Evolution of Prosocial Behavior through Preferential Detachment and Its Implications for Morality

by

Aaron L Bramson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Political Science and Philosophy)
in The University of Michigan
2012

Doctoral Committee:

Professor Scott E. Page, Co-Chair
Professor Peter A. Railton, Co-Chair
Professor Allan F. Gibbard
Professor Ken W. Kollman
Associate Research Scientist Rick L. Riolo

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES vii

CHAPTER

I. Theory of Preferential Detachment 1

1.1 General Theory 3
1.1.1 Assumptions 4
1.2 Brief Examples and Applications 8
1.3 Population Size 10
1.4 Direct Social Connections 12
1.5 Strategic Contexts 13
1.6 Determining Agent Behavior 14
1.6.1 Neighbor Preferences 15
1.6.2 Realized Preferences 18
1.6.3 Game Categorization 19
1.6.4 Configurations 21
1.6.5 Revealed Preferences 22
1.7 Learning and Population Dynamics 25
1.7.1 Success-Based Imitation 26
1.7.2 Fitness-Based Birth and Death 27
1.8 The Game Library 28
1.8.1 Prisoners' Dilemma 29
1.8.2 Hawk and Dove 32
1.8.3 Battle of the Sexes 34
1.8.4 Coordination Game 35
1.8.5 Lichen 36
1.8.6 Stag Hunt 38
1.8.7 Commensal 39
1.8.8 Matching Pennies 41
1.8.9 Lane Choice 42
1.8.10 Biased Lane Choice 44

1.9 Explaining Prosocial Outcomes 45
1.9.1 Assortative Mixing 46
1.9.2 Games of Life 46
1.9.3 Why Preferential Detachment Works 47
1.10 Summary of Preferential Detachment 47

II. Markov Model of Preferential Detachment 50

2.1 Random Connections as Bernoulli Trials 51
2.1.1 Example of Random Connection Probabilities . . . 55
2.2 Probabilities of Detachment by Other Agents 57
2.2.1 Example Detachment Probabilities 59
2.3 Agent Action Probabilities by Strategic Context 61
2.3.1 Example Action Transition Probabilities 62
2.4 Combined Probabilities 64
2.4.1 Example Combined Transition Matrices 64
2.5 Stationary Distributions 67
2.5.1 Example Calculation for Cooperative Games 68
2.5.2 Stationary Distributions for Each Game Form . . . 71
2.6 Generalizing the Results 74
2.6.1 Generalizing over the Maximum Degree 74
2.6.2 Generalizing over the Population Size 78
2.6.3 Effects on the Strategic Contexts 81
2.7 Summary of the Markov Model 86
2.8 Next Steps 87

III. Agent-Based Model of Preferential Detachment 89

3.1 Model Procedure Overview 90
3.2 Parameters 91
3.3 Model Setup and Time 0 92
3.4 Model Iterations 92
3.4.1 Detachment Step 93
3.4.2 Learning Step 93
3.4.3 Random Connection Step 94
3.4.4 Population Dynamics Step 94
3.4.5 Visualization Update Step 94
3.4.6 Measures 95
3.4.7 Halting Conditions 96
3.5 Experiments 98
3.5.1 Branch One: Sweeping Population Size 98
3.5.2 Branch Two: Sweeping Maximum Degree 99
3.5.3 Branch Three: Sweeping Strategic Contexts 100
3.6 Analysis Techniques 102
3.6.1 Measures of System Properties 102

3.6.2 Statistical Treatment 103
3.6.3 Visualization 106
3.7 Branch One Results: Sweeping Population 108
3.7.1 Prosociality of Preferential Detachment Alone . . . 108
3.7.2 Prosociality with Imitative Learning 112
3.7.3 Prosociality with Population Dynamics 118
3.7.4 Prosociality with Learning and Population Dynamics 123
3.7.5 Stable Heterogeneous Groups 127
3.7.6 Clustering and Group Formation 129
3.7.7 Summary of Population Sweep Results 131
3.8 Branch Two Results: Sweeping Maximum Degree 133
3.8.1 Prosociality of Preferential Detachment Alone . . . 134
3.8.2 Prosociality with Imitative Learning 137
3.8.3 Prosociality with Population Dynamics 141
3.8.4 Prosociality with Learning and Population Dynamics 145
3.8.5 Summary of Maximum Degree Sweep 149
3.9 Branch Three Results: Sweeping Strategic Contexts 151
3.9.1 Prisoners' Dilemma and Hawk and Dove 152
3.9.2 Battle of the Sexes and Coordination Game 156
3.9.3 Lichen 163
3.9.4 Stag Hunt 171
3.9.5 Commensal 175
3.9.6 Matching Pennies 182
3.9.7 Lane Choice 188
3.9.8 Biased Lane Choice 196
3.9.9 Summary of Game Sweep Results 204
3.10 ABM Conclusions 210
3.10.1 Sensitivity to Parameters 210
3.10.2 Adding Learning 213
3.10.3 Adding Population Dynamics 215
3.10.4 Adding Both Learning and Population Dynamics . . 217
3.10.5 Efficiency and Intervention 220
3.10.6 Network Properties 221
3.10.7 Mechanism Comparison and Complementarity . . . 222
3.10.8 Evolutionary Approach 225
3.10.9 Future Work for Preferential Detachment 227
3.10.10 Conclusions Summary 229

