4-22-2009

# Simulating Realistic Social and Individual Behavior in Agent Societies

Jason Leezer
*Trinity University*

# Simulating Realistic Social and Individual Behavior in Agent Societies

Jason Leezer

## Abstract

While the value of simulations as a tool in the natural sciences has been realized for quite some time, its potential in the social sciences is only beginning to be explored. A class of simulations used to study social behavior and phenomena is known as *social simulations.* One particular type of social simulation is known as agent based social simulation. Here agents are used to model social entities such as people, groups and towns. A purpose of these models is to reproduce realistic behavior in the simulation which is then used to draw conclusions about the corresponding real world entities. However reproducing realistic behavior is a difficult task. This is in part due to the fact that human actions and interactions do not adhere to well defined rules. A successful solution to this problem must reproduce realistic individual decision making as well as realistic social interactions.

We propose two models. First, a model for producing realistic decision making is based off human intuition and deliberation. This model is tested in the Iterative Ultimatum Game and Bargaining Game. It is shown that when agents use both intuitive and deliberative decision making they make decisions similar to those of human subjects.

Next we propose a realistic model for social interactions. Our agents remain selfish and are able to break relationships in order to maximize their utility. It is shown that when agents are able to break unrewarding relationships that a Pareto-optimum strategy arises as the social convention. In addition we conclude the rate and amount of Pareto-optimum strategy that arises is dependent on the network structure when the networks are dynamic and the rate is independent of the network structure when the networks are static.

# Acknowledgements

This thesis represents the culmination of four years of work and research. During this period there were many people who influenced me, taught me and helped me. The following people played an integral role in my education and thus this body of work.

**Dr. Yu Zhang** for first exposing me to Artificial Intelligence and for her guidance and support over the past two years as a research partner.

**Dr. Mark Lewis** for exposing me to research early in my education and sharing his enjoyment of solving hard problems.

**Dr. Berna Massingill** for always helping me with the countless problems I came to her with.

Dr. Gerald Pitts for his helpful comments and guidance while writing this thesis.

**Mike Pellon** and **Phillip Coleman** for their support as collaborators and friends.

And **my parents** for their constant support and encouragement.

# Simulating Realistic Social and Individual Behavior in Agent Societies

Jason Leezer

# Simulating Realistic Social and Individual Behavior in Agent Societies

Jason C. Leezer

A departmental thesis submitted to the
Department of Computer Science at Trinity University
in partial fulfillment of the requirements for graduation.
April 22, 2009

_____ _____

Thesis Advisor Department Chair

_____

Associate Vice President for
Academic Affairs

# Table of Contents

# List of Figures

# List of Equations

# List of Tables

# 1. Introduction

## 1.1 Motivation

While the value of simulations as a tool in the natural sciences has been realized for quite some time, its potential in the social sciences is only beginning to be explored. A class of simulations used to study social behavior and phenomena is known as *social simulations.* One particular type of social simulation is known as *agent based social simulation*. Here agents are used to model social entities such as people, groups and towns. One purpose of these models is to reproduce realistic behavior in the simulation which is then used to draw conclusions about the corresponding real world entities. If realistic behavior can be reproduced then researchers in the social sciences can be given virtual laboratories from which they can experience the same benefits received by the natural sciences. Such a framework for modeling realistic behavior needs to reproduce both internal decision making as well as social interactions. This is certainly not an easy task as entire fields are founded in researching both of these questions.

The applications of human behavioral simulations could be applied to investigate social phenomena. They could be used to make predictions about how people will act in complex situations. For example, these simulations could be used to investigate emergency evacuation plans, or for military purposes where they can test various strategies in a safe environment (Brooks, et al. 2004) (Christensen and Sasaki 2008). They could even be used for entertainment purposes, as in various simulation based video games (Aylett, Louchart and Pickering 2004).

Currently most research into agent decision-making has been performed with the goal of creating an agent who acts rational (Kim 1999). Many advances in this field have been made that have led to practical applications. However, experimental economics has shown that human beings do not always make rational decisions; therefore these agents cannot reproduce realistic behavior (Erve and Roth 1998).

Decision theory finds its roots as a formal topic of research in the fields of economics and psychology (Excelente-Toledo and Jennings 2004), (Gmytrasiewicz and Noh 2002). These two fields were the first to conceptualize the rational agent as employing the principle of *maximum expected utility* (EUT)*,* i.e. an agent should take actions that maximize its expected measure of reward and the first to view the process of decision-making as a series of logical steps and implications (Von-Neumann and Morgenstern 1944). However, the common human decision-making attitudes cannot be captured by EUT (Myers 2002). Cognitive psychology attributes the reason to the domain of human *intuitions* – thoughts or preferences that come to mind quickly and without much reflection (Myers 2002) (Norling 2004).Today, neuroscientists, philosophers, biologists, computer scientists and other scholars are all making contributions to contemporary decision theory as they seek to understand the nature of real decision making where cognitive agents are at work. The growing numbers of disciplines involved not only deepens our understanding of decision-making processes but also creates an emerging field of research where various descriptions of decision making (neuronal, cognitive, formal, behavioral and evolutionary) all intersect.

On the basis of social-psychological developments Kahneman and Tversky point to the existence of two generic modes of cognitive decision-making: an *intuitive* mode in which

decisions are made automatically and rapidly, and a *deliberative* mode, which is deliberate and slower (D. Kahneman 2002). The *deliberative* mode was already well represented by formal descriptions like EUT but any approach that hoped to fully capture human decision-making would also have to account for the more intuitive judgments of humans as well. The *intuitive* mode, Kahneman and Tversky believed, was a process responsible for evaluating decision outcomes in a reference-dependent fashion which was, as they noted, incompatible with the standard interpretation of EUT, a deficiency that can be traced all the way back to Bernoulli's first essay on the topic (Bernoulli 1954).

Another difficulty in producing social simulations lies in the problem of modeling the emergence of social norms. This is in part due to the difficulty of simulating systems that don't abide by well defined rules. While empirical evidence has provided insight into how human relationships are organized, the way in which those relationships are used to produce cooperative behavior where each agent only seeks to maximize its own utility is not well defined.

We begin by formally defining our problem, then the direction and contribution of our research.

## 1.2 Background

An *agent* is defined as anything that perceives its environment through sensors and acts upon its environment through actuators (Russel and Norvig 2003). An agent resides in its environment and behaves purposefully to achieve its goal. An agent is said to be *rational* if at all times it perform actions in an attempt to maximize its utility (Russel and Norvig 2003). An agent

who performs “satisfying” actions or actions that are “good enough” operates under *bounded rationality* (Simon 1957). In order to be rational in a dynamic environment, an agent’s internal representation of their environment must change as they gain more knowledge. This process is known as *learning* (Buchanan and Mitchell 1978).

When an agent must learn optimal behavior about a dynamic environment through trial and error, they employ what is known as *reinforcement learning* (Kaelbling 1996). It has been shown that reinforcement learning has the ability to model realistic behavior in problems of cooperation and coordination (Erve and Roth 1998).

In mutiagent simulations, when agents communicate with each other or work together on a common goal, agents are often organized into *networks*. For a survey on networks, see (Newmann 2003) . A network is a set of items called *vertices* and connections between them called *edges* (Newmann 2003). A network will be called *static* if edges are never created or removed after the generation of the graph. A *dynamic network* is one in which the edges are created and removed as the network evolves. Many types of relationships can be modeled as networks. For example, vertices could represent people and the edges represent friendship, or vertices could be web pages and the edges links. In this case agents represent the vertices and the edges represent a relationship between the agents.

Many classic network structures exist and are used to model relationships, however recent research has shown that human social networks are not properly modeled with classic network structures but instead can be modeled with various complex networks. The group of complex networks used to model social relationships is known as *social networks*. One type of social network used in this work is known as the *small-world network*. This network and models

for its generation were proposed by Watts and Strogatz (Watts 1999). The small-world network is characterized as having a average shortest path length, where the path length is the smallest number of edges between two vertices. Small-world networks are significant because they appear in many real world social networks. A *social network* is defined as a network in which vertices represent people or groups of people and edges represent some sort of interaction between them (Redner 1998). Examples of such networks include the networks of movie actors, where edges symbolize that two actors performed in the same movie, or the friendship networks of high school students (Watts 1999), (Fararo and Sunshine n.d.).

A second significant network is the *scale-free* network. In such a network, the degrees of nodes follow a power law distribution (Newmann 2003). A vertex's *degree* refers to the number of connecting edges. In the scale-free network, a small group of agents have a high degree, that is, they are connected to a large number of agents. These networks are also observed in real world social settings such as the network of citations between scientific papers, links between web pages on the World Wide Web and network of human sexual contact (Redner 1998) (Barb'asi and Albert 2000) (Liljeros, et al. 2001).

## 1.3 Research Goals

One purpose of this work is to study how realistic decision making can be reproduced. A solution to this problem must rely heavily on the field of psychology. Previous attempts at modeling realistic decision making have three major disadvantages:

- They disregard human intuition
- Agents act perfectly rational
- Agents do not operate under bounded rationality

Another purpose of this work is to study the emergence of social norms in systems of social agents. In multiagent simulations, a *social norm* is defined as a regular behavior that is a solution to a recurrent or continuous social cooperation problem (Dignum 1999). Social cooperation problems are also known as *social dilemmas*. These problems are dilemmas because in such settings, everyone benefits from joint cooperation but an individual has the potential to benefit more by defecting. The problem of how cooperative behavior can emerge in such a setting has fueled research in multiple disciplines such as Sociology, Game Theory, Economics and Artificial Intelligence. Existing attempts to ask the question of how social norms emerge have two major disadvantages:

- In existing models, agents do not learn about their environment; instead they serve as vessels to spread influence. This gives no insight into the adoption of social norms in selfish autonomous agents.
- Only static networks are modeled, which is unrealistic in a real world setting.

## 1.4 Our Approach and Its Contributions

The goal of this research is to reproduce both realistic individual and social behavior. This is to be done in two parts. First a model of the intuitive and deliberative decision making processes is to be created. It is shown that when combining both intuitive and deliberative decision making one can reproduce behavior that is irrational and realistic.

In the second part we attempt to explain the emergence of social norms in a dynamic model setting. This is to be executed in two phases. First an investigation into the emergence of social norms of learning agents in static complex networks is to be done. In the second phase we employ a dynamic network modeled after human relationship networks.

The central thesis of this research is that

> *In multiagent systems, agents are organized in networks. In these networks agents compete with neighbors in order to maximize their utility. When learning agents, operating under bounded rationality, compete in these networks, cooperative behavior emerges even though agents are selfish and attempt to only maximize their own utility. This is because agents are able to break unrewarding relationships and therefore are able to maintain mutually beneficial neighborhoods. This leads to a Pareto-optimum social convention.*

This research has multiple contributions. By employing agents who learn from their environment we provide a more complex setting in which to analyze the spread of influence in social networks and the evolution of the social system. This work also contributes to the fields of Social Simulation by researching the emergence of social norms in networks of selfish agents. In addition, this research contributes to the field of multi-agent learning by showing how networks of agents can evolve to work together in order to maximize their utility.

# 2 Related Work

## 2.1 Individual Decision Making

Many of the great leaps in understanding how humans make decisions were made by the psychologists Daniel Kahneman and Amos Tversky. One of the more notable contributions made by Kahneman was his development of a two-phase human decision making model (Kahneman 2002). When faced with a problem humans either use their intuition to solve it, or they reason about the problem. These two methods of solving problems are very different, are typically used in different situations, and can produce different results. Intuition is "fast, automatic, effortless, associative, and difficult to control or modify", while reasoning is "slow, serial, effortful, and deliberately controlled" (Kahneman 2002). Intuitive thoughts come to mind spontaneously (Kahneman 2002), whereas reasoning about a problem must be done deliberatively.

Because intuition is highly associated with the thoughts that come to mind when faced with a problem, the results of intuition are dependent on what is accessible in a given situation. "Accessibility is the ease of which particular mental thoughts come to mind" (Kahneman 2002). When asked to approximate the height and width of different block configurations (see (Kahneman 2002)) we find an example of accessibility that acts similar to the reasoning and intuitive process. "These perceptual examples serve to establish a dimension of accessibility. At one end of this dimension we find operations that have the characteristics of perception and of the intuitive System 1: they are rapid, automatic, and effortless. At the other end are slow, serial and effortful operations that people need a special reason to undertake. Accessibility is a continuum, not a dichotomy, and some effortful operations demand more effort than others"

(Kahneman 2002). Many factors influence what information is accessible and what isn't. Factors include: stimulus salience, physical salience, selective attention, high emotion or motivation, priming and skill. The acquisition of skill selectively increases the accessibility of useful responses and of productive ways to organize information.

