Dynamic Network Formation with Reinforcement Learning

W. Graham Mueller
Department of Economics, Texas A&M University
June 19, 2012

Abstract

I examine a dynamic model of network formation in which individuals use reinforcement learning to choose their actions. Typically, economic models of network formation assume the entire network structure to be known to all individuals involved. The introduction of reinforcement learning allows us to relax this assumption. Q-learning is a reinforcement learning algorithm from the artificial intelligence literature that allows for state-dependent learning. Using Q-learning, one may allow for varying degrees of information available to the agents. I determine what networks, if any, the model may converge to in the limit.

JEL classification: D85, D83, C73
Keywords: Networks, Reinforcement Learning, Q-Learning, Network Formation, Multi-agent Learning, Simulation

1 Introduction

In many network formation models, such as Watts (2001), Jackson and Watts (2002), Johnson and Gilles (2000), and Bala and Goyal (2000), the assumption of complete information plays a key role in determining what types of networks will form. The interested reader may refer to Jackson (2008) or Jackson (2003) for surveys on models of network formation. In Watts (2001), the network formation process is deterministic with agents myopically modifying their links to improve their current payoff. Using the payoff structure and stability notion defined in the Connections Model proposed by Jackson and Wolinsky (1996), Watts finds that, depending on parameter values,
this network formation will always converge to “stable” network. Jackson and Watts (2002) extend this model to a stochastic setting by allowing the possibility that agents choose the wrong action with some small probability. As the probability of making a mistake approaches zero, two scenarios arise: the network formation process converges to a single network or forms a closed cycle of networks. In each model, agents know the current state of the network, as well as how each action would affect their payoff, and make decisions based on this knowledge. Bala and Goyal (2000) proposed a game-theoretic model of network formation where agents possess complete information about the network structure and choose their link-formation strategies based on this information. The authors examine what types of networks may emerge and show that for many different parameter values, the predicted equilibrium network is a star-network even though agents are indistinguishable with respect to payoffs and costs. It is unlikely that the structure of any economic network would be known to all those involved. It is easy to imagine a scenario that one would have limited information regarding network structure, for example, privacy settings on Facebook may prohibit an individual from viewing the friends of a friend. If the assumption of complete information is removed, one might be able to understand how information drives network formation.

Another contribution of Bala and Goyal (2000) was the introduction of learning dynamics to the study of network formation. The learning dynamics examined were based on myopic best responses where agents exhibit some inertia when choosing their strategies. The assumptions made for the learning process may not be realistic for large networks, since each agent must observe the strategies of all other agents. On the other-hand, reinforcement learning allows the agents to merely observe their own payoff and adjust the probabilities associated with each action, which seems more plausible for situations where observing all agents who make up the network is not realistic. Q-learning, first proposed by Watkins and Dayan (1992a), is a reinforcement learning algorithm which allows for state-dependent learning, and therefore will allow agents to observe some information about the network. Network formation with reinforcement
learners was examined by Skyrms and Pemantle (2000). However, in their model, the payoffs were independent of network structure. In this paper, I will explicitly capture the trade-off between the cost of maintaining direct links to others and the informational benefit of being well connected by using the Connections Model presented by Jackson and Wolinsky (1996).

My approach is to model network formation as a dynamic process driven by reinforcement learning. See Busoniu et al. (2008) for a comprehensive survey of multi-agent reinforcement learning in non-network contexts. Agents meet over time and choose to maintain or delete existing connections, or add a new connection or maintain non-existent connections. These actions are chosen according a probability distribution and these probabilities adjust over time based on experience. This experience changes in response to the payoffs received after choosing a particular action. It may be helpful to imagine the network which evolves over time as a communication network in which a link between two agents indicates they have chosen to share information. After the active agents have chosen their actions, the communication network changes accordingly. As time passes, agents choose the action which has yielded the highest payoff increasingly often. I examine three informational settings in this paper. I first examine what happens when agents are unaware of who they are attempting to establish social ties with, the structure of the network, and how a certain network structure would benefit them. They merely observe their current and past payoffs. In the second scenario, agents observe the individuals they are attempting to form connections with, as well as, the payoff received each period. A third informational setting is examined in which agents observe the number of connections they have.

In each of the models described above, I first characterize which networks may be absorbing networks for the process and then further examine the convergence and stability properties of these absorbing networks using simulations. Each of the models are simulated and the long-run behavior is discussed. My results suggest that certain networks, which “pairwise stable” in the Connections Model, are not absorbing net-
works if agents do not observe any information beyond their payoff. However, if one increases the amount of information available to agents and reformulates the learning rule to be state dependent, these pairwise stable networks become absorbing networks for the network formation process.

This paper contributes to the theory of network formation in two ways. First, reinforcement learning is introduced to a model of network formation in which the agents’ payoffs depend explicitly on the network structure. The second contribution is the use of state-dependent reinforcement learning to allow the agents access to limited information during the network formation process.

The rest of this paper is organized as follows. Section B presents some preliminary definitions for networks. Section C defines the utility function specified in the Connections Model and defines pairwise stability and efficiency for networks. Section D introduces the basic model and Section E allows for limited information by incorporating a state-dependent learning rule. Section 6 presents the results of simulations carried out in order to examine the convergence properties of each model as well as the effect the various parameters have on the long-run behavior of each model.

2 Representing Networks

Before each model is introduced, some preliminary definitions and notation are needed. This notation follows the conventions established in the previous economic literature on networks and network formation whenever possible.

2.1 Nodes and agents

A set $N = \{1, 2, \ldots, n\}$ is the set of nodes belonging to a network. Nodes can represent individuals, firms, countries, or objects such as webpages and may be referred to as “nodes”, “agents”, “individuals”, or “players”.

4
2.2 Graphs and networks

The traditional representation of a network is an undirected labeled graph. Nodes are either connected or not. A directed graph can be used to represent networks where one-sided relationships may be present. Trading relationships or friendship networks are usually represented by undirected graphs, as both sides must establish and maintain the relationship. The formal model of networks is given below:

A network \((N, g)\) consists of a set of nodes \(N = \{1, 2, \ldots, n\}\) and \(g\), a set of links, or edges, present in the graph. Links may also be called called edges or connections. For instance, \(g = \{\{1, 2\}, \{2, 3\}\}\) or simply, \(g = \{\{12\}, \{23\}\}\) indicates links between nodes 1 and 2, as well as nodes 2 and 3. The notation \(ij\) will be used to represent the edge connecting node \(i\) and \(j\). An example of this notation is shown in Fig. 1. Furthermore, I will abuse notation at times and write \(ij \in g\) if nodes \(i\) and \(j\) are connected in network \((N, g)\). I will take \(g + ij\) to mean the network obtained by adding edge \(ij\) to the set of connections of \(g\). More precisely, \(g + ij = g \cup \{ij\}\). Similarly, \(g - ij\) will represent the network obtained by deleting edge \(ij\) from the set of connections of \(g\) or mathematically, \(g - ij = g \setminus \{ij\}\).

3 The Connections Model

The Connections Model was first proposed by Jackson and Wolinsky (1996). In the model, links between nodes or agents, represent social relationships that present both
benefits to the agents and a require cost in order to maintain the link. Agents also benefit from their indirect connections or the “friends of a friend”, so that there is a trade-off between maintaining direct links and the benefits received from the overall connectivity of the network. Formally, the model is defined as follows: The utility or payoff agent \( i \in \{1, 2, 3 \ldots, n\} \) receives from network \( g \) is

\[
 u_i(g) = w_i + \sum_{j \neq i} \delta_{ij}(g) - k_i(g)c_{ij} 
\]  

(1)

where \( d_{ij}(g) \) is the length of the shortest path distance between nodes \( i \) and \( j \), \( \delta_{ij} \) represents the value received by \( i \) from \( j \), \( k_i(g) \) is the agent \( i \)'s degree in the network, \( c_{ij} \) is the cost of maintaining the link between \( i \) and \( j \), and \( w_i \) is constant. Furthermore, it is assumed that \( \delta_{ij} \in (0, 1) \) so that indirect links are of less value than direct links. If there is no path connecting nodes \( i \) and \( j \), then \( d_{ij}(g) = \infty \). The Symmetric Connections Model assumes \( \forall i, j \in N \quad \delta_{ij} = \delta, w_i = w_j, \) and \( c_{ij} = c \). I will assume the individual payoffs are symmetric throughout the rest of this paper.

