Network Formation with Adaptive Agents

Stephan Schuster

University of Surrey, UK

2010

Network Formation with Adaptive Agents

Stephan Schuster, Department of Economics, University of Surrey

December 12, 2010

Abstract

In this paper, a reinforcement learning version of the connections game first analysed by Jackson and Wolinsky [25], is presented and compared with benchmark results of fully informed and rational players. Using an agent-based simulation approach, the main finding is that the pattern of reinforcement learning process is similar, but does not fully converge to the benchmark results. Before these optimal results can be discovered in a learning process, agents often get locked in a state of random switching or early lock-in.

1 Introduction

Networks are an important paradigm for modelling social and economic relationships. Such relationships play a role, for example, in the transmission of information, e.g. about job opportunities, the spread of diseases or similar events that vary with the underlying interaction structure of groups or societies.

A characteristic feature of economic network analysis has been the strategic perspective, i.e. how networks form and are sustained. Usually some direct or indirect costs are involved in establishing and maintaining these relationships. Strategic network formation can be modelled as a game in which players decide to link or not to link to each other, depending on some value function of the network and an allocation rule that distributes the value among the players. The question of interest is what equilibria can develop, what stable states the process can converge to, and how this compares to the social optimal network that maximises the benefit of the group as a whole.

In this paper, a stylized game of communication network formation is treated. Communication networks model relationships among individuals which exhibit some benefit to the members of the network. One can think of contact networks among persons through which valuable information is exchanged. The value a participant experiences depends on the number of other persons he is linked to; the more persons there are in the network, the higher its value to the individual. In its simple form the utility of an individual is a linear function of the size of other players in the network. A more realistic form assumes decay in value the more distant the source of information is. Establishing and maintaining direct links is costly as it involves some effort, so that individual utility depends

on the relationship between costs and benefits, which finally determines the connectivity and shape of the network that can be formed.

In the game theoretic literature only few attempts have been made to model network formation processes with bounded rational agents. In most approaches, bounded rationality is modelled as limits in memory and computing capacity of agents or 'errors' or 'random noise' in the decision making process. The contribution of this paper is the application of a simple reinforcement learning (RL) model to network formation. Simulations are used to analyse and compare the model to the perfect rationality and full information case that has been extensively treated in the literature. The simulation approach is based on a computational framework that has been applied to similar problems before [35].

This paper is structured as follows: In section 2 notation and definitions are given. Sections 3 and 4 shortly discusses the related learning and network literature. Section 5 describes the model, and explains how it brings together the two perspectives reinforcement learning and network games. Section 6 describes the game theoretic results. These results serve as the benchmark to compare the reinforcement learning results. The results or the learning model are obtained using simulations, which are described and analysed in section 7.

2 Definitions and Notation

Graphs

Definition 1. *Graphs. A graph g, $g \subseteq G$, consists of a nonempty set of elements, called vertex and denoted $v_i, v \subseteq V$, and a list of pairs of vertices, called edges. Edges connecting two vertices v_i and v_j directly are denoted ij. A weighted graph is a graph in which weights are attached to the edges. The cardinality of a graph is the number of edges it contains, and is denoted with c_g.*

$\mathcal{N}$ denotes the set of all possible graphs that can be generated from V.

g+ij denotes the graph that can be obtained by adding the edge ij to graph g. Conversely, g-ij denotes the graph obtained by deleting this link.

Graphs that are obtained by adding or deleting links are called 'adajacent'.

For simplification of describing networks, an undirected, unweighted graph can be defined as follows:

Definition 2. *Network patterns. Let the vector a as the ordered in- or out-degree of all vertices. The in-degree is the number of edges arriving at vertex i, the out-degree is the number leaving from it, the sum of both is called in-out degree. In an undirected graph the in-degree equals the out-degree, since for all edges arriving at i, there must be one leading back. If the labels of the nodes are interchangeable, a describes the structure of the network completely.*

For example the structure 1,1,1,1,4 represents a star with 5 vertices, four vertices having one link, denoted by '1', and one vertex having four links to all other vertices, denoted by '4'.

Definition 3. *Network density. Network density measures how strongly the vertices of a graph are interconnected by dividing the number of existing edges by the number of possible edges. In the directed graph it is defined as* $D = \frac{1}{n*(n-1)} \sum_{i=0}^{n} \sum_{j=0}^{n} ij$*, for the undirected graph it simplifies to* $D = \frac{1}{0.5(n*(n-1))} \sum_{i=0}^{n} \sum_{j>i}^{n} ij$*. The fully connected graph has a density of 1, the empty graph a density of 0.*

Definition 4. *Shortest path. Let* P_{xy} *be a nonempty path in a weighted graph g from vertex x to vertex y, consisting of k edges* $xv_1, v_1v_2...v_{k-1}y$*. The weight of* P_{xy}*, denoted as* $W(P_{xy})$*, is the sum of the weights,* $W(xv_1), W(v_1v_2), ...W(v_{k-1}y)$*. If x=y, the empty path is considered to be a path from x to y. The weight of the empty path is zero. If no path between x and y has weight less than* $W(P_{xy})$*, then* P_{xy} *is called a shortest path between x and y, and is denoted as* SP_{xy}*.*

Definition 5. *Average path length. The average path length is the average of all shortest paths in the graph g and denoted as L:* $L = \frac{1}{n} \sum_{i \neq j}^{n} SP_{ij}$

Games on graphs In a network game, the vertices v_i represent players, and the edges the relationships they can engage in.

Network games further include value and allocation functions on the set of possible graphs G. Value functions specify what total utility is generated by the network, and the allocation rule defines how this value is distributed among the individual players.

Definition 6. *Value functions [25].*

(i) A value function vf is a mapping $vf : \{g | g \subset g^{\mathcal{N}}\} \to \mathbb{R}$
(ii) The value function cvf *is defined as the sum of individual utilities of the players:* $cvf(g) = \sum_i u_i(g)$

Definition 7. *Allocation function [25]. An allocation function* $Y : \{g \mid g \subset g^{\mathcal{N}}\}$ *distributes the value generated by* vf*. The 'equal split rule' distributes the value evenly among the players and is defined as:* $Y_e(g, v) = cvf(g)/n$*.*

Stability definitions

Definition 8. *Pairwise Stability [25]. A network is pairwise stable if*

(i) for all edges $ij \in g$*,* $Y_i(g, v) \geq Y_i(g - ij, v)$ *and* $Y_j(g, v) \geq Y_i(g - ij, v)$ *(ii) for all edges* $ij \notin g$*,* $Y_i(g, v) < Y_i(g + ij, v)$ *then* $Y_j(g, v) > Y_i(g + ij, v)$

In words: If a link between two players is stable, then there cannot be an adjacent network with higher value obtainable by deleting this link. Conversely, for any player not being part of the network, the value that can be added by

this player must be smaller than the current value, otherwise the link would be formed.

The concept of pairwise stability requires is defined for the case that at most two players act at the same time, and that the players look only one step ahead. The concept of strong stability extends pairwise stability to coalition of players:

Definition 9. *Strong Stability [23]. A network g is strongly stable with respect to Y and vf if for $H \subseteq V$ and g' obtainable from g via deviations by H, and $v_i \in H$ such that $Y_{v_i}(g', vf) > Y_{v_i}(g, vf)$, there exists $j \in S$ such that $Y_j(g', vf) < Y_j(g, vf)$.*

That is, a network can only be stable if a subset H of players has no incentive to alter it.