IV. Evolution of Prosociality and the Moral Experience 231

4.1 Evolution of Prosocial Behavior 232
4.1.1 The Primacy of Behavior 233
4.1.2 Prosociality in Collectives 238
4.1.3 Intergroup Competition 241

4.1.4 Implications of Evolved Prosociality 248
4.2 Evolution of the Moral Experience 252
4.2.1 Whence the Moral Attitudes 253
4.2.2 Universality of Social Problems 260
4.2.3 Continuity in Nature 264
4.3 Summary and Concluding Remarks 276
4.3.1 Implications 277

BIBLIOGRAPHY 278

LIST OF FIGURES

Figure

1.1 Prisoners' Dilemma 29
1.2 Hawk and Dove 33
1.3 Battle of the Sexes 34
1.4 Coordination Game 35
1.5 Lichen 37
1.6 Stag Hunt 38
1.7 Commensal 40
1.8 Matching Pennies 41
1.9 Lane Choice 43
1.10 Biased Lane Choice 44
2.1 Configuration Dynamics in Cooperative Games 70
2.2 Probabilities of random connections for $N_A = N_B = 0$ to 2000. . . . 79
2.3 Probabilities of random connections for N_A/N_B. 80
3.1 Empirical CDFs of a measure at one time slice. 105
3.2 Behavior Plot Reading Guide 106
3.3 Summary Plot Reading Guide 107
3.4 Stable Heterogeneous Groups in the Prisoners' Dilemma 129
3.5 Stable Heterogeneity in Stag Hunt 173

LIST OF TABLES

Table

1.1 2×2 Payoff Matrix 15
1.2 Possible neighbor preferences for 2×2 games 16
1.3 Distinct Game Forms and Categories 16
1.4 Game Library 17
1.5 Realized Pairwise Payoffs: What Row-Type Gets from Column-Type 19
1.6 Categorized Games 20
1.7 Prosocial Outcomes by Category 20
1.8 The Three Possible Revealed Preference Relations for K=4 23
1.9 Agents' Revealed Preferences in the Lane Choice Game 43
1.10 Summary of Game Types and Prosocial Outcomes 49
2.1 The Three Possible Revealed Preference Relations for K=4 50
2.2 Structure of Bernoulli Probabilities of Random Connections for $K = 4$ 53
2.3 $\mathcal{R}_A$: A-Type Random Connection Transition Probabilities for $N_A = N_B = 100$, $K = 4$ 56
2.4 $\mathcal{R}_B$: B-Type Random Connection Transition Probabilities for $N_A = N_B = 100$, $K = 4$ 56
2.5 Categories of Configurations 58
2.6 Categories Detachment Matrices 59
2.7 $\mathcal{Y}\alpha$ Detachment Transition Probabilities for $K = 4$ 60
2.8 $\mathcal{Y}\beta$ Detachment Transition Probabilities for $K = 4$ 60
2.9 $\mathcal{Y}\delta$ Detachment Transition Probabilities for $K = 4$ 60
2.10 $\mathcal{X}_{\phi_1}$: Action Transition Probabilities if $B \prec A$ 63
2.11 $\mathcal{X}_{\phi_2}$: Action Transition Probabilities if $A \prec B$ 63
2.12 $\mathcal{X}_{\phi_3}$: Action Transition Probabilities if $B \approx A$ 63
2.13 Combined Transition Matrices for Each Category 64
2.14 $\mathcal{A}$: Combined Transition Probabilities for A-Types in Cooperative Games, $N_A = N_B = 100$, $K = 4$ 66
2.15 $\mathcal{B}$: Combined Transition Probabilities for B-Types in Cooperative Games, $N_A = N_B = 100$, $K = 4$ 66
2.16 Stationary Distributions in Cooperative Games 69
2.17 Stationary Distributions for All Game Categories 73
2.18 Combined Transition Matrices for Each Category 82

3.1 Summary of agent-based model parameters. 91
3.2 Population size sensitivity experiments. 99
3.3 Maximum degree sensitivity experiments. 99
3.4 Game Library 101
3.5 Agent type color encoding. 106
3.6 Results from Prisoners' Dilemma Across Population Size 133
3.7 Results from Prisoners' Dilemma Across Maximum Degree 151
3.8 Results Across Strategic Contexts 209
3.9 Assortativity levels for the Markov and agent-based models. 212

CHAPTER I

Theory of Preferential Detachment

The evolution of prosocial behavior is both a wide and varied field. It is wide insofar as it covers many different behaviors and cognitive attitudes including altruism, reciprocity, promise-keeping, cooperation, coordination, morality, parental care, fairness, trust, justice, and others. It is varied in that it is approached from within many different disciplines each using distinct assumptions, methodology, and interpretive scheme. The topic draws upon research in philosophy, economics, sociology, political science, mathematics, biology, primatology, computer science, robotics, anthropology, ethology, archeology, and law (and certainly others). Furthermore, the science involved requires more than just a combination of insights from these fields, it requires novel and unique assumptions, methods, and interpretations that do not fall squarely within any of those domains. There is already a sizable body of research on prosociality and its evolution; and it rests on a foundation of research from the supporting fields. Yet, despite these efforts, there is not yet a consensus on what mechanisms are responsible for producing or maintaining prosocial behaviors, nor an obvious convergence of ideas. In this sense the field is still young. The current project adds to this research by introducing a general theory with supporting models that offers a plausible explanation and viable mechanism for generating and perpetuating prosocial behavior.