All contained information can come to mind; what does depend on the attributes of the particular situation. "Salience can be overcome by deliberative attention, if you tell yourself to focus and search for a specific object it will enhance the accessibility of its features" (Kahneman 2002). "Absent a system that reliably generates appropriate canonical representations, intuitive decisions will be shaped by the factors that determine the accessibility of different features of the situation. Highly accessible features will influence decisions, while features of low accessibility will be largely ignored. Unfortunately, there is no reason to believe that the most accessible features are also the most relevant to a good decision" (Kahneman 2002).

Another great contribution by both Kahneman and Tversky was their development of Prospect Theory. Prospect Theory embraces the idea that preferences are reference-dependent, and includes the extra parameter that is required by this assumption" (Kahneman 2002) (Kahneman and Tversky 1979). During their experiments they discovered situations which produced a trend of humans acting irrationally. They determined that the models had failed. "We therefore proposed an alternative theory of risk, in which the carriers of utility are gains and losses changes of wealth rather than states of wealth. This theory states that "perception is reference-dependent: the perceived attributes of a focal stimulus reflect the contrast between that stimulus and a context of prior and concurrent stimuli." " The reference

value to which current stimulation is compared also reflects the history of adaptation to prior stimulation" (Kahneman 2002).

Learning in a changing environment requires that agents are able to identify similar states. This means such an agent will have to use a form of pattern recognition. Pattern recognition is a subfield of machine learning where information from the environment is used in conjunction with prior knowledge to infer further information about the environment or the state an agent is in. For a detailed description of pattern recognition see (Bishop 2006).

Prevalent work in the field of reproducing human behavior was conducted by Steven Kimbrough and Ming Lu (Kimbrough 2005). They conducted experiments in which agents using the Q-Learning reinforcement learning algorithm played a number of iterative 2x2 normal form games such as Prisoners Dilemma, Stag Hunt and Chicken. Their results showed that when exploring rationally, agents tended to make decisions that maximized their total wealth extracted as opposed to making economical rational decisions that led to a smaller total wealth extracted. Their experiments showed that in games of risk, such as the Prisoners Dilemma and Stag Hunt, the lower the risk, the more agents were likely to play the Pareto optimal profile. This result is quite profound in that it presents us with an agent that seems to be averse to risk, an agent that seems to be weighing the possible rewards received against the risk. While this is likely not true, it does reproduce human like behavior of aversion to risk (Kahneman and Tversky 1979). Also important is how this work shows reinforcement learning's potential to divert agents from the rational economic solution to the problem, the Nash-equilibrium.

Another prevalent paper that showed reinforcement learning's ability to simulate aversion to risk was written by J. Neil Bearden (Bearden 2001). In his experiments, he evolved

populations of Q-Learning agents using genetic algorithms. The agents competed against each other in the repeated Stag Hunt game. The agents competed in different version of the Stag Hunt with varying degrees of risk. The author found that the number of Pareto-optimum profiles played directly corresponded with the degree of risk.

## 2.2 Agent Based Social Simulations

Research into reproducing human social behavior tends to fall into two categories: one in which different models that superimpose behavior are presented and tested and a second in which researchers study the ways in which a society of agents can converge onto a solution to a common problem. In the first category, models that superimpose behavior are often solutions to very specific problems and don't transition well into different domains. In addition, the models often superimpose the behavior without regard to the agent's payoff.

One example of work in which a social structure is superimposed is work done by Jiang and Ishida (Jiang and Ishida 2007). In this paper the authors present a unique model for representing the way in which strategies are diffused through a network. The authors include the idea of social rank and the power of numbers. While their model presents a new unique way to model realistic social law evolution, the adoption of strategies is guided solely by the rules of their model and not by utility maximization. In addition, their model super imposes social rank; a better solution would be to let such an influential position emerge, for example as a high degree node in a scale free network.

In an experiment performed by Stephen Younger, agents followed a constant rule set governing exchanges which affected their reputation and opinion of others (Younger 2004). From this simple model, he was able to produce agents that reproduced the human social

characteristics of reciprocity of exchange. While the results obtained match human social characteristics, the behavior is governed by a set of rules intended to reproduce the behavior, not to maximize the agent's utility.

Other models have attempted to create agents that behave "realistically" both at the individual and the social level. In a paper published by Paola Rizzo et al., a model is proposed that does such a thing (Paoloa Rizzo 1999). Their model creates believable agents by superimposing personalities and preferences both at the individual level and the social level that affect the agent's goal. While the results obtained are impressive, I believe a better solution is to obtain such results by not superimposing human characteristics, but instead letting the believable behavior emerge from the model.

The second category of research into human social behavior is the study of evolutionary networks. Here researchers investigate the ways in which populations of agents may converge onto a particular strategy. While the research is merited, assumptions are often made that make the experiments unrealistic. For example agents often have no control over whom they play with. Also, agents don't employ selfish reward maximizing decision making but instead often imitate their neighbors. Lastly, agents are often able to see the actions and rewards of their neighbors, which is unrealistic in many social settings.

One model that does incorporate a network in which agents seek to maximize their reward by discontinuing disliked relationships and searching for new ones was defined by Zimmermann and Eguiluz (Zimmermann and Eguiluz 2005). Through this model they have been able to show that cooperative behavior can be sustained when agents may evolve their local network. However, their agents seek to maximize their reward not by exploring their

environment and the consequence of their actions but instead by mimicking the highest rewarding action of their neighbors. This makes an unrealistic assumption about what knowledge is commonly available to agents. In many environments agents are not given this information. Also, their model for discontinuing local interactions is specific to the Prisoners Dilemma Domain and it also makes assumptions that a cooperator would be unwilling or unable to discontinue their relationship with a defector. In addition the authors only use random networks and do not study the emergent behaviors in more realistic complex networks.

Studies that looked at the properties of realistic social networks include work done by Abramson and Kuperman. In their paper they study how evolutionary agents act in small world networks playing the Prisoners Dilemma (Abramson and Kuperman 2001). Their work shows that the emergence of cooperative behavior is greatly dependent on the level of risk in the game and the topology of the network. However, their agents are unable to evolve their local networks and their agents mimic behavior.

Delgado presents us with another work that tests the network role in the emergence of social conventions (Delgado 2002). Here he shows that complex networks are much more efficient and therefore converge onto a social convention faster. However, Delgado employs simple evolutionary agents that are able to observe all of the payoffs of their network. Also, his agents are forced to continue play with their neighbors and are unable to adapt their network to their strategy.

# 3 Model Specifications

## 3.1 Individual Decision Making

Our decision model incorporates Kahneman and Tversky's two-phase decision theory. The editing phase starts from *framing* where an agent decides evaluation criteria based on its attitude toward potential risk, reward and proactive behavior. A classic example would be a man purchasing an automobile; he will frame his criteria on year, model, mileage, color, fuel economy, warranty, and safety record.

Framing leads to another phenomenon referred to as *anchoring*. Anchoring is a psychological term used to describe the human tendency to overly or heavily rely (*anchor*) on some particular information when making decisions. In the above automobile example, the client tends to "anchor" his decision on the odometer reading and year of the car rather than the condition of the engine or transmission. Agents anchor by building selective attention on information. The salience of information *i*, $i \in I$, is determined by Equation 1.

$$\Delta_i = \frac{\sum_c i \text{ was used}}{cardinality(I \text{ was used})}, \; i \in I$$

Equation 1: Salience of information.

where $\Delta_i$ is the frequency that *i* was used under the context *c*. If the salience of *i* is higher than a predefined threshold, *i* becomes the anchored information. *I** denotes a set of anchor information defined by Equation 2.

$$I^* = \{i \mid \Delta_i > \text{threshold}\}$$

Equation 2: Set of anchored information.

*Accessibility* is the ease with which particular information come to mind. In our model, every agent saves past states in memory and clusters them in patterns. When making decisions,

the agent determines the similarity between states only with *I** establishing the relation. This is described by Equation 3.

$$S_t \sim S_m \text{ if } d(S_t, S_m) < \mathrm{D}$$

Equation 3: State similarity.

where $S_t$ is the current state, $S_m$ is a state pattern in memory, $d(S_t, S_m)$ is the distance between $S_t$ and $S_m$, and D is the distance threshold. Those states most similar to the current one are said to be more *accessible* than others.

The next phase is the evaluation phase. There exist two modes of cognitive function: an *intuitive* mode in which decisions are made automatically, and a *deliberative* mode, which is effortful. Our way of modeling intuitive behavior is to assume that, if $S_t$, the current state, is close to a state in memory, $S_m$, then the optimal policy $\pi^*(S_t)$ and $\pi^*(S_m)$ should be close too. Hence, the agent uses an optimal policy that it has employed before in a similar state and updates its state memory by adding the current state. If the policy the agent employed was successful then the reward associated with that policy and its accessibility will be increased.

The slower process of deliberation determines the state similarity across all information available to the agent, *I*, not just that which is anchored, *I**. Traversing its memory an agent attempts to re-optimize a previously used policy stored in memory using Equation 4:

$$\pi^*(S_m) = \arg\max E[\sum_{t=0}^{\infty} \gamma^t R(S_t) \mid \pi], \ 0 < \gamma < 1$$

Equation 4: Updating optimal policy when using deliberation.

where $\pi^*(S_m)$ denotes the optimal policy of the current state $S_m$, *E* means the expected value, $\gamma$ is the time discount factor, $R(S_t)$ is the reward the agent receives when it arrives at state $S_t$, $\pi$ is any possible policy that the agent can choose on the current decision-making point.

In keeping with the notions of *satisficing choice* under their intuitive mode, our agents, do not compute an optimal policy to use in the current $S_t$ if there is a state in the agent's memory $S_m$ that is similar and the policy utilized under that state can be used once again.

## 3.2 Social Network Evolution

### 3.2.1 Network Structure

Consider a fixed population of agents, $P$. Each agent $p \epsilon P$ has a set of agents $N(p)$ referred to as the agent's neighborhood. The size of each agents neighborhood is referred to as their *degree*. The neighborhood is such that if $q \epsilon N(p)$ then $p \epsilon N(q)$. Therefore if $q$ is removed from $N(p)$, then $p$ is also removed from $N(q)$. We employ an undirected graph but the ideas can be easily extended to a directed graph. The network that contains all agents is known as the population graph $G$. The *average shortest path* of the graph G is defined as the smallest number of links between a pair of agents, averaged over all pairs of agents. In this study the structure of G is varied. Four different network structures are used in this work.

(1) The first and perhaps simplest is the complete graph $k_n$ where $n$ is the number of nodes in the graph. In this graph every agent is linked to every other agent. Thus the average shortest path is 1 and the degree of each node is $n$. While simple and easy to implement the complete graph is unrealistic and doesn't properly model human social relationships. Instead it is used to serve as a basis for comparison. A complete network $k_{10}$ is depicted by Figure 1.

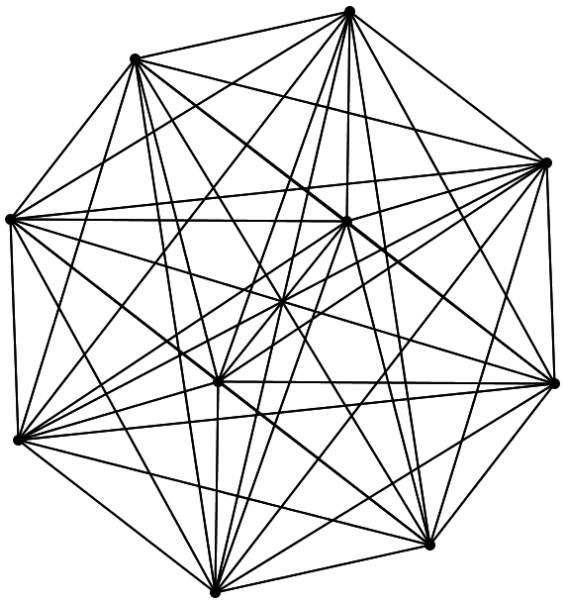

Figure 1: Complete network.

In order to more closely approximate the structure of a real world social network we employ the small-world graph and the scale free graph. In order to construct a small-world graph, we must first understand another simple network structure known as the lattice.