3.1 Stability and efficiency of social networks

In order to characterize the networks that arise when formation of links requires consent of both agents involved, Jackson and Wolinsky (1996) introduced a notion of network stability which they called pairwise stability. As implied by the definition, no one individual wishes to make any changes to the connections in the network.

**Definition** A network \( g \) is pairwise stable if for each \( i, j \) the following hold:

i) \( ij \in g, u_i(g) \geq u_i(g - ij) \) and \( u_j(g) \geq u_j(g - ij) \)

ii) \( ij \notin g, \) if \( u_i(g) < u_i(g + ij) \), then \( u_j(g) > u_j(g + ij) \)

This notion of stability is independent of the the network formation process and therefore, provides a good benchmark for my model. Although the concept of pairwise
stability is somewhat weak, it makes some strong predictions for the stability of networks. This is illustrated in the following theorem which addresses the existence and uniqueness of pairwise stable networks.

**Theorem** Jackson and Wolinsky (1996) In the symmetric connections model, a pairwise stable network exists for all $N$ and is given by:

i) The unique pairwise stable network is the complete graph, $g^N$ if $c < \delta - \delta^2$.

ii) A star encompassing all players if $\delta > c$ and $\delta - c \leq \delta^2$.

iii) The empty network (no links) if $\delta \leq c$

In the parameter range $\delta < c$, the notion of pairwise stability rules out any network in which an agent possesses one link. In particular, the star network cannot be stable.

Another property of networks one may be concerned with is efficiency. Jackson and Wolinsky use the following definition for efficiency:

**Definition** A network is *efficient* if $\sum_i u_i(g) \geq \sum_i u_i(g')$ for all $g' \in \mathcal{G}$

Here $\mathcal{G}$ denotes the set of all networks of size $n$. Here the concept of efficiency is equivalent to the idea of social welfare for a given network structure. For the following parameter ranges, the efficient networks were shown to be unique and are characterized as follows:

**Theorem** Jackson and Wolinsky (1996) In the symmetric connections model, a unique efficient network exist for all $n$ and is given by:

i) the complete graph $g^N$ if $c < \delta - \delta^2$.

ii) a star encompassing everyone if $\delta - \delta^2 < c < \delta + \frac{n-2}{2}\delta^2$

iii) the empty network if $\delta + \frac{n-2}{2}\delta^2 < c$

---

1 A star network consists of one central agent and periphery agents, each connected only to the central agent.
These results on stability and efficiency have used as a benchmark for the models presented in Watts (2001) and Jackson and Watts (2002) and will be used as a comparison for the models presented in the next two sections.

4 The RL Model

Agents meet over time and decide whether or not to form connections or sever existing connections. Time is represented by a countably infinite set $T = \{1, 2, 3, \ldots \}$. The network structure at time $t$ is represented by an undirected graph which is denoted as $g_t$. The initial network $g_0$ is a parameter of the model. The network evolves slowly enough so that only two agents may alter their shared connection at a given point in time. These two agents meet randomly with a probability of $\frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$. Once agent $i$ and agent $j$ have met, they choose actions $a_i$ and $a_j$ from the set $A = \{\text{add}, \text{delete}\}$ according to distributions $(p_i(t), 1-p_i(t))$ and $(p_j(t), 1-p_j(t))$, respectively, which are defined below. It is assumed that agents must both agree to add or maintain an existing connection, however, individuals may unilaterally sever a connection. Therefore, if both agent $i$ and agent $j$ choose to add a link and $ij \notin g_t$, then $g_{t+1} = g_t + ij$. If $ij \in g_t$, then $g_{t+1} = g_t$. Likewise, if either agent $i$ or agent $j$ chooses to delete link $ij$ and $ij \in g_t$ then $g_{t+1} = g_t - ij$. If $ij \notin g_t$ and one agent chooses not to add the link, then $g_{t+1} = g_t$.

Let $Q_i^0 = [Q_i^0(a), Q_i^0(d)]$ denote the initial assessments agent $i$ has for each action. After the agents have been randomly matched, they choose their actions probabilistically based on their current assessments. The probability that agent $i$ chooses to add a connection is denoted by $p_i(t)$ and the probability that agent $i$ chooses to delete a connection is given by $1 - p_i(t)$. Throughout the rest of the paper I will denote $p_i(t)$ by $p_{i,t}$ for compactness. This probability is defined by:
$$p_{i,t} = \frac{e^{Q_i(a)/\tau_i^t}}{e^{Q_i(a)/\tau_i^t} + e^{Q_i(d)/\tau_i^t}}$$  \hspace{1cm} (2)$$

This choice rule is known by various names including logistic choice or softmax selection. Here $\tau_i^t$ is sometimes called the temperature parameter\(^2\) or rationality parameter. If $\tau_i^t$ is very large, all actions are chosen with approximately equal probabilities. If $\tau_i^t$ is very close to zero, the action with the highest payoff assessment is chosen with probability close to one. I will assume that for each agent $i$, $\tau_i^t \to 0$ as $t \to \infty$, so that, as time increases, each agent $i$ chooses the action with highest Q-value more and more often. One typical temperature function is given by $\tau(t) = \eta^t$ where $\eta < 1$.

Since agents may not unilaterally establish a connection with another agent, the probability that $ij \in g_{t+1}$ is $p_{i,t}p_{j,t}$. After choosing their actions at time $t$, the network transitions according to the following probabilities:

$$g_{t+1} = \begin{cases} 
    g_t + ij \text{ with probability } p_{i,t}p_{j,t} \\
    g_t - ij \text{ with probability } 1 - p_{i,t}p_{j,t}
\end{cases}$$

Agents then receive a payoff, which they are able to observe. The payoff received by agent $i$ given agent $i$ chooses to add a connection is:

$$\pi_{i,t+1}(a) = \begin{cases} 
    u_i(g_t + ij) \text{ with prob } p_{j,t} \\
    u_i(g_t - ij) \text{ with prob } 1 - p_{j,t}
\end{cases}$$

If agent $i$ chooses to delete a connection the payoff is:

---

\(^2\)This choice rule is related to a distribution in classical statistical mechanics describing the velocities of particles in a gas.
\[
\pi_{i,t+1}(d) = u_i(g_t - ij)
\]

The agents then update their payoff assessments using the following update rule:

\[
Q_{i,t+1}^i(a_i) = \begin{cases} 
(1 - \alpha)Q_{i,t}^i(a_i) + \alpha \pi_{i,t+1}(a_i) & \text{if } a_i \text{ is chosen.} \\
Q_{i,t}^i(a_i) & \text{otherwise.}
\end{cases}
\]

This reinforcement learning rule has been studied in Sarin and Vahid (2001) and Kirman and Vriend (2001), among others. The parameters include the initial assessments, \(Q_{i,0}^i\), and the learning rate, \(0 < \alpha < 1\). The learning rate determines how much previous payoffs are weighted in the current assessments. If \(\alpha\) is close to one then payoffs in the distance past hold little weight in the current assessment. In fact, if \(\alpha\) is close to one, the current assessment is very close to the last observed payoff.

Combining these transition probabilities, the each agent’s Q-values change with the following probabilities:

\[
Q_{t+1}^i(a) = \begin{cases} 
(1 - \alpha)Q_{t}^i(a) + \alpha u_i(g_t + ij) & \text{with prob } p_{i,t}p_{j,t} \\
(1 - \alpha)Q_{t}^i(a) + \alpha u_i(g_t - ij) & \text{with prob } p_{i,t}(1 - p_{j,t}) \\
Q_{t}^i(a) & \text{with probability } 1 - p_{i,t}
\end{cases}
\]

\[
Q_{t+1}^i(d) = \begin{cases} 
(1 - \alpha)Q_{t}^i(d) + \alpha u_i(g_t - ij) & \text{with prob } 1 - p_{i,t} \\
Q_{t}^i(d) & \text{with prob } p_{i,t}
\end{cases}
\]
Given these transition probabilities, the next Lemma will give an upper and lower bound for the payoff assessments as time grows large.