For dynamic models of network formation, Jackson and Watts [24] have adapted the concept of stochastic stability. In the dynamic version of the game, at each time step two randomly selected players decide to form or sever a link. The players act myopically, and base their decision on whether they are better of with the alteration in t+1, that is, they do not consider the possible consequences that may follow by changing the utility of other players. After the action is taken, with some probability $\epsilon > 0$ the alteration is applied, or with $1 - \epsilon$ not. This is a Markov chain with the states being the respective networks that are formed during the process. With $\epsilon \to 0$ the stationary distribution converges to a unique limiting stationary distribution. From this follows the next definition:

Definition 10. *Stochastic Stability [24]. A network in the support of the limiting stationary distribution of the dynamic process is stochastically stable.*

3 Reinforcement learning

Humans learn through a variety of sources, such as own experience, observation or imitation. Reinforcement learning models represent a very simple form of stimulus-response learning. Agents in such models have no explicit representation of the domain and learn by trial and error. Successful actions are rewarded, unsuccessful punished. Reinforcement learning rules are functions that induce agents to choose actions that were successful in the past more often, while they avoid actions that led to unsatisfactory outcomes. This if often referred to as the 'Law of effect'.

The main components of simple RL models typically are:

- An action set A from which an action a is chosen from, and payoffs π associated with them;
- An action strength function that updates the experience over time. The typical function is introduced in [33]:

$$q_k(t+1) = q_k(t) + \pi(t) \tag{1}$$

  which updates the strength q of the k-th action with the current reward.

- A selection function that selects successful actions based on the q_k. The selection function is usually based on Luce's choice theorem [28]:

$$p_k = \frac{q_k}{\sum q_j} \tag{2}$$

  This function computes the choice probability of action k relative to its strength q_k.

A major interest in RL models stems from experimental game theory, which discovered RL as a method that matches experimental data well (e.g., [33, 14, 8, 26, 30]. Many authors have tested different variants of the simple RL model given in equation 1, some with more cognitive capabilities than others. For example, Erev and Roth [33, 14] consider a model integrating fictitious play, adding this way some simple form of expectation formation. After fitting data of a broad range of experimental data on ultimatum games, bargaining and simplified best-shot games, they however find that the simple model fits the data best; but this might not always coincide with the equilibrium prediction. Other authors find similar results - for example, Mookerjee and Sopher [30] for constant sum games or Chen and Tang [10] for public good provision games.

Camerer and Ho [8] however argue that there are two fundamental types of learning which a realistic learning model has to integrate, experience- and belief-based learning, and that belief-based learning can add to explaining actual behaviour better when applied properly. They propose a more complex approach combining fictitious play with reinforcement learning in the experienced-weighted attraction model (EWA). They find that EWA can predict data better, however it depends on many parameters that are often not available in experimental game data sets. Later, in [7] they develop a simplified version of EWA with just two variables which predicts the data not as good, but still very accurately. Similarly, also Chen and Khoroshilov [9] find that EWA predicts better than simple RL; only Sarin's and Vahid's Payoff Assessment model [34] fits data better. In the Payoff Assessment model, players players assess expected payoffs myopically by estimating the expected payoff using average returns per actions, and choose the action with the expected maximum payoff always.

Many authors have analysed the properties of learning rules and try to establish conditions under which the actions of players converge to the optimal action in single-player decision problems or equilibrium in games. Typically these proofs rely on stochastic approximation theory. Early work has established results for limited classes of games or only simple single-agent decisions (e.g. [1, 5, 6] or for the prisoner's dilemma [26]). Only more recently more general results for the boundary behaviour of the process and larger classes of games were established (e.g. [27, 4, 20, 17]). In stationary environments in single-agent decision problems, it is found that simple ER learning leads to the selection of the optimal action with probability 1. [27] show that this happens in 2x2 games with positive probability; however it is demanding to prove that this holds with probability 1. For example, Beggs [4] can prove this for 2x2 constant sum games, but Hopkins and Posch [20] find that this cannot be generalised to any class of games.

4 Network games

In a static setup, Jackson and Wolinsky [25] show that stable and efficient networks can exist under certain conditions. Subsequent work based on this model [37, 24, 21, 12] as well as similar connection models [2, 3] provide a dynamic perspective. For a complete overview of the current state of economic network research, see [22, 19].

In [25] players are fully informed, perfectly rational and myopic. Two players can choose at a time to link to each other or not. After their decision, the network value is computed, and the value distributed among the agents according to the equal-split allocation rule. Direct links are costly, and both agents bear the costs of the link. Then the next two players are selected, who then make their decisions based on the current value of the network and the value that would result by their respective actions. As they are myopic they only consider the next state of the network. This process goes on until pairwise equilibrium is reached. Depending on the cost of links, three different equilibria can be sustained: the fully connected network, a sparsely connected network, and the empty network. However, there not all stable networks are efficient. Details are given in section 6.

Watts [37] analyses the process of forming a network in the connection model. In this dynamic process, two players are selected randomly and given the opportunity to form a link. Players are myopic, and thus anticipate in their decision only the utility of the network that forms in the next step. The process only stops if it ends in a stable network. She finds two main attractors the formation of a pairwise stable network, or a cycle of adjacent networks without any sustainable equilibrium. In the first case, there exists a path of adjacent networks that leads to a state where no player can be better off by severing a link. In the latter case, the stable network is not reachable over such a path, and the process has to cycle along the feasible networks on that path. More details are given in section 6.

Jackson and Watts [24] generalise this approach by modelling it as a stochastic process: As in the previous model, two players are selected randomly, but their decision to form or not form a link is only carried out with a certain probability $1 - \epsilon$, whereas with probability ϵ the opposite happens. The parameter ϵ may be thought of as errors individuals make in doing their calculations, or deliberate deviations in order to explore different paths. The smaller ϵ the more likely the results converges to that of Watts [37]. However, with larger random perturbations, the myopic nature of the players can be overcome by visiting networks that would not result by rational decisions. Thereby it is possible to reach a new path of adjacent networks that can lead to a pairwise stable network. As already outlined in definition 10, the dynamics can be formalised as a Markov process on the random variable ϵ. As $\epsilon \to 0$, stable networks that cannot be reached are excluded, thus selecting only those pairwise stable networks that can actually be reached. An application to the co-author model [25] demonstrates that the complete network is selected as the unique stochastically stable network out of several possible solutions. However, as they also demonstrate, there are

examples where all pairwise stable networks are also stochastically stable, so that the stochastic process does not always help to improve the predictability of outcomes in a network formation game.

Hummon [21] uses the same model specification as Watts, but simulates the model computationally to obtain his results for $n = 3$, 5 and 10 (see also [12] for a detailed, but purely descriptive follow up for $n = 5$ and $n = 6$). The most important observation in this context is that on average in all cost ranges either a star or a ring emerges as the most frequent solution. Which formation occurs depends solely on n and the order in which actors meet. As Watts derived later theoretically [37], the simulations confirm that with increasing n the frequency of the star decreases. Only in the lower cost ranges the star as the efficient network can still form, where maintaining any link is profitable.

In the noncooperative version of the game of [2] links can be formed unilaterally. Agents who initiate the forming of links have to bear all the costs, but benefits are shared. From a static perspective, there is a large number of stable networks (Nash networks), depending on the cost settings. In any setting where the benefits exceed the costs, it is a best response to link to at least one other player. Any minimal connected network is shown to be a Nash equilibrium, i.e. no player can be better off by deleting a link. In the dynamic game is modelled as a repetition of one-shot games with a random start network where all players decide simultaneously. All players observe the resulting network as well as the strategy played. Players remain with their last strategy with a probability p, or decide to play a new action with probability $1 - p$. In the latter case they decide on a best-response given the actions played by the other players in the previous round of the game. Bala and Goyal then identify limiting cases of strict Nash equilibria by looking at the changes that are induced when exactly one player adapts his strategy. Simulations are used to test whether the repeated game converges to these limits for different p and to determine the speed of convergence. They find that complete networks result in the low cost range; in the intermediate any type of star is stable; and in the high cost ranges the empty network is stable. As in the JW model, the trade-off between distance and link-minimisation is the driving force. Another observation is that simulations typically result in 'flower networks': In such networks most players are linked in the form of a ring with one or a few shortcuts between some players, contracting the distance among all players.

