types of behaviors.

Given that future institutions are introduced to a vibrant, already-existing political environment, whether a new institution (or altered institution) supports the intended behavior may not be a foregone conclusion.

**Applications of the Model**

There are different directions this project could go and I'm struggling with this portion. Right now the model is largely theoretical, explaining what might happen given some broadly characterized institution and some 'change'. Ideally, I would like to evaluate whether actual changes would be successful or not and I am not certain which direction I would like to pursue. I present several options and would appreciate feedback. Right now, I am considering the creation of institutions, such as anti-corruption commissions, or the stability of regime transitions. The
biggest hurdle will be determining how to map something as complex as the institution or regime onto a simple two-by-two game.

**Regimes & Failed States:**
Geddes emphasizes the different nature of institutional frameworks, arguing that different institutional structures require different actions to overthrow them. (Geddes 1999) These different institutional structures will have different types of revolutions, or regime changes, and will encourage a different type of society than other types. My model allows us to consider the success of different institutional structures absent from different mechanisms for change. While this is clearly an important component part of institutional success, as explained previously, I argue that institutional change is often endogenous to the current institutional environment and is thus secondary to the existing institutional structures. Future works will consider the importance of this concept, but for now it is tabled.

There is no universal answer or system to understanding why, how and when people act the way they do independent of their environment or context. Sometimes institutions strongly influence outcomes³ (Jackman 1987), other times culture strongly influences institutional design (Greif 1994). However, at times culture and institutions coevolve, jointly influencing each other (among others: Calvo et al 2004; Tsai 2006; Schelling 1978; Henrich et al 2005).

Cultural explanations of behavior often implicitly rely upon institutional structures and their relationships to behavioral outcomes. Similarly, many institutional explanations of behavior rely upon cultural frameworks. Institutional explanations sometimes ignore other factors that contribute to institutional success and longevity. Studying the connections between institutions and culture can address this relationship directly and incorporate the feedback between culture and institutions.

Coupling the simplification gained from using formal models with existing and continued research of different cultures will enable future scholars to organize and explain the random or

---

³ Citations relevant to motivating examples will go here.
puzzling relationships they observe. These solutions may, in turn, influence or shape the design of formal institutions. Any study that evaluates only one side of influence, either culture or institutions, can miss the interplay that develops over time as culture and institutions jointly influence behavior. Approaches considering this joint influence can explain why and how social perceptions are shaped and altered and create a clearer picture of the political world. The model I present offers a means to analyze the role of context on behavior and may help to explain institutional success and failure.

**Agent Based Model**

I use an agent-based model to evaluate the implications of different regime structures. Agents are represented by automata with finite memory (currently with a maximum of 4 states) and the ability to evolve strategies, which are contextual rules for behavior. Agents are able to form strategies that can account for other agents' behavior, although this is not necessary—they can have a strategy as simple as 'all C' or something that is contingent upon multiple plays in history.

**The Games**

Currently there are six different games that represent different institutional structures. The suite of games may be altered at some point in the future, depending upon the specific cases I evaluate. The games represent different ways behaviors could be incentivized—to induce coordination on a single outcome (Pick C and Pick D), or to present asymmetrical payoffs that could lead to different equilibria (Battle Of the Sexes).
Typical Run of the Model

In each period, agents will be matched with every other agent in a round-robin tournament. Agents are individuals who interact randomly with each other using their strategy, choosing to cooperate or defect depending upon the choice of their last opponent based on the strategy they employ. Institutions structure the agents' environment such that under a different institutional structure, the agents will receive different payoffs for the same behavioral outcomes. At this point, I only evaluate which outcome occurs, which tracks the choices of both agents (CC, CD, DC, DD).

The agent can only 'choose' their behavior, not whom they interact with or whether they want to interact with someone or not. Agent memory is externally set and strategies can use up to the maximum allowable memory, although they may use less. At the end of the tournament, agents are randomly matched with (only) one other agent and the agent with the higher payoff is 'cloned' (with some probability of mutation) in the next generation while the weaker agent is removed from the population. Strategies persist but memory (whether their opponent cooperated or defected in the last round) is reset between games (this may be revisited in robustness checks).

After these duels, play continues until an externally-set moment, at which the rules change.

