# Network formation by reinforcement learning: the long and medium run

Robin Pemantle [1,2]
Brian Skyrms [3]

**ABSTRACT:**
We investigate a simple stochastic model of social network formation by the process of reinforcement learning with discounting of the past. In the limit, for any value of the discounting parameter, small, stable cliques are formed. However, the time it takes to reach the limiting state in which cliques have formed is very sensitive to the discounting parameter. Depending on this value, the limiting result may or may not be a good predictor for realistic observation times.

Keywords: absorption, discount, urn model, Friedman, Stag Hunt, urn, stochastic approximation, meta-stable, trap, three-player game, potential well, exponential time, quasi-stable.

Subject classification: Primary: 60J20

[1]Research supported in part by National Science Foundation grant # DMS 0103635
[2]Department of Mathematics, The Ohio State University, 231 W. 18 Ave., Columbus, OH 43210
[3]Corresponding author. Trl.: +1-949-8246495; fax: +1-949-8248388.
E-mail address: bskyrms@uci.edu (B. Skyrms)
School of Social Sciences, University of California at Irvine, Irvine, CA 92607

# 1 Introduction

Each day, each member of a small group of individuals selects two others with whom to interact. The individuals are of various types, and their types determine the payoff to each from the interaction. That is to say that the interaction is modeled as a symmetric 3-person game. Probabilities of selecting individuals evolve by reinforcement learning, where the reinforcements are the payoff of the interaction. We consider two games. The first is a degenerate game, "Three's Company". Here there is only one type and everyone gets equal reinforcement of for every interaction. The analysis of "Three's company" is then used in the analysis of a second game, a three-person Stag Hunt. Here there are two types, Stag Hunters and Hare Hunters. Hare Hunters always get a payoff of 3, but a Stag Hunter gets a payoff of 4 if he interacts with two other Stag Hunters, otherwise he gets nothing.

We do not pretend to be giving a realistic account of the friendship formation or of the small group dynamics of hunting. But we hope that our modeling exercise has a significance that is more than purely philosophical. This hope is based on three considerations. First, we use a model of reinforcement learning that is backed by a large body of laboratory data for both animals and humans, including interactive human learning in games. In analyzing the models we focus on ranges of parameter values of the learning dynamics that have been found in the experimental literature. Second, the notion of modeling the co-evolution of interaction networks and strategies was set out in Skyrms and Pemantle (2000). This calls for the formulation and analysis of the most basic stochastic network models. The idea is that basic, simple models can serve as modules for the construction of more complex models. To this end, we provide an analysis of two simple models. As the second model evolves, the first model appears as a module. Stag Hunters end up playing "Three's company" among themselves. Finally, we aim at a robust analysis. Our analysis of "Three's Company" and "Stag Hunt" is sufficiently robust so as to shed light on similar models, which share a common mathematical description. We aim to provide a collection of rigorous results on this class of stochastic models, which will help scientists to understand the similarities and differences between their long- and medium-term behavior.

# 2 Reinforcement learning and reinforcement processes

*Reinforcement learning*, as the term is used in psychology, and *reinforcement models*, as used in applied mathematics, are not coextensive. We will model reinforcement learning by a certain kind of reinforcement process, following Herrnstein (1970) and Roth and Erev (1995). In a reinforcement process, as defined in the mathematical literature on reinforced random walks Coppersmith and Diaconis (1997); Davis (1990, 1999); Pemantle (1988, 1992) the current probability can depend on the entire history of choices and payoffs, via summary statistics or *propensities* associated to the possible actions. The possibility of using such processes to model reinforcement learning was introduced by Luce (1959).

Luce considered a range of models for the evolution of the propensities. The payoffs for an action taken might modify its propensity multiplicatively, additively, or in some combination. He computes the probabilities of actions by simply normalizing the propensities. This is Luce's *linear response rule*. The separation of the questions of propensity evolution and response rule opens the possibility of other alternatives such as the *logistic response rule* used by Busemeyer and Townsend (1993) and by Camerer and Ho (1999).

Herrnstein (1970) quantified Thorndike's "Law of Effect", as the "Matching Law": the probability of choosing an action is proportional to the accumulated rewards. Let propensities evolve by adding the payoff of the action chosen to its propensity. If we follow the Luce's linear response rule, we obtain Herrnstein's matching law. Herrnstein reports data from laboratory experiments with humans as well as with animals, which support the broad applicability of the model.

There is a special case whose limiting behavior is well known. If each action is equally reinforced, the process is mathematically equivalent to Pólya's urn process, Eggenberger and Polya (1923), with each action represented by a different color of ball initially in the urn. The process converges to a random limit, whose support is the whole probability simplex. In other words, any limiting state of propensities or probabilities is possible.