**Lemma 1** Provided $0 < \alpha < 1$, the Q-values are asymptotically bounded between the minimum and maximum possible payoffs received for choosing action $a_i$, that is,

**Proof.** First, notice the following observation which is obtained by induction.

\[
Q_T^i(a_i) = (1 - \alpha)^T Q_0^i(a_i) + \sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} u_i(g_t + 1)
\]

To see this, begin with the following:

\[
Q_1^i(a_i) = (1 - \alpha) Q_0^i(a_i) + \alpha r_1
\]

Now assume that $Q_{T-1}^i(a_i) = (1 - \alpha)^{T-1} Q_0^i(a_i) + \sum_{t=1}^{T-1} \alpha (1 - \alpha)^{T-1-t} r_t$ holds. The following expression is obtained for $t = T$:

\[
Q_T^i(a_i) = (1 - \alpha) Q_{T-1}^i(a_i) + \alpha r_T
\]

\[
= (1 - \alpha)^T Q_0^i(a_i) + \sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} r_t + \alpha r_T
\]

\[
= (1 - \alpha)^T Q_0^i(a_i) + \sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} r_t
\]

As $T \to \infty$, the first term approaches zero. The second term becomes a geometric series with ratio $(1 - \alpha)$ where each term is multiplied by $\alpha r_t$. For any values of $N$, $\delta$ and $c$ in the connections model, there is a maximum and minimum value for $u_i$. Therefore, we must have for all $t$, $\min_{g} u_i(g) \leq r_t \leq \max_{g} u_i(g)$. Using this fact,

\[
\sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} \min_{g(a_i)} u_i(g(a_i)) \leq \sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} r_t \leq \sum_{t=1}^{T} \alpha (1 - \alpha)^{T-t} \max_{g(a_i)} u_i(g(a_i))
\]
Where $g(a_i)$ denotes the set of possible networks formed after action $a_i$ has been chosen by agent $i$. This distinction must be provided because in the connections, links cannot be formed unilaterally, the set of networks obtained when delete is chosen differ from those formed after add is chosen. Using the fact that $0 < \alpha < 1$ and the convergence of the geometric series, the following is true:

\[
\min_{g(a_i)} u_i(g(a_i)) \leq \lim_{T \to \infty} Q^i_T(a_i) \leq \max_{g(a_i)} u_i(g(a_i))
\]

Rewriting this in a more convenient form yields:

\[
\min \pi_t(a_i) \leq \lim_{t \to \infty} Q^i_t(a_i) \leq \max \pi_t(a_i) \tag{3}
\]

\[\square\]

In reference to the symmetric connections model, we have the following asymptotic bounds for the Q-values. If $\delta > c$ and $\delta - c > \delta^2$, then

\[
0 \leq \lim_{t \to \infty} Q^i_t(d) \leq (N - 2)(\delta - c) + \delta^2
\]

\[
0 \leq \lim_{t \to \infty} Q^i_t(a) \leq (N - 1)(\delta - c)
\]

Whenever $\delta - c > \delta^2$, direct connections net cost are worth more than indirect connections. Therefore, the maximum payoff occurs when an agent maintains the maximum number of direct connections. For a network of size $N$, the maximum number of connections an individual can maintain is $N - 1$ and therefore, the maximum payoff is $(N - 1)(\delta - c)$. However, this payoff may only be obtained after choosing to add a connection, since connections may not be unilaterally added. The maximum payoff obtained from deleting a connection is $(N - 2)(\delta - c) + \delta^2$. The minimum payoff occurs
when an agent maintains has no connections and this payoff equal to zero.

If $\delta > c$, $\delta - c < \delta^2$, then the following bounds hold for the Q-values.

\[
0 \leq \lim_{t \to \infty} Q_t^i(d) \leq (\delta - c) + (N - 2)(\delta^2)
\]

\[
0 \leq \lim_{t \to \infty} Q_t^i(a) \leq (\delta - c) + (N - 2)(\delta^2)
\]

Whenever $\delta - c < \delta^2$, indirect connections may be worth more than direct connections net cost. Since $0 < \delta < 1$, the most valuable indirect connections give a payoff of $\delta^2$ since $\delta^2 > \delta^k$ for any $k > 2$. However, to receive any benefits from an indirect connection, agents must maintain at least one direct connection. Therefore, the maximum possible payoff is $\delta - c + (N - 2)\delta^2$. Note that this payoff may be received after adding or deleting a connection. Again, the minimum payoff occurs when an agent maintains no connections and is zero.

If $\delta < c$ and $c < \delta + (N - 2)\delta^2$, then

\[
(N - 2)(\delta - c) \leq \lim_{t \to \infty} Q_t^i(d) \leq (\delta - c) + (N - 2)(\delta^2)
\]

\[
(N - 1)(\delta - c) \leq \lim_{t \to \infty} Q_t^i(a) \leq (\delta - c) + (N - 2)(\delta^2)
\]

Whenever $\delta < c$ and $c < \delta + (N - 2)\delta^2$ direct connections always yield a negative payoff. The minimum payoff occurs when an individual maintains the maximum number of direct connections without any indirect connections. Since connections may not be added unilaterally, the minimum possible payoff received after deleting a connection is larger than the minimum possible payoff received after adding a connection. These two minimum payoffs are given by $(N - 2)(\delta - c)$ and $(N - 1)(\delta - c)$, respectively. Since $c < \delta + (N - 2)\delta^2$, the maximum payoff is positive and occurs when individuals maintain one direct connection with $(N - 2)$ indirect connections of length 2.
If $\delta < c$ and $c > \delta + (N - 2)\delta^2$, then

$$(N - 2)(\delta - c) \leq \lim_{t \to \infty} Q_t^i(d) \leq 0$$

$$(N - 1)(\delta - c) \leq \lim_{t \to \infty} Q_t^i(a) \leq 0$$

Whenever $\delta < c$ and $c > \delta + (N - 2)\delta^2$, the minimum possible payoffs are the same as the previous case. However, the maximum possible payoff is now zero since the cost of maintaining a direct connection is higher than any possible benefit that would be received from indirect connections.

In order to characterize which networks form over time, the following definition will be useful.

**Definition** A network is an absorbing network if there exist values $Q_t^i(a), Q_t^i(d)$ where $i \in N$ and $\tau_t$, such that $g_s = g_t$ with probability 1 for all $s \geq t$.

Let us examine the evolution of the overall network as follows, consider at time $t$ the total number of connections in the network, which we will denote by $|g_t|$. The total number of edges can only increase if a connection $ij \notin g_t$ is added at time $t$. That is, for some $ij \notin g_t$, $g_{t+1} = g_t + ij$, which occurs with probability $p_{i,t}p_{j,t}$. Furthermore, the probability that any two agents $i$ and $j$ are matched is $\frac{2}{n(n-1)}$. The probability that $ij \in g_{t+1}$, given that $ij \in g_t$ is $\frac{2}{n(n-1)} \sum_{ij \in g_t} p_{i,t}p_{j,t}$. Similarly, the probability that $ij \notin g_{t+1}$, given $ij \notin g_t$ is $\frac{2}{n(n-1)} \sum_{ij \notin g_t} p_{i,t}p_{j,t}$. Thus, the total probability that $|g_{t+1}| = |g_t|$ is given by $\frac{2}{n(n-1)} \sum_{ij \in g_t} p_{i,t}p_{j,t} + \frac{2}{n(n-1)} \sum_{ij \notin g_t} p_{i,t}p_{j,t}$. The other cases may be derived similarly.

Given the current network, $|g_{t+1}|$ evolves according to the following:
\[
|g_{t+1}| = \begin{cases} 
|g_t| + 1 & \text{with probability } \sum_{ij \notin g_t} \left( \frac{2}{n(n-1)} \right) p_{i,t} p_{j,t} \\
|g_t| & \text{with probability } \sum_{ij \in g_t} \frac{2}{n(n-1)} p_{i,t} p_{j,t} + \sum_{ij \notin g_t} \frac{2}{n(n-1)} (1 - p_{i,t} p_{j,t}) \\
|g_t| - 1 & \text{with probability } \sum_{ij \in g_t} \frac{2}{n(n-1)} (1 - p_{i,t} p_{j,t}) 
\end{cases}
\]

These transition probabilities can help characterize which, if any, network may be an absorbing network.

**Proposition 1** The only possible absorbing networks for the RL model are the complete network, the empty network, or a network with both a completely connected component and an empty component. In addition, a necessary condition for an absorbing network is \( \tau^i_s = 0 \) for all \( i \in N \) and all \( s \geq t \).