---

4 There are two populations 'row' and 'column' to account for different incentives agents may face in asymmetric games. Evaluating the dominant strategies in these two populations over time may also prove interesting.
Agents then play a new game with the selection procedures identical to the first game. This switch represents the change in institutional structure—the old institution has been replaced by a new institution that rewards behavior with different payoffs. For example, agents may begin with 500 periods of 'Battle of the Sexes' but then the regime changes and agents find themselves in the 'Prisoners' Dilemma' or 'ALLC'. They then play this second game for a pre-determined number of periods, after which the model run is complete.

Current measures only include the resulting action combinations for pairs (how many pairs were at CC, CD, DC, DD). Eventually I will determine the strategy profile that emerges within different regime environments although this first pass will only examine population-wide phenomena. I hope that I'll be able to examine and classify different predominant strategy types that emerge. There are no set period lengths and the overall time to equilibrium varies by the number of other participants (more agents means more plays per round and a slower time-to-equilibrium (e.g. 40 agents hit equilibrium around 30 ticks while 100 agents hit equilibrium around 75)).

Robustness

Although there are no robustness measures in place just yet, I will experiment with different period lengths for institutions (taken from average regime lengths for transitions from Huntington and others). I'll also determine sensitivity to specific payouts from the games to determine whether there is a relationship between the payoffs and behavior (and how sensitive that relationship is).

<table>
<thead>
<tr>
<th>Regime Type</th>
<th>Average Length of Rule (years)</th>
<th>Average Age of Surviving Regimes</th>
<th>Percent of Regimes Surviving in 1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>Military</td>
<td>8.8 (31)</td>
<td>7.3 (4)</td>
<td>11.4%</td>
</tr>
<tr>
<td>Military/Personal</td>
<td>10.3 (3)</td>
<td>12.3 (2)</td>
<td>19.8</td>
</tr>
<tr>
<td>Personal</td>
<td>15.1 (43)</td>
<td>18.8 (8)</td>
<td>15.7</td>
</tr>
<tr>
<td>Single-Party/Personal</td>
<td>15.0 (8)</td>
<td>39.0 (3)</td>
<td>27.0</td>
</tr>
<tr>
<td>Single-Party (stringent criteria)</td>
<td>22.7 (17)</td>
<td>35.1 (17)</td>
<td>50.0</td>
</tr>
<tr>
<td>Single-Party (less stringent criteria)</td>
<td>25.7 (22)</td>
<td>33.5 (11)</td>
<td>33.3</td>
</tr>
<tr>
<td>Single-Party/Military</td>
<td>23.8 (4)</td>
<td>— (0)</td>
<td>0.0</td>
</tr>
<tr>
<td>Single-Party/ Military/Personal</td>
<td>31.0 (2)</td>
<td>37.3 (3)</td>
<td>60.0</td>
</tr>
</tbody>
</table>

5 How, or whether, these strategies are reset is still an open question.
The table above (from Geddes, 1999) lists the average length of different regime types which I will also attempt to incorporate in the model. It may be difficult to find a proper 'year' length to adequately correspond to number of rounds. However, if the relative duration is preserved among regime types, we may be able to learn something about the behaviors and strategies that develop. I also imagine that the number of states (which influences the complexity of strategies) will also become important to a point.

Note that the agents do not anticipate regime change in any meaningful way—one day they are living happily under institution A and the next institution B is in place. One might argue that in reality agents may change their behavior in anticipation of the changed payoffs under the new institution. I do not yet incorporate this anticipated alteration of behavior because the extent to which individuals are able to correctly adjust their behavior is not clear. Including this potential adjustment would require additional assumptions for the model that may not be justified. Additionally, this current analysis may still prove useful as a base case of sorts—how does behavior change for those (the 'wait-and-see-ers') who are uncertain how other will adjust?

---

6 I am not wedded to Geddes, but it seems to be a good starting point.
7 Thanks to Ali Sanaei for a helpful conversation on this point.
Initial Summaries/Evaluations

These are my initial summary results from some basic runs. For all the games below, the mutation rate was set at 0.30 and the number of states was 4 (strategy length). In the single-institution results, the results are summarized for 300 separate runs of 100 agents each who played 10 rounds per player with all other players, for 500 time-steps. When there was a game switch, agents interacted for 1000 time-steps. The first 500 were under the first game with a switch at 500 to the second game. All other parameters remained the same.