The proposed mechanism is *preferential detachment*[1]: The behavior of agents classifies them into types, and they interact with a limited number of other agents

[1] As will be discussed in more detail later, the theory and models depend purely on individual behavior, so the use of "preferential" in the name may be misleading. The individuals need not be the kinds of objects that we typically attribute preferences to, though actual or revealed preferences do satisfy the requirements for behavioral contingency. Thus "selective detachment", "differential detachment", or even "interest-based detachment" may be more accurate names, but "preferential detachment" has other benefits that compensate for requiring the no-preferences-actually-required caveat. Among these benefits is that we can then use "prefer" to indicate the behavioral tendencies of individuals regardless of the mechanism driving them.

every period. The benefit that agents receive from each neighbor depends only on the type of the focal agent and its neighbors. Agents will remove (stop interacting with) a neighboring agent of a less preferred type if and only if it also has a preferred neighbor; i.e., agents have no global information about the available types and hence base their behavior only on current actual neighbors. If agents have fewer than the maximum capacity of interaction partners and haven't disconnected this period, then agents randomly attach to another agent. These behaviors are enacted simultaneously by the agents. Finally, extensions of the base model include learning and/or population dynamics so that agents may imitate a more successful neighbor and/or be selected for reproduction or elimination respectively.

As an example of such a situation consider local utility companies sharing access to underground water supplies [Ostrom 1990]. Each company must apply for a pumping license from the regional authority and shares infrastructure with neighboring utilities. The system is only efficient when companies openly share infrastructure access with their neighbors, but they can make greater profits from their customers by prioritizing throughput to them. The situation is similar to a tragedy of the commons, but the effects of each company's behavior are localized to its neighbors (the ones it has contracts with). Faced with such a situation a compliant company can decide to shut out any non-compliant partners by preventing water flow to it. To make up for the lost volume the same compliant company can seek out other utility companies open to new water-share partnerships. Companies that get shut out of too many compliant partnerships will not be able to service their customers sufficiently and will thus perform relatively poorly: possibly losing their pumping license, going bankrupt, changing management, etc.

The theory proposed is that agents utilizing preferential detachment will sort themselves into social arrangements such that the agents who contribute a benefit to the members of their group also do better for themselves in the long run. Agents can do this with minimal information about their environment, the other agents, the future, and with minimal cognitive/computational ability. Because the mechanism is simple, the theory yields powerful results that can be applied generally. The conclusion is that self-organizing into groups that maintain prosocial behaviors may be simpler and more robust than previously thought. *The primary contribution of this research is that a single, simple mechanism operating in different contexts generates the conceptually distinct prosocial behaviors achieved by other models, and in a manner that is more amenable to evolutionary explanations.*

The results from this research may help explain many instances of behavioral patterns promoting social benefit without making many assumptions about the characteristics of the agents involved or restrictions on the situation in which it will happen. Because the mechanisms and theory apply broadly across the animal kingdom we can use research from one species' prosocial and moral tendencies as models for other species – including our evolutionary predecessors. All of this can improve our understanding of the evolution of human prosocial behavior and moral experiences as well as spur new questions and research paths in many of the realted fields previously mentioned.

This work will proceed as follows. The first step is to describe in detail the theory of preferential detachment: what it assumes, what the salient features are, and what sorts of explanations/predictions it produces. In chapters II and III I present formal models that each aims to capture the preferential detachment theory's features. If these models succeed in generating results that can be reasonably interpreted as prosocial behavior, then we will be in a good position for the fourth and final chapter. The concluding chapter addresses using formal models of prosociality as a reasonable explanation for the evolution of context-appropriate prosocial behaviors, and through them an explanation of the origins of moral experiences.

1.1 General Theory

The name "preferential detachment" describes a mechanism, but that mechanism is grounded in a larger idea of what decisions and processes shape interaction structures and properties of societies. The proposed mechanism and supporting assumptions are general and applicable to (though not equally insightful for) a great diversity of interacting objects and social properties. This section describes the foundational assumptions of the theory and the details of the mechanism. Though this statement of the theory and mechanism is distinct from, and more general than, the later formal models of them, the theory itself will be stated through a fair amount of formalization with modeling in mind.

Here I provide a highly abstract description of the theory and some examples of familiar phenomena that match the theory's framework to facilitate understanding. Then I go into a discussion with a greater amount of detail with respect to the application to human and protohuman societies to focus our attention on the sorts of phenomena we will be addressing in later. The high-level description is nevertheless essential to some of the evolutionary arguments made later, as well as useful for future applications of the theory to entirely different phenomena.