In a pioneering study, Suppes and Atkinson (1960) used a different (Markovian) model of reinforcement learning to model learning behavior in games. A number of players choose between alternatives as before, but the payoffs to each player now depend on the acts chosen by all players. Players modify their choice probabilities by the learning dynamics.

In 1995, Roth and Erev (1995). proposed a multi-agent reinforcement model based on Herrnstein's linear reinforcement and response. Here and in subsequent publications, Erev and Roth (1998); Bereby-Meyer and Erev (1998), they show a good fit with a wide range of empirical data. Limiting behavior in the basic model has recently been studied by Beggs (2002).

In Skyrms and Pemantle (2000), both basic and discounted versions of Roth-Erev learning are applied to social network formation. Individuals begin with prior propensities to interact with each other, and interactions are modeled as two-person games. Individuals have given strategies, and interactions between individuals evolve by reinforcement learning. The analysis begins with a series of results on "Making Friends", a network formation model in the special case where the game interaction is trivial. Nontrivial strategic interaction is then introduced, and it is shown that the co-evolution of network and strategy depends on relative rates of evolution as well as on other features of the model.

The present work is a natural outgrowth of the investigations begun in Skyrms and Pemantle (2000). In the richer context of multi-agent interactions, more phenomena arise, namely clique formation and a meta-stable state of high network connectivity for an initial epoch whose length depends dramatically on the discounting parameter. In Section 5.3 we discuss the implications of these features for a wide class of models.

# 3 Mathematical background

Our ultimate goal is to understand qualitative phenomena such as clique formation, or tendency of the interaction frequencies toward some limiting values. The mathematical literature on reinforcement processes contains results in these directions. It will be instructive to review these, and to examine the mathematical classification of such processes, although we will need to go beyond this level of analysis to explain the behavior of network models such as Three's Company on timescales we can observe.

Reinforcement processes fall into two main types, trapping and non-trapping. A process is said

to be trapping if there are proper subsets of actions for each player such that there is a positive probability that all players always play from this subset of actions. For example, if the repetition of any single vector $(\mathbf{i})$ of actions (action $i_j$ for player $j$) is sufficiently self-reinforcing that it might cause action $\mathbf{i}$ to be perpetuated forever, then the process is trapping. The specific dynamics investigated by Bush and Mosteller in 1955 are trapping, as are most logistic response models. By contrast, models that give all times in the past an equal effect on the present, such as Herrnstein's dynamics and Roth-Erev dynamics, tend not to be trapping.

One of several modifications suggested by Roth and Erev to maximize agreement of their model with the data is to introduce a discounting parameter $x \in (0, 1)$. The past is discounted via multiplication by a factor of $(1 - x)$ at each step. It is known from the theory of urn processes that discounting may cause trapping. For example, it follows from a theorem of H. Rubin, reported in Davis (1990), that if Pólya's urn is altered by discounting the past, there will be a point in time beyond which only one color is ever chosen. This holds as well with Roth-Erev type models: the discounted Roth-Erev model is trapping, while the undiscounted model is not. In Skyrms and Pemantle (2000), discounted and nondiscounted versions of several games are studied, and equilibria examined for stability. Again, discounting causes trapping, and we investigate the robustness of the trapping when the discounting parameter is close to negligible. In a related paper, Bonacich and Liggett (2003) investigate Bush-Mosteller dynamics in a two-person interaction representing gift giving. Their model has discounting, and they find a set of trapping states.

It is in general an outstanding problem in the theoretical study of reinforcement models to show that trapping must occur with probability 1 if it occurs with positive probability. This was only recently proved, for instance, for the reinforced random walk on a graph with three vertices, via a complicated argument in Limic (2001). Much of the effort that has gone into the mathematical study of these models has been directed at these difficult limiting questions. In the non-trapping case, even though the choice of action does not fixate, the probabilities for some of the actions may tend to zero. A series of papers in the 1990's by Benaim and Hirsch (1995); Benaim (1998, 1999) establishes some basic tests for whether in undiscounted Roth-Erev type models, probabilties will tend toward determinstic vectors.

From the point of view of applications, in a situation where it can be proven or surmised that trapping occurs, we are mainly interested in characterizing the states in which we may become trapped and in determining how long will it be before the process becomes trapped. Recalling our initial discussion of modeling goals, we are particularly interested in results that are robust as parameters and modeling details vary, or when they are not robust, of understanding how these details of the model affect observed qualitative behavior.

The present work is a natural outgrowth of the investigations begun in Skyrms and Pemantle (2000). In the richer context of multi-agent interactions, more phenomena arise, namely clique formation and a meta-stable state of high network connectivity for an initial epoch whose length depends dramatically on the discounting parameter. In Section 5.3 we discuss the implications of these features for a wide class of models.