**Proof.** Using the transition probabilities given in the previous section, the absorbing networks may be described by characterizing the total number of connections. In this case, \( g_{t+1} = g_t \) with probability 1 and thus \( |g_{t+1}| = |g_t| \) with probability 1. Therefore, the probability that \( |g_{t+1}| = |g_t| + 1 \) and the probability that \( |g_{t+1}| = |g_t| - 1 \) must both equal zero. In terms of the current network, one must have:

\[
\sum_{ij \notin g_t} \left( \frac{2}{n(n-1)} \right) p_{i,t} p_{j,t} = 0 \tag{4}
\]

and

\[
\sum_{ij \in g_t} \frac{2}{n(n-1)} (1 - p_{i,t} p_{j,t}) = 0 \tag{5}
\]
There are three cases which satisfy both of these equations:

Case 1: If $|g_t| = \frac{n(n-1)}{2}$ then $g_t$ is the complete network. Furthermore, $\sum_{ij \in g_t} (1 - p_{i,t}p_{j,t}) = 0$ which implies $p_{i,t} = p_{j,t} = 1$ for all $i,j \in N$.

Case 2: If $|g_t| = 0$, then $g_t$ is the empty network. Furthermore, $\sum_{ij \notin g_t} p_{i,t}p_{j,t} = 0$ which implies for each $i,j \in N$ either $p_{i,t} = 0$ or $p_{j,t} = 0$ or both. In particular, we may have at most one agent $k$, with $p_{k,t} \neq 0$.

Case 3: If $\sum_{ij \in g_t} (1 - p_{i,t}p_{j,t}) = 0$ and $\sum_{ij \notin g_t} p_{i,t}p_{j,t} = 0$. Then if $ij \in g_t$ but $ik \notin g_t$ for some $i$, we must have $p_{i,t} = p_{j,t} = 1$ and $p_{k,t} = 0$. Therefore, there is no connection involving $k$ in $g_t$, otherwise $\sum_{ij \in g_t} (1-p_{i,t}p_{j,t}) > 0$. Thus, $k$ has no links. We can therefore partition the network into a connected component and an empty component. □.

In each case we have $p_{i,t} = 1$ or 0, which implies $\tau^i_t = 0$. Given this fact, the behavior of the network formation process when $\tau^i = 0$ for each $i \in N$ should be examined. □

**Proposition 2** Suppose $\delta > c$ in the connections model. If $\tau_t = 0$ and $Q^i_t(d) > Q^i_t(a)$, then $Q^i_s(d) \leq Q^i_s(a)$ for some $s \geq t$ unless $Q^i_t(a) < 0$.

**Proof.** If $Q^i_t(a) \geq Q^i_t(d)$, then $s = t$. Otherwise, suppose at time $t$, $\tau^i_t$ is zero. If agent $i$ has $k_i(g_t) = k$ connections and $Q^i_t(d) > Q^i_t(a)$ then with probability 1, for some $T \in \mathbb{N}$, $k_i(g_{t+T}) = 0$ (provided the $Q^i_s(d) > Q^i_s(a)$ for all $s \in [t, t+T]$, otherwise the result is immediate). Agent $i$ will have zero connections at time $t + T$ and $u_i(g_{t+T}) = 0$, which implies that $\Delta Q^i_{t+T+1} = -\alpha Q^i_{t+T}$. Furthermore, $Q^i_s(d) = (1 - \alpha)^s Q^i_{t+k}(d)$ for any $s > t + T$. As $s$ grows large, $Q^i_s(d)$ approaches zero. Therefore, $Q^i_s(d) > Q^i_s(a) = Q^i_t(a)$ only if $Q^i_t(a) < 0$. In other words, if $\tau^i_t = 0$, agent $i$ will only choose delete forever, if $Q^i_t(a) < 0$. This in turn implies agent $i$ cannot have zero connections in the underlying network indefinitely. □

**Corollary** If $\delta > c$ in the connections model, a network in which an agent has zero connections cannot be absorbing for the RL model.
This is an immediate consequence of the above proposition and Lemma 1, which states
\[ 0 \leq \lim_{t \to \infty} Q_i^t(a) \text{ for all } i \in N. \]

**Proposition 3** Suppose \( \delta < c \) in the connections model. If \( \tau_i^t = 0 \) and \( Q_i^t(a) > Q_i^t(d) \) for each agent \( i \) at some time \( t \), then \( \exists s \geq t \) such that \( Q_s^i(d) > Q_s^i(a) \) unless \( Q_i^t(d) < (N - 1)(\delta - c) \) for all \( i \in N \).

**Proof.** The payoff to agent \( i \) for maintaining \( N - 1 \) connections (\( k_i = N - 1 \)) is \( (N - 1)(\delta - c) \). If agent \( i \) has \( k_i(g_t) \) connections and then with probability 1, for some \( T \in \mathbb{N} \), \( k_i(g_{t+T}) = N - 1 \) for each agent \( i \) (provided the \( Q_i^t(a) > Q_i^t(d) \) for all \( s \in [t, t+T] \) and all \( i \in N \) otherwise the result is immediate). Furthermore, given that \( p_{i,t} = 1 \), the change in \( Q_i^t(d) \) is zero and since \( u_i(g_{t+i}) = (N - 1)(\delta - c) \), we have \( \Delta Q_i^t(a) = \alpha((N - 1)(\delta - c) - Q_i^t(a)) \). These relationships hold for all agents in the network. Therefore, if \( p_{i,t} = p_{j,t} = 1 \) for any \( i, j \in N \), then \( s \) periods later we have \( Q_i^{t+s}(a) = (1 - \alpha)^s Q_i^t(a) + \sum_{l=1}^{s} \alpha(1 - \alpha)^{s-l}(N - 1)(\delta - c) \). Thus, as \( s \) grows large we have \( Q_i^{t+s}(a) \approx (N - 1)(\delta - c) \) for each agent \( i \). Therefore, \( Q_i^s(a) > Q_i^s(d) = Q_i^t(d) \) only if \( Q_i^t(d) < (N - 1)(\delta - c) \) for all agents \( i \in N \). In other words, if \( \tau_i^t = 0 \) and \( \delta < c \) each agent in the network will only choose add forever if \( Q_i^t(d) < (N - 1)(\delta - c) \) for all agents \( i \in N \). This implies that the complete network cannot occur indefinitely unless the conditions in the proposition are met. \( \square \)

**Corollary 1** If \( \delta - \delta^2 > c \) in the connections model, the complete network will form and remain in the RL model.

This is an immediate consequence of Proposition 3 and Lemma 1, which states \( 0 \leq \lim_{t \to \infty} Q_i^t(d) \leq (N - 2)(\delta - c) \) for all \( i \in N. \) \( \square \)

**Corollary 1** If \( \delta < c \) in the symmetric connections model, the complete network is not a absorbing network for the RL model.

Again, this is an immediate consequence of the above proposition and Lemma 1, which states whenever \( \delta < c \) then \( (N - 2)(\delta - c) \leq \lim_{t \to \infty} Q_i^t(d) \leq 0 \) for all \( i \in N. \) \( \square \)
**Proposition 4** If $\delta > c$ in the connections model, the complete network may be absorbing for the RL model.

**Proof.** By Proposition 2, if $Q^i_t(d) > Q^i_t(a)$, then for some time $s \geq t$, $Q^i_s(a) > Q^i_s(d)$ provided $Q^i_t(a) \geq 0$. By Proposition 3, the condition $Q^i_t(a) > Q^i_t(d)$ for all agents $i \in N$ will hold for all $s \geq t$ if $Q^i_t(d) < (N-1)(\delta - c)$. By Lemma 2, both of these bounds are met in the limit $t \to \infty$. Note that the all the absorbing networks other than the complete network have at least one agent with no connections. This implies the complete network is the only “absorbing” network in the limit whenever $\delta - \delta^2 > c$. □

Putting these propositions, we see that whenever $\delta > c$ the only possible absorbing network is the complete network. Whenever, $\delta < c$ we can only rule out the complete network. To examine whether the process converges to a absorbing network, a different approach must be taken. This question will be examined in the next section as well as through simulations in section 6.

## 5 Partial Observability

In the section, I will propose two informational settings. In the simple model considered above, agents had no information. However, agents may have access to some information. For instance, they may recognize the individuals they have been randomly matched with or be aware of the number of connections they maintain. I will use a state-contingent reinforcement learning algorithm (similar to Q-learning) to capture both these ideas. The model is the as same as the one presented in the previous section with one distinction: agents observe the individual they have been randomly matched with or the number of connections they currently have. In both situations, each agent maintains two $Q$-values (one for each action) for each agent they may be matched with (or number of connections they have) for a total of $2(n-1)$ $Q$-values. The learning rule is now given by the following:
\[
Q^i_{t+1}(j, a) = \begin{cases} 
(1 - \alpha)Q^i_t(j, a) + \alpha(u_i(g_{t+1})) & \text{if action } a \text{ is chosen} \\
Q^i_t(j, a) & \text{otherwise}
\end{cases}
\]

where \( j \in \{1, \ldots, n\} \) denotes the agent matched with agent \( i \) at time \( t \) and \( \alpha \) is the learning rate.