1.1.1 Assumptions

The first assumption is that there are individuals (called "agents") and they can be sorted exhaustively into types. Agent types capture any feature upon which agent *behavior* may be contingent. The agents can be entities that do not have agency in the sense of being conscious decision makers; it only matters that their behavior is contingent in the appropriate way. Agents are related to other agents over time; i.e., the relation is persistent or repeating. The relation may be an interaction, communication, connection, or anything that the agents can unilaterally form and break and that informs each agent of the types of its relationship partners. It will be convenient for us to represent the relationship structure as a network, but it could be geographic location, transaction patterns, discussions, contract partners, etc. I will use "interaction" and "relation" interchangeably.

Agents connect to new agents without any information about them except that they are available for connection. Thus no information regarding the set of potential partners' type or their other current interaction partners is used. Agents' behavior is contingent upon

C1 their own type,

C2 the number of their current interaction partners, and

C3 the types of their current interaction partners.

Depending on the values of C1, C2, and C3 agents will perform one of the following behaviors.

B1 End the relationship/interaction with certain partners.

B2 Connect to new partners.

B3 Do nothing.

The specifics of which types detach which other types depend on the particulars of the situation. Applying the theory consists in determining/deciding the map from C1-3 to B1-3. The assumptions for the base preferential detachment theory can be summarized in the following list:

A1 There exist agents partitioned into types.

A2 Agent can assess C1-3 every period.

A3 Agents perform an action among B1-3 depending on C1-3 every period.

A4 There are multiple periods in succession.

1.1.1.1 Types

Typing agents is a common tool for organizing a theory's subject matter based on observed or purported properties or behaviors. In formal models similar to those presented below, types have been used to identify agents' differing strategies, utility functions, information sets, endowments, opportunity sets, etc. Types in the current theory categorize agents by their contingent behavior – how those agents behave vis-à-vis other agents. Thus each agent of the same type has the same *mapping of* C1-3 *to* B1-3 and treats other agents of identical type identically. With agent types specified in this way, there are several reasons why type would be constant; e.g., cultural stickiness, habit, short term behavioral memory, or cognitive limitation to name a few relevant to human agents.

When we apply preferential detachment theory to a particular domain or problem, the strategic properties of the situation determine the mappings agents use. Returning to the water distribution example from the introduction we can identify the compliant and non-compliant utility companies as the relevant agent types. We would then need to specify four mappings: how both compliants and non-compliants behave toward compliants and toward non-compliants. From the example we can see that compliant companies would detach from non-compliant partners but not from compliant partners. They would also pursue new partners until their water needs are reliably met. Non-compliant companies would also pursue new partners until their water needs are met. To specify more of the behavior we will need to investigate the situation in closer detail.

1.1.1.2 Choosing Edges Not Actions

Preferential detachment changes the focus of agent behavior compared with much of the previous literature on the evolution of prosociality by having agents choose their interaction partners. The mechanisms of kin selection, reputation, experience, signaling, punishment, and hierarchy typically influence what *action* or *strategy* an agent performs given their partners and/or partners' previous actions. Social network models in this field (e.g., [Sobkowicz 2003, Arapaki 2009, Bergin and Bernhardt 2009, Yang et al. 2009]), as well as cellular automata models (e.g., [Nowak and May 1992, Grim et al. 2006]) and many other agent-based models, typically use a static interaction structure as well.

According to preferential detachment theory, an agent's behavior (action/strategy) is fixed according to its type. Rather than choosing actions *in the game*, agents act to alter their environment through choosing to remove certain interaction partners. The switch from fixed partners to a dynamic social network is not unprecedented; it appears in models employing homophily, preferential attachment [Tesfatsion 2001], geographic location, and a few others (see [Ashlocka et al. 1996, Skyrms and Pemantle 2000, Sobkowicz 2004, Zimmermann and Eguíluz 2005, Santos et al. 2006, Szolnoki and Perc 2009, Tomassini et al. 2010, Wu et al. 2010]). Though dynamic interaction structures are still quite uncommon in the literature, this feature is increasingly mentioned as a topic of future work.

1.1.1.3 Connecting without Information

The assumption that connections are made without information over potential partners is in contradistinction to those few models in which kin selection, reputation, experience, signaling, hierarchy, and homophily do drive partnership formation. However note that most differential equation (and some simulation) models assume either a fully connected interaction structure or randomly pull interacting pairs of agents from the agent pool (e.g., [Axelrod 1984, Binmore 1998, Skyrms 2004, Boyd and Richerson 2005]). Both of these techniques likewise utilize no information about potential partners in making connections (though in some random mixing models the agents consider the *expected* partners), but there is no explicit interaction structure. Excluding such information from the determination of agent action allows preferential detachment to incorporate the network structure while still employing a simple mechanism.

1.1.1.4 Behavior-Focused Approach

The namesake behavior for the preferential detachment theory is that agents disconnect from certain current neighbors. Because this general formulation of the theory is described in terms of contingent behavior rather than preferences, the name is a bit of a misnomer. The individuals need not be the kinds of objects that we typically attribute preferences to, so “selective detachment” or “differential detachment” may be more accurate names. I will nonetheless continue to use “preferential detachment” with the overt caveat that the preferences of which I speak are whatever mechanism produces behavior actual. The action space is over current connections and the choice is which, if any, to remove. Though the theory can apply to entities that don’t have agency (e.g., the agents could be just reacting through physical laws or stimulus-response), their behavior must be contingent in the way specified. This behaviorist approach is fully compatible with thinking that for some choice of agents, the agents really do have the underlying preferences that make them behave this way – and that the preferences *are* the proximate cause.