# 4 Three's Company: a ternary interaction model

## 4.1 Specification of the model

The game "Three's Company" models collaboration of trios of agents from a fixed population. At each time step, each agent selects two others with whom to form a temporary collusion. The two selected always agree to participate. And at each time step of the process -you can think of it as a day - there is ample opportunity for each individual to initiate a trio. Thus, an agent may be involved in multiple collusions during a single time step: one that she initiates, and zero or more initiated by the other agents. Analogously to the basic game "Making Friends", introduced in Skyrms and Pemantle (2000), Three's Company has a constant reward structure: every collaboration results in an identical positive outcome, so every agent in every temporary collusion increases by an identical amount her propensity to choose each of the other two agents in the trio. The choice probabilities follow what could be called *mulitlinear response*. The probability of an agent choosing to form a trio with two other agents $i$ and $j$ is taken to be proportional to the product of her propensity for $i$ with her propensity for $j$. In addition to providing a model for self-organization based on a simple matching law type of response mechanism, this model is meant to provide a basis for the analysis of games such as the three person stag hunting game discussed in the next section. We now give a more formal mathematical definition of Three's Company.

Fix a positive integer $N \geq 4$, representing the size of the population. For $t \geq 0$ and $1 \leq i, j \leq N$, define random variables $W(i,j,t)$ and $U(i,t)$ inductively on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows. The $W$ variables are positive numbers, and the $U$ variables are subsets of the population of cardinality 3. One may think of the $U$ variables as random triangles in the complete graph with a vertex representing each agent. The variable $U(i,t)$ is equal to the trio formed by agent $i$ at time $t$. The $W$ variables represent propensities: $W(i,j,t)$ will be the propensity for player $i$ to choose player $j$ on the time step $t$. The initialization is $W(i,j,0) = 1$ for all $i \neq j$, while $W(i,i,0) = 0$). We write $W(e,t)$ for $W(i,j,t)$ when $e$ is the edge (unordered set) $\{i,j\}$ (note that the evolution rules below imply that $W(i,j,t) = W(j,i,t)$ for all $i, j$ and $t$). The inductive step, for $t \geq 0$, defines probabilities (formally, conditional probabilities given the past) for the variables $U(i,t)$ in terms of the variables $W(r,s,t)$, $r,s \leq N$, and then defines $W(i,j,t+1)$ in terms of $W(i,j,t)$ and the variables $U(r,t)$, $r \leq N$. The equations are:

$$\mathbb{P}(U(i,t) = S \,|\, \mathcal{F}_t) \;=\; \frac{\mathbf{1}_{i \in S} \prod_{r,s \in S, r<s} W(r,s,t)}{\sum_{S' : i \in S'} \prod_{r,s \in S', r<s} W(r,s,t)}\,; \tag{4.1}$$

$$W(i,j,t+1) \;=\; (1-x)W(i,j,t) + \sum_{r=1}^{N} \mathbf{1}_{i,j \in U(r,t)}\,. \tag{4.2}$$

Here $(1-x)$ is the factor per unit time by which the past is discounted, and the $\sigma$-field conditioned on is the process up to time $t$,

$$\mathcal{F}_t := \sigma\left\{W(i,j,u) : u \leq t\right\}\,.$$

The following alternative statement of the evolution equation (4.1) is useful for those familiar with the analytic machinery, cf. Pemantle (1992), that is typically used to reduce such a process to

a stochastic approximation. Think of the normalized matrix

$$\mathbf{W}_t := \frac{1}{\sum_{i,j} W(i,j,t)} W(\cdot,\cdot,t)$$

as the state vector. This is then an asymptotically time-homogeneous Markov chain, with an evolution rule

$$\mathbb{E}\left(\mathbf{W}_{t+1} - \mathbf{W}_t \,|\, \mathcal{F}_t\right) = g(t)\left[\mu(\mathbf{W}_t) + \xi_t\right] \tag{4.3}$$

where $g(t) = x + O(1/t)$, the drift vector field $\mu$ maps the simplex of normalized matrices into its tangent space and may be explicitly computed, and $\xi_t$ are martingale increments of order 1. In the non-discounted case, $g(t) = O(1/t)$, and much information about the long term behavior of this process can be discovered by an analysis of the the flow $d\mathbf{X}/dt = \mu(\mathbf{X})$ Benaim (1999). In the discounted case, $g(t)$ does not go to zero and an alternate analysis is required.

## 4.2 Analysis of the model

Equations (4.1) and (4.2) completely specify the model for the given parameters $N$ and $x$. Simulations for a population of size 6 ($N = 6$) showed the following behavior.

When $x = .5$ (a rather steep discount rate, though not unheard of in psychological laboratory experiments Busemeyer and Stout (2002)), all 1,000 trials broke up into two cliques of size 3, with no interactions across clique boundaries. In larger populations, with the same discount rate, again decomposition into cliques occurs, this time of sizes 3, 4 and 5, whose members interact exclusively with other members of the same clique.