(2) A lattice, $C_{n,k}$ where $n$ is the number of nodes, is a ring of agents where each agent is connected to its $k$ nearest neighbors (Newmann 2003). In this network, the degree of each agent is $k$ and the average shortest path shrinks as $k$ is increased. Figure 2 depicts a lattice ring $C_{100,5}$.

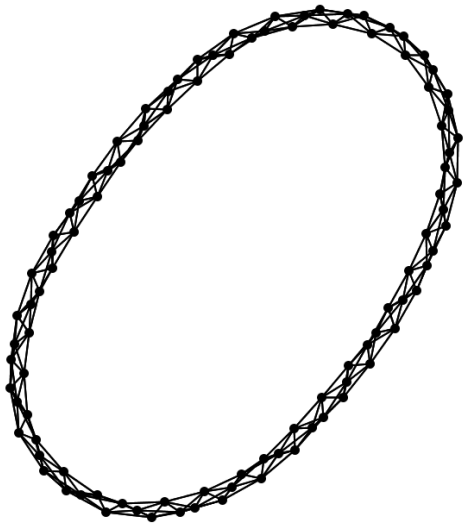

Figure 2: A Lattice Ring

(3) The small-world network modes, $W_{n,k,p}$, was proposed by Watts and Strogratz begins by creating a ring lattice of size $n$ with a small $k$ (Watts 1999). Then the graph is modified by taking a small percentage $p$ of the edges and moving one of the ends (changing one of the connected vertices). The small-world graph is characterized as having a large clustering coefficient and a small average shortest path between pairs of agents. The clustering coefficient is defined as the probability that two neighbors of an agent will also be neighbors of each other. A small-world $W_{100,6,0.05}$ network is depicted by Figure 3.

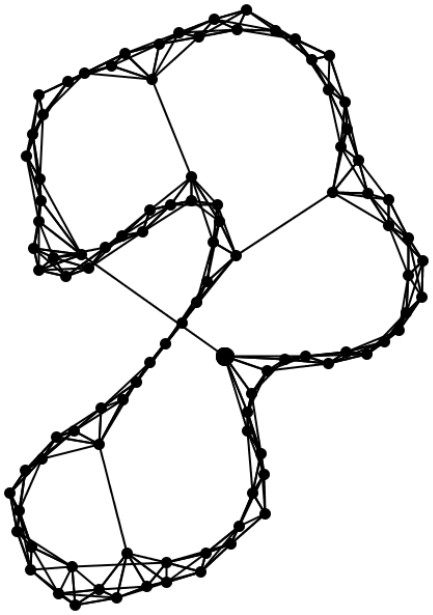

Figure 3 A Small World Network

(4) The last structure explored in this work is known as a scale-free graph, $S_{n,m_0,m,p,q}$. These are networks whose degree distribution follows a power law. This means that a small percentage of vertices contain a large percentage of the edges. This network is also characterized as having a small average shortest path. The most well established model for generating scale-free networks is the Barabasi-Albert extended model (Albert and A.L. 2002).

(a) Begin with $m_0$ isolated nodes.

(b) With probability p, add $m < m_0$ links. The starting point is chosen uniformly, the ending point is chosen according to the probability distribution described in Equation 5.

$$p(d_i) = \frac{d_i}{\sum_j d_j}$$

Equation 5: Probability of adding agent to local network in Scale-Free network

(c) With probability q, m edges are rewired. This is done by first selecting a node at random with equal probability. Second select a link at random with equal probability. Remove one end of the link and rewire it to the selected node.

(d) With probability 1-p-q, add a node with m links. The links connect to nodes with the probability distribution described in Equation 5.

(e) Repeat until the network contains $n$ nodes.

Figure 4 depicts the scale-free network $S_{100,3,1,0,0}$.

Figure 4: A Scale-Free network

### 3.2.2 Highest Rewarding Neighborhood

We begin by defining a social norm as a regular behavior accepted by a majority of the population that is a solution to a recurrent or continuous social cooperation problem (Dignum

1999). Therefore a social norm will be reached if a large majority of agents repeatedly choose the same action (Delgado 2002). Here we wish to study the conditions in which the social norm adopted is the Pareto-optimum solution. We begin by describing how an agent updates their strategy then end by defining the Highest Rewarding Neighborhood rule which is used in the best self interest of the agents and maintains a neighborhood that fosters cooperative behavior.

(1) At each time-step, every agent must choose an action $a$ to perform where $a \epsilon A$. Which action the agent chooses is governed by the agent's strategy $\rho$, where $\rho$ is a probability distribution over all possible actions. After performing action $a$, each agent transitions to a new state and receives a reward $r$. An agent's reward is dependent on the actions of both itself and its neighbors where the reward received by agent $i$ with neighbor $j$ is $r_{ij} = u(a_i, a_j)$. This is referred to as the payoff matrix shown in Table 1. In this model agents are unable to view the actions of their neighbors. In addition, agents don't know the payoff matrix and must learn it through exploration. Therefore, agents can't infer the actions of their neighbors.

| Agent A's Actions/ Agent B's Actions | β1 | β2 |
|---|---|---|
| α1 | $u(\alpha1, \beta1), u(\beta1, \alpha1)$ | $u(\alpha1, \beta2), u(\beta2, \alpha1)$ |
| α2 | $u(\alpha2, \beta1), u(\beta1, \alpha2)$ | $u(\alpha2, \beta2), u(\beta2, \alpha2)$ |

Table 1: Payoff Matrix

(2) Thus the total reward agent $i$ receives is $r_i = \sum_j u(a_i, a_j)$. Lastly after transitioning from state $s$ to state $s$', an agents updates their strategy with their learning algorithm. In the experiments presented, we test employ the Q-Learning algorithm. An agent updates their

strategy by increasing the probability of selecting the chosen action if it results in a high reward and decreasing the probability of selecting the chosen action if it resulted in a low reward.

(3) At the end of each timestep the agents review their relations. Here we employ the Highest Rewarding Neighborhood rule. In an attempt to maximize their own utility, each agent only wishes to maintain relationships that are rewarding. Therefore each agent computes the average reward received from play with each neighboring agent using Equation 6.

$$\bar{r}_{ij} = {\sum r_{ij}}/{t}, \text{ where t is the number of timesteps, } j \in N(i)$$

Equation 6: Average reward earned from play with neighbor.

Where $N(i)$ is the neighborhood of agent $i$. This is compared to the average reward received from play with every opponent computed with Equation 7.

$$r_{total} = {\sum_{i \in N(i)} \bar{r}_{ij}}\Big/{|N(i)|}$$

Equation 7: Average reward earned from play with every neighbor.

If the average reward is lower than the total average reward by a predetermined threshold value $\tau$, then the relationship is deemed unrewarding and is broken. This is determined with Equation 8.

$$\bar{r}_{ij/} r_{total} < \tau$$

Equation 8: Relationship evaluation.

Each agent is removed from the others neighborhood. Then, the agent who chose to end the relationship is linked to a new agent, preserving the total number of edges in the graph. The identity of the newly chosen neighbor agent is determined by the structure of the network. In order to maintain the structure and characteristics of the complex networks, a network specific function is used to generate the probability of each agent becoming the new

neighbor. For the scale-free networks we use a probability, $p(d_i)$ calculated with Equation 5, that corresponds to the degree of the agent (Albert and A.L. 2002).

$$p(d_i) = \frac{d_i}{\sum_j d_j}$$

When using a small world network, we use a probability, $p(b)$, calculated with Equation 9, that corresponds to the number of mutual friends in each agent's network. This has been shown to maintain the features of the small-world network (Jin, Girvan and Newman 2001). Here $M$ represents the number of mutual neighbors in the two agents' local networks and $|G|$ represents the total number of agents in the graph.

$$p(b) = \frac{M(a,b)}{|G|}$$

Equation 9: New neighbor probability in small-world network

Finally we are able to define the Highest Rewarding Neighborhood Rule.

**Definition HRN**: According to the Highest Rewarding Neighborhood rule, and agent will only maintain a relationship iff the average reward earned from that relationship is no less than a specified percentage of the average reward earned from every relationship.

This rule essentially states that agents will only maintain a high rewarding neighborhood. This results in a cooperative neighborhood in which Pareto-optimum strategies are played.

# 4 Individual Model Experiment

## 4.1 Individual Decision Making

In order to implement the proposed model, two sub-problems must first be solved. First, how does an agent find states that are similar to the one the agent is currently in? Second, how does the agent learn from past experiences? This requires an implementation of a pattern matching algorithm and a learning algorithm.

The pattern matching algorithm chosen to be used is based upon the k-nearest-neighbor algorithm. Each state is represented as a point in an n-dimensional space, where each dimension corresponds to an attribute of the state. For example, if the state was represented as an agent's x and y coordinates then the state would be represented as a point in a two dimensional space. To find similar states, an area is constructed that centers on the point in space to search for while extending in each dimension a specified distance where a shorter distance will require a more similar state and a larger distance a less similar state.

The learning algorithm selected is the Q-Learning algorithm. This was selected due to previous work that has shown it is effective in reproducing human like behavior (Newmann 2003). This learning algorithm is a form of reinforcement learning where weights are modified to more closely reproduce desired responses.

This algorithm was first proposed by Watkins (Watkins 1989). Q-Learning works by assigning each state and action pair a Q-Value. The likelihood of an agent choosing a particular action at a state is dependent on the value of the Q-Value, the higher the more likely. Equation 10 states how Q-Values are updated.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t(s_t, a_t)[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

Equation 10: Q-Value Update Function

In this formula $Q(s_t, a_t)$, represents the Q-value of action $a$ at state $s$, $Q(s_{t+1}, a)$ represents the Q-Value in the new state, $r$ is the reward of performing $a$ at $s$, $\alpha(0 < \alpha \leq 1)$ is the step size and $\gamma(0 \leq \gamma \leq 1)$ is the time discount factor. If each action is executed an infinite number of times on each state with a *decaying* $\alpha$ value, then the Q values converge to $Q^*$, *where* $Q^*$ is the expected discounted reward of taking action $a$ in state $s$ and taking all following actions greedily (Watkins 1989).

When Q values are near $Q^*$ it is efficient if the agent chooses actions greedily, meaning the agent chooses the action with the highest Q-Value. However, this is inefficient if the agent is still learning. While learning, the Q Values will not be near $Q^*$, therefore, if the agent was to choose actions greedily, the agent would likely converge onto a suboptimal policy. Therefore when learning we are faced with the same problem as the one armed bandit problem (Robbins 1952). This is a problem of whether or not one should exploit the current best known payoff or explore for a higher payoff. Many solutions exist to this problem. These are known as exploration functions. These define the probability of choosing action $a$ at state $s$.

One such exploratory function is a frequency based exploratory function. In this exploratory function the number of times each State Action pair is used is recorded. State Action pairs that have been used less than a set number of times are given the highest possible reward. If two State Action pairs tie then the one used less frequently is chosen. If two State Action pairs tie for Q-value and frequency then an action is randomly chosen. This randomness

is what differs from each run.

A second well established exploratory function is known as the Cooling ε -Greedy function (R. S. Barto 1998). Here the agent chooses a random action with probability ε while choosing the action with the highest Q-Value all other times. At each step the probability ε is decreased by a percentage equal to the defined cooling rate. This exploration is effective because by decreasing the ε value it allows for the possibility of convergence upon a specific strategy.

In order to incorporate Intuition and Deliberation with Q-Learning, two separate Q-Learning tables are used, one for making deliberative decisions and one for making intuitive decisions. The deliberative Q-Learning table has been updated after every state transition the agent has undergone. This allows the agent to make deliberative decisions that take into account everything the agent knows about their domain. The second Q-Learning table has only been trained on selected information. At each state the agent searches their memory base for memories in which the state that the memory occurred in differs from the current state by no more than a distance D. All Q-Values in the Intuition table are reset to 0. This subset of memories is then used to train the Intuition table. This allows an agent to make a decision based only upon a selected amount of information. This process is illustrated by Figure 5.