Such a modeling choice is common among agent-based models that are often looking (as this one is) for patterns in the structure and dynamics of agent interactions, and assumes that whatever the underlying forces are, they are reliable and consistent in producing the modeled behaviors. Peter Blau captures this idea in his theory of organizational differentiation of 1970 in which he states, “The theory centers attention on the social forces that govern the interrelations among differentiated elements in a formal structure and ignores the psychological forces that govern individual behavior. Formal structures exhibit regularities that can be studied in their own right without investigating the motives of the individuals in organizations [Hedstrom 2005].”

The motivation here goes beyond modeling convenience. The focus of this project is to understand to what degree preferential detachment can stand as an explanation for the *evolution* of prosocial behavior. Natural selection exerts pressure on behaviors – not intentions, beliefs, preferences, emotions, or other mental states. Furthermore, we typically can't observe the underlying mechanism. We observe behavior, test what it is contingent upon, and build our action theories from this evidence. So all that matters here is that some or other mechanism exists to facilitate the necessary contingent behavior. In chapter IV I will present evidence that humans, most mammals, and many other animals possess the capacities and capabilities to have their behavior contingent in the ways hypothesized by the theory of preferential detachment.

The dominant evolutionary explanations for prosocial behavior are mutualism, reciprocal altruism, and kin selection; and these can be described purely in terms of reproductive fitness [Trivers 1971, Binmore 1998, Bekoff and Pierce 2009]. This fitness-based interpretation is important because we see cooperation, coordination, task specialization, and other prosocial behaviors in creatures as simple as bacteria and amoeba. An important ingredient in the evolutionary explanation of prosociality in humans is its evolutionary roots in predecessor organisms and so the capacity to achieve prosocial outcomes must have been available to these predecessors. Explanations involving punishment, policing, reputation, signaling, expectations, experience, hierarchy, etc. require the entities to possess significant cognitive capabilities. Even kin selection and reciprocal altruism require agents to detect, recall, and recognize who is kin or who owes/deserves reciprocity. Yet we also see prosocial behaviors in organisms without these capabilities. There are evolutionary models that succeed in achieving cooperation without the need for detailed memory [Riolo et al. 2001], and we will see that preferential detachment offers another mechanism, the success of which can be described in terms of the reproductive fitness.

1.1.1.5 Other Mechanisms

Considering the above set of assumptions we can see that preferential detachment imposes much weaker requirements on the capabilities of the agents compared to previous work in the field. This feature is a benefit of explaining procociality with preferential detachment. It doesn't preclude that more sophisticated cognitive mechanisms are at work, but it does require that they produce behavior that is appropriately contingent. Since many of the other mechanisms used to generate the prosocial outcome assume or require more of agents, they couldn't be translated into preferential detachment. However, success-based imitation is added by giving agents' access to neighbors' utility, and fitness-based population dynamics are added using aggregated agent utility (see 1.7).

1.1.1.6 Groups

It is worth pointing out some things about *groups* in the current research. I define a group here as a contiguous collections of agents with some particular property (e.g., of the same type or being fully connected). Group membership is implicit in this theory rather than explicit; i.e., groups are not social-level entities to which agents belong. Individuals' behavior generates this social phenomena, but the individuals themselves are unaware that there are groups and which they are in. So while we can see patterns in groups and explain group-level phenomena by describing group-level dynamics, it is important to keep in kind that the agent-level behavior is driving all the group-level dynamics. This implicit form of groups is more valuable for examining any non-human agents because they likely lack the capability to explicitly consider themselves as part of a group. It also benefits an evolutionary approach because whatever social phenomena we identify we can unambiguously explain it with aggregated individual behavior.

1.2 Brief Examples and Applications

The theory of preferential detachment encapsulates certain ways the world might be and what is entailed if it is. Before moving on to further refinements, and in order to facilitate the readers' understanding of my formal models, I now briefly present a few examples that fit the preferential detachment framework (in addition to the introductory water management example).

The are many straightforward applications in the social sciences. For example, we can interpret the agents as people and types as adherence to the views of political parties or religions. Both are affiliations which are difficult to change in the short term and have implications for the benefit gained when interacting with agents of the same or other type. The problem could be organizing a social event, fund raising, or matching mates – in which assortativity is beneficial. Or the problem could be arranging a debate or uniting a fractured community – in which case disassortativity is beneficial. The context in which the agents operate will determine the appropriate mapping of C1-3 to B1-3; i.e., who would detach whom.

Agents can be interpreted as countries and their types as whether that country supports some particular policy (e.g., tariffs, deportation, or arms reduction). If the type represented a trade policy, then groups of like-typed agents would be something like membership in a trade organization, and the interactions would represent actual trading between partners. Distinct trade policies offer benefits differentially to countries depending on their own policy; the relationships among them could be equal, cooperative, competitive, or exploitative. Using preferential detachment we could determine likely patterns of trade and reasonable candidates for new trade agreements (similar to [Axelrod and Bennett 1993], but using a different mechanism).