Furthermore, the probability with which the agent chooses his action is now state dependent and we will denote the probability that agent \( i \) chooses to add a connection with agent \( j \) by \( p_{i,t}(j) \), which is defined as follows:

\[
p_{i,t}(j) = \frac{e^{Q^i_t(j,a)/\tau^i_t}}{e^{Q^i_t(j,a)/\tau^i_t} + e^{Q^i_t(j,d)/\tau^i_t}}
\]

where \( \tau^i_t \) is same the rationality parameter as in the previous model. As before, we will assume that \( \tau^i_t \to 0 \) as \( t \to \infty \).

**Lemma 2** Provided \( 0 < \alpha < 1 \), the Q-values are asymptotically bounded between the minimum and maximum possible payoffs received for choosing action \( a_i \), that is,

\[
\min \pi_t(j, a_i) \leq \lim_{t \to \infty} Q^i_t(j, a_i) \leq \max \pi_t(j, a_i)
\]

**Proof.** Using the same argument as Lemma 2, it can shown that limiting values of the payoff assesments for each state must fall between the minimum and maximum possible payoffs occurring in that state. □

As before, the transition probabilities related to the underlying network can be used to determine the set of possible absorbing networks.
\[
|g_t + 1| = \begin{cases} 
|g_t| + 1 & \text{with probability } \sum_{ij \notin g_t} \frac{2}{n(n-1)} p_{i,t}(j)p_{j,t}(i) \\
|g_t| & \text{with probability } \sum_{ij \in g_t} \frac{2}{n(n-1)} p_{i,t}(j)p_{j,t}(i) + \sum_{ij \notin g_t} \frac{2}{n(n-1)} (1 - p_{i,t}(j)p_{j,t}(i)) \\
|g_t| - 1 & \text{with probability } \sum_{ij \in g_t} \frac{2}{n(n-1)} (1 - p_{i,t}(j)p_{j,t}(i)) 
\end{cases}
\]

**Proposition 5** Any network may be an absorbing network for the RLA model.

**Proof.** Let \(g_t\) be a network at time \(t\). Then \(g_{t+1} = g_t\) with probability 1 if and only if \(|g_{t+1}| = |g_t|\) with probability 1. This implies that \(\sum_{ij \notin g_t} \frac{2}{n(n-1)} p_{i,t}(j)p_{j,t}(i) = 0\) and \(\sum_{ij \in g_t} \frac{2}{n(n-1)} (1 - p_{i,t}(j)p_{j,t}(i)) = 0\). In this case, since the probabilities are state dependent, any network as a absorbing network by required \(p_{i,t}(j) = p_{j,t}(i) = 1\) for \(ij \in g_t\) and either \(p_{i,t}(j) = 0\) or \(p_{j,t}(i) = 0\) or both for \(ij \notin g_t\). \(\square\)

**Proposition 6** If \(\tau^i_t = 0\), \(\delta > c\), and \(Q^i_t(j,d) > Q^i_t(j,a)\) for some \(i\) and all \(j \in N\), then \(Q^s_t(j,d) \leq Q^s_t(j,a)\) for some \(s \geq t\) unless \(Q^i_t(j,a) < 0\).

**Proof.** If \(Q^i_t(j,d) > Q^i_t(j,a)\) for some \(i\) and all \(j \in N\) and \(\tau^i_t = 0\), then \(p_{i,t}(j) = 0\). If agent \(i\) has one or more connections at time \(t\), they will sever all their connections with probability 1 and continue to maintain zero connections. Their payoff assessments for each state, \(Q^s_t(j,d)\) will decrease towards zero and, unless \(Q^i_t(j,a) < 0\), \(Q^s_t(j,d) \leq Q^i_t(j,a)\) for some time \(s\). \(\square\)

**Corollary** If \(\delta > c\) in the connections model, a network in which an agent has zero connections is not an absorbing network.

This corollary is an immediate consequence of Lemma 3 and Proposition 18. \(\square\)

**Proposition 7** If \(\tau^i_t = 0\) and \(Q^i_t(j,a) > Q^i_t(j,d)\) for each agent \(i,j \in N\), then \(Q^s_t(j,d) > Q^s_t(j,a)\) for some \(s \geq t\) unless \(Q^i_t(j,d) < (N-1)(\delta - c)\) for all \(i,j \in N\).
Proof. If \( Q^i_t(j,a) > Q^i_t(j,d) \) for each pair of agents \( i,j \in N \) and \( \tau^i_t = 0 \) then \( p_{i,t}(j) = p_{j,t}(i) = 1. \) Therefore, at some time \( s \geq t \) each agent will have \((N-1)\) connections. Using the same argument as before, each payoff assessment, \( Q^i_s(j,a) \), will approach \((N-1)(\delta - c)\) as \( s \) increases. Therefore, unless \( Q^i_t(j,d) < (N-1)(\delta - c) \) for each pair of agents \( i,j \), \( Q^i_s(j,a) > Q^i_s(j,a) \) for all \( s \geq t. \) \( \square \)

Corollary If \( \delta < c \) in the symmetric connections model, the complete network is not an absorbing network.

This corollary is an immediate consequence of Proposition 19 and Lemma 3. \( \square \)

Corollary If \( \delta - \delta^2 > c \) in the symmetric connections model, the complete network is an absorbing network.

This corollary is also an immediate consequence of Proposition 19 and Lemma 3. \( \square \)

Proposition 8 The set of absorbing networks for the RLC model is the same as the RL model.

Proof. Let \( g_t \) be a network at time \( t. \) Then \( g_{t+1} = g_t \) with probability 1 if and only if \( |g_{t+1}| = |g_t| \) with probability 1. This implies that number of connections each agent maintains is constant. Suppose agent \( i \) currently has \( deg_i \) connections, if \( p_i(deg_i, t) = 0, \) then eventually \( deg_i \) will decrease provided \( deg_i \neq 0. \) Therefore, for any absorbing network \( p_i(deg_i, t) = 1 \) or \( deg_i = 0. \)

If \( P_j(deg_i, t) = 1 \) and there exists an agent \( j \) with \( p_j(deg_j, t) = 1 \) such that \( ij \notin g_t, \) eventually these agents will meet and both agents will choose to add the connection. Therefore, for any absorbing network each pair of agents \( i,j \) such that \( p_i(deg_i, t) = 1 \) and \( p_j(deg_j, t) = 1, \) \( ij \in g_t. \) Together with the previous statement, this implies that a absorbing network for the RLC must consist of a completely connected component and an empty component. This set of networks coincides with the set of absorbing networks for the RL model. \( \square \)
**Proposition 9** If \( \tau^i_t = 0, \delta > c, k_i = 0, \) and \( Q^i(0,d) > Q^i(0,a) \) for some \( i, \) then \( Q^i_s(0,d) \leq Q^i(0,a) \) for some \( s \geq t \) unless \( Q^i_t(0,a) < 0. \)

**Proof.** If \( \tau^i_t = 0, \delta > c, k_i = 0, \) and \( Q^i_t(0,d) > Q^i_t(0,a) \) for some \( i, \) then agent \( i \) will continue to maintain zero connections in the network and the assessment \( Q^i_s(0,d) \) will approach zero. If \( Q^i_t(0,a) > 0, \) then for some time \( s, Q^i_s(0,d) \leq Q^i(0,a). \)

**Corollary** If \( \delta > c \) in the connections model, a network in which an agent has zero connections is not an absorbing network.