When $N = 6$ and $x = .4$ we found that 994 out of the 1,000 trials had decomposed into two cliques of three (we allowed the process to continue for 1,000,000 time steps). When $x$ was decreased to .3, only 13 of the 1,000 trials showed decomposition into cliques, while in the remainder of the trials all six members of the population remained well connected through the 1,000,000 time steps. Finally, when $x = .2$, a reasonable discount rate for individuals though still steeper than in most economic models, none out of 1,000 trials had broken into cliques. All six members of the population remained well connected after 1,000,000 time steps.

To put these results in perspective, we note that a number of laboratory experiments have been used to estimate the discount rate that best fits the aggregate data for Roth-Erev learning in games played in the laboratory. The best estimates of x in the data discussed in Bereby-Meyer and Erev (1998) put x less than .01. There seems, however, to be a great deal of individual variability in discount rates with some individuals having a value of x near .5, and for humans in other settings or for other biological systems the discount rate could be quite different.

To summarize the simulation data, high discount rates lead to trapping, with each agent restricting her choices to members of a clique of size 3 (or, in larger populations, size 4 or 5). Less steep discount rates lead to less trapping or no trapping at all. Interestingly, the simulation data is contradicted by the following theorem, proved in the appendix.

**Theorem 4.1** *In Three's Company, with any population size $N \geq 6$ and any discount rate $x \in (0,1)$, with probability 1 the population may be partitioned into subsets of sizes 3, 4 and 5, such that each*

*member of each subset chooses each other with positive limiting frequency, and chooses members outside the subset only finitely often. Every partition into sets of sizes 3, 4 and 5 has positive probability of occurring.*

In other words, despite the simulation data, trapping always occurs. The set of traps is the set of all ways of decomposing into cliques of sizes 3, 4 and 5. The apparent contradiction between the simulation and the theorem is resolved by Theorem 4.2, whose proof is given in the companion paper Pemantle and Skyrms (2004). The theorem states that the time for the population to break into cliques increases exponentially in $1/x$ as the discount rate $1-x$ increases to 1.

**Theorem 4.2** *For each $N \geq 6$ there is a $d > 0$ and numbers $c_N > 0$ such that in Three's Company with $N$ players and discount rate $1-x$, the probability is at least $d$ that each player will play with each other player beyond time* $\exp(c_N x^{-1})$.

# 5 The three player stag hunt

## 5.1 Specification of the model

We now replace the uniformly positive reward structure by a nontrivial game, which is a three player version of Rousseau's Stag Hunt. For the purposes of our model, agents are divided into two types, hare hunters and stag hunters. That is, we model strategic choice as unchanging, at least on the time scale where network evolution is taking place. No matter which other two agents a hare hunter goes hunting with, the hare hunter comes back with a hare (hares can be caught by individuals). A stag hunter, on the other hand, comes home empty-handed unless in a trio of three stag hunters, in which case each comes home with one third share of a stag. One third of a stag is better than a whole hare, but evidently riskier because it will not materialize if any member of the hunting party decides to play it safe and focus attention on bagging a hare. In the three player stag hunting game, as in Three's Company, at each time step each agent chooses two others with whom to form a collusion. The payoffs are as follows. Whenever a hare hunter is a member of a trio, his reward is 3. A stag hunter's reward is 4 if in a trio of three stag hunters and 0 otherwise. A formal model is as follows.

Let $N = 2n$ be an even integer representing the size of the population and let $x \in (0,1)$ be the discount parameter. The variables $\{W(i,j,t), U(i,t) : 1 \leq i,j \leq N; t \geq 0\}$ are defined again on $(\Omega, \mathcal{F}, \mathbb{P})$ with the $W$ variables taking positive values and representing propensities and the $U$ variables taking values in the subsets of $\{1,\ldots,N\}$ of cardinality 3 and representing choices of trios. We initialize the $W$ variables by $W(i,j,0) = 1 - \delta_{ij}$, just as before, and we invoke a linear response mechanism (4.1) just as before. Now, instead of the trivial reward structure (4.2), the propensities evolve according to the hunting bounties

$$\begin{aligned} W(i,j,t+1) &= (1-x)W(i,j,t) + 3\mathbf{1}_{i\leq n} \sum_{r=1}^{N} \mathbf{1}_{i,j\in U(r,t)} \\ &\quad + 2\mathbf{1}_{i>n} \sum_{q,r,s=n+1}^{N} \mathbf{1}_{i\in U(q,t)=\{q,r,s\}}\,. \end{aligned} \tag{5.4}$$

The factor in front of the last sum is 2 because the sum counts the trio $\{q, r, s\}$, chosen by agent $q$, exactly twice, as $(q, r, s)$ and as $(q, s, r)$.