Two reoccurring problems across much of the animal kingdom (including humans) are 1) the portioning without destruction of a common-pool resource or 2) completing a relatively large-scale project – both of which have similar formalisms. The scale of the problems would determine whether it is appropriate for agents to be individuals, tribes, regional utilities, corporations, or governments. Types would then represent a disposition to act a certain way; e.g., through habit, culture, sticky preferences, business plan, policy, or mandate. The interpretation would be that within the time frame covered by the theory the agents are focused on selecting with whom to share the resource or tasks. To use the theory for a collective action application some assumptions must be made about the ability to exclude agents and monitor sharing/tasking compliance, but these features would be represented by the relative utility of different outcomes rather than explicitly included in the mechanism. So it would take considerable domain knowledge to match a situation to the appropriate strategic context, but nothing beyond what is required to apply other approaches to a particular problem.

Though created to address problems of humans and protohumans, the theory of preferential detachment is general enough to apply to entirely different sorts of agents: ants, worms, amoeba, bacteria, proteins, and even atoms. If agent types are elements, which have different energy profiles and thus act differentially with other elements, the behavior of the atoms can be contingent in a way that corresponds to a preferential detachment situation. For example, in a mixture of carbon and fluoride the carbon-fluoride bonds are stronger than the carbon-carbon bonds, which are themselves stronger than the fluoride-fluoride bonds due to differing levels of electronegativity. Stated differently, carbon will detach carbon (when fluoride is present), but not detach fluoride. Fluoride will detach fluoride (when carbon is present), but will not detach from carbon atoms. These detachments can be considered unilateral because (for example) when a fluoride pair encounters a carbon atom one of those fluoride atoms *will bond* to the carbon atom, and nothing about the other fluoride atom can prevent this. The situation can be represented as a miscoordination/specialization situation in which disassortativity is the prosocial outcome (representing situations as games is discussed below). It is "prosocial" because both types of atoms have an affinity for such an arrangements, and groups composed in that way are more stable. Because more complicated agents are capable of more sophisticated form of contingency, identifying examples using more sophisticated agents is even easier.

The four assumptions of preferential detachment set a low bar on environment sensing and information processing, thus many particular systems are compatible with this general framework. But generality of applicability does not ensure breadth of usefulness. The assumption that connections can be made and broken unilaterally will not be satisfied in some problems. There are also situations in which non-random partner finding is an aspect of the actual dynamics, and perhaps an essential part. These differences can be tested with minor expansions of the mechanism, but the base model put forward in this project suffices to address a large body of interesting situations.

We will return to issues of general applicability when supporting the arguments for the evolution of morality in chapter IV, but much of the discussion focuses on explaining prosocial behaviors in humans and protohumans. For this purpose I will now introduce further assumptions and refinements to the base model.

1.3 Population Size

The number of individuals in a system can have specific quantitative as well as qualitative effects on its social (higher-level) properties. An important (though obvious) example is minimum bound on population sizes sufficient for the emergence of social phenomenon to be even possible. This is certainly expected to be the case here: populations too small to embody certain network features and complexities will also fail to deliver the high-order phenomena we wish to study. In very large populations the complex social properties examined here may get washed out by larger social phenomena or "law of large numbers" effects.

Though the mechanism applies for populations as low as two agents, it is inappropriate for populations below a few dozen (about the size of a wolf pack, elephant herd, or hunter-gatherer tribe). The reasoning is that agents disconnect from a low-payoff partner because they would rather gamble to get a better partner (though they don't calculate this likelihood). In very small populations we might expect everyone to interact with everyone out of necessity, or for the informationless connection assumption to be unreasonable, or for the gamble that one can do better to be non-beneficial in expectation.

The preferential detachment plus random connection process utilized here is certainly going to be path dependent, as well as sensitive to some stochastic elements in the model. Achieving certain ending conditions may depend on some "seed" configurations being reached early in the social development. The greater the population, the greater the probability that these seeds may form; and below some threshold population size it will simply be impossible for them to form. There may be other differences in the system behavior as the population increases. For example, it may be the case that at some population levels only one final group of each type is stable, but at much larger populations separately seeded groups will not merge and distinct groups will persist.

In the mathematical model, and for the cross-game battery of agent-based model experiments, I use an initial population of two-hundred agents (i.e., one-hundred agents of each type for 2 × 2 games). Anthropologist Robin Dunbar estimates that, based on the neocortical size of human brains compared to other primates, the largest group that an anatomically modern human should be able to maintain is between 100 and 230 [Dunbar 1993; 1998]. Actual population sizes for hunter-gatherer groups, neolithic villages, and Hutterite settlements agree with these figures. Numbers in this range are above the threshold generally needed to achieve the prosocial outcomes, but still in the range at which it is not guaranteed. A population of this magnitude was also chosen because it is sufficiently large for kinship, mutualism, and reciprocal altruism to fail as viable explanations for prosociality. These mechanisms do not generally scale well with the number of participants [Fehr and Henrich 2003] without also requiring capabilities beyond what can reasonable be attributed to our evolutionary ancestors. Yet a population of two hundred individuals is still small enough to represent protohuman and social animal populations.