Beal and Querou [3] base their model on bounded rational, cooperative players. Bounded rationality is here represented as limited memory in a repeated game. Furthermore, players incur costs for *offering* the link; consequently players only offer links if they know that their opponent does the same. Under full rationality, this results in the empty network as unique Nash equilibrium. In the dynamic game, players have limited memory, but are otherwise perfectly informed about other players' past actions. The game is repeated over a finite number of time steps larger than players' memory. Players maximise their average payoff. Beal and Querou show that with this form of bounded rationality, non-empty networks can exist. Because of full information, deleting a link might be harmful as other players never offer to link to this player (expecting

that they only incur costs, but no benefits), until they forget the deviation. As a result, the costs of establishing new links cannot be too high or the potential value gained from a link must be large enough before any link can emerge.

Several more recent papers are based on the non-cooperative version of the game Many of these models look at the role that heterogeneity plays for equilibrium networks, which is however not the focus of the model presented here, but some experimental results from this branch of research is insightful. For example, McBride [29] focuses on value heterogeneity (the value of connections is different among players) and partial information. In his model, partial information is defined by observing only player's action with a distance ≤ 1. In such cases inefficient outcomes might develop, whereas under perfect information the efficient minimal connected networks are also equilibrium networks. Other authors look at heterogeneous cost for establishing links (e.g., [15]). They find that in equilibrium state, cost-heterogeneous players form either empty or star networks where the centre agent bears the costs; if values varies as well, a strict equilibrium is either the empty network or a minimal connected network with components being connected in the form of centre-sponsored stars. Experimental evidence is provided by Goeree et al. [16]. In their experiments they find that with homogeneous agents, equilibrium networks are almost never achieved. Introducing cost heterogeneity leads to development of equilibrium networks (in the form of minimal connected or star-networks), but with a lower frequency than when agents receive different value from linking.

Also only remotely related is work on other types of communication network models. These are models with farsighted players (e.g. [38, 11]), or coalition formation [13, 23, 36]. When players are allowed to form coalitions, conditions for equilibrium are stronger and thus reduce the number of possible equilibria since deviations require the consent of all concerned players in the coalition. Using definition 9, Jackson and Nouweland[23] show that strongly stable networks are efficient. When players are farsighted, situations where the costs of links formed during the process exceeds the benefits but the resulting (non-empty)network has a positive payoff for the connected players, can be overcome. However, although efficient networks could be formed in such cost ranges, this does not happen because each player wants to prevent to become the centre of a star-like structure. As a result, circle networks distributing costs and benefits equally are more likely to form.

There have been no applications of RL to strategic network formation games in particular. In their approach, only [31, 32] are remotely related. They are interested in the differences between long- and medium term behaviour of clique formation game such as Stag Hunt with ER models. Comparing the impact of different levels of the time discounting factor, they find with the help of simulations that trapping in stable states in simplified network games occur. However, while this is theoretical true for any time discounting parameter, convergence may simply never be seen in the very long run if the past is discounted heavily. Two interesting implications from these results follow: First, for any simulation approach of networks (representing usually the medium-run behaviour of RL models), the occurrence of stable states is more likely to be observed if dis-

counting is limited; second, the more complex the game the less likely trapping into stable states. Pemantle and Skyrms find this result for even very simplified network games.

Quite similar as Pemantle and Skyrms the question asked in this paper is what limiting behaviour can be observed in a reinforcement learning network formation process in the medium run. It is a purely explorative simulation study applying a RL process to a more complex game with a non-constant, non-linear reward structure. Its contribution with respect to the game literature is the application of RL to network formation games, which so far has only received very limited attention (only few authors treat learning in a very general way , e.g. [18]).

5 Model

The basis of any network model is the utility function of the players derived from a certain network structure. In the connections model, this function is:

$$u_i(g,t) = w_{ii} + \sum_{j \neq i} \delta^{t_{ij}} w_{ij} - \sum_{j:ij \in g} c_{ij} \tag{3}$$

t_{ij} is the number of links in the shortest path between individuals i and j. Links between players have a certain value w_{ij}, plus a constant 'intrinsic' value w_{ii} that each player perceives. $0 < \delta < 1$ is a decay factor by which the value of connections may decrease. $\delta^{t_{ij}}$ captures the fact that the longer the path between the two nodes, the smaller its benefit becomes. If i is not connected to j, δ is set to 0. But although direct links may be the most valuable, they come at a cost: c_{ij} denotes the costs of maintaining direct relationships (e.g. time and effort); for all indirect connections, it is set to 0.

In most network models bounded rationality was described as an injection of 'irrationality' into otherwise perfect rational players - for example an error term ϵ as in [24] or [2], or a limited memory as in [3].

In this model, limited rationality of the players is represented by reinforcement learning. Agents start with no information at all, and learn by trial and error about the game and the application of the appropriate actions. Players know only the name of the other players in the game, and may choose for each player i they encounter from the action set $A_i = \{a_0 \dots a_i \dots a_n\}$ given by {offer link, not-offer link}.

Analogous to the discussed dynamic models, this model also proceeds by picking two agents j and k at a time. Each agent decides whether to offer or not offer a link to the other agent at the same time. Afterwards, the new network is computed, and both agents are given their payoffs p based on the utility function as in equation 3.

Action strengths are updated not cumulative, as in equation 1, but by taking averages:

$$q(a_t) = q(a_{t-1}) + \gamma(p(t) - q(a_{t-1})) \tag{4}$$

Depending on the parameter γ, the action-value function updates the strength of the current action based on the weight γ of previous experiences and the current reward. For a value $\gamma = 0.5$, for example, and reasonably large t this function approximates the average payoff generated with action a. The smaller γ, the stronger the impact of past experiences; conversely, for $\gamma = 1$ only the reward of the last action is considered, and all previous experiences discarded. That is, a simple form of expectation formation about average rewards is used.

As selection function, an exponential form is used:

$$pr(a_{i,t+1}) = \frac{e^{q(a_i)\alpha}}{\sum_{j,j\neq i} e^{q(a_j)\alpha}} \tag{5}$$

This averaging mechanism deviates from many typical ER models, and insofar as agents try to approximate a true value, they mimic an expected utility maximiser. This choice is deliberate; it is assumed here that non-trivial games require more attention to changes in the payoff structure, otherwise agents will become trapped easily in inferior outcomes for long periods of time. While - at least for simple games, as [32] show - it can be assumed that equilibria are attainable for any time discounting parameter, using an average speeds up the convergence of the game to such stable states. On the other hand, trapping in the sense that the equilibrium action will be played infinitely is unlikely, as experimenting is more likely to result in changes in the player's action propensities. For the selection function, the reason for its form is similar. The more complex the environment of a reinforcement learner, the more important it becomes to react reasonably fast - hence averaging in the update function - and discriminate efficiently between actions in a short period of time - this is implemented by the exponential form. Intuitively, cumulative RL seems a feasible choice for the low cost range, as payoffs are always positive; in more sophisticated scenarios as in the medium cost range, slow learning agents can be expected to perform badly because they possibly experience rewards so lately that it cannot be associated with any network structure.

6 Benchmark model

The Static Connections Model with perfect rational players Setting w_{ii} to zero and w_{ij} to one, Jackson and Wolinsky [25] use a simplified setup of the connections network game for their analysis: $u_i(g,t) = \sum_{j\neq i} \delta^{t_{ij}} - \sum_{j:ij\in g} c_{ij}$. They prove the following properties about the stability of the possible networks if players are fully informed, rational and myopic:

- $c < \delta - \delta^2$: The complete graph is the only unique stable solution. Since benefits always exceeds costs, severing a link will result in smaller utility.
- $\delta - \delta^2 < c < \delta$: All solutions benefiting from indirect links are stable. One of the stable solutions is star, as this structure minimises the number of links and the distance among the nodes.

- $\delta < c$: The only feasible solution is the empty network no player would be willing to create a connection, even if there exists a network that yields positive payoffs.