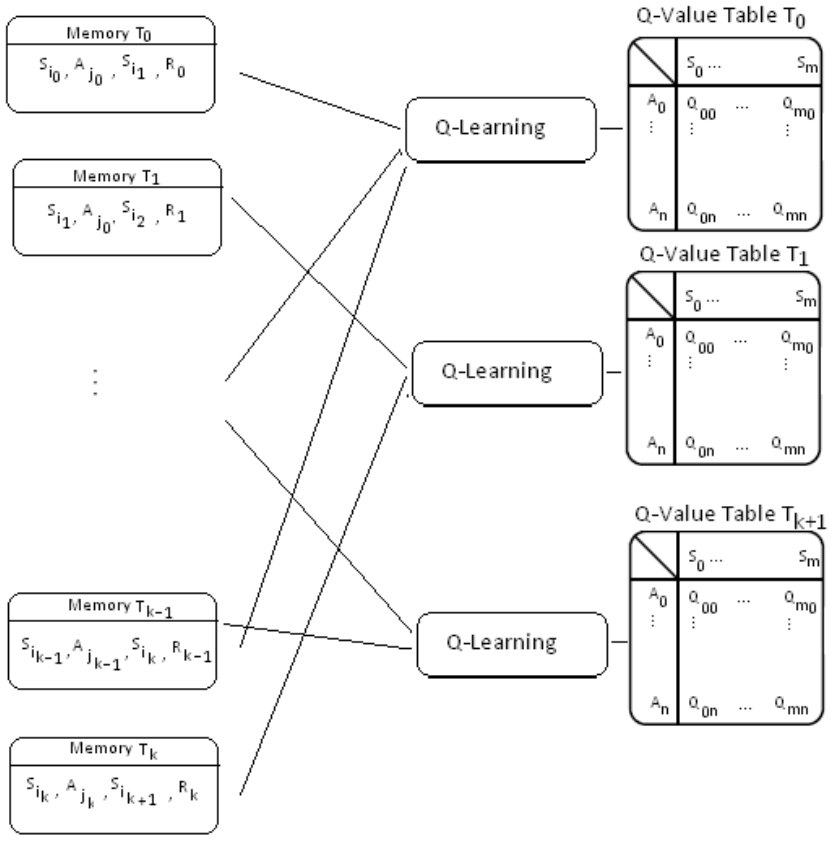


Figure 5: Illustration of the intuitive decision making process

A problem to be solved is the external verification of the proposed method for human decision-making (Keiki Takadama 2008). This crucial step provides many other benefits as well as allowing for us to say with assurance that our model demonstrates human-like decision making, external verification also allows for a quantifiable measurement of the success of the proposed model. Also a measurement will allow for the comparisons of our model with other models as well as provide measurements of any improvements we make to the model. Assuming that our model does correctly model human like decision-making, verification will assure us that our implementation is correct.

The experiment domain to be used for the verification of our decision making model will be the Ultimatum Bargaining Game. There are many reasons for choosing this domain. First, the game has been well researched (Prasnikar and Roth 1992) (Guth and Tietz 1990). There is a great deal of experimental data available of tests run on human subjects. This is crucial for our method of validation by external verification. Lastly, experimental data from the game shows a strong trend in which the human test subjects consistently perform differently from the rational agent. Not only do the human test subjects perform differently but their decisions tend to fall within a narrow well defined range. This provides a very quantitative way for us to measure the success of our human-like decision making model.

The Ultimatum Bargaining game is an extension to the Bargaining Game in which there is only one iteration. There is a quantity of money Q that is to be divided among two agents where both agents wish to maximize their amount of Q. Agent 1 decides on a number x ($x <= Q$) where Agent 1 will receive amount x and Agent 2 will receive amount $Q – x$. Agent 2 then decides whether or not they will accept the division. If the agent accepts then they receive the specified amounts, if not then both agents receive 0 (Guth and Tietz 1990).

A Nash-equilibrium exists in this game when x is the largest possible amount. If we assume that both agents are rational then Agent 2 will accept any amount over nothing. Therefore Agent 1 should pick an amount x that is as large as possible where $Q – x > 0$ in order to maximize the amount of Q they receive (Guth and Tietz 1990).

The second domain used to test the effectiveness of Intuition Deliberation is the Bargaining Game. This domain was chosen because it offers a more complex and natural domain than the Ultimatum Game. The Bargaining Game is very similar to the ultimatum game

the only difference is that the negotiation goes back and forth until either agent *a* or agent *b* accepts an offer or until they run out of steps. In the latter case, both agents receive 0 . An iterative implementation of the Bargaining Game is used in the following experiments. In each iteration the agents are given a set number of steps, known as the maximum negotiation size, to come to an agreement. The agents do not know how many iterations they are to play nor how many negotiation steps they have to play each iteration.

## 4.1 Individual Decision Making Experimental Results

Experimental data from tests where human subjects are used has shown that humans often act irrational (Prasnikar and Roth 1992), (Guth and Tietz 1990). These tests have shown that human subjects consistently pick an amount x that is 40% - 60% of the amount of Q (Guth and Tietz 1990). This provides us a narrow range of results to attempt to reproduce that would be unattainable with rational agents. This is depicted by Figure 6.

Figure 6: Human player vs rational player in the Ultimatum Game

An iterative implementation of the Ultimatum Bargaining Game is used in the following experiments. In this implementation each agent's states are dependent on the other agent's previous action. Agent  begins in either an accept state or in a not accepted state. This is determined by whether or not their previous split offer was accepted by Agent  . At the beginning of the experiment, time 0, agents are assigned a random state. Agent a can perform 11 different actions. Their actions are the split values 0 – 10. Agent  's state is determined on the split value chosen by Agent  . They can be in states 0 – 11. Their actions are to accept or not accept the split value.

```
TwoPhaseDecesion(State S)
{
   anchoredMemories = null;
   for all memories m
      if (m.S_it − S)> D
         add m to anchoredMemories;
   if size of anchoredMemories > threshold
   {  // make intuitive decision
      I_T = null; // Create a new blank q-value table
      for all anchredMemories am
           update I_T with am using Formula 5;
      action a = Boltzmann(I_T);
   }
   else  // Deliberative Decision
      action a = Boltzmann(D_T);
   create new memory to store S_it, a, S_it+1 and reward;
   update D_T with new memory using Formula 5;
   add new memory to set of all memories;

   return action a;
}
```

Figure 7: Two Phase Decision Making Algorithm

Figure 7 describes the two-phase decision algorithm. An agent begins each step by creating a set of anchored memories. The anchored memories are selected if their state $S_i$ is within a threshold *D* to the agent's current state *S*. In order for an agent to make an intuitive decision, it is required that there be equal to or more than a specified number of anchored memories, this number is domain specific. Because of this constraint, agents begin each experiment by making deliberative decisions until enough memories are created. If the agent has enough to make an intuitive decision the learning algorithm, in this case Q-Learning, updates the Q-Values of a newly created *I_T* (intuition table) by applying the data stored in the memories to the update function. These memories are applied in the order they occurred in time. With this new matrix, the agent uses its exploratory function, in this case Frequency Based exploration function to choose a new action. If the agent does not have enough memories the agent uses deliberation. The agent will use *D_T*, the deliberative Q-value table, to choose a new action. This deliberative table has been updated with every state transition the agent has undergone. After the new

action is performed the agent transitions to a new state and receives a reward. A new memory is then created that stores this information which is then used to update the deliberation table and then stored in the agents set of memories.

An example of an iteration of play is as follows. Consider two agents $a$ and $b$ who have played $t-1$ iterations and thus are at timestep $t$. Agent $a$'s state is dependent on the action agent $b$ selected at timestep $t-1$. For this example we'll say agent $b$ selected "accept". Agent $a$ begins timestep $t$ by searching their knowledge base for each memory in which their initial state was "accept". If the number of memories found is greater than the number required to make an intuitive decision then a new Q-Table is created. All values in this table begin at 0. Then the information contained in each memory is used with the Q-Learning update formula, Equation 10, to update the values in the Q-Learning table. Next agent $a$ chooses an action that corresponds to the highest Q-Value, for example "5". Agent $b$ now transitions to state "5". Here Agent $b$ create a new memory where the initial state is the state they were in at timestep $t-1$, their action is accept, the reward, since they selected accept, is equal to agent $a$'s action at timestep t-1, and their new state is "5". After this memory is created, the information it contains is used to update the deliberation Q-Table. Agent $b$ then searches their knowledge base for all memories in which their initial state was "5". For this example, we'll assume that the agent doesn't have enough of these memories to make an intuitive decision. Then the agent uses the deliberation Q-Table to select their action. This action corresponds to the highest Q-Value in the column with initial state "5", we'll assume agent $b$ selects accept. Agent $a$ then transitions to state accept, creates a new memory and then timestep t+1 begins.

In all experiments, the anchored memories, or memories that are used when making intuitive decisions, differ from the current state with distance D = 0. This means that all intuitive decisions are based off of past experiences from when the agent was in the exact same state. In these experiments the number of possible states is small; therefore, with a D > 0, the number of memories used for Intuitive decisions would be near or equal to the number used in Deliberative decisions.

Presented are the results of a set of experiments where the Q-Learning learning rate and discount factor are varied as well as the number of intuitive and deliberative decisions. It is important to note that in this experiment and in all experiments presented, the agent's do not know the payoff table initially and must learn it by exploring their environment. In all experiments the two agents play the game for 1500 iterations. All of the results shown below are taken from the last 500 iterations of play. This gives the agents 1000 iterations to learn and gather information about the domain. The results presented below are the average of 100 runs.

The first experiment shows the effect the number of intuitive and deliberative decisions has. In Figure 8 the effect of a large number of intuitive decisions can be seen. This experiment was run with a learning rate of 0.1 and a discount factor of 0.75. The number of intuitive and deliberative decisions are varied by changing the number of memories required to make an intuitive decision. With no memories required to make an intuitive decision the agent always makes intuitive decisions that are limited in the amount of information they use. This results in a scattered distribution of split convergences. However, when requiring agents make both intuitive and deliberative decisions the agent is able to learn about their domain before making intuitive decisions resulting in a higher reward. This results in a more consistent convergence

around the values 4 to 8. This convergence range is similar to the results from human ultimatum game experiments (Prasnikar and Roth 1992) (Guth and Tietz 1990) (Von-Neumann and Morgenstern 1944). Analysis of the results using all deliberation and the result using both intuition and deliberation, conducted with the T-Test show a standard deviation 21.297 and a P-value of 0.9731, thus the differences are insignificant. This is due to the fact that both experiments employ a Q-Learning algorithm with a non-decreasing learning rate, which makes so that agents making deliberative decisions are unable to make rational decisions.

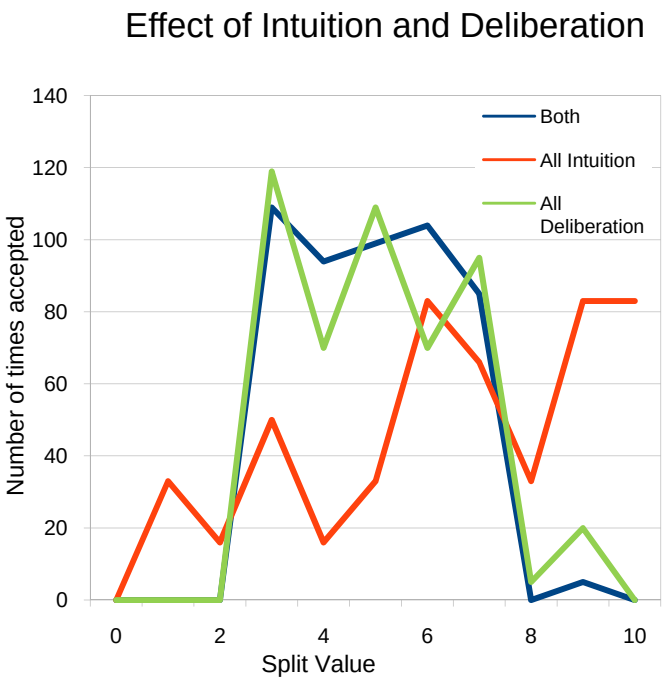


**Figure 8: Simulation that shows the effect of a large number of intuitive decisions**

In Figure 9 the results of two experiments are shown that demonstrate the effect of a high number of deliberative decisions. In both experiments the required number of memories in which the agent was in the same state as the present state required to make a intuitive decision is set very high. This means an agent will make more deliberative decisions. In one

experiment the minimum number of memories is set so high that the agent never makes an intuitive decision. As represented in Figure 9, when the required number of memories is set high the agents tend to split among the same range as when making intuitive decisions, however as the number of deliberative decisions increases the agents tend to converge onto specific values more often. Note that these values are inconsistent with the value a rational player would split on. The reason for this difference is that in these experiments, in order to better compare our work with previous work, the learning rate remains constant. It has been shown that the Q-Values will only converge onto the actual rewards when the learning rate is slowly lowered (Watkins 1989). Without Q-Values that accurately reflect the true rewards, an agent is unable to make a perfectly rational decision. While the difference between the two experiments may seem significant, analysis shows that the two results have a standard deviation of 21.840 and a P-value of 0.9705. Thus the results obtained from agents employing a high number of deliberative decisions are very similar to agents employing all deliberative decisions. Again this is due to the fact that agents are implemented with a Q-Learning algorithm that has a constant learning rate.