## 5.2 Analysis of the model

The propensities for stag hunters to choose hare hunters remain at their initial values, whence stag hunters choose other stag hunters with limiting probability 1. The stag hunters are never affected by the hare hunters' choices, so the stag hunters mimic Three's Company among themselves precisely except for the times, numbering only $O(\log t)$ uniformly in $x$ by time $t$, when they choose hare hunters. We know therefore, that eventually they fall into cliques of size 3, 4 and 5, but that this will take a long time if the discount parameter is small.

Hare hunters may form cliques of size 3, 4 and 5 as well, but because they are rewarded for choosing stag hunters, they may also attach to stag hunters. The chosen stag hunters have cliques of their own and ignore the hare hunters, except during the times that they are purposelessly called to hunt with them. These attachments can be one hare hunter continually calling on a particular pair of stag hunters or two hare hunters continually calling on a single stag hunter. In either case the one or two hare hunters are isolated from all hunters other than their chosen stag hunters.

What matters here is not the details of the trapping state but the time scale on which the trap forms and the likelihood of a hare hunter ending up in a sub-optimal trap[4]. This likelihood decreases as the discount rate becomes small for the following reason. Hare hunters choosing to hunt with stag hunters are getting no reciprocal invitations, whereas any time they choose to hunt with other hare hunters, their mutual success creates a likelihood of future reciprocal invitations. These reciprocal invitations are then successful and increase the original hunter's propensity for choosing the other hare hunter. Thus, on average, propensity for a hare hunter to form a hunting party with other hare hunters will increase faster than propensity to call on stag hunters, and the relative weights will drift toward the hare-hare groupings. The smaller the discount parameter, $x$, the more chance this has to occur before a chance run of similar choices locks an agent into a particular clique.

Simulations show that stag hunters find each other rapidly. With 6 stag hunters and six hare hunters and a discount rate of .5, the probability that a stag hunter will visit a hare hunter usually drops below half a percent in 25 iterations. For 50 iterations of the process this always happened in 1000 trials, and this remains true for values of x between .5 and .1. For x=.01, 100 iterations of the process suffices for stag hunters to meet stag hunters at this level and for 200 iterations are enough when x=.001. Hare hunters find each other more slowly, except when they are frozen into interactions with stag hunters. When the past is heavily discounted the latter possibility a serious one. At x=.5, at least one hare hunter interacted with a stag hunter (after 10,000 iterations) in 384 of 1,000 trials. This number dropped to 217 for x=.4, 74 for x=.3, 6 for x=.2, and 0 for x=.1. Reliable clique formation among stag hunters is much slower in line with results of the last section, taking about 100,000 iterations for x=.5 and 1,000,000 iterations for x=.6.

[4]In this model, since any number of collusions is permitted for an agent on each time step, the sub-optimality is manifested not through wasted time on the stag hunter's part. Instead, it is a societal opportunity cost, borne by all the hare hunters passed over in favor of the stag hunter.

### 5.3 General Principles

The two models discussed in this paper are highly idealized. But from these, we can learn some general principles as to how to analyze a much wider class of models.

When $x$ is near zero, the process should for a long time behave similarly to the non-discounted process ($x = 0$). To understand non-discounted process, following Benaim (1999), one must find equlibria for the flow $d\mathbf{X}/dt = \mu(\mathbf{X})$, and classify these as to stability. Unstable equilibria, in general, do not matter. The non-discounted process cannot converge to an unstable equilibrium of the flow. (However, for cases in which the effects of unstable equilibria may last quite a while see Pemantle and Skyrms (2003).)

Stable equilibria of the flow may or may not be possible trapping states of the discounted stochastic process. Perhaps the most interesting case is that illustrated by Three's Company, where a stable equilibrium for the flow is not a possible trapping state for the discounted process. In Three's Company with six individuals, the state in which each individual chooses each possible pair of companions with equal probability is a stable equilibrium of the flow. As we have seen, it is not a trapping state of the discounted process. The discounted process may get pseudo-trapped there, that is, may remain there for a very long time.

This is a general phenomenon. When there are stable states of the flow that are non-trapping for the discounted process such psuedo-trapping may occur. Regarding this phenomenon, Theorem 4.2 extends rather robustly to a broader class of linearly stable states (for the flow corresponding to the non-discounted process) that are non-trapping for the discounted process, Pemantle and Skyrms (2004)

## 6 Conclusion

Our analysis reinforces the emphasis of Suppes and Atkinson, and of Roth and Erev, on the medium run for empirical applications. Long run limiting behavior may simply never be seen. It is useful to quantify the time scale on which we can expect medium run behavior to persist, and Theorem 4.2 is meant to serve as a prototypical result in this direction. Indeed, Theorem 4.2 is proved via a stronger result Pemantle and Skyrms (2004, Theorem 4.1), which applies to many trapping models as the discount rate becomes negligible. As to the nature of the medium run behavior, analyses tend to be model-dependent.