This corollary is an immediate consequence of Lemma 3 and Proposition 21. \( \square \)

**Proposition 10** If \( \tau^i_t = 0, Q^i_t(N-1,a) > Q^i_t(N-1,d), \) and \( \text{deg}_i = N-1 \) for each agent \( i, \) then \( Q^i_s(N-1,d) > Q^i_s(N-1,a) \) for some \( s \geq t \) unless \( Q^i_t(N-1,d) < (N-1)(\delta - c) \) for all \( i \in N. \)

**Proof.** If \( \tau^i_t = 0, Q^i_t(N-1,a) > Q^i_t(N-1,d), \) and \( \text{deg}_i = N-1 \) for each agent \( i, \) then \( p^i(N-1,a) = 1 \) for all agents \( i \in N. \) The payoff assessment \( Q^i_s(N-1,a) \) will approach \( (N-1)(\delta - c) \) as \( s \) grows. Therefore, unless \( Q^i_t(N-1,d) < (N-1)(\delta - c) \) for all \( i \in N, \) eventually \( Q^i_s(N-1,d) > Q^i_s(N-1,a). \)

**Corollary** If \( \delta < c \) in the symmetric connections model, the complete network is not an absorbing network.

This corollary is an immediate consequence of Lemma 3 and Proposition 22. \( \square \)

**Corollary** If \( \delta - \delta^2 > c \) in the symmetric connections model, the complete network is an absorbing network.

This corollary is also an immediate consequence of Lemma 3 and Proposition 22. \( \square \)

Traditionally, the long-run behavior of reinforcement learning has been analyzed using stochastic approximation, by approximating the learning dynamics with a continuous time limit, or using properties of Markov chains. See, for example, Beggs
(2005), ?, and Borgers and Sarin (1997). Unfortunately, these approaches encounter difficulty whenever the payoffs are state-dependent, as is the case for each of the models discussed so far. In order to analyze the behavior of these models, I will turn to simulations.

6 Simulations

In this section, I examine the long-run behavior of the RL, RLA, and RLC models. The number of possible networks increases exponentially with the number of agents and limits my simulations to a relatively low number of agents. I examine the complete distribution over networks for simulations with \( n = 4 \) and \( n = 5 \) agents. In order to simulate each model the following need to be specified: the initial Q-values for each agent, the parameters of the connections model, \( \delta \) and \( c \), the learning rate \( \alpha \), choice parameter \( \tau_t \), and the initial network \( g_0 \). Given this information and the payoff function defined in the connections model, the model may be simulated. 1000 simulations were performed for each case, so that no initial Q-values were overweighted. The rationality parameter \( \tau_t \) used in the simulations gradually decreased over time according to the following rule:

\[
\tau_t = 0.9999^t
\]

This ensures the rationality parameter is close to one at the beginning of the simulations and very close to zero as \( t \) grows large. A similar rule has been used in other studies on Q-learning, including, for example, Waltman and Kaymak (2008) and Sandholm and Crites (1996).
7 Results of the Simulations

7.1 Effect of the parameters $\delta$ and $c$

Together the parameters $\delta$ and $c$ determine which networks are pairwise stable and efficient. As discussed in Section 3, there are three regions with different networks exhibiting stability and efficiency. The regions of efficiency and the associated network are shown in Fig. 2 and the different regions of pairwise stability and the associated networks are given in Fig. 3.

The simulations used the following parameter ranges $\delta \in [1, 2, \ldots, 9]$ and $c \in [1, 2, \ldots, 9]$. The initial Q-values were uniformly distributed on the interval $[u_{\text{min}}, u_{\text{max}}]$ where $u_{\text{min}}$ and $u_{\text{max}}$ represent the minimum and maximum payoff an agent receives under the symmetric connections model, respectively. A second framework was considered in which the agents were initially unbiased with all the initial Q-values equal to zero, but the results of the simulations were not affected. Two frameworks for the starting network were also considered. As in Watts (2001), in the first scenario, the initial network was the empty network. In the second case, the initial network was chosen randomly with all possible networks being equally likely. Again, the results of the simulations were not affected. In what follows, I report the results of the simulations in which the initial Q-values and the starting network were random as described above.
The simulations were repeated 1000 times and the relative frequency for each network was recorded for the each model. The relative frequency or empirical likelihood of a network, $g^*$, at time $t$ is defined the number of times $g_t = g^*$ divided by the number of simulation runs.

### 7.1.1 RL Model

Fig. 4 shows the relative frequency of the empty network at $t = 10,000$ for the various combinations of $\delta$ and $c$, while Fig. 5 gives the relative frequency for the complete network.

![Figure 4: RL: Rel. freq. of the empty network](image1)

![Figure 5: RL: Rel. freq. of the complete network](image2)

The simulations for the RL model imply that the likelihood of the network being complete network is highest whenever $\delta > c$ and the probability of empty network when $c > \delta$. The relative frequency of these networks increases as the difference between $\delta$ and $c$ increases. The result of the simulations support the analytical results for the RL model. The relative frequencies of all other networks were very small compared to the empty and the complete networks.
7.1.2 RLA Model

The following figures show the relative frequency for the various combinations of $\delta$ and $c$ for the RLA model. When the relative frequencies of the various networks are compared to the regions illustrated in Figures 2 and 3, a pattern emerges. The empirical likelihood for the empty network, shown in Fig. 6, is very close to one whenever $c > \delta + \delta^2$. In this case, the empty network is not only pairwise stable, but also efficient. The empirical probability of complete network, shown in Fig. 7, is largest whenever $c < \delta - \delta^2$. In this region, the complete network is both pairwise stable and efficient. One interesting observation regarding the simulation of the RLA model is the fact that the empirical likelihood of the empty network is very close to one whenever $c > \delta + \delta^2$, however, the empirical likelihood of the complete network is relatively low in the region in which it is both pairwise stable and efficient. The empirical likelihood for the star network (shown in Fig. 8, Fig. 9, Fig. 10, and Fig. 11) is low, in general, but a star network is most likely to form when $\delta - \delta^2 < c < \delta + \delta^2$.

![Figure 6: RLA: Rel. freq. of the empty network](image1)

![Figure 7: RLA: Rel. freq. of the complete network](image2)

7.1.3 RLC Model

The following figures show the relative frequency for the various combinations of $\delta$ and $c$ for the RLC model. Patterns similar to those the simulations of the RLA model are
seen for the RLC model. The empirical probability of the empty network (Fig. 12) is largest whenever \( c > \delta + \delta^2 \). The empirical probability of the complete network (Fig. 13) highest whenever \( c < \delta - \delta^2 \). In contrast to the RLA model, the empirical probability of the star network (Fig. 14 and Fig. 15) is very low when \( \delta - \delta^2 < c < \delta + \delta^2 \).

### 7.2 Effect of the learning rate \( \alpha \)

The learning rate \( \alpha \) determines how much past payoffs are weighted. High alpha imply the last observed payoff carry much more weight than the distant past. My simulations show that low values of \( \alpha \) may slow the rate of convergence of the network formation process. The following figures show the effect of varying alpha on the relative frequency
7.2.1 RL Model

The simulations show that as the learning rate alpha increases, the probability of forming a network other than the complete or empty network decreases significantly. In Fig. 16 the relationship between the relative frequency of the empty network and the learning rate is examined in the region in which $\delta \leq c$. In general, there seems to be a positive relationship between the two. The relationship between the relative frequency of the complete network and the learning rate in the region in which $\delta - \delta^2 > c$ is shown in Fig. 17. Again, there is a positive relationship between the two variables. Fig. 18
shows that the learning rate has a significant effect on the relative frequency of the complete network whenever \( \delta - \delta^2 < c \). For low learning rates, the relative frequency of the complete network is also low. However, as the learning rate increase the relative frequency of the complete network increases in a nearly linear relationship. Intuitively, this makes sense because whenever the learning rate is low, agents give a larger weight to past payoffs. If an agent deletes a direct link with an agent, but is still indirectly connected to the same agent, their payoff will increase. Agents with “long memories” will remember this and adjust their probabilities more slowly than agents with “short” memories. If an agent with a short memory performs the same action and increases their probability of deleting links, they will delete more links and remember the payoffs associated with these actions which may be lower than the payoffs they received in the complete network. Hence, agents with a high learning rate may sever their links and observe the drop in payoff and quickly return to adding links with a high probability. Therefore, one might expect the relative frequency of the complete network to increase as alpha increases.

7.2.2 RLA Model

For the simulations of the RLA model, the empty network is observed to have the highest empirical likelihood whenever \( \delta < c \). In this region, the effect of the learning
rate on the empirical likelihood of the empty network is illustrated in Fig. 19. In general, low values of $\alpha$ correspond lower empirical likelihood for the empty network than ceterus paribus high values of $\alpha$. There is no clear relationship between the learning rate and the empirical likelihood of the complete network in the region in which $\delta - \delta^2 \leq c$ (Fig. 20) or $\delta - \delta^2 > c$ (Fig. 21).