Bridging the gap from small populations to larger ones is a research goal of this project and is achieved by demonstrating cooperation, cooperation, specialization, etc. in a range of population sizes. In the battery of agent-based model experiments focused on exploring preferential detachment in the Prisoners' Dilemma game, I use even initial population values from quite small (10) to very large (1,000) to test how sensitive the dynamics are to population scale. Furthermore, for both the mathematical model and for other payoff structures of the ABM I show that the prosocial results scale for increasing population sizes without an upper bound.

For actual populations, population size is not a fixed figure over significant time scales; birth, death, and migration processes constantly perturb the total population. Over longer periods populations can grow or expire (sometimes very quickly). In the models presented below I make the simplifying assumption that the total population is constant (though the proportions of each type can change through learning and/or population dynamics). Future work specifically aimed at exploring societal viability will relax this assumption and more closely model natural population fluctuations.

1.4 Direct Social Connections

In addition to anthropological and ethological data on the sizes of groups, there is also research revealing that people, non-human primates, and other animals have uppers bounds for the number of direct social ties they can maintain. The number of people in a person's "core" social network is typically 5 or 6 with as many 12 to 16 people who qualify as "close friends" [Dunbar 1998]. This upper bound results from a combination of time and cognitive constraints and also applies to primate groups. The models of preferential detachment reflect this feature of social networks by imposing a maximum number of interaction partners.

The theory does not necessitate any particular value for this maximum degree constraint. The mathematical model provides example calculations for a maximum degree of four, and demonstrates that the pattern of results hold for any larger value. The agent-based model runs the cross-game simulations using a maximum degree value of five. I explore the effect of varying the maximum degree in the Prisoners' Dilemma battery of experiments. The maximum degree is assumed homogeneous across all individuals; an assumption that will be relaxed in future work.

Though the maximum degree is set, there is no assumed minimum number of connections; in all models agents are initialized without any connections and may find themselves unable to maintain any connections. Because connections are made and broken unilaterally, some agents may gain or lose a large number of new connections in one turn (up to the maximum). A drastic increase can happen if many agents happen to randomly connect to some one individual. And it is also possible that a large number (possibly all) of one's neighbors will disconnect from the individual in a very short time (e.g., one simulated time period). So even though an agent can itself only add or remove social connections slowly through its own effort, the actual number of connections one has from turn to turn can still be quite volatile.

1.5 Strategic Contexts

In game theory agents are confronted with a strategic situation in which they choose an action believed to yield the greatest benefit, but the realized outcome of their action depends also on the choices made by some or all of the other agents. This dependency of the outcome on the collective actions is what makes the situation strategic rather than just a decision. In game theoretic modeling one specifies the utility for every agent for each possible outcome. Agents can have full or partial knowledge of the outcomes and of other agents' preferences over them, and they choose the action that they expect will produce the most preferred outcome given their beliefs and preferences.

A simple example of a game is two agents choosing an action in a Prisoners' Dilemma situation under full information. Its infamous equilibrium outcome of mutual defection is the result of choosing preferred actions in that context. But rarely are situations so simple. Deviations and extensions from this basic version abound: limited information, iterated games, heterogeneous interaction structures, changing preferences, and signaling are just a few of the alternatives.

Bob Axelrod's famous computer tournament [Axelrod 1984] required submissions to be in the form of a algorithm that specifies what choice to make given any possible set of input (within the confines of the model) – a complete strategy. An agent using the Tit-for-Tat strategy does not choose an action based on the expected utility of that action given its beliefs. Rather, agents behave according to a rule that converts observed play in the last round of the iterated game into an action for the current round. This is done even though agents are rewarded (reproduced or kept from elimination) based on the utility they receive in the game. The salient feature is that, even though the agents receive utility according to the collective outcome achieved in a Prisoners' Dilemma situation, they do not "game" the situation to achieve maximum individual benefit.

Because behavior is based on a rule that does not explicitly consider the expected actions of the other agents, the theory is no longer game theoretic, though it continues to share many features with game theory. Many (if not most) models in the evolution of cooperation literature have this quasi-game theoretic flavor. Another example is when agents choose *strategies* from a library of possible strategies [Grim et al. 2006]. The agents in agent-based models typically operate on simple contingent behavior rules, while still assigning agent rewards from a payoff matrix. Rather than explicitly optimizing utility over this payoff matrix, they operate on fast and frugal heuristics which perform well despite being very simple [Gigerenzer 2000]. Preferential detachment is similar in this regard: the rewards received by each agent are contingent upon the actions of all those involved but behavior is not the result of optimizing over that payoff matrix.

Because the rewards for each agent depend on the behaviors of all involved agents, the *context* in which the agents behave is still strategic; even though the agents are not strategic in their action. Different situations (problems, environments, circumstances) have different sets of relevant behaviors, and the combinations of agent behaviors yield differing payoffs for each type of agent. The set of payoffs received by each agent for each potential set of collective behaviors is what I refer to as a *strategic context*.