They also show that for all n, a unique efficient network exists:

- If $c < \delta - \delta^2$ then the complete network is efficient, as for each player the utility of any direct link exceeds the benefit of an indirect link.
- for $\delta - \delta^2 < c < \delta + (n-2)/2 * \delta^2$ the star is efficient. It minimises the number of direct links while connecting all players with a minimal distance.
- for $\delta + (n-2)/2 * \delta^2 < c$ only the empty network is efficient; that is, for any situation where costs exceed the value that can be generated by the star.

The Dynamic Connections Model with perfect rational players As described above, Watts [37] analyses the process of network formation from a dynamic perspective. The main results are:

- $\delta - c > \delta^2 > 0$: Also in the dynamic model the fully connected network forms. In each period utility strictly increases for any two players not yet directly connected. Since breaking any link an agent already has reduces his payoff, no links will ever be broken, as in the static model.
- $0 < \delta - c < \delta^2$: Stable non-empty networks can form in the dynamic process. The star is most efficient and is also a pairwise stable network, although not the unique one. The condition for the star to form is that there exists a centre agent. If this agent meets another agent not yet linked to her, the centre agent will only establish the connection if this agent is not linked to anyone else. If she did, she would loose the advantage of being indirectly connected to the unconnected player. Thus, the star can only form if all agents meet the centre agent first. The larger n, the more likely it is that unconnected players meet each other before meeting the centre agent. As a consequence, the process is likely to converge to a network with only one path connecting every pair of players (i.e. a 'line network').
- $\delta - c < 0$: No link is formed. Myopic agents cannot form any links, since there is no benefit in establishing the first link, even if connected networks with a utility > 0 exist.

7 Simulations

The model has four parameters of interest, α and γ, c and δ. As in the original model, w_{ij} is set to 1, and w_{ii} to 0. Again, cost and value are the same for all players and written c and δ. δ is fixed at 0.5. For each cost range, the values for

c are drawn randomly in order to obtain some samples within each cost range. The following three cost ranges the following values result:

- Low cost range $c < \delta - \delta^2 : c < 0.25$
- Medium cost range $\delta^2 < c \leq \delta : 0.25 < c \leq 0.5$
- High cost range $c > \delta : c > 0.5$

The focus of the simulations is on variations of cost c and the learning parameter α. The greater α the more likely exploration in the action selection process; the smaller the faster the agents stick to the first best solution. In the context of this article, α can also be interpreted as the degree of rationality - the larger α, the more 'irrational' players act in the sense of random noise. α is incremented in steps of 0.01 beginning from 0.1 and ends at 1.

For the main analysis, the time discounting parameter γ is held constant. It is mainly seen as a tuning parameter to find the best fitting models for a detailed analysis of α. Some experiments with different γ were used to select the best model for a more detailed analysis of α. A short overview of these experiments is given at the end of the next section. As indicated in the previous section, time discounting can mainly be interpreted as only influencing the speed of discovering true values. To investigate the learning properties of the model, only the relationship between exploration/exploitation and stable states of the game are considered. The concrete γ values are chosen from a set of {0.1, 0.25, 0.75, 1}, depending on the fitness of the solution in the respective cost range.

Additionally to the definitions introduced at the beginning, three additional measures are defined here:

Definition 11. *Stability.* $S_t = 1 - \frac{1}{2}\frac{n_{(g,t-1)} - n_{(g,t-1)}}{(n(n-1))}$

Stability is simply the difference in the number of links between two time steps, divided by the number of maximum possible links to standardise the measure. For a single simulation step, the value can be either 0 or 1. Over a sample of simulations, S_t can be interpreted as the probability that a link changes at t. It thus varies between 0 and 1 and the closer it is to 0 the more stable the network is.

To compare the results with the game-theoretic prediction, a fitness measure is defined as follows:

Definition 12. *Fit / Efficiency. Let the vector* g_{stable} *be the stochastic stable network (efficient network), and* g_{actual} *a simulated network. Let* $steps_{max}$ *be the maximum number of modifications starting from any network to* g_{stable}*, and* $steps_{actual}$ *the number of modifications to reach* g_{stable} *from* g_{actual}*. Define the fitness at time t as:* $fit_t = \frac{1}{2}(\frac{steps_{actual,t}}{steps_{max}} + \frac{steps_{actual,t}}{steps_{max}} S_t)$

The resulting measure varies between 0 and 1 and tends towards 1 the closer the network structure to the stochastic stable network and the more stable the simulation result (multiplying the distance with S_t and adding it in the enumerator has the effect that stable states are weighted higher as $S_t = 0$ if a

linked changed, 1 otherwise). Computing stochastic stability according to [24] revealed that only the star is stochastic stable, and was therefore chosen as the benchmark (however, this result is strong; recall [37]'s finding that the likelihood that stars actually develop depends on the order how players meet).

8 Results

8.1 Structural properties of the RL process

Simulations were run for at least 10.000 time steps with several repetitions per α and γ value for each cost range, giving a reasonably large sample. A single simulation was typically run for 2000 steps.

A comparison of the fitness measure in each cost range revealed that the γ and α combinations maximising fitness are: $\alpha = 0.1$ and $\gamma = 0.75$ for the low, $\alpha = 0.04$ and $\gamma = 0.25$ for the medium, and $\alpha = 0.02$ and $\gamma = 0.75$ for the high cost range (in the high cost range several combinations achieve a fit of 1. Out of the top 20 results the simulation belonging to the most frequent γ value and the highest α was chosen). Some more simulations with different cost values for these specific values were simulated. Using network density as an indicator, figure 1 illustrates connectivity as a function of cost.

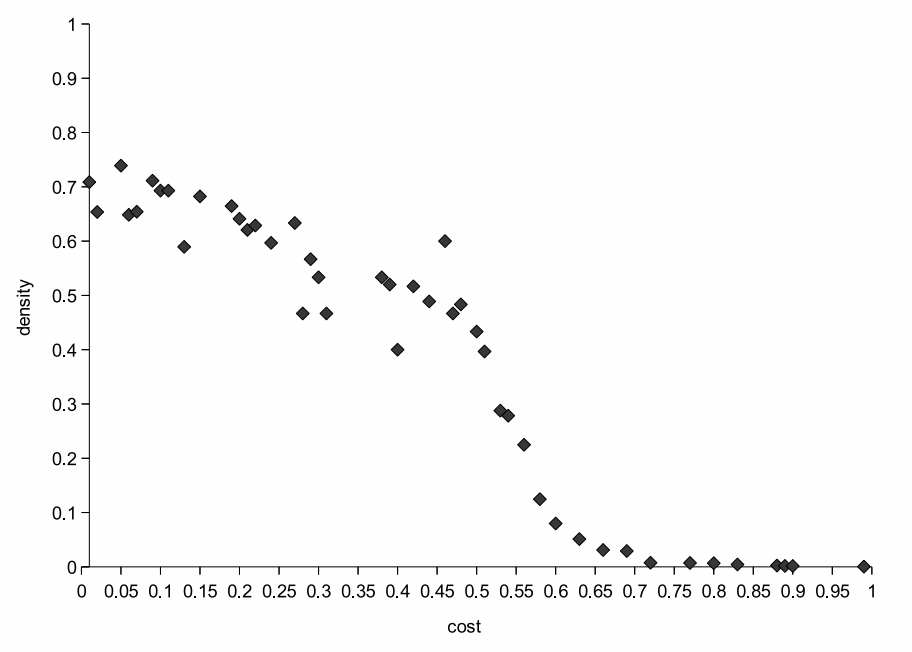

Figure 1: Network density over all cost samples.

In the high cost range ($c > 0.5$) the empty network emerges as solution. In the low cost range ($c < 0.25$), connectivity is high (almost fully connected structures). In the medium cost range ($0.25 < c \leq 0.5$) networks become gradually more sparse the higher cost becomes (density between 0.4 and 0.5). For the 'border' regions between low and medium as well as medium and high cost range connectivity changes gradually. In the RL process no threshold function between cost ranges emerges as would be expected from perfect rational, fully informed players.