Although I haven't done any systematic evaluation of the specific parameters chosen, I have worked with different values individually—mutation rate, as it increases, time to equilibrium also increases; states, more states means 'smarter' agents who take less time to reach equilibrium. The next phase of the project will look at the different values and try to determine explicit relationships. I'd also like to start looking at both strategies and dig deeper into the existing data I have. The graphs I've included show the proportion of population-wide behaviors that fall in one of four possible outcomes with 95% confidence intervals surrounding each estimate. The charts I present show bounds on the outcomes so you can see how widely the outcomes range across the different games. For BOS, there are two equally-frequent outcomes, so both CC and DD happen with almost any frequency. In PickC, in contrast, the agents quickly learn the optimal outcome for all and then rarely stray from it.

---

8 This is imprecise language—what I mean is that 95% of the runs fall within the band around the estimate. The bands were calculated by multiplying 1.96 by the standard deviation for the each of the four outcomes at each timestep.
**Battle Of the Sexes:**

Notice that the outcomes are what we would expect—that CC and DD are statistically indistinguishable from one another. However, they have wide confidence interval bands, meaning that behavior is a bit erratic even over 500 timesteps. Future experimental runs should include a larger sample and perhaps should allow for longer runs, such as 700 or 1000 timesteps.

There is a bit of a bug in the code for asymmetric games and the graph below represents the asymmetric balance of power that emerges between the players. This is under investigation.

<table>
<thead>
<tr>
<th>BOS</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1,3</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>3,1</td>
</tr>
</tbody>
</table>

![Outcomes in BOS Over Time](image)
Coord on C:
This looks about right, some bumps along the way but nothing dramatic or surprising. Notice how thin the confidence intervals are here, I think I might be able to use them somehow to create a measure of game difficulty (agents can learn optimal behavior and then do not deviate in 'easy' games with thin bands).

<table>
<thead>
<tr>
<th>Coord</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,3</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>1,1</td>
</tr>
</tbody>
</table>

Outcomes in CoordC Over Time

![Outcomes in CoordC Over Time](image)
Switching Games:

Notice that both scenarios feature a slow shift (~250 timesteps after the switch) before the majority of outcomes are those 'encouraged' by the setup. It looks like the confidence intervals become distinct at about t=700 or so in both BOS to C and to D.
The two 'before' lines for CC and DD should be equal (need a larger sample size) but notice that the curve in the average game for 'Coord C' is less steep than for 'Pick C' because there is less incentive to change behavior (DD yields (1,1) rather than (0,0). Thus, the severity of institutional change may influence the speed of adaptation.

<table>
<thead>
<tr>
<th>BOS</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,1</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>1,3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pick C</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,3</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>0,0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coord C</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,3</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>1,1</td>
</tr>
</tbody>
</table>
C to BOS vs D to BOS

<table>
<thead>
<tr>
<th>BOS</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,1</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pick C</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,3</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>0,0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pick D</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>3,3</td>
</tr>
</tbody>
</table>

These results are

A gent behaviors/strategies work under both the original infrastructure and under the second so while there has been a change in the payoffs there appears to be a small change in behaviors. Notice that to switch from CC to DD (or vice versa) agents would have to coordinate (CD and
DC have payoffs of zero\footnote{Stag Hunt might be useful because it can offer an incentive to switch rather than BOS which requires full coordination.}) We do see a drop after the switch, which may be driven by the mutation rate but the decline seems pretty steep...I'm interested in doing more runs to see if this effect is intensified.

**Model Analysis: MLN(?)**

There are three primary ways to evaluate the data—first, to determine whether there is path dependence in the data, secondly whether the 'desired' behavioral outcome emerges (is the outcome one of the known equilibria for the game?), and finally how long it takes before a stable outcome is reached. My real question is whether this makes the most sense for how to present and analyze this data.

**Outcomes**

We can analyze the results of the model using a multinomial logit because the outcome variables take on one of four different outcomes, which correspond to the outcomes in a two-player game. Typically, when estimating multinomial logit we assume that the outcomes are categorical, exclusive, and cannot be ranked. I am not entirely certain that these conditions are satisfied, for example, depending upon the different games, one outcome may be said to be 'preferred' or 'optimal'. However, in this very rough first pass, we will put this concern to the side, particularly because across all games, there is no possible ranking for all games.

A second assumption of multinomial logit is the independence of irrelevant alternatives (IIA). IIA means that the ratio of any two probabilities are independent of the remaining choices, so if CC, CD, DC, and DD are the outcome possibilities, the removal of one option (eg DC) should not change the probability distribution across the other outcome possibilities. However, it is unclear whether this is satisfied because the removal of that option could influence the distribution of probabilities unequally. I think this may be satisfied, but it really depends upon how we conceptualize the removal of the strategy and how we consider the other outcomes (in 'should' or actual values).