7.2.3 RLC Model

As with the RLA model, in the region in which the empty network is pairwise stable there is a positive relationship between the learning rate and the empirical likelihood of the empty network. Fig. 22 shows this relationship for the various values of $\delta$ and
\( c \) which fall in this region. In the region \( \delta - \delta^2 < c \), there is a very noisy relationship, as seen in Fig. 23, between the learning rate and the empirical likelihood of the wheel network, which had the largest empirical likelihood in many of the simulations within that region. In the region in which \( \delta - \delta^2 > c \), the learning rate does not exhibit a consistent relationship effect with the empirical likelihood of the complete network, as shown in Fig. 24.

Figure 21: RLA Model: Rel. freq. of \( g^N \) vs \( \alpha \)

Figure 22: RLC Model: Rel. freq. of \( g^e \) vs \( \alpha \)  
Figure 23: RLC Model: Rel. freq. of \( g^W \) vs \( \alpha \)
7.3 Networks most likely to form

As seen above, the empirical likelihood of the efficient network for the various simulations was quite low for some regions in the parameter space over $\delta$ and $c$. The network with largest empirical likelihood for each simulation is reported below.

For the RL model, the empty network has the largest empirical likelihood whenever $\delta \leq c$. The complete network has the largest empirical likelihood whenever $\delta > c$ and the empty network has the largest empirical likelihood when $\delta \leq c$. In Fig. 25, $\emptyset$ represents the empty network and $g^N$ denotes the complete network. These results are supported by the analytical results proven in section 4.

The networks which are most likely to form for the RLA model illustrated in Fig. 26 where $g^{\text{star}}$ denotes the star network. In the blank points, there was no clear pattern as to which network was the most likely to form and the networks were neither pairwise stable nor efficient. The empty network was the network most likely to form whenever $\delta < c$ and in some cases when $\delta = c$. The complete network was the mostly likely to form when $\delta - \delta^2 > c$, however, this relationship did not always hold. In the region in which $\delta - \delta^2 < c$ no clear pattern was observed.

Fig. 27 shows the networks that were most likely to form during the simulations of the RLC model. As with the RLA model, the empty network had the highest observed
empirical likelihood when \( \delta < c \) and for some cases when \( \delta = c \). The complete network had the largest empirical likelihood whenever \( \delta - \delta^2 > c \). Furthermore, when compared to the RLA model, this pattern was stronger in the sense that for only one simulation \((\delta = .3, c=.2)\) did the complete network not have the highest observed frequency. An interesting pattern emerged in the region for which \( \delta - \delta^2 < c < \delta \). Two network topologies dominated this region, namely the wheel\(^3\) and the chain network\(^4\), denoted by \( g^W \) and \( g^C \), respectively. As with the RLA model, there were regions in which no clear pattern emerged with regards to the network with the highest observed frequency.

\(^3\)The wheel network is a network which \( n \) links in which in agents may be labeled such that \( g^W = \{12, 23, \ldots, n1\} \).

\(^4\)A chain network is a network with \( n - 1 \) links in which agents may relabeled such that \( g^C = \{12, 23, \ldots, (n-1)n\} \).
Although it appears that each model produces different long-run empirical distributions with regards to networks, this needs to be quantified in some way. I will use the two-sample $\chi^2$ test (see ?, Chapter 2) to test the hypothesis that the long-run behavior of the each model is the same. Under the two-sample $\chi^2$ test, the null hypothesis is that the two mutually independent random samples are drawn from the same distribution with the alternative that the populations differ in some way. The $\chi^2$-statistic for two samples of identical size is given by:

$$Q = \sum_{i=1}^{r} \frac{(f_{1i} - f_{2i})^2}{f_{1i} + f_{2i}}$$
where \( f_{ij} \) denotes the number of observations classified as category \( i \) in the sample \( j \) and \( r \) is the total number of categories. It is known that this test-statistic has a \( \chi^2 \) distribution with \( r - 1 \) degrees of freedom under the null hypothesis. In what follows, \( n = 4 \), which means there 64 different networks and hence 63 degrees of freedom. The critical value for the the \( Q \)-statistic at the 1% confidence level is \( Q_c = 92.01 \). Table 1 tests the empirical distributions of the RLA model and the RLC model. Table 2 evaluates the relationship between the RL and RLC models and Table 3 compares the RL and RLA models.

The results of the \( \chi^2 \) two sample tests seem to suggest that the long-run behavior of the RL and RLC is the same whenever \( c > \delta \). The long-run behavior of each model is different whenever \( \delta > c \). In the region in which \( \delta - \delta^2 > c \), the empirical likelihood...
Table 1 $\chi^2$ two-sample test: RLA vs RLC

<table>
<thead>
<tr>
<th></th>
<th>$\delta = .1$</th>
<th>$\delta = .2$</th>
<th>$\delta = .3$</th>
<th>$\delta = .4$</th>
<th>$\delta = .5$</th>
<th>$\delta = .6$</th>
<th>$\delta = .7$</th>
<th>$\delta = .8$</th>
<th>$\delta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = .1$</td>
<td>66.21</td>
<td>141.16</td>
<td>357.57</td>
<td>481.57</td>
<td>710.76</td>
<td>582.12</td>
<td>601.17</td>
<td>560.82</td>
<td>538.48</td>
</tr>
<tr>
<td>$c = .2$</td>
<td>169.63</td>
<td><strong>62.81</strong></td>
<td><strong>66.87</strong></td>
<td>334.54</td>
<td>842.36</td>
<td>914.86</td>
<td>666.67</td>
<td>707.09</td>
<td>621.75</td>
</tr>
<tr>
<td>$c = .3$</td>
<td>434.57</td>
<td>865.82</td>
<td>1468.4</td>
<td>1976.3</td>
<td>1978.3</td>
<td>1984.7</td>
<td>1724.2</td>
<td>1380.2</td>
<td>728.10</td>
</tr>
<tr>
<td>$c = .4$</td>
<td>119.98</td>
<td>304.50</td>
<td>193.60</td>
<td>93.58</td>
<td>451.02</td>
<td>707.70</td>
<td>780.58</td>
<td>1092.1</td>
<td>967.51</td>
</tr>
<tr>
<td>$c = .5$</td>
<td><strong>44.50</strong></td>
<td>139.48</td>
<td>308.23</td>
<td>337.03</td>
<td>523.11</td>
<td>669.18</td>
<td>873.94</td>
<td>1213.0</td>
<td>1133.7</td>
</tr>
<tr>
<td>$c = .6$</td>
<td>22.83</td>
<td>50.83</td>
<td>97.15</td>
<td>198.49</td>
<td>179.50</td>
<td>940.94</td>
<td>855.89</td>
<td>981.04</td>
<td>1223.8</td>
</tr>
<tr>
<td>$c = .7$</td>
<td>26.29</td>
<td>45.70</td>
<td>53.16</td>
<td>156.67</td>
<td>273.63</td>
<td>608.34</td>
<td>1312.4</td>
<td>1110.8</td>
<td>1016.6</td>
</tr>
<tr>
<td>$c = .8$</td>
<td>31.49</td>
<td>21.15</td>
<td>30.34</td>
<td>50.23</td>
<td>120.20</td>
<td>197.27</td>
<td>1660.0</td>
<td>1358.3</td>
<td>1084.0</td>
</tr>
<tr>
<td>$c = .9$</td>
<td><strong>48.13</strong></td>
<td><strong>37.62</strong></td>
<td><strong>38.59</strong></td>
<td><strong>15.81</strong></td>
<td><strong>47.96</strong></td>
<td>134.10</td>
<td>189.62</td>
<td>1874.1</td>
<td>1312.9</td>
</tr>
</tbody>
</table>