The following paragraphs analyse the behaviour of the simulations as a function of α. γ is hold constant at the value maximising the fitness in each cost range.

Summary measures per α Figures 2, 3 and 4 show how density, stability and fit develop in the low, medium and high cost ranges.

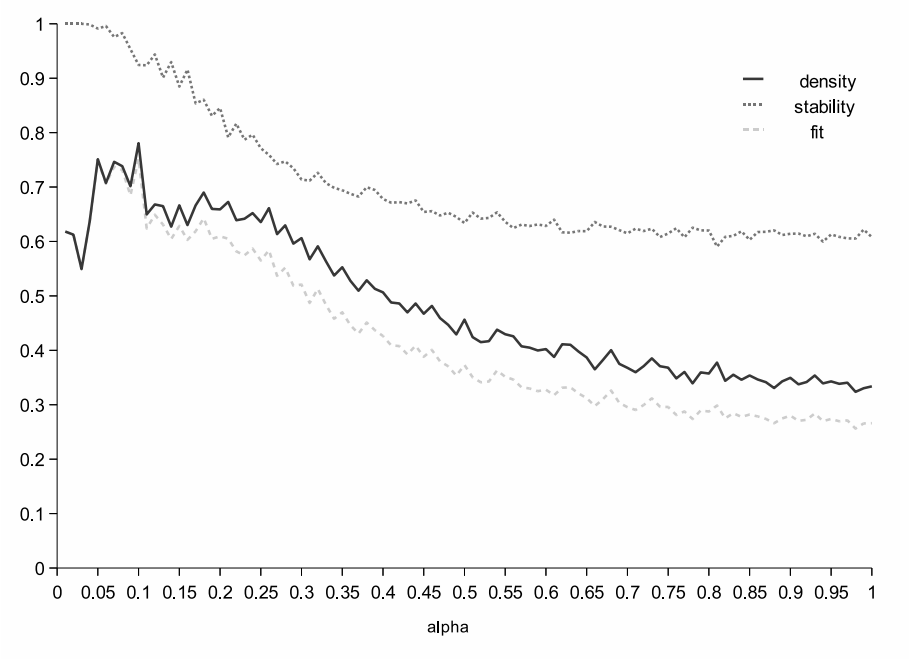

Figure 2: Network density, stability and fit for the low cost range.

In the *low cost range* ($c < \delta - \delta^2$) the optimal solution is the fully connected network ($D = 1$). Figure 2 shows that for small $\alpha(<\approx 0.07)$ the network is strongly connected ($D \approx 0.7$) without reaching the complete network, and stability tends towards 1. As could be expected, for small α, agents tend to stick to first-best solutions, which are those providing the largest increase in marginal utility. With α increasing towards ≈ 0.11, the network is developing towards the

fully connected network ($D \approx 0.8$). However, this comes on the cost of stability, i.e. some agents keep switching. Finally, for $\alpha > 0.6$ connectivity and variability approach a limit in an asymptotic manner with density about 0.35. The random limit is given by the probability that offered links are accepted. Assuming total randomness, the chance of offering a link is 0.5, the chance that the other players offers a link at the same time is equally 0.5. Thus, the probability that a link can actually be formed by pure chance is 0.25. This indicates that RL performs better than randomness, even for the smallest degree of rationality possible.

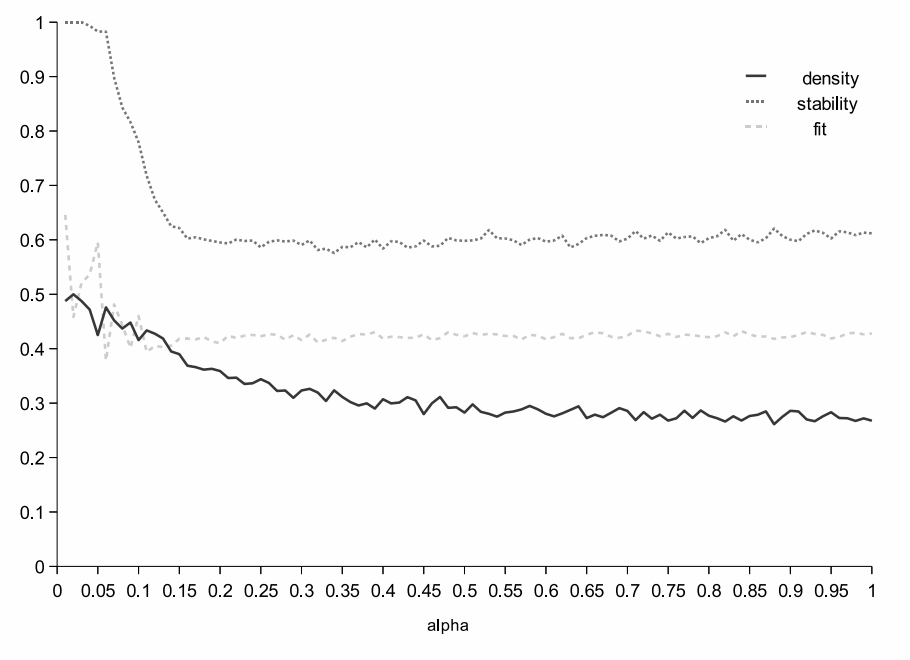

Figure 3: Network density, stability and fit for the medium cost range.

According to the benchmark model, in the *medium cost range* ($\delta^2 < c \leq \delta$) minimal connected networks should form (i.e. $D \approx 0.5$). Computations showed that the star is the efficient as well as stochastic stable pairwise network. In the simulations, agents end up very close to a minimal connected network ($D \approx 0.5$) for $\alpha < 0.06$. These networks are very stable. For $0.07 < \alpha < 0.15$ there is a decrease in density to ≈ 0.4, with a sharp drop in stability and corresponding decreases in fit. For $0.15 \leq \alpha < 0.3$ density decreases further. For $\alpha > \approx 0.3$ the connectivity of the network settles asymptotically near to the random limit; similar to the low cost range the RL process performs also here (slightly) better than random. The density of ≈ 0.4 in the range $0.07 < \alpha < 0.15$ indicates that networks are not overconnected. The sharp decrease in stability points however to random matching rather than reinforcements. This means that the observed

network structures may develop simply because they are closer to a random outcome.

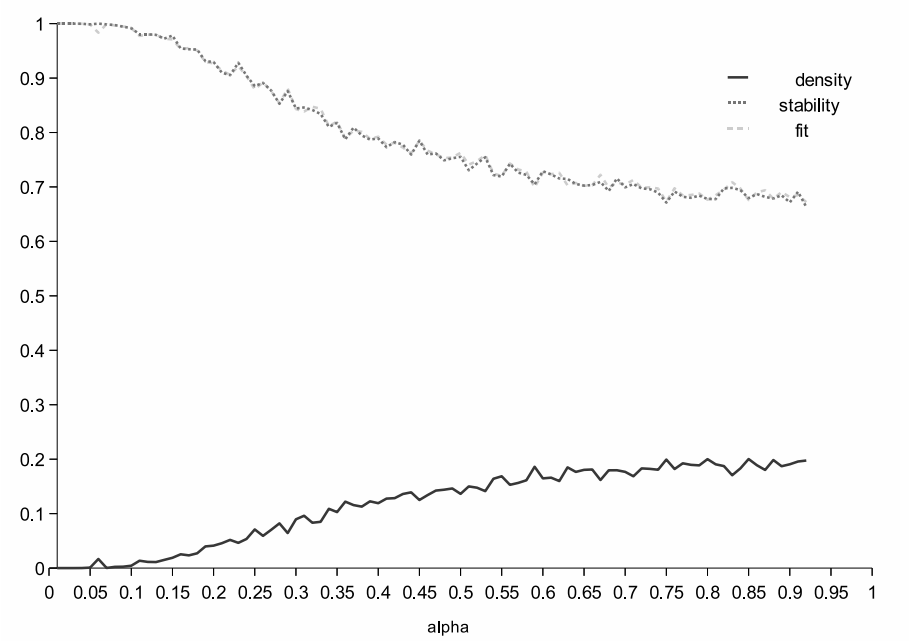

Figure 4: Network density, stability and fit for the high cost range.