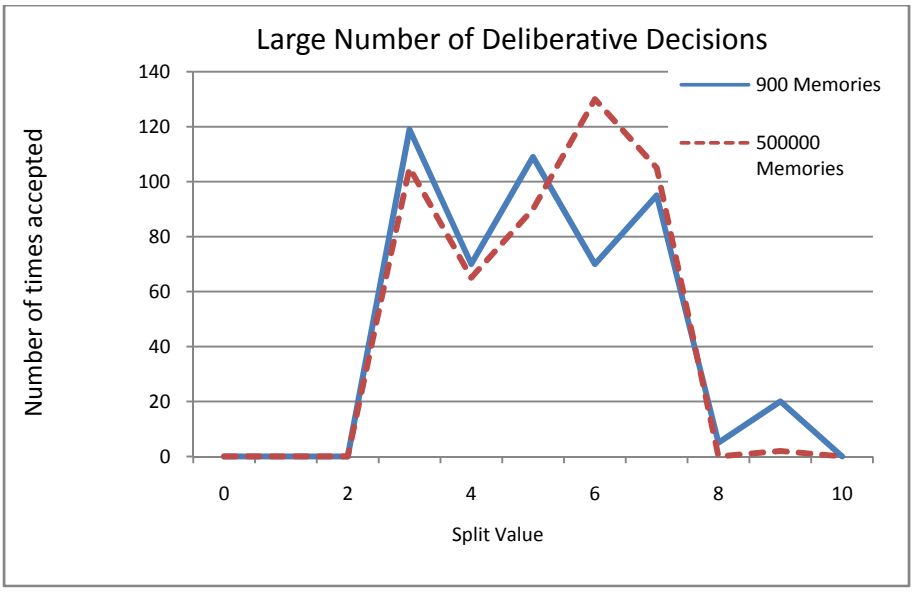


Figure 9: Simulation that shows the effect of a large number of deliberative decisions

The next experiment demonstrates the effect of the discount rate. In these experiments the learning rate is set to 0.1 and the minimum number of memories is set to 500. Once again the results are the average of 100 runs each with 1500 iterations where the results of the last 500 iterations are recorded and averaged.

When making intuitive decisions, the amount of information about states that are not similar to the state in which the agent currently resides in is restricted. Therefore with a high discount factor, when the Q-Value of a state is updated it puts a lot of weight in the Q-Value of a future state without knowing much about it. This results in a wide spread of accepted split values. As the discount factor decreases and the Q-value as well as the reward of the current state action are weighed heavily the agent converges more consistently on a single split value.

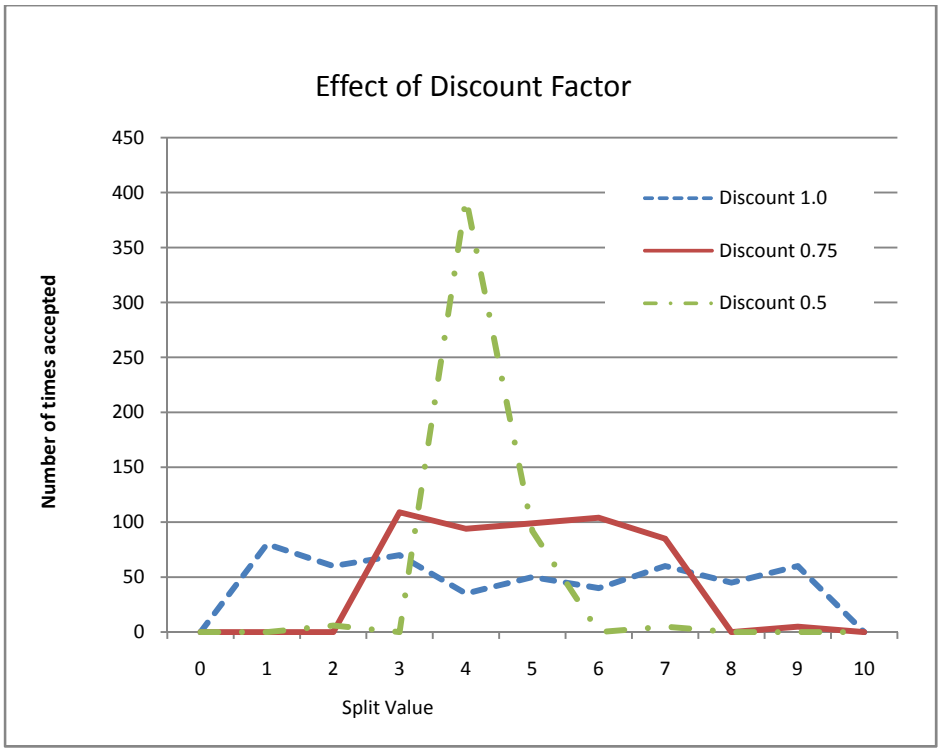


Figure 10: Simulations where the discount factor is varied

The final experiment tests the effect that learning rate has on the performance of the agents. In this experiment the number of memories used is 500 and the discount factor is 0.75.

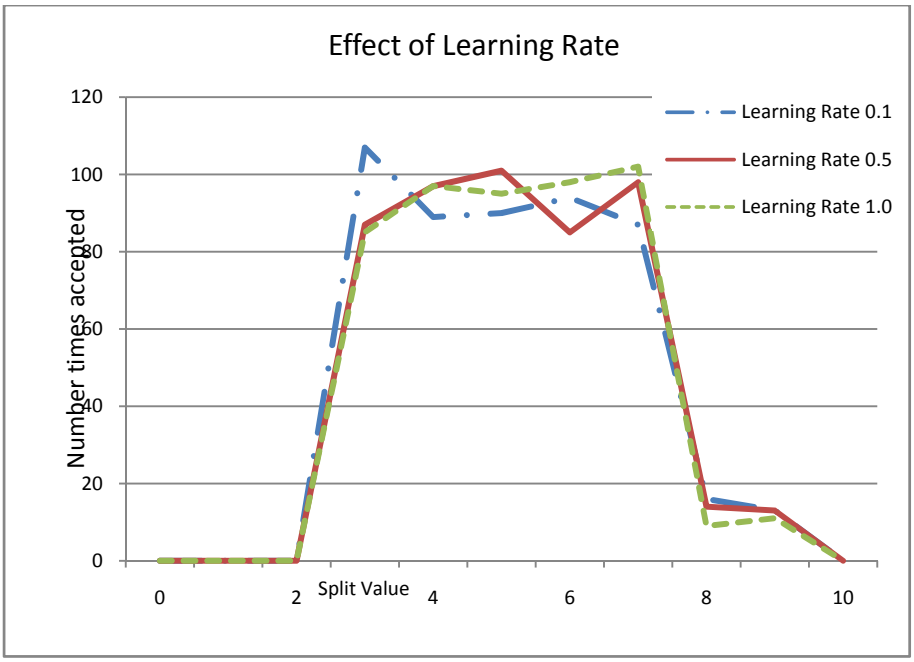


Figure 11: Simulations that show the effect of learning rate on intuitive and deliberative decisions

Figure 11 demonstrates that the learning rate has a negligible effect on the overall performance of the agents. The reason for this is likely due to the 1000 iterations the agents play to learn about the domain before their actions are recorded. This 1000 iterations gives the agents enough time to learn about their environment even with a small learning rate. Figure. 12 shows the human subject results (Takadama 2008).

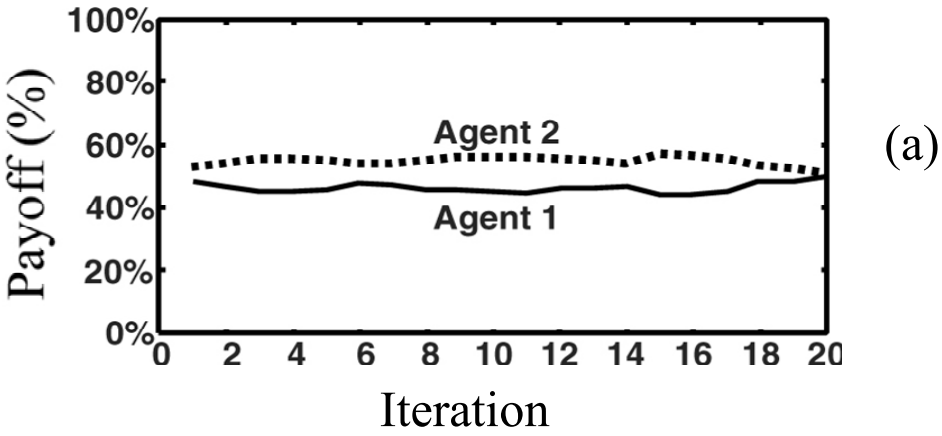


Figure 12: Human subject results from the Bargaining Game

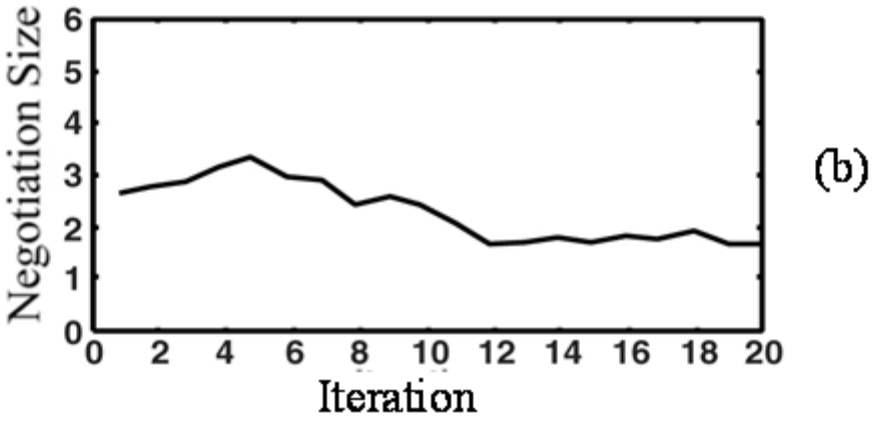


Figure 13: Human subject results from the Bargaining Game

We next present our Bargaining Game results, displayed in Figure 14 and Figure 15. In our experiments, the max negotiation size is 6, the learning rate is 0.1, temperature T=0.5 and change_rate=0.000001. If the agent accepts an offer or has a offer accepted or runs out of negotiation steps before coming to an agreement, a time discount factor of 0 is used since the agent will not be transitioning to another state, in all other cases a time discount factor of 1.0 is used. In these experiments 300 memories are required to make an intuitive decision. Two agents play 5000 games against each other maintaining roles, meaning that agent *a* always gives the first offer. All results presented are the average of 30 runs.