## 7 Appendix: proof of Theorem 4.1

Let $G(t)$ be the graph whose edges are all $e$ such that $e \subseteq U(i,t)$ for some $i$, that is, the set of edges whose weights are increased from time $t$ to time $t+1$. The following two easy lemmas capture some helpful estimates.

**Lemma 7.1** *(i)*

$$\sum_e W(e,t) \to 3Nx^{-1}$$

*exponentially fast as* $t \to \infty$.

*(ii) If* $e \in G(t)$ *then*

$$W(e,t+k) \geq (1-x)^{k-1}\,.$$

PROOF: The first part is a consequence of the equation for the evolution of the total weight:

$$\sum_e W(i,t+1) = (1-x)\sum_e W(e,t) + 3N\,.$$

The second part follows from the first, and from the fact that when $e \in G(t)$ then $W(e,t+1) \geq 1$ and hence $W(e,t+k) \geq (1-x)^{k-1}$.

Let $\overline{G}$ denote the transitive (irreflexive) closure of a graph $G$; thus $\overline{G}$ is the smallest disjoint union of complete graphs that contains $G$.

**Lemma 7.2** *There are constant* $c$*, depending on* $N$ *and* $x$*, such that with probability at least* $c$*, every edge* $e \in \overline{G(t)}$ *satisfies*

$$W(e,t+N^2) \geq (1-x)^{N^2}\,. \tag{7.1}$$

PROOF: Let $H$ be a connected component of $G(t)$. Fix a vertex $v \in H$ and let $w$ be any other vertex of $H$. There is a path from $v$ to $w$ of length at most $N$; denote this path $(v = v_1, v_2, \ldots, v_r = w)$. If $r = 2$ then the inequality (7.1) for $e \in G(t)$ follows from Lemma 7.1. If $r \geq 3$, we let $E(H,v,w,1)$ be the event that for every $2 \leq j \leq r-1$, the edge $\{v_{j-1}, v_{j+1}\}$ is in $G(t+1)$. Since this event contains the intersection over $r$ of the events that $U(v_j,t) = \{v_j, v_{j-1}, v_{j+1}\}$, since Lemma 7.1 bounds each of these probabilities from below, and since the events are conditionally independent given $\mathcal{F}_t$, we have a lower bound on the probability of $E(H,v,w,1)$. In general, for $1 \leq k \leq r-2$, let $E(H,v,w,k)$ be the event that for every $2 \leq j \leq r-k$, the edge $\{v_{j-1}, v_{j+k}$ is in $G(t+k)$. We claim that conditional on $E(H,v,w,l)$ for all $l < k$, the conditional probability of $E(H,v,w,k)$ given $\mathcal{F}_{t+k-1}$ can be bounded below: inductively, Lemma 7.1 bounds from below the product of $W(v_j, v_{j-1}, t)W(v_j, v_{j+k}, t)$, and hence the probability that $U(v_j,t) = \{v_j, v_{j-1}, v_{j+k}\}$; these conditionally independent probabilities may then be multiplied to prove the claim, with the bound depending only on $x$ and $N$.

From this argument, we see that the intersection $E(H,v,w) := \bigcap_{1 \leq k \leq r-2} E(H,v,w,k)$ has a probability which is bounded from below. Sequentially, we may choose a sequence of values for $w$ running through all vertices of $H$ at some distance $r(w) - 1 \geq 2$ from $v$, measured in the metric on $H$. For each such $w$, we can bound from below the probability that in $r-2$ more time steps the path from $v$ to $w$ will be transitively completed. We denote these events $E'(H,v,w)$, the prime denoting the time shift to allow events analogous to $E(H,v,w)$ to occur sequentially. Summing the time to run over all $w \in H$ yields at most $N^2$ time steps. Let $E(H,v)$ denote the intersection of all the events $E'(H,v,w)$. Inductively, we see that the probability of $E(H,v)$ is bounded from below by a positive number depending only on $N$ and $x$.

Finally, we let $(H,v)$ vary with $H$ exhausting components of $G(t)$ and $v$ a choice function on the vertices of $H$. The events $E(H,v)$ are all conditionally independent given $\mathcal{F}_t$, so the probability of

their intersection, $E$, is bounded from below by a positive constant which we call $c$. By Lemma 7.1 once more, on $E$, we know that (7.1) is satisfied for each $e \in \overline{G(t)}$.