In the *high cost range* ($c > \delta$) the empty network is expected. Although for some agents positive utility by indirect links could be generated in the high cost range, there is always at least one agent for which the costs exceeds the value it receives and thus motivates the deletion of direct links. Figure 4 shows that the simulation converges to the equilibrium prediction if agents explore little ($\alpha <\approx 0.15$). For $\alpha > 0.25$, at least two agents are linked ($D = 0.1$). The random limit is approached for α values >0.3, a situation where at least two agents are linked. At $\alpha \approx 0.6$ the simulation converges to the random limit ($D \approx 0.2$). Here again, RL clearly performs better than a random process.

Network structures The following paragraphs show summary measures of the most frequent network structures. To group the results a cluster analysis for α on the variables D and S was performed. These variables have been chosen because they characterise both dimensions network structure (D) as well as the more time-dependent variable stability. For the resulting clusters, the emerging networks are characterised by network structure, density D, average path length L, fitness fit, stability S and efficiency E. The choice of 3 cluster centres reflects roughly the main dynamics observed in figures 2, 3 and 4: A good fit in the

Cluster		*Network*	*D*	*L*	*S*	*Fit*	*Share*
α	$0.01 - 0.24$	1,2,3,3,3	0.6	1.88	0.92	0.58	0.07
avg(D)	0.68	1,2,2,2,3	0.5	2.05	0.92	0.48	0.09
avg(L)	1.67	2,3,3,3,3	0.7	1.63	0.93	0.67	0.1
avg(S)	0.93	3,3,4,4,4	0.9	1.38	0.93	0.87	0.11
avg(Fit)	0.66	1,2,2,3,4	0.6	1.75	0.93	0.58	0.12
avg(E)	0.66	2,3,3,4,4	0.8	1.5	0.92	0.77	0.15
		2,2,2,3,3	0.6	1.75	0.94	0.58	0.17
		2,2,3,3,4	0.7	1.63	0.93	0.68	0.18
α	$0.25 - 0.46$	1,1,2,2,2	0.4	2.22	0.66	0.33	0.08
avg(D)	0.59	2,3,3,4,4	0.8	1.5	0.75	0.7	0.1
avg(L)	1.84	1,1,2,3,3	0.5	2	0.7	0.42	0.1
avg(S)	0.71	1,2,3,3,3	0.6	1.88	0.71	0.51	0.11
avg(Fit)	0.51	1,2,2,3,4	0.6	1.75	0.72	0.52	0.14
avg(E)	0.51	2,2,2,3,3	0.6	1.75	0.73	0.52	0.15
		2,2,3,3,4	0.7	1.63	0.74	0.61	0.15
		1,2,2,2,3	0.5	2.06	0.69	0.42	0.18
α	$0.47 - 1.0$	1,2,2,2,3	0.5	2.06	0.62	0.4	0.07
avg(D)	0.26	1,1,1,2,3	0.4	2.25	0.62	0.32	0.08
avg(L)	0.51	1,1,1,1,2	0.3	0	0.61	0.24	0.08
avg(S)	0.63	1,1,2,2,2	0.4	2.18	0.62	0.32	0.09
avg(Fit)	0.21	0,1,1,1,1	0.2	0	0.63	0.16	0.1
avg(E)	0.21	0,1,1,2,2	0	0	0.62	0.24	0.17
		0,0,0,1,1	0.1	0	0.65	0.08	0.19
		0,0,1,1,2	0.2	0	0.64	0.16	0.21

Table 1: Low cost range results

lower α regions, then a decreasing region, finally approximation of the random limit. The tables show the most network architectures that occurred during this process. For readability, only the upper quartile is represented. The share of each network is based on the frequency of its occurrence in the quartile.

For the *low cost range* ($c < \delta - \delta^2$), table 1 shows the following: In the first cluster, the most common visited networks are 2,3,3,4,4; 2,2,2,3,3 and 2,2,3,3,4 with a relatively high connectivity ($D = 0.6$ - 4 missing links to the complete network; and $D = 0.8$ - 2 missing links to the complete network). The path lengths of 1.5-1.75 indicate that most networks are connected in a way that each other player can be reached directly or with one intermediary at maximum. In the second α range, the most frequent networks 2,2,2,3,3; 2,2,3,3,4 and 1,2,2,2,3 are still connected more densely than spares networks, but are also quite unstable ($S \approx 0.7$ as compared to ≈ 0.9 in the first cluster). Finally, cluster 3 illustrates that with $\alpha \to 1$, network density approaches its random limit 0.25, with frequent unconnected networks (i.e, $L = 0$).

Cluster		*Network*	*D*	*L*	*S*	*Fit*	*Share*
α	$0.01 - 0.2$	1,2,2,3,4	0.6	1.75	0.85	0.62	0.07
avg(D)	0.46	1,1,2,3,3	0.5	2	0.78	0.44	0.07
avg(L)	1.78	2,2,2,2,2	0.5	1.88	0.94	0.16	0.09
avg(S)	0.83	2,2,2,3,3	0.6	1.75	0.85	0.31	0.14
avg(Fit)	0.44	0,1,1,2,2	0.3	0	0.72	0.43	0.14
avg(E)	0.44	1,1,2,2,2	0.4	2.37	0.81	0.3	0.15
		1,1,1,2,3	0.4	2.25	0.88	0.63	0.17
		1,2,2,2,3	0.5	2.04	0.85	0.46	0.22
α	$0.21 - 0.4$	1,2,2,2,3	0.5	2.06	0.54	0.39	0.09
avg(D)	0.24	1,1,1,2,3	0.4	2.25	0.57	0.52	0.09
avg(L)	0.43	1,1,1,1,2	0.3	0	0.59	0.4	0.09
avg(S)	0.5	0,1,1,1,1	0.2	0	0.62	0.27	0.1
avg(Fit)	0.38	1,1,2,2,2	0.4	2.16	0.56	0.26	0.11
avg(E)	0.38	0,0,0,1,1	0.1	0	0.66	0.41	0.14
		0,1,1,2,2	0.3	0	0.6	0.4	0.19
		0,0,1,1,2	0.2	0	0.63	0.54	0.19
α	$0.41 - 1.0$	0,1,2,2,3	0.4	0	0.56	0.52	0.07
avg(D)	0.25	1,1,1,2,3	0.4	2.25	0.56	0.52	0.07
avg(L)	0.32	1,1,2,2,2	0.4	2.15	0.55	0.26	0.08
avg(S)	0.62	1,1,1,1,2	0.3	0	0.59	0.34	0.08
avg(Fit)	0.35	0,1,1,1,1	0.2	0	0.63	0.27	0.11
avg(E)	0.35	0,1,1,2,2	0.3	0	0.6	0.4	0.17
		0,0,0,1,1	0.1	0	0.68	0.42	0.19
		0,0,1,1,2	0.2	0	0.64	0.55	0.22

Table 2: Medium cost range results

In the *medium cost range* ($\delta^2 < c <= \delta$) , relatively stable networks close to minimal connected networks form in the first cluster. The network 1,2,2,2,3 is the most common, with an average path length of 2.04, meaning that now often at least one intermediary connects two different players. This is close to a ring, the structure minimising the costs while at the same time distributing them evenly so that no incentives for deviation exist. This coincides with results of [37], and - for the noncooperative game - of [2]. The ring itself has a share of 0.09. In this range, more efficient structures (1,1,2,2,2; 1,1,1,2,3) are more common. Also in the other clusters, the only connected network are also those 'flower networks', that is ring-structure with a few shortcuts that decrease the distance between the players. However many networks are unconnected. Whereas D indicates a relatively close match with pairwise stable networks (these are: 1,1,1,1,4; 1,2,2,3,4; 1,3,3,3,4; 2,3,3,3,3; 2,2,2,3,3; 1,1,2,2,4; 2,2,2,2,2 for cost close to the low cost limit, plus the more sparse structures 1,2,3,3,3; 1,1,2,3,3; 1,2,2,2,3; 1,1,1,2,3 for costs close to the high cost range), the distance to the unique stochastic stable network 1,1,1,1,4 is larger as in the low cost range. That is, while rational myopic players according to the stochastic process of [24] are most likely to end up with a star network, the RL process does not converge to this result, but remains somewhere between ring and star.