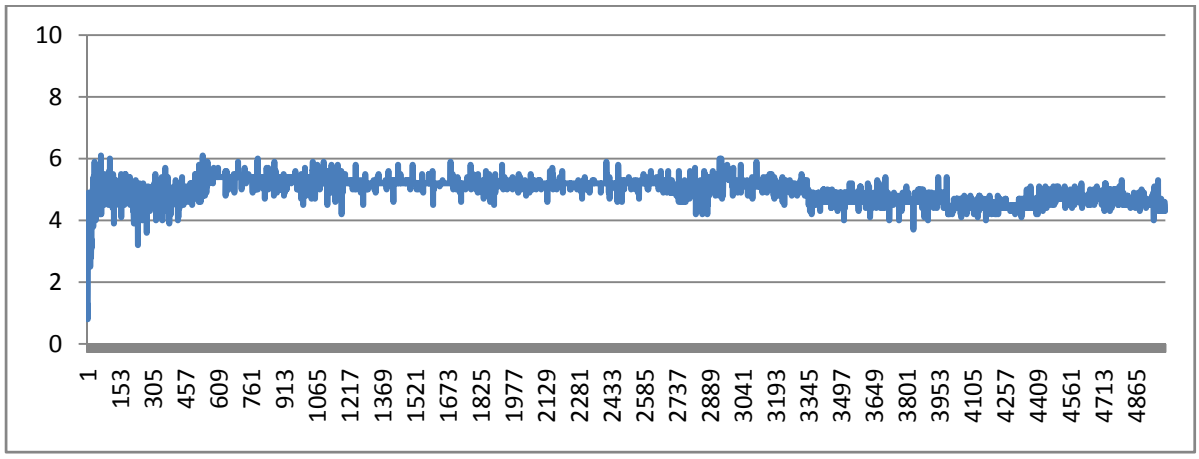


Figure 14: Payoff received by Agent A in Bargaining Game

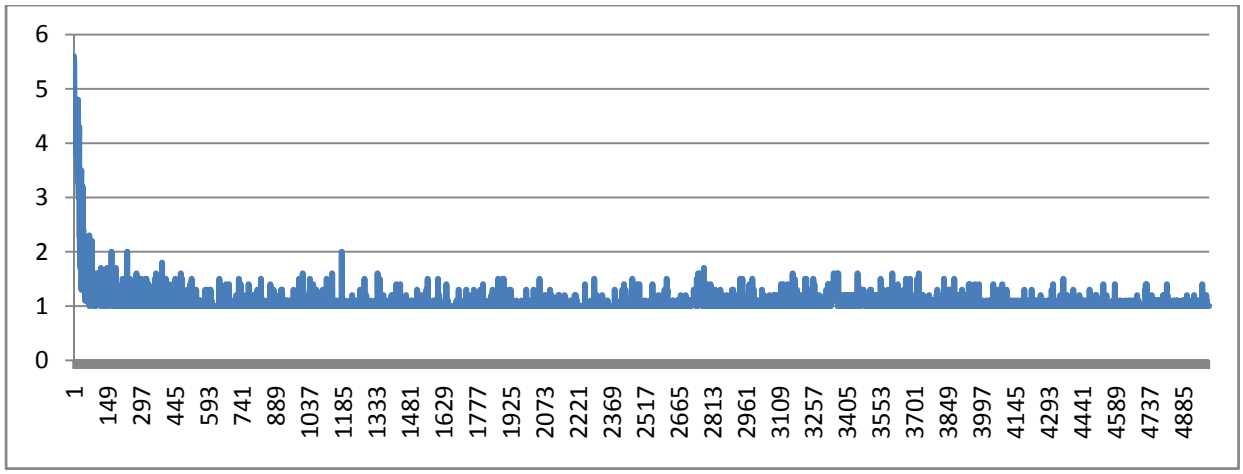


Figure 15: Negotiation size

Figures 14 and 15 are both the results of the simulation. Figure 14 shows a quick convergence of the amount earned by agent *a* per time step to 5. Figure 15 shows a quick drop in negotiation size to about 2 and the negotiation size slightly decreases as iteration increases. Both results agree with the result from the human subject experiment.

# 5 Social Model Experiment

We begin by first describing each agents update phase in Figure 16. This function is run on every agent at each time step.

```
Social Update()
{
    if (random < e)
        Pick random action with uniform probability;
    else
        choose the action with the highest Q value;
        if there are multiple actions with maxQ
            randomly choose one of maxQ actions;

    Move to new state;
    Receive reward;
    Update Q-Value with previous state-action transition;
    Evaluate relationships and update local network with the HRN rule (only in the
        case of dynamic network);
}
```

Figure 16: Social Update Function

Three different domains are used to test the emergence of social norms. One, the Pure Coordination Game, is used to test the effect of the network structure on the rate at which information flows through it and norms evolve (Lewis 1969). These experimental results are then compared with the results of past work (Delgado 2002). The Pure Coordination game is a simple game in which agents receive a reward in the event they choose the same strategy and a penalty in the event they choose different strategies. Table 2 shows the payoff matrix for the Pure Coordination Game used.

| | Cooperate | Defect |
|---|---|---|
| Cooperate | 1,1 | -1,-1 |
| Defect | -1,-1 | 1,1 |

Table 2: Payoff Matrix for Pure Coordination Game

Notice that there are two equal Nash Equilibrium, (Cooperate, Cooperate) and (Defect Defect). Notice that the optimal strategy depends on the strategy of an agent's neighbor. However, when this game is played with multiple neighbors, the optimal strategy is the strategy adopted by the majority of an agent's neighbors. The rate at which this majority strategy is adopted is studied for different network types.

In order to test social norms we employ social dilemma games. These are games in which the Pareto-Optimum solution is also a risky solution. The two games used are The Prisoner's Dilemma and Stag-Hunt.

Prisoner's Dilemma is one of the most classic and well studied problems in Game Theory (Kaminski 2004). The story goes like this: Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated both prisoners, visit each of them to offer the same deal. If one testifies ("defects") for the prosecution against the other and the other remains silent, the betrayer goes free and the silent accomplice receives the full 10-year sentence. If both remain silent, both prisoners are sentenced to only six months in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner

must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act? While the story of two prisoners being offered such a deal is fairly unrealistic, real world equivalents of the Prisoner's Dilemma do occur. Examples include the ways in which natural resources are used (Kaminski 2004).

The game is interesting because the Nash-equilibrium strategy is the Pareto-suboptimum solution, (Defect, Defect). The Pareto-optimum strategy is risky because any agent who plays it could be given a low reward and their opponent a high reward in the event that the opponent defects. Table 3 shows the pay-off matrix The Prisoner's Dilemma that is used in our experiment.

| | Cooperate | Defect |
|---|---|---|
| Cooperate | 3,3 | 0,5 |
| Defect | 5,0 | 1,1 |

Table 3: Payoff Matrix for The Prisoners Dilemma

The second game used is Stag-Hunt (Skyrms 2004). Stag Hunt is another two player normal form game. The story goes that two hunters are out to hunt a stag. They are to both stand and guard an area waiting for the stag. If they both stand and wait they will definitely catch the stag and receive the greatest reward. However, if one of them defects and goes after a hare then that player will receive a smaller reward and the other player, if cooperating, will

receive nothing. If both players defect and go after the hare, they will both catch a hare and receive a smaller reward.

| | Cooperate | Defect |
|---|---|---|
| Cooperate | 4,4 | 1,3 |
| Defect | 3,1 | 3,3 |

Table 4: Payoff Matrix for The Stag Hunt

In this game there are two pure strategy Nash Equilibrium, (Cooperate, Cooperate) and (Defect, Defect). (Cooperate, Cooperate) is the Pareto efficient solution; this differs from the Prisoner's Dilemma where the Pareto efficient strategy is not a Nash Equilibrium. Here the best of the two pure strategies is determined by the players approach to risk and maximum benefit. If a player chooses to hunt the hare, then they are guaranteed a playoff of 3 regardless of the other players choice, however if they choose to hunt the stag then they have an opportunity to maximize their payoff, but run the risk of getting the minimum payoff.

For the Pure Coordination Game, we test our agents in a complete network, a lattice ring network, a scale-free network and a small world network. This is done using static networks, or without allowing agents to change their local neighborhood. For the two

Cooperation games we test static complete, lattice ring, scale free and small world. In addition we also employ dynamic versions of the small world and scale free networks.

Each network is generated with the network analysis program ORA[1] with the exception of the Scale-Free network, which is generated with Pajek[2]. These programs are used in an attempt to both speed development time as well as to insure validation of the network structures. Each network generated contains 1000 agents. We employ a Complete network with $n$=1000. We also employ a lattice ring network with $k = 12$, that is each agent is linked to their 12 nearest neighbors. We use a small world network generated with the Watts-Strogatz algorithm with $p = 0.1$ and $k = 12$, meaning a lattice ring of $k = 12$ is generated, then with probability $p = 0.1$, links are rewired to a distant agent. Lastly a scale-free network is generated with the Albert-Barabasi extended algorithm using a $m_0 = 4, m = 2$ and $p = q = 0.4$. In addition agents employ the Q-Learning algorithm in conjunction with the $\varepsilon$-Greedy exploration function.

| | |
|---|---|
| Network Type | Dynamic Scale-Free, Dynamic Small World,<br>Static Scale-Free, Static Small World, Static Lattice-Ring, Static Complete |
| Normal Form Game | Prisoners Dilemma, Stag-Hunt, Pure Coordination |

Table 5: Experiment Settings

[1] ORA: http://www.casos.cs.cmu.edu/projects/ora/
[2] Pajek: http://pajek.imfm.si/doku.php

To test the social model we have our agents play different 2x2 normal form games with their local neighbors. The Pure coordination game is used to test whether or not the agents can coordinate their actions together and to validate the system. The cooperation games are to test whether or not the agents can work together to maintain a cooperative behavior and therefore maximize the networks long term average payoff. Our agents employ the Q-Learning algorithm with the cooling $\varepsilon$-greedy exploration function. For the experiments presented we use the $\varepsilon$-greedy exploration function with a starting ε of 0.2 with a cooling rate of 0.001. Also, we use the Q-Learning algorithm with a learning rate of 0.1 and a time discount factor of 0.5. In the experiments presented agents do not know the payoff table initially and must learn it by exploring their environment. In addition agents do not view the actions or rewards of their neighbors.

At the beginning of each step, every agent simultaneously chooses their strategy for that turn. Then each agent receives a reward that is the summation of all rewards earned from play with their neighbors that time step. It is important to note here that agents are only aware of their own actions and the rewards they receive. This separates our work from a majority of the work on evolutionary networks in which agents are often able to witness the actions and rewards of their neighbors and then use this information to update their own strategy (Abramson and Kuperman 2001) (Zimmermann and Eguiluz 2005). Next agents update their Q-Value tables. Then, when using a dynamic network, agents use the HRN rule to update their neighborhood. If they find a relationship they value unrewarding, that is the average payoff received from that relationship differs from their total average payoff received by a predetermined threshold value, they end the relationship. That is, the agent removes the

unrewarding agent from their neighborhood then replaces it with a new agent. This new agent is chosen by the system and their identity is based upon the characteristics of the network used. This is done to preserve the unique characteristics of the complex networks. Therefore, since every broken link is replaced, the total number of links in the system remains constant.

The first experiment performed tested the effect static network structure has on the rate of social norm adoption in the Pure Coordination Game, the results of which are presented in Table 6. The Pure Coordination Game is described in Table 2. The first results presented are the number of time steps until 90% of the population adopts the same strategy in the Pure Coordination game domain, referred to as T90% (Kittok 1995). Each result is the average of 10 runs. Our results are compared with the results obtained by Jordi Delgado (Delgado 2002). In his experiments he employs the same networks; however his agents use the Highest Current Reward rule to update their strategies (Shoham and Tennenholtz 1997). In this update rule, agents keep track of the payoff received from the last play of each strategy. Then at each time step the agent selects the action that previously earned them the largest reward. Therefore, an agent will only change their strategy in the event that it earns them a payoff that is less than the previous reward earned from another strategy.

| | T90% | |
|---|---|---|
| Networks | Our Results | Delgado |
| Complete | 10 | $10^4$ |
| Scale-Free | 25 | $10^4$ |
| Small World | 200 | $10^5$ |
| Lattice | 2000 | $10^8$ |

Table 6 Timesteps till 90% of Population adopts the same strategy in static networks

Notice that while our agents were able to adopt the social norm faster than the agents using the Highest Current Reward rule, the relative rate at which the norms are adopted remains the same across the networks. This is because our agents employ the Q-Learning algorithm and thus are encouraged to explore their environment and adopt the optimal strategy much faster. The learning algorithm, Highest Current Reward Rule, employed by Delgado is much simpler. The HCR algorithm picks a strategy based solely off of the previous received reward and doesn't consider past experiences nor future rewards, nor does it employ an exploratory function thus agents.

Next, the results for agents playing the social dilemma games are presented. We first test the effect of static networks then next dynamic networks. Every network type is tested in the Prisoner's Dilemma and the Stag Hunt. Figure 17 shows the rate at which (Defect, Defect) is adopted as the social norm for each of the four static networks used. In Figure 16 the x-axis represents the timestep and the y-axis represents the total number of agents who chose Cooperate that timestep. Each line represents a different static network type. In Figure 17 agents play the Prisoner's Dilemma; here the Nash-equilibrium is (Defect, Defect) and the Pareto Optimum strategy profile is (Cooperate, Cooperate).