PROOF OF THEOREM 4.1: For any subset $V$ of agents, let

$$E(V,t) := \bigcap_{s \geq t} \bigcap_{v \in V, w \in V^c} \{\{v,w\} \notin G(s)\}$$

denote the event that from time $t$ onward, $V$ is isolated from its complement. If $V$ is the vertex set of a component of $G(t)$, then the conditional probability given $\mathcal{F}_t$ of the event $E(V,t)$ may be bounded from below as follows. For any $v \in V, w \in V^c$, and for any $s \geq t$, if the edge $e := \{v,w\}$ is not in $G(r)$ for any $t \leq r < s$, then by part 1 of Lemma 7.1, its weight $W(e,s)$ is at most $(1-x)^{s-t}3Nx^{-1}$. Since $\sum_z W(v,z,s) \geq 2$ for all $v, z, s$, it follows from the evolution equations that

$$\mathbb{P}(e \in G(s) \,|\, F_s) \leq \frac{(1-x)^{s-t}3Nx^{-1}}{2 + (1-x)^{s-t}3Nx^{-1}} \,.$$

It follows that

$$\mathbb{P}\left(\exists v \in V, w \in V^c \,:\, \{v,w\} \in G(s) \,|\, \mathcal{F}_s\right) = O\left(Nx^{-1}(1-x)^{s-t}\right)$$

uniformly in $N, x$ and $t$ as $s - t \to \infty$ (though the uniformity in $N$ and $x$ is not needed). By the conditional Borel-Cantelli Lemma, it follows that

$$\mathbb{P}(E(V,t) \,|\, F_t) > c(N,x) \tag{7.2}$$

on the event that $V$ is the vertex set of a component of $\overline{G(t)}$.

By the reverse direction of the Conditional Borel-Cantelli Lemma, the event $E(V,t)$ occurs for some $t$ with probability 1 on that event that $V$ is a component of $\overline{G(t)}$ infinitely often. Let $e = \{v,w\}$ be any edge. If $e \notin \overline{G(t)}$ infinitely often, then since there are only finitely many subsets of vertices, it follows that $v \in V$ and $w \in W$ for some disjoint $V$ and $W$ that are infinitely often components of $G(t)$. This implies that $e \in \overline{G(t)}$ finitely often. We have shown that, almost surely, the edges come in two types: those in $\overline{G(t)}$ finitely often and those in $\overline{G(t)}$ all but finitely often. This further implies that $\overline{G(t)}$ is eventually constant. Denote this almost sure limit by $G_\infty$. It remains to characterize $G_\infty$.

It is evident that $G_\infty$ contains no component of size less than three, since $G(t)$ is the union of triangles $U(i,t)$. Suppose that $\overline{G(t)} = H$ for some $H$ of cardinality at least six. By Lemma 7.2, conditional on $\mathcal{F}_t$ and $\overline{G(t)} = H$,

$$W(e, t+N^2) \geq \frac{(1-x)^{N^2}}{3N}$$

for every $e \in H$. Write $H$ as the disjoint union of sets $J$ and $K$, each of cardinality at least three. Then with probability at least

$$\left(\frac{(1-x)^{N^2}}{3N + (1-x)^{N^2}}\right)^{|J|+|K|}$$

$U(i, t+N^2) \subseteq J$ for every $i \in J$ and $U(i, t+N^2) \subseteq K$ for every $i \in K$. In this case, $\overline{G(t+N^2)}$ has components that are proper subsets of $H$. By the martingale convergence theorem,

$$\mathbb{P}(H \text{ is a component of } G_\infty \,|\, \mathcal{F}_t)$$

converges with probability 1 to the indicator function of $H$ being a component of $G_\infty$. From the above computation, it is not possible for $\mathbb{P}(H \text{ is a component of } G_\infty \,|\, \mathcal{F}_t)$ to converge to 1 when $H$ has cardinality six or more. Therefore, every component of $G_\infty$ has cardinality 3, 4 or 5.

The rest of the proof is easy. Let $V_1, \ldots, V_k$ be any partition of $[N]$ into sets of cardinalities 3, 4 and 5. The derivation of (7.2) shows that

$$\mathbb{P}\left(\bigcap_{j=1}^{k} E(V_j, 1)\right) > 0\,,$$

in other words, with positive probability $G_\infty$ has $k$ components which are precisely the complete graphs on $V_1, \ldots, V_k$. It is elementary that a coupling may be produced between the Three's Company processes on populations of sizes $N$ and $K < N$ (with the same $x$ value), so that if $\{\tilde{W}(i,j,t), \tilde{U}(i,t)\}$ are the weight and choice variables for the smaller population, then $\tilde{U}(i,t) = U(i,t)$ and $\tilde{W}(i,j,t+1) = W(i,j,t+1)$ for all $t < \tau$ where $\tau$ is the first time, possibly infinite, at which $U(i,t)$ contains an edge between $[K]$ and $\{K+1, \ldots, N\}$. In general, coupling methods show that if $\mathbb{P}(G_\infty = G_0 \,|\, t) > 1 - \epsilon$ then the conditional distribution of the Three's Company process from time $t$ onward given $\mathcal{F}_t$ and $G_\infty = G_0$, shifted back $t$ time units and restricted to a component $H$ of $G_\infty$, is within $\epsilon$ in total variation of the distribution of the Three's Company process on $H$ started with initial weights $W'(i,j,0) := W(i,j,t)$.