For example, to remove (only) DC as an option may mean that when the second player cooperates, the first player must also cooperate, that when the first player defects, the second player must also defect, or that somehow when the first and second players select D and C, respectively, that this selection is transformed to some other outcome. This final option does not make much sense given how the games are constructed. In the first two options, it depends on how we understand the comparison of proportions: if we think about the games and consider what an optimal strategy would be, then the equilibria would share the 100\% allocation while the other outcomes would have 0\%. For example, removing one of the equilibria from a game that has two (such as battle of the Sexes) would encourage all outcomes to be on the remaining outcome, which would have a much higher numerical outcome but its proportion (given that the other outcomes would be zero) would remain unchanged.
Supposing, for now, that these assumptions are satisfied\(^\text{10}\), which they may well be, based on discussions we had in class\(^\text{11}\). Assuming IIA we would create a model, as follows

\[
L = \text{PRODUCT}(i=1 \text{ to } N) [\Pr(CC)]^{(\text{if CC})} \cdot [\Pr(CD)]^{(\text{if CD})} \cdot [\Pr(DC)]^{(\text{if DC})} \cdot [\Pr(DD)]^{(\text{if DD})}
\]

This can be rewritten as

\[
\begin{align*}
\mathcal{L} = \prod_{i=1}^{N} \left( \frac{e^{X_i \beta_{DC}}}{e^{X_i \beta_{CD}} + e^{X_i \beta_{DD}}} \right)^{ijCC_i} \cdot \left( \frac{e^{X_i \beta_{CD}}}{e^{X_i \beta_{CC}} + e^{X_i \beta_{DD}}} \right)^{ijCD_i} \cdot \left( \frac{e^{X_i \beta_{DC}}}{e^{X_i \beta_{CC}} + e^{X_i \beta_{DD}}} \right)^{ijDD_i}
\end{align*}
\]

To analyze the data, I created a categorical variable, 'EQ' which is either 0, 1, 2, or 3, depending upon whether CC, CD, DC, or DD is the 'most common' outcome for a given round. Model 1 (in the appendix) uses different data but has similar results to the experimental data used in Model 2.

**Model 2:**

Time plays a very small role in the final outcome (the coefficient is less than 0.0009 on this data), so I did not consider it as a variable in this model. Because time does not seem to play as much of a role as the period for which the two games are played. To incorporate this, I include the point at which the institutions switch (Switch). I also create a measure of the disparity in payoffs between the row and column players, as the difference indicates asymmetrical payoffs and CD, DC will be more likely outcomes. I also include an interaction term (\(startGameXEQ^2\)) for the length to equilibrium and the game being played to incorporate whether there is a difficulty associated with particular games beyond how long it takes players to equilibrate.

Here I present the relative likelihoods, although the coefficients from the mlogit estimation are also in the Appendix). Because this is preliminary, and exploratory in a sense, I present the marginal effect of different variables to give a sense to how they contribute to the four different outcomes being more or less likely. Remember that 0 and 3 are coordination outcomes (CC, DC) while outcomes 1 and 2 are where players coordinate on different actions (CD, DC).

Note that I have a rather large log likelihood (-6939192.1), which is probably due to the size of the dataset at hand. This value refers to the likelihood of observing the data and will be useful when I estimate different (similar) models.

What the values under 'RRR' mean, is that for a unit increase in the variable at left, it makes the given outcome that much more likely than the base outcome. For example, for outcome 1, we see AveFitDiff | 120.0669, which means that for every unit increase in the absolute difference between the average Row and Column player payoffs at the end of a round, outcome 1 is 120% more likely than outcome 0. This makes sense, the more asymmetrical payoffs are, the more

\(^{10}\) If we did not believe IIA held, we could try conditional logit, but this might be an even bigger mess because there would be four overlapping nests for the outcomes (CC and CD, DC and DD; CD and DD, CD and DC).