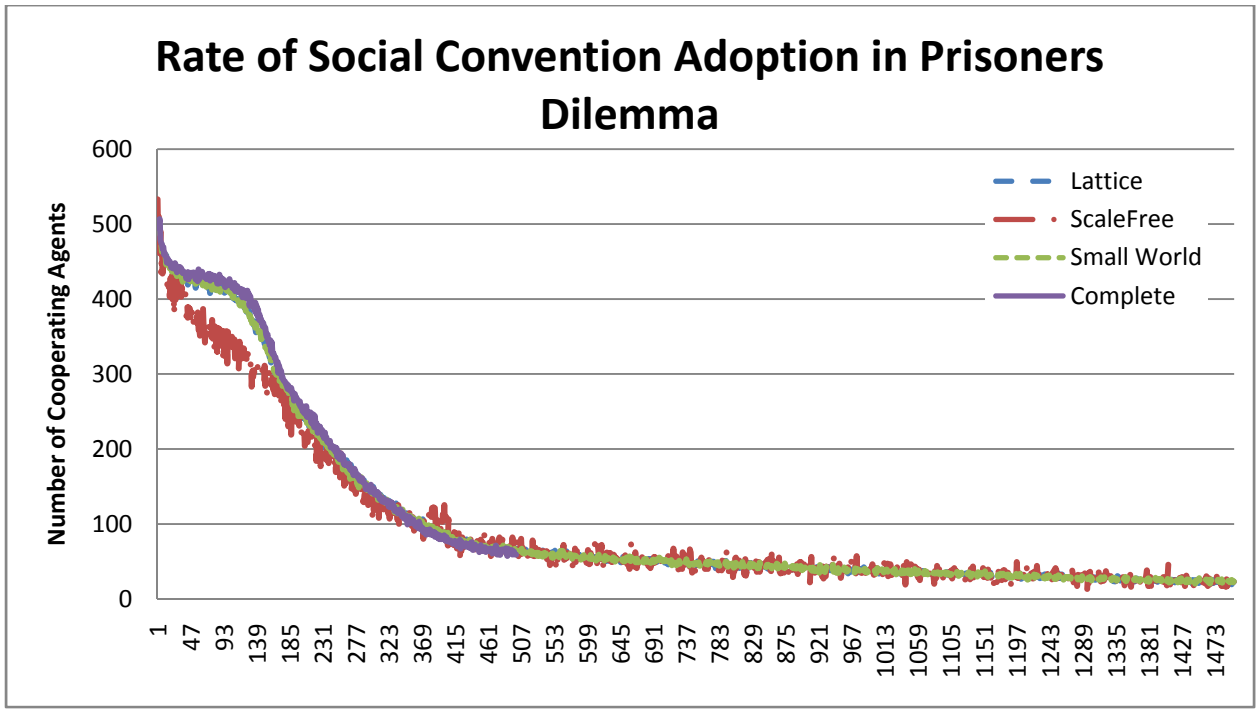


Figure 17: Rate of Norm Adoption per Static Network Type in the Prisoners Dilemma

What is important to note here is that in each network type, (Defect, Defect) is adopted as the social norm. In these experiments all networks are static, therefore every agents is forced to play with a set neighborhood. Thus a cooperative agent can't escape from an exploiting defecting agent and thus must eventually also defect to achieve a higher reward. Therefore it can be said that when agents are unable to protect themselves from defectors cooperate is unable to emerge as the social norms. This isn't surprising as (Defect, Defect) is the Nash equilibrium. However, networks in which agents are unable to end relationships and are forced to play with the same people are unrealistic. Also important to note is that each network converges to (Defect, Defect) at the same rate. After the results from the Pure Coordination game in which the network structure had a great impact on the rate of convergence, this result

may seem to indicate an error. However, it will be shown that in the Prisoners Dilemma, the best strategy for an agent to perform is independent of the agents in its network.

Consider an agent playing the Pure Coordination game, defined in Table 2, who has ten neighbors. If seven of them play Cooperate and three of them play Defect then the best strategy to adopt is Cooperate. This would result in the agent earning 7 points from the cooperating agents and -3 points from the defecting agents. Now consider an agent with ten neighbors playing the Prisoners Dilemma, defined in Table 3. If 7 of the neighbors adopt Cooperate, to maximize payoff the agents should choose Defect, earning a payoff of 35, versus 21 which the agent would earn playing Cooperate. If the remaining 3 neighbors play Defect, then the agent should maximize their payoff playing Defect, earning them 3, versus 0 playing Cooperate. Therefore it can be seen that regardless of the strategies played by their neighbors, an agent's best strategy is Defect. Thus, the network structure has no effect on the rate at which the social norm is adopted.

Figure 18 shows the results for agents playing the Stag Hunt. Here again the x-axis represents the timestep and the y-axis represents the total number of agents who chose cooperate that timestep. Each line represents a different static network type. The results of this experiment are very similar to the results obtained from agents playing the Prisoners Dilemma in static networks. This is because both games are very similar. In both games, (Defect, Defect) is the less risky strategy. Thus a similar argument holds for why (Defect, Defect) is the adopted social norm and for why the rates of convergence are the same across each network.

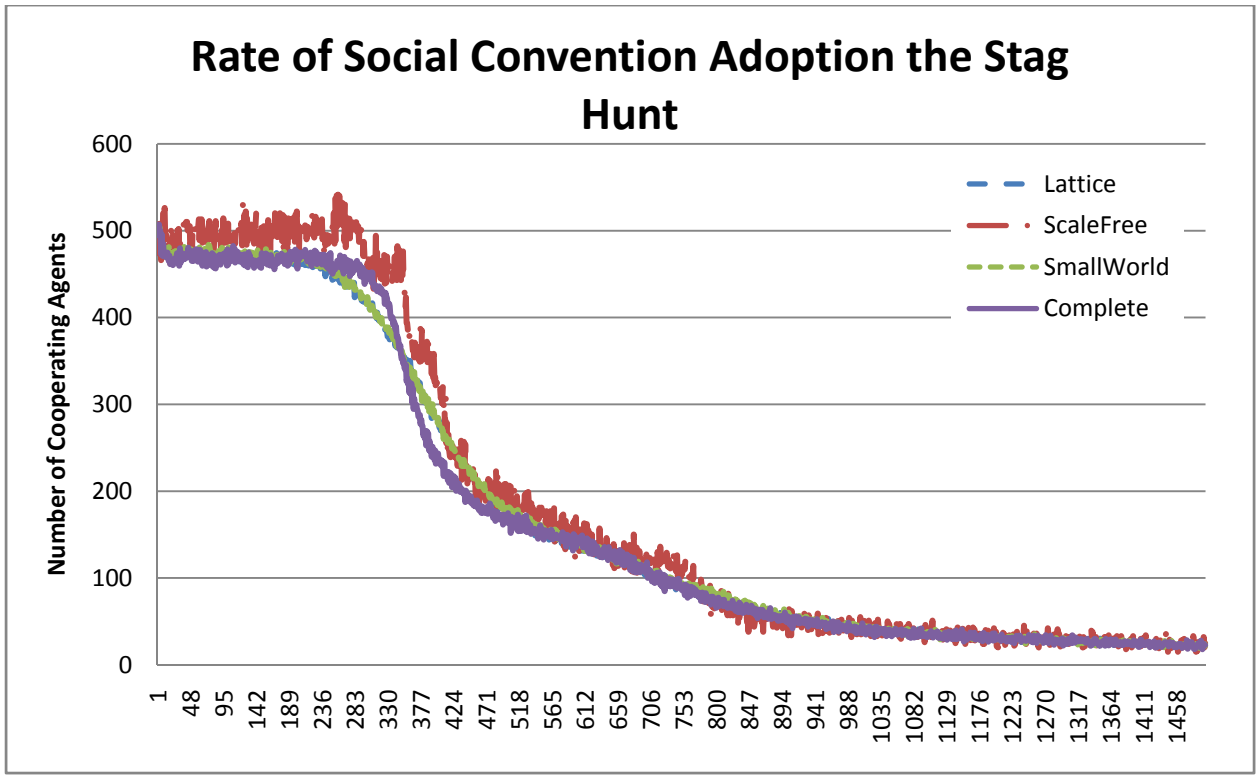


Figure 17: Rate of Norm Adoption per Static Network Type in the Stag Hunt

Next, results from play in dynamic networks in which agents end relationships that are deemed unrewarding are shown in Figure 19. Here for the network update function we use a threshold value of 1. This means that any relationship in which an agent receives a reward that is less than the average reward they earn is ended and replaced. Figure 18 has the same x-axis and y-axis as the previous grap. It shows the number of cooperators at each timestep per network type.

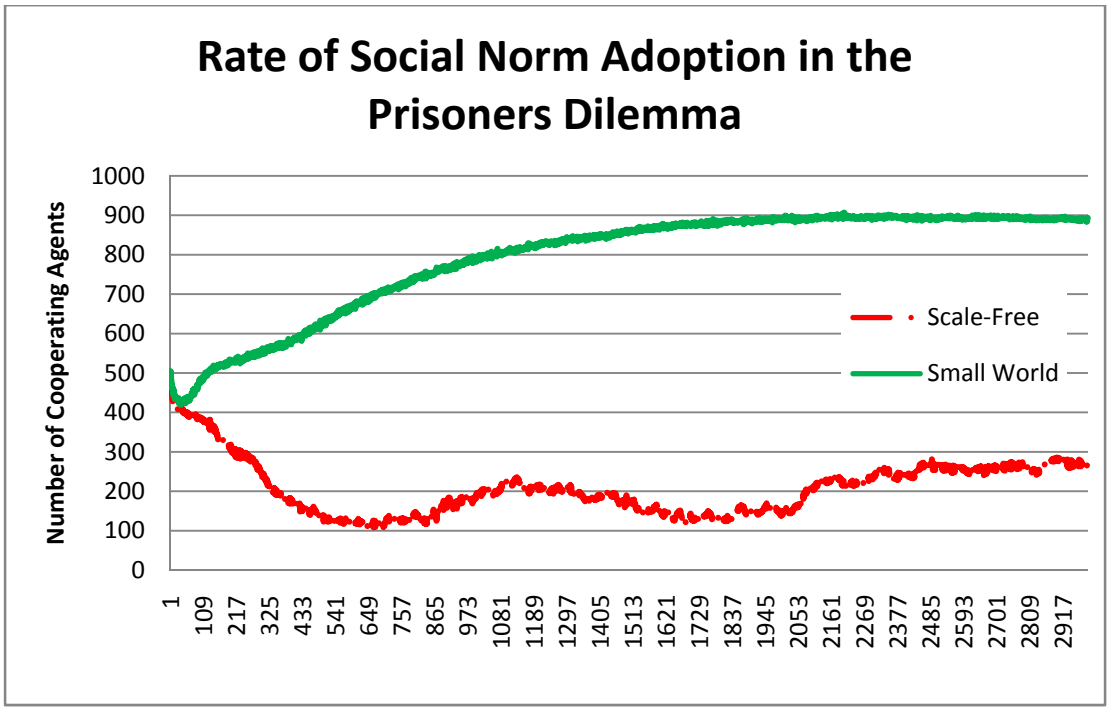


Figure 189: Rate of Norm Adoption per Dynamic Network Type in the Prisoners Dilemma

Here it is shown that unlike in the static networks, when using a dynamic network the structure has a large impact not only on the rate of convergence but to where the system converges to. While neither network converges to over T90%, both perform far better than the static networks. It is important to note that when using a dynamic small world network, the system is able to adopt a much higher percentage of (Cooperate, Cooperate) than the scale-free network. One of the most prevalent differences between the scale-free network and the small world network is that the small world contains a high clustering coefficient. That is, two neighbors of the same agents are likely to be neighbors of each other. Therefore in the small-world networks, clusters of agents playing cooperate are able to emerge and sustain themselves because they are protected from defectors. However, in the scale-free networks a large population of the agents are neighbors with a very small population of the agents. If this

small population chooses to adopt defect as their strategy not only will they earn a high payoff but they will also influence a great many agents to also adopt defect as their strategy.