In the first cluster of the*high cost range* ($c > \delta$), the most frequent network is the empty network with a share of 0.73. In the most frequent non-empty network only two players are connected. In the other clusters, non-empty networks are more frequent. In the other clusters, at least two players connect most of the time - in the second cluster, only 30% of the networks remain unconnected, in the third cluster only 15%.

8.2 Memory effects

To round up this discussion, summary measures are reported for simulation runs with different γ values while holding α constant. For each cost range, the optimal α values were chosen - 0.1 in the low, 0.04 in the medium, and 0.02 in the high cost range.

Table 4 shows that in the low cost range fit and connectivity are best for the higher γ value. Moreover, a γ value of 1 increases connectivity as compared to smaller values, but also has an effect on the stability of the network, as the probability of deviations is the highest. $\gamma = 0.75$ seems to compromise well between exploration on the one hand, and stability on the other.

In the medium cost range $\gamma = 0.25$ is optimal. Also here higher γ values, but also $\gamma = 0.1$, induce higher density - which is inefficient in this scenario. Furthermore, $\gamma = 0.1$ and $\gamma = 0.25$ both maximise path length, which means that here are networks that connect the players in the most sparse way. Note that the better fit of $\gamma = 0.25$ is due to the choice of stochastic stability as benchmark. When looking at the shorter path length at $\gamma = 0.75$ this actually points to a prevalence of circle networks as predicted by [37].

Although in the high cost range there seems the fitness score is very high irrespective of γ, also here it seems that - analogous to the low cost range

Cluster		*Network*	*D*	*L*	*S*	*Fit*	*Share*
α	$0.01 - 0.27$	1,1,1,2,3	0.4	2.25	0.41	0.42	0.01
avg(D)	0.05	0,1,1,1,1	0.2	0	0.76	0.7	0.02
avg(L)	0.09	0,1,1,2,2	0.3	0	0.7	0.59	0.03
avg(S)	0.91	1,1,2,2,2	0.4	2.39	0.65	0.5	0.04
avg(Fit)	0.9	0,0,1,1,2	0.2	0	0.77	0.71	0.04
avg(E)	0.9	0,0,0,1,1	0.1	0	0.88	0.85	0.14
		0,0,0,0,0	0	0	0.99	0.99	0.72
α	$0.28 - 0.51$	0,1,1,1,3	0.3	0	0.65	0.58	0.02
avg(D)	0.08	1,1,1,1,2	0.3	0	0.64	0.57	0.02
avg(L)	0	0,1,1,2,2	0.3	0	0.64	0.57	0.04
avg(S)	0.75	0,1,1,1,1	0.2	0	0.72	0.69	0.07
avg(Fit)	0.76	0,0,1,1,2	0.2	0	0.72	0.69	0.16
avg(E)	0.76	0,0,0,0,0	0	0	0.89	0.95	0.31
		0,0,0,1,1	0.1	0	0.81	0.81	0.38
α	$0.51 - 1.0$	1,1,2,2,2	0.4	2.13	0.55	0.47	0.03
avg(D)	0.16	0,1,1,1,3	0.3	0	0.62	0.57	0.04
avg(L)	0.05	1,1,1,1,2	0.3	0	0.61	0.56	0.05
avg(S)	0.71	0,1,1,2,2	0.3	0	0.62	0.67	0.1
avg(Fit)	0.72	0,1,1,1,1	0.2	0	0.68	0.91	0.11
avg(E)	0.72	0,0,0,0,0	0	0	0.82	0.57	0.15
		0,0,1,1,2	0.2	0	0.68	0.67	0.22
		0,0,0,1,1	0.1	0	0.75	0.79	0.32

Table 3: High cost range results

- shorter memory favours efficiency and stability. $\gamma = 0.75$ provides the most stable solutions. Also here, $\gamma = 1.0$ induces too fast switching, increasing density while at the same time reducing stability.

8.3 Discussion

The networks in *low cost range* ($c < \delta - \delta^2$) almost never approach full connectivity. At the beginning of the process, the first links provide the highest utility and reinforce link offers on both sides of a player pair. Once the network is complete, the marginal utility of exchanging an indirect for a direct link is small. This holds especially for flower networks, which contract the distance between players the most. The selection probabilities become thus very similar for both forming and severing a link. Only scenarios with little experimentation can become trapped in a stable state. In most other scenarios, probability switching occurs for the players acting last. The larger α, the greater the chance that a link is severed, and the more likely that the marginal agent's switch between linking and not linking. This could result in a cycle where most of all players

Cost range	γ	D	L	S	*fit*
$(c < \delta - \delta^2)$	0.1	0.54	1.95	0.99	0.54
	0.25	0.54	1.85	0.99	0.54
	0.5	0.59	1.78	0.95	0.58
	0.75	0.67	1.65	0.91	0.65
	1	0.63	1.74	0.94	0.61
$(\delta^2 < c <= \delta)$	0.1	0.5	1.76	1	0.47
	0.25	0.46	2.01	0.99	0.56
	0.5	0.48	2.08	0.99	0.42
	0.75	0.45	1.82	0.97	0.5
	1	0.46	2.01	0.98	0.43
$(c > \delta)$	0.1	0.21	1.25	0.99	0.99
	0.25	0.002	0	1	0.99
	0.5	0.0005	0	1	1
	0.75	0	0	1	1
	1	0.004	0	1	0.99

Table 4: Simulation results for various γ

are at some stage the marginal agent that is not worth linking at some point in time. The networks that are being formed during the process are thus very similar, but not fully stable. This can also be inferred from the trends in density and stability: For the smallest α values stability is highest, but not density. As α increases, stability decreases stronger than density increases. Also the distribution of visited network structure does not change very much, which means that similar network structures exist at any particular point of time, but with more frequently changing links.

Up to the level where the utility of not being linked is smaller than being linked, the learning process in the *medium cost range* $(\delta^2 < c <= \delta)$ follows a similar dynamic as in the low cost range. Once utility becomes negative, the average rewards decrease strongly and prevents further linking. Thus the cost settings act as a natural cut-off to the reward perceived by the agents. In the low cost range there is no such bound, but the additional utility becomes very small, leading to random switching. The closer cost to $\delta - \delta^2$, the more similar behaviour in the medium cost range becomes to behaviour in the lower cost range - density increases. The optimal γ is with 0.25 lower than the optimum in the other cost ranges, which means that agents are more tolerant to deviations; this is plausible since utility is not strictly increasing. Allowing some tolerance for deviating behaviour ensures that the network does not collapse quickly as a consequence of a deviating agent. Note that the stochastic stable network is a star, which is a strong assumption. The RL networks rarely match this prediction. Since most networks here are minimally connected, the networks that form are similar as in [37] ring-like structures.

The results for the *high cost range* ($c > \delta$) largely reflects the equilibrium prediction. Here the learning task for the agents is the simplest because there are very few non-empty networks in which an agent can experience a positive reward. The only deviation is induced by the increased randomness in the action probabilities with increasing α. Also, the optimal γ value of 1 shows that the optimal agents react very quickly with no memory at all to alterations in the network structures. This is plausible, since independent of the history, any addition of a link has always a negative impact for at least one agent - which was also stated in the dynamic benchmark model.