Table 2 $\chi^2$ two-sample test: RL vs RLC

<table>
<thead>
<tr>
<th></th>
<th>$\delta = .1$</th>
<th>$\delta = .2$</th>
<th>$\delta = .3$</th>
<th>$\delta = .4$</th>
<th>$\delta = .5$</th>
<th>$\delta = .6$</th>
<th>$\delta = .7$</th>
<th>$\delta = .8$</th>
<th>$\delta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = .1$</td>
<td>61.87</td>
<td>335.15</td>
<td>333.39</td>
<td>209.07</td>
<td>384.70</td>
<td>756.06</td>
<td>1226.3</td>
<td>1469.6</td>
<td>1582.1</td>
</tr>
<tr>
<td>$c = .2$</td>
<td>47.02</td>
<td>77.84</td>
<td>499.22</td>
<td>760.86</td>
<td>760.69</td>
<td>1118.6</td>
<td>1353.4</td>
<td>1594.8</td>
<td>1644.4</td>
</tr>
<tr>
<td>$c = .3$</td>
<td>28.05</td>
<td>43.48</td>
<td>84.83</td>
<td>515.68</td>
<td>841.56</td>
<td>1155.8</td>
<td>1504.4</td>
<td>1541.4</td>
<td>1642.0</td>
</tr>
<tr>
<td>$c = .4$</td>
<td>19.71</td>
<td>25.26</td>
<td>29.55</td>
<td>123.47</td>
<td>506.89</td>
<td>957.56</td>
<td>1244.6</td>
<td>1423.6</td>
<td>1499.0</td>
</tr>
<tr>
<td>$c = .5$</td>
<td>15.44</td>
<td>17.81</td>
<td>33.26</td>
<td>55.01</td>
<td>171.91</td>
<td>643.85</td>
<td>1023.9</td>
<td>1230.6</td>
<td>1380.6</td>
</tr>
<tr>
<td>$c = .6$</td>
<td>17.12</td>
<td>16.78</td>
<td>17.22</td>
<td>28.88</td>
<td>58.81</td>
<td>229.66</td>
<td>672.10</td>
<td>1164.8</td>
<td>1335.7</td>
</tr>
<tr>
<td>$c = .7$</td>
<td>21.71</td>
<td>21.98</td>
<td>25.41</td>
<td>13.05</td>
<td>29.30</td>
<td>71.86</td>
<td>234.93</td>
<td>741.26</td>
<td>1164.5</td>
</tr>
<tr>
<td>$c = .8$</td>
<td>39.78</td>
<td>20.34</td>
<td>21.27</td>
<td>13.91</td>
<td>18.04</td>
<td>38.16</td>
<td>88.61</td>
<td>274.44</td>
<td>765.98</td>
</tr>
<tr>
<td>$c = .9$</td>
<td>39.78</td>
<td>23.29</td>
<td>27.14</td>
<td>30.86</td>
<td>9.92</td>
<td>11.89</td>
<td>50.13</td>
<td>68.15</td>
<td>252.79</td>
</tr>
</tbody>
</table>

Table 3 $\chi^2$ two-sample test: RL vs RLA

<table>
<thead>
<tr>
<th></th>
<th>$\delta = .1$</th>
<th>$\delta = .2$</th>
<th>$\delta = .3$</th>
<th>$\delta = .4$</th>
<th>$\delta = .5$</th>
<th>$\delta = .6$</th>
<th>$\delta = .7$</th>
<th>$\delta = .8$</th>
<th>$\delta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = .1$</td>
<td><strong>84.55</strong></td>
<td>482.71</td>
<td>820.94</td>
<td>883.38</td>
<td>1191.5</td>
<td>1309.4</td>
<td>1522.7</td>
<td>1642.6</td>
<td>1666.8</td>
</tr>
<tr>
<td>$c = .2$</td>
<td>151.66</td>
<td><strong>78.54</strong></td>
<td>545.82</td>
<td>1160.2</td>
<td>1475.2</td>
<td>1623.0</td>
<td>1653.5</td>
<td>1729.6</td>
<td>1734.3</td>
</tr>
<tr>
<td>$c = .3$</td>
<td>201.91</td>
<td>129.20</td>
<td>120.56</td>
<td>684.95</td>
<td>1221.8</td>
<td>1576.8</td>
<td>1774.2</td>
<td>1753.7</td>
<td>1768.0</td>
</tr>
<tr>
<td>$c = .4$</td>
<td>129.20</td>
<td>272.74</td>
<td>192.64</td>
<td>92.92</td>
<td>883.52</td>
<td>1321.3</td>
<td>1574.2</td>
<td>1771.9</td>
<td>1728.8</td>
</tr>
<tr>
<td>$c = .5$</td>
<td><strong>49.65</strong></td>
<td>139.19</td>
<td>322.48</td>
<td>330.43</td>
<td>564.98</td>
<td>1251.4</td>
<td>1576.5</td>
<td>1740.3</td>
<td>1784.6</td>
</tr>
<tr>
<td>$c = .6$</td>
<td><strong>37.11</strong></td>
<td><strong>64.41</strong></td>
<td>116.02</td>
<td>219.60</td>
<td>209.49</td>
<td>1074.0</td>
<td>1482.1</td>
<td>1638.7</td>
<td>1787.2</td>
</tr>
<tr>
<td>$c = .7$</td>
<td><strong>39.40</strong></td>
<td><strong>70.08</strong></td>
<td><strong>63.60</strong></td>
<td>165.56</td>
<td>302.91</td>
<td>642.58</td>
<td>1302.7</td>
<td>1387.5</td>
<td>1653.4</td>
</tr>
<tr>
<td>$c = .8$</td>
<td><strong>70.03</strong></td>
<td><strong>44.02</strong></td>
<td><strong>53.01</strong></td>
<td><strong>51.32</strong></td>
<td>140.34</td>
<td>207.46</td>
<td>1692.4</td>
<td>1421.9</td>
<td>1479.5</td>
</tr>
<tr>
<td>$c = .9$</td>
<td><strong>86.03</strong></td>
<td><strong>55.73</strong></td>
<td><strong>64.03</strong></td>
<td><strong>28.23</strong></td>
<td><strong>52.49</strong></td>
<td>129.84</td>
<td>259.67</td>
<td>1873.4</td>
<td>1359.0</td>
</tr>
</tbody>
</table>
of the complete model is large for each model. This suggests the rate of convergence may differ for each model.

7.5 Summary of the simulations

The simulations discussed above suggest the following characteristics of the long-run behavior of the RL, RLA, and RLC models.

In the RL model, the network formation process converges to either the empty network or the complete network depending on whether $\delta > c$ or $\delta \leq c$. The rate of convergence is related to the magnitude of $|\delta - c|$. The learning rate $\alpha$ has a strong effect on the probability the complete network forms whenever $\delta > c$ and in particular when $\delta - \delta^2 < c$.

In the RLA model, the empty network forms with high probability whenever $\delta + \delta^2 < c$. The probability of the complete network forming is largest whenever $\delta - \delta^2 > c$ and the probability of the star network is largest whenever $\delta - \delta^2 < c < \delta + \delta^2$. However, the probability of the star network forming is very low in this region, due to the coordination problem of each agent choosing the same central agent.

In the RLC model, the empty network forms with high probability whenever $\delta < c$. As with the RLA model, the probability of the complete network forming is largest in the region $\delta - \delta^2 > c$. In the region, $\delta - \delta^2 < c < \delta + \delta^2$ the wheel network was observed to have the highest probability of forming. Within this region, the wheel network alleviates the coordination problem since it has a larger degree of symmetry than the star network.

The long run behavior of each model differs in the region $\delta > c$. The two-sample $\chi^2$ test cannot distinguish between the behavior of the RL and RLC model whenever $\delta < c$. Furthermore, the behavior of the RLA is not found to be different than the RL or RLC model in regions in which $c$ is much larger than $\delta$. As noted in Watts (2001) and Jackson and Watts (2002), in the region $\delta < c$, agents may experience strong
economies of scale, in that the cost of the first link exceeds its benefits to the agents involved. However, the benefits of further links may exceed the cost.

8 Conclusion

In this paper, I develop a model of network formation in which agents decide whether or not to form links with others based on past experience and with varying degrees of additional information. I have shown that reinforcement learning agents need little information to form pairwise stable networks within the symmetric connections model, whenever the expected payoff of adding or deleting a link is always higher than the expected payoff of deleting or adding a link, respectively. However, for certain parameter ranges, the action with the highest expected payoff will depend on the network structure. More information is needed for agents using reinforcement learning to form these networks. My simulations yield evidence to suggest the network formation process will converge to either the empty network or the complete network depending on the parameters in the connections model. If I allow agents to observe the agent they are matched with and condition their actions based on this information, the network formation process may now converge to the pairwise stable networks ruled out in the first informational setting. This convergence is investigated through the use of simulations. The simulations suggest the network convergence to pairwise stable network occurs only when one action strictly dominates the other. A coordination problem arises when the star network is both pairwise stable and efficient.

The main assumption of this model is behavioral in nature. I assume a specific behavior for agents in order to allow for network formation with limited information. Testing this assumption in an economic laboratory is something that could be pursued in future research. Secondly, the assumption of symmetric costs and benefits in the connections model could be relaxed in future work.
References