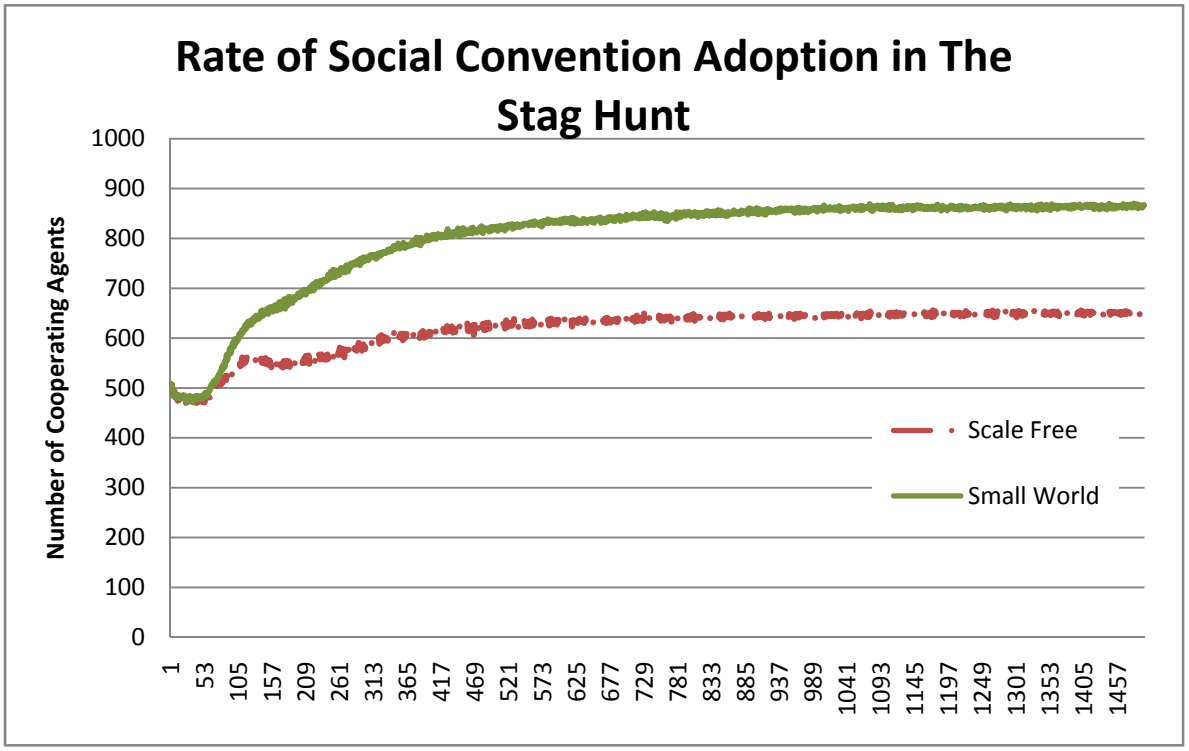


Figure 190: Rate of Social Norm Adoption per Dynamic Network type in the Stag Hunt

It can be seen in Figure 20 that the relationship between the two different complex dynamic networks remains the same with regards to their effect on the emergence of (Cooperate, Cooperate) as the social norm. Small-World still fosters cooperative behavior better than scale-free. What is different is the rate and the level of cooperative behavior. The population using the dynamic small-world network adopt (Cooperate, Cooperate) as the social norm much faster when playing Stag Hunt than when playing the Prisoners Dilemma. In addition, Scale-Free is able to achieve a much higher level of cooperative behavior in the Stag Hunt than in the Prisoners Dilemma. This is attributed to the characteristics of the game. In the Stag Hunt, both (Cooperate, Cooperate) and (Defect, Defect) are Nash-equilibriums. However,

Cooperate is a risky strategy and therefore an agent isn't likely to play it when they are forced to continuously play with a defecting agent.

# 6. Conclusion and Future Work

## 6.1 Conclusion

Simulations have long eluded the social sciences as a viable tool. This is in part due to the fact that human actions and interactions do not adhere to well defined rules. However recent advances in studying both human decision making and human social organizations have provided insight into the mechanics behind human behavior. In order to provide the social sciences with a viable tool, one must divert from the classical economic models that present humans as perfectly rational utility maximizers.

This work represents a contribution to this goal. By employing realistic models based off of recent advances in human decision making and human social interaction we are able to reproduce behavior more realistic than the perfectly rational agent.

## 6.2 Further Work

There are many potential extensions onto the work presented. First and perhaps most relevant would be a combination of the two models presented. As the social model only requires an agent to use only the information gained by affecting their environment and the individual decision making algorithm was implemented with the same constraint, no contradictions exist between the two models.

### 6.1.1 Individual Model

There are many options for further exploration in the individual decision making model presented. An important and perhaps necessary addition to the model would be a formal state description. As much of the algorithm depends on the representation of the state and

consequently how it's compared to other states, a formal description would provide a domain independent representation and therefore a more flexible model. In addition questions about what should determine the threshold of states searched, should more recent memories hold higher weight than old memories, should memories that are accessed frequently hold a higher weight or should the overall number of memories stored be limited, are all valid questions and merit research.

### 6.2.2 Social Model

Many avenues of potential research exist as extensions to the social model presented. Perhaps most important, a formal comparison of specific human social behaviors, such as reciprocity of exchange should be performed in an attempt to compare the low level interactions of the agents with human social behavior. In addition, the model should be implemented with other reinforcement learning algorithms as well as tested in other domains to determine the models flexibility.

## Works Cited


Abramson, Guillermo, and Marcelo Kuperman. "Social Games in a Social Network." *Physical Review*, 2001.

Albert, R., and Barb'asi A.L. "Statistical Mechanics of Complex Networks." *Modern Physics*, 2002: 47-97.

Aylett, R, S Louchart, and J Pickering. "A mechanism for acting and speaking for empathic agents." *Autonomous Agents and Multi-Agent Systems Workshop.* 2004.

Barb'asi, A.-L., and R. Albert. "Scale-free characteristics of random networks: The topology of the World Wide Web." *Physical Review*, 2000: 69-77.

Barto, R.S. *Reinforcement Learning: An Introduction.* Cambridge: MIT Press.

Barto, Richard S. Sutton and Andrew G. *Reinforcement learning : an introduction.* Cambridge, Mass: MIT Press, 1998.

Batagelj, Vladimir. *Pajek.* http://pajek.imfm.si/doku.php?id=pajek (accessed 04 12, 2009).

Bearden, J. Neil. "The evolution of inefficiency in a simulated stag hunt." *Behavior Research Methods, Instruments, & Computers*, 2001: 124-129.

Bernoulli, D. "Exposition of a New Theory on the Measurement of Risk." *Econometrica*, 1954: 22-36.

Bishop, Christopher M. *Pattern Recognition and Machine Learning.* Springer, 2006.

Brooks, H, et al. "Using agent-based simulation to reduce collateral damage during military operations." *Systems and Information Engineering Design Symposium.* 2004. 71 - 77.

Buchanan, B.G., and T.M. Mitchell. "Model Directed Learning of Production Rules." In *Pattern-Directed Inference Systems*, by D.A. Watermann and F. Hayes-Roth, 297-312. New York: Academic Press, 1978.

Carely, Kathleen. *ORA Software.* http://www.casos.cs.cmu.edu/projects/ora/software.htm (accessed Febuary 15, 2009).

Christensen, K, and Y Sasaki. "Agent-Based Emergency Evacuation Simulation with Individuals with Disabilities in the Population." *Journal of Artificial Societies and Social Simulation*, 2008.

Delgado, Jordi. "Emergence of Social Convenetions in Complex Networks." *Artificial Intelligence*, 2002: 171-185.

Dignum, F. "Autonomous Agents with Norms." *Artificial Intelligence and Law*, 1999: 1-15.

Erdos, P, and A Renyi. *On Random Graphs.* Debrecen: 290-297, 1959.

Erve, I., and A.E. Roth. "Predicting how people play games: reinforcement learning in experimental games with unique mixed equilibria." *American Economic Review*, 1998: 848-881.

Excelente-Toledo, C.B., and N.R. Jennings. "The Dynamic Selection of Coordination Mechanism." *Journal of Autonomous Agents and Multiagent Systems*, 2004: 55-85.

Fararo, T.J., and M.H. Sunshine. *A Study of a Biased Friendship Net.* Syracuse, NY: Syracuse Univ. Press.

Gmytrasiewicz, PJ., and S. Noh. "Implementing a Decision-Theoretic Approach to Game Theory for Socially Competent Agents." In *Game Theory and Decision Theory in Agent-Based Systems*, by S. Parsons, PJ. Gmytrasiewicz and M. Woolridge, 97-118. Kluwer Academic Publishers, 2002.

Guth, Werner, and Reinhard Tietz. "Ultimatum Bargaining Behavior, A survey and comparison of experimental results." *Journal of Economic Psychology*, 1990: 417-449.

Hamill, Lynne. "A Simple but more realistic Agent based model of a social network." *Center for Research in Social Simulation*, 2006.

Jiang, Yichuan, and Toru Ishida. "A Model for Collective Strategy Diffusion in Agent Social Law Evolution." *IJACAI*, 2007.

Jin, Emily, Michelle Girvan, and M. E. J. Newman. "The Structure of Growing Social Networks." *Physical Review*, 2001.

Kaelbling, Leslie. "Reinforcement Learning: A Survey." *Journal of Artificial Intelligence*, 1996: 237-285.

Kahneman, D, and A Tversky. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 1979: 263-292.

Kahneman, D. "Maps of Bounded Rationality: A Perspective on Intuitive Judgement and Choice." *The Nobel Prizes Lecture*, 2002.

Kaminski, Marek M. *Games Prisoners Play.* Princeton: Princeton University Press, 2004.

Keiki Takadama, Tetsuro Kawai, Yuhsuke Koyama. "Micro - and Macro-Level Validation in Agent-Based Simulation: Reproduction of Human-Like Behaviors and Thinking in a Sequential Bargaining Game." *Journal of Artificial Societies and Social Simulation*, 2008.

Kim, S. "Intelligent Software Agents and buisness oriented application scenarios." Hidelberg, Germany: AOIS, 1999.

Kimbrough. "Simple reinforcement learning agents: Pareto beats Nash in algorithmic game theory study." *Information Systems and E-Buisness Management*, 2005.

Kittok, J.E. "Emergent conventions and the structure of multi-agent systems." In *Lectures in Complex Systems*, by L Nadel and D Stein. MA: Addison-Wesley, 1995.

Lewis, D.K. *Convention: A Philosophical Study.* Cambridge: Harvard Univ. Press, 1969.

Liljeros, F., C.R. Edling, L.A.N. Amaral, H.E. Stanely, and Y. Aberg. "The web of human sexual contacts." *Nature*, 2001: 907-908.

Lynne Hamill, Nigel Gilbert. "A Simple but More Realistic Agent-based Model of a Social Network." *Centre for Research in Social Simulation*, 2006.

Myers, DG. *Intuition: Its Powers and Perils.* New Haven: Yale University, 2002.

Newmann, M.E.J. "The Structure and Function of Complex Networks." *Society for Industrial and Applied Mathematics*, 2003.

Norling, E. "Folk Psychology for Human Modeling: Extending the BDI Paradigm." *AAMAS.* New York, 2004.

Paoloa Rizzo, Manuela Veloso, Maria Miceli, Amedeo Cesto. "Goal-based personalities and social behaviour in believable agents." *Applied Artificial Intelligence Journal*, 1999: 239-271.

Prasnikar, Vesna, and Alvin Roth. "Considerations of Fairness and Strategy Experimental Data from Sequential Games." *The Quarterly Journal of Economics*, 1992: 865-888.

Redner, S. "How popular is your paper? An emprical study of the citation distribution." *Eur. Phys.*, 1998: 131-134.

Robbins, H. "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society*, 1952: 527-535.

Russel, Stuart, and Peter Norvig. *Artificial Inteligence.* New Jersey: Pearson Education, 2003.

Shoham, Y, and M Tennenholtz. "On the emergence of social conventions: Modeling, analysis and simulations." *Artificial Intelligence*, 1997: 139-166.

Simon, H. "A Behavioral Model of Rational Choice." In *Models o Man, Social and Rationl: Methematical Esays on Rational Human Behavior in a Social Setting*. New York: Wiley, 1957.

Skyrms, Brian. *The Stag Hunt and Evolution of Social Structure.* Cambridge: Cambridge University Press, 2004.

Steven Kimbrough, Ming Lu. "Simple reinforcement learning agents: Pareto beats Nash in algorithmic game theory study." *Information Systems and E-Business Management*, 2005.

Takadama, Keiki. "Mico and Macro Level Validation in Agent-Based Simulation: Reproduction of Human-Like Behaviors and Thinking in a Sequential Bargaining Game." *Artificial Societies and Social Simulation*, 2008.

Von-Neumann, J., and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton: Princeton University Press, 1944.

Watkins, C.J. *Learning from Delayed Rewards.* Ph.D Thesis, Kings College, 1989.

Watts, D.J. *Small Worlds.* Princeton: Princeton University Press, 1999.

Younger, Stephen. "Reciprocity, Normative Reputation, and the Development of Mutual Obligation in Gift Giving Societies." *JASSS*, 2004.

Zimmermann, Martin, and Victor Eguiluz. "Cooperation, Social Networks and the Emergence of Leadership in a Prisoners Dilemma with Adaptive Local Interactions." *Physical Review*, 2005.