The networks evolving from the learning model differ with the exception of the high cost range quite considerably from the equilibrium prediction. A closer look at the data showed that for the optimal α and γ values (0.1/0.75, 0.01/0.25 and 0.07/1), only 1 % of all networks in the low range were the stochastic stable outcome (4,4,4,4,4 the only pairwise/strongly and stochastic stable network) as compared to 13 % of the most frequent network 2,2,3,3,4; in the medium cost range 51% of all visited networks were pairwise stable, but only 19% stochastic stable; only in the high cost range 71% of all networks were the predicted empty network. Looking at the structure of the networks that evolved, it is more accurate to speak of two characteristic cost ranges one with $c < \delta$ and one with $c > \delta$. In the ranges where positive utility is achievable, agents form sparsely connected networks, adding some shortcuts contracting the distance among them (flower networks). The smaller the cost, the closer the resulting networks are to the complete network, the higher the cost the more sparse the resulting network will be - independent of whether the cost is in the low or medium range. The shorter the distance between agents in the network, the more undecided agents become whether to connect to some other player directly or not. If $\delta^2 - \delta < c < \delta$, the RL process matches pairwise stable networks more often because utility is increasing with the first additional links, but later with too many direct links decreasing. Another factor is simply chance - sparse networks are simply closer to the random limit of 0.25.

9 Conclusion

In this paper a reinforcement-learning version of Jackson/Wolinsky's connections model was studied with simulations. Using the concept of stochastic stability as developed in [24], the results have been compared with the equilibrium predictions of the full rationality benchmark model. Results indicate the patterns (high connectivity in the low, medium connectivity in the medium, and low connectivity in the low cost ranges) are similar, but that there is some considerable distance between the equilibrium and learning predictions.

The outcome of the RL process is similar as in the original model driven by the shape of the utility function, which has very different forms depending on the cost range. Thus, in the low range utility is convex, but always positive; in the medium cost range it slopes downwards after a certain density of the network is reached; in the high cost range it is strictly negative. For a probabilistic

choice model this results in random switching in the low cost range the more connected the network becomes, low rates of experimentation in the medium range once utility starts to decrease; and punishment of any links in the high cost range. Whereas the full rationality model expects three different regions, divided sharply by full / sparse / empty networks, the RL model exhibits a gradual decrease in connectivity from almost complete to empty networks. There are no apparent thresholds once the characteristic cost range boundaries are crossed. Furthermore, there is a tension between stability of the learning process and the pairwise as well as stochastic stability. Coordination in the learning model requires slow updates, which results in suboptimal exploration of the state space.

This model focused on a limited set of parameters. Variations on some of the other parameters might be interesting for future work. Examples are the size of the network n or the value δ of the network. As the most fully connected networks were those with the lowest cost, it seems likely that increasing the difference between δ and c will result in denser networks in lower cost ranges. Furthermore, connectivity in low cost ranges is likely to decrease with larger n as it can be expected that the relative difference of rewards between linking and not linking will increase at a similar rate as in smaller networks. As stable states are reached quickly once the network becomes sparsely connected, this will typically be at lower density values the larger n becomes.

References

[1] W. Brian Arthur. On designing economic agents that behave like human agents. *Journal of Evolutionary Economics*, 3(1):1–22, 1993.

[2] V. Bala and S. Goyal. A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229, 2000.

[3] Sylvain Beal and Nicolas Querou. Bounded rationality and repeated network formation. *Mathematical Social Sciences*, 54:71–89, 2007.

[4] A. W. Beggs. On the convergence of reinforcement learning. *Journal of Economic Theory*, 122:1–36, 2005.

[5] T. Boergers and R. Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77:1–14, 1997.

[6] Tilman Boergers and Rajiv Sarin. Naive learning with endogenous aspirations. *International Economic Review*, 41(4):921–950, 2000.

[7] Colin F. Camerer, Juin-Kuan Chong, and Teck H. Ho. Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133:177–198, 2007.

[8] Colin F. Camerer and Teck H. Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

[9] Yan Chen and Yuri Khoroshilov. Learning under limited information. *Games and Economic Behavior*, 44:1–25, 2003.

[10] Yan Chen and Fang-Fang Tang. Learning and incentive-compatible mechanisms for public goods provision: An experimental study. *Journal of Political Economy*, 106(3):633–662, 1998.

[11] Frederic Deroian. Farsighted strategies in the formation of a communication network. *Economic Letters*, 80:343–349, 2003.

[12] Patrick Doreian. Actor network utilities and network evolution. *Social Networks*, 28(2):137–164, 2006.

[13] B. Dutta and S. Mutuswami. Stable networks. *Journal of Economic Theory*, 76:251–272, 1997.

[14] I. Erev and A. Roth. Predicting how people play games: reinforcement learning in experimental games with unique mixed-strategy equilibria. *American Economic Review*, 88, 1998.

[15] Andrea Galeotti, Sanjeev Goyal, and Jurjen Kamphorst. Network formation with heterogeneous players. *Games and Economic Behavior*, 54:353–372, 2006.

[16] Jacob K. Goeree, Arno Riedl, and Aljaz Ule. In search of stars: Network formation among heterogeneous agents. *Games and Economic Behavior*, 67:445–466, 2009.

[17] Nicholas M. Gotts, Luis R. Izquierdo, Segismundo S. Izquierdo, and J. Gary Polhill. Transient and asymptotic dynamics of reinforcement learning in games. *Games and Economic Behavior*, 61:259–276, 2007.

[18] Sanjeev Goyal. Learning in networks: a survey. University of Essex, 2003.

[19] Sanjeev Goyal. *Connections*. Princeton University Press, 2007.

[20] Ed Hopkins and Martin Posch. Attainability of boundary points under reinforcement learning. *Games and Economic Behavior*, 53:110–125, 2005.

[21] N.P. Hummon. Utility and dynamic social networks. *Social Networks*, 22:221–249, 2000.

[22] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[23] Matthew O. Jackson and Anne van den Nouweland. Strongly stable networks. *Games and Economic Behavior*, 51:420–444, 2005.

[24] M.O. Jackson and A. Watts. The evolution of social and economic networks. *Journal of Economic Theory*, 106:265–295, 2002.

[25] M.O. Jackson and A. Wolinsky. A strategic model of social and economic networks. *Journal of Economic Theory*, 71:44–74, 1996.

[26] Rajeeva Karandikar, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo. Evolving aspirations and cooperation. *Journal of Economic Theory*, 80:292–331, 1998.

[27] Jean-Franois Laslier, Richard Topol, and Bernard Walliser. A behavioral learning process in games. *Games and Economic Behavior*, 37:340–366, 2001.

[28] Robert D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

[29] Michael McBride. Imperfect monitoring in communication networks. *Journal of Economic Theory*, 126:97–119, 2006.

[30] D. Mookherjee and B. Sopher. Learning and decision costs in experimental constant-sum games. *Games and Economic Behaviour*, 19:97–132, 1997.

[31] R. Pemantle and B. Skyrms. A dynamic model of network formation. *Proceedings of the National Academies of Science*, 97:9340–9346, 2000.

[32] R. Pemantle and B. Skyrms. Network formation by reinforcement learning: the long and medium run. *Mathematical Social Sciences*, 48(3):315–327, November 2004.

[33] A. Roth and I. Erev. Learning in extensive form games: Experimental data and simple dynamic models in the intermediate run. *Games and Economic Behaviour*, 6, 1995.

[34] Rajiv Sarin and Farshid Vahid. Payoff assessments without probabilities: A simple dynamic model of choice. *Games and Economic Behavior*, 28:294–309, 1999.

[35] Stephan Schuster. Bra: An algorithm for simulating bounded rational agents. *Computational Economics*, pages 1–19, 2010. 10.1007/s10614-010-9231-1.

[36] Marco Slikker and Ann van den Nouweland. Network formation models with costs for establishing links. *Review of Economic Design*, 5:333–362, 2000.

[37] A. Watts. A dynamic model of network formation. *Games and Economic Behaviour*, 34:331–341, 2001.

[38] Alice Watts. Non-myopic formation of circle networks. *Economic Letters*, 74:277–282, 2002.