\(^{11}\) I think we need to talk more about this because I'm not sure I understand why this is likely true.
likely an outcome that requires coordination on different actions is.

| EQ | RRR | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|----|-----|-----------|---|---|---------------------|
| 0  | (base outcome) |   |   |   |                     |
| 1  | AveFitDiff 120.0669 | .6133089 | 937.35 | 0.000 | 118.8708 121.275 |
|    | mutationRate | (omitted) |   |   |                     |
|    | Switch 1.000106 | 5.53e-06 | 19.22 | 0.000 | 1.000095 1.000117 |
|    | EQ2 | .1916838 | 0.008544 | -111.34 | 0.000 | .1863807 1971689 |
|    | startGameX2 1.012637 | .0040376 | 3.20 | 0.001 | 1.004954 1.020782 |
|    | startGame 2.406523 | .004214 | 498.27 | 0.000 | 2.398225 2.414851 |
| 2  | AveFitDiff 166.0033 | .882229 | 961.89 | 0.000 | 164.2831 167.7415 |
|    | mutationRate | (omitted) |   |   |                     |
|    | Switch 1.000035 | 5.64e-06 | 6.27 | 0.000 | 1.000024 1.000046 |
|    | EQ2 | .2166243 | 0.003363 | -100.53 | 0.000 | .2122391 2252016 |
|    | startGameX2 0.004115 | 1.86 | 0.064 | 0.9843027 1.000434 |
|    | startGame 2.658548 | .0049082 | 529.62 | 0.000 | 2.648945 2.668185 |
| 3  | AveFitDiff 1.635805 | .0036823 | 218.62 | 0.000 | 1.628603 1.643038 |
|    | mutationRate | (omitted) |   |   |                     |
|    | Switch 0.999974 | 3.42e-06 | -0.76 | 0.445 | 0.999907 1.000004 |
|    | EQ2 | 6.464888 | .0288222 | 578.18 | 0.000 | 6.414206 6.495829 |
|    | startGameX2 0.007424 | -165.79 | 0.000 | .8663986 0.8692187 |
|    | startGame 1.035294 | .005522 | 68.79 | 0.000 | 1.034271 1.036318 |

While this might not make sense for all games, with our subset, this seems very reasonable because there are two situations in which we might see these off-diagonal choices emerge: when the game has asymmetric payoffs for CC and DD and is difficult to learn (BOS), and if it has asymmetrical payoffs on the off-diagonal that happen frequently (PD). One possible cause for concern is the difference in relative likelihoods between CD and DC (an increase in the Fitness Difference, for example, makes an outcome 120 vs 166% more likely). There may be some bias in the model, so I need to investigate this before moving forward with my analysis.

Two variables that show promise but need to be reconfigured: both EQ2 (my confusingly-named attempt at quantifying whether equilibrium has been reached) and startGame (the first institution) seem to have an influence on the resulting outcome, but a small one. I think this is in part due to the way they are both created (the appendix also contains how these variables were constructed). For example, EQ2 is a dummy variable for whether a system is in equilibrium or not (I had a difficult time figuring out how to create a variable that returned the timestep at which equilibrium was reached). As such, I think the marginal effect is both difficult to interpret (moving to equilibrium makes different outcomes more likely in what respect?) and doesn't
capture the measure of game difficulty as I had hoped. Similarly, startGame is important in that it shows difference between games, but I need to reconsider how it's created. I had initially created a categorical variable for the different types of games (one equilibrium, two equally likely equilibria, and other) that I might return to instead.

In all, this is informative in a sense, but I think I need to strongly reconsider how I organize the data.

Alternate Model

In contrast, if I were focused upon the time to stability, I would consider whether this variable would be binary (like the current dataset) or numerical (reflecting the time at which equilibrium was reached). For the dichotomous measure of equilibrium, I would begin with a simple logit regression with stability as the dependent variables (something like this: $EQ2 \ AveFitDiff \ Switch \ startGame$). If the outcome were numerical, which I may try once I create this variable, I would perform a logistic regression, similar to the one just mentioned. However, because the variables of interest are both the outcome and the time it takes to reach equilibrium, I need to consider how to use my data to create these two separate models. I think similar factors influence both outcome variables of interest, and will continue working on designing the models.

Conclusion: Underwhelming Results and Future Plans

Unfortunately, I still have a lot of work to do before continuing my analysis. Using the MNL setup was incredibly useful for demonstrating the relative role different parameters play in the model. Something to consider is my overabundance of data (for example, the dataset I reference in this paper has 7.5 million observations from 300 runs for each of the parameter combinations). Certain 'rare' outcomes that occur once in every 100 outcomes will occur, on average, three times for every run of the sweeps, potentially.

I am also working on a solution for aggregating the data in a meaningful way that can be analyzed using statistical software. I think that what I might want is data by runs rather than all the timesteps per a given run.
Bibliography


Democracies.” American Political Science Review 81(2): 405-424.