The Three's company process on a population of size 3, 4 or 5 with any discount rate $1 - x < 1$ is ergodic: to see this just note that the Markov chain whose state space is the collection of $W$ variables is Harris recurrent as a consequence of Lemma 7.2. The invariant measure gives positive weight to each edge, so each agent chooses each other with positive frequency, finishing the proof of Theorem 4.1.

# References


Beggs, A. 2002. On the convergence of reinforcement learning. *Preprint.*

Benaim, M. 1998. Recursive algorithms, urn processes and chaining number of chain-recurrent sets. *Ergodic Th. Dynam. Systems* **18**, 53–87.

Benaim, M. 1999. Dynamics of Stochastic approximation algorithms. *Seminaires de Probabilités XXXIII, Springer Lec. Notes in Math.* **1709**, 1–68.

Benaim, M. and Hirsch, M. 1995. Dynamics of Morse-Smale urn processes. *Ergodic. Th. Dynam. Systems* **15**, 1005–1030.

Bereby-Meyer, Y. and Erev, I. 1998. On learning to become a successful loser: a comparison of alternative abstractions of learning processes in the loss domain. *Journal of Mathematical Psychology* **42**, 266–286.

Bonacich, P. and Liggett, T. 2003. Asymptotics of a matrix-valued Markov chain arising in sociology. *Stochastic Processes and their Applications* **104**, 155-171.

Borgers, T. and Sarin, R. 1997. Learning through reinforcement and the replicator dynamics. *Journal of Economic Theory* **74**, 235–265.

Busemeyer, J. and Stout, J. 2002. A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara Gambling Task. *Psychological Assessment* **14**, 253–262.

Busemeyer, J. and Townsend, J. 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* **100**, 432–459.

Bush, R. and Mosteller, F. 1955. *Stochastic models for learning.* John Wiley & Sons: New York.

Camerer, C. and Ho, T. 1999. Experience-weighted attraction in games. *Econometrica* **64**, 827–874.

Davis, B. 1990. Reinforced random walk. *Prob. Th. Rel. Fields* **84**, 203–229.

Davis, B. 1999. Reinforced and perturbed random walks. *Random Walks (Budapest, 1998* 113–126, *Bolyai Soc. Math. Studies* **9**.

Coppersmith, D. and Diaconis, P. Unpublished manuscript.

Eggenberger, F. and Pólya, G. 1923. Uber die Statistik verketter vorgäge. *Zeit. Angew. Math. Mech.* **1**, 279–289.

Erev, I. and Roth, A. 1998. Predicting how people play games: reinforcement learning in experimental games with unique mixed-strategy equilibria. *American Economic Review* **88**, 848–881.

Estes, W. K. 1950. Toward a Statistical Theory of Learning. *Psychological Review* **57**, 94–107.

Herrnstein, R. J. 1970. On the Law of Effect. *Journal of the Experimental Analysis of Behavior* **13**, 243–266.

Iosifescu, M. and Theodorescu, R. 1969. *Random processes and learning.* Springer-Verlag: New York.

Limic, V. 2001. Attracting edge property for a class of reinforced random walks. *Preprint.* http://www.math.cornell.edu/ limic/

Luce, D. 1959. *Individual choice behavior.* John Wiley & Sons: New York.

Macy, M. 1990. Learning theory and the logic of critical mass. *American Sociological Review* **55**, 809–826.

Macy, M. 1991. Learning to cooperate: stochastic and tacit collusion in financial exchange. *American Journal of Sociology* **97**, 808–843.

Maynard Smith, J. 1982. *Evolution and the theory of games.* Cambridge University Press: Cambridge.

Norman, F. 1972. *Markov processes and learning models.* Academic Press: New York.

Othmer, H. and Stevens, A. 1997. Aggregation, blowup and collapse: the ABC's of taxis and reinforced random walk. *SIAM J. Appl. Math.* **57**, 1044–1081.

Pemantle, R. 1988. Phase transition in reinforced random walk and RWRE on trees. *Ann. Probab.* **16**, 1229–1241.

Pemantle, R. 1992. Vertex-reinforced random walk. *Prob. Theor. and Rel. Fields* **92**, 117–136.

Pemantle, R. and Skyrms, B. 2004. Time to absorption in discounted reinforcement models. *Stochastic Processes and their Applications.* **109**, 1–12.

Pemantle, R. and Skyrms, B. 2003. Reinforcement schemes may take a long time to exhibit limiting behavior. *In preparation.*

Roth, A. and Erev, I. 1995. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* **8**, 164–212.

Skyrms, B. and Pemantle, R. 2000. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the USA* **97**, 9340–9346.

Suppes, P. and Atkinson, R. 1960. *Markov learning models for multiperson interactions.* Stanford University Press: Palo Alto.