Strategic contexts contain all the same information as game payoff matrices, so the formal models of preferential detachment presented later use payoff matrices (including many that are common from game theory) to represent strategic contexts. Doing so fosters comparison of the results here to results from previous game theoretic, mathematical, and computational models.

1.6 Determining Agent Behavior

The preferential detachment theory assumes agents are making choices about with whom they interact rather than how they act. We can then see that the game used to capture a strategic context will naturally partition the agents into types by the the available actions in the payoff matrix. This is so because agent types are differentiated by 1) how agents map their conditions C1-3 to their behaviors B1-3, and 2) how other agents react to them. And the payoff matrices capture precisely this information. If there were an action with no associated type then no actual behavior would be contingent upon it, thus violating (1). If there were a type with no corresponding action then agents would have to treat it identically to another type, thus violating (2).

1.6.1 Neighbor Preferences

To determine which types of agents detach which other types we next consider the space of pairwise comparisons. First we consider 2 × 2 games, i.e., strategic contexts with two agent types. If each type must have a strict preference over neighbor types (i.e. $a \neq b \neq c \neq d$ and $w \neq x \neq y \neq z$ in 1.1), then there are just four distinct ways that neighbor preferences could be arranged. With respect to preferential detachment, these four possibilities capture the necessary information from all 78 possible strict ordinal preference 2 × 2 games [Rapoport and Guyer 1966].

		player2	
		A	B
player1	A	a,w	b,x
	B	c,y	d,z

Table 1.1: 2 × 2 Payoff Matrix

If we allow for ties in the preference rankings then there are nine cases for preferential detachment compared to the 726 possible 2 × 2 games [Kilgour and Fraser 1986]. Because agents here are fixed in their type, they are only comparing the order of two figures rather than the order of all four entries of the payoff matrix. What this means is that every one of those 726 possible strategic contexts corresponds to one of the nine patterns presented in table 1.2 ("$x \prec y$" means "x is less preferred than y"; "$x \approx y$" means "x is equally preferred to y"; and "$x \succ y$" means "x is more preferred than y").

Pattern1	$A : A \prec B$	$B : A \prec B$
Pattern2	$A : A \prec B$	$B : B \prec A$
Pattern3	$A : A \prec B$	$B : A \approx B$
Pattern4	$A : B \prec A$	$B : B \prec A$
Pattern5	$A : B \prec A$	$B : A \prec B$
Pattern6	$A : B \prec A$	$B : B \approx A$
Pattern7	$A : A \approx B$	$B : B \prec A$
Pattern8	$A : A \approx B$	$B : A \prec B$
Pattern9	$A : A \approx B$	$B : A \approx B$

Table 1.2: Possible neighbor preferences for 2 × 2 games

Any game created by switching both the rows and columns of a symmetric game is isomorphic to the original game. For example, whether the cooperate-cooperate outcome is in the top-left or bottom-right corner of the payoff matrix makes no difference to the dynamics of Prisoners' Dilemma situations – it only changes which type is called the cooperator type. Thus Pattern1 and Pattern4 are isomorphic. Appreciating such isomorphisms further condenses these nine patterns into the six forms presented in table 1.3

Form	Patterns		Relationship		Category
Form1	Pattern1	Pattern4	$A : B \prec A$	$B : B \prec A$	Cooperative
Form2	Pattern2		$A : B \prec A$	$B : A \prec B$	Coordinative
Form3	Pattern5		$A : A \prec B$	$B : B \prec A$	Specialized
Form4	Pattern6	Pattern8	$A : B \prec A$	$B : A \approx B$	Contributive
Form5	Pattern3	Pattern7	$A : A \prec B$	$B : A \approx B$	Commensal
Form6	Pattern9		$A : A \approx B$	$B : A \approx B$	Undifferentiated

Table 1.3: Distinct Game Forms and Categories

Each of the six distinct game forms represents a particular preference relationship between the types of 2 × 2 game players. Thus any 2 × 2 game matches one of these six forms with respect to preferential detachment behavior. We can also categorize the forms by what social phenomenon the game structure represents.

The next step is to locate each of the games used in the current project in one of these categories. For this purpose I present a table of the included payoff matrices in table 1.4. Further details for each game are delayed until section 1.8.

Prisoners' Dilemma

		player2	
		A	B
player1	A	3,3	1,4
	B	4,1	2,2

Hawk and Dove

		player2	
		A	B
player1	A	3,3	2,4
	B	4,2	1,1

Battle of the Sexes

		player2	
		A	B
player1	A	4,3	2,1
	B	1,2	3,4

Coordination Game

		player2	
		A	B
player1	A	4,4	2,2
	B	2,2	4,4

Lichen

		player2	
		A	B
player1	A	1,1	3,3
	B	3,3	1,1

Stag Hunt

		player2	
		A	B
player1	A	4,4	1,2
	B	2,1	2,2

Commensal

		player2	
		A	B
player1	A	1,1	4,2
	B	2,4	2,2

Matching Pennies

		player2	
		A	B
player1	A	3,1	1,3
	B	1,3	3,1

Lane Choice

			player2	
		A	B	C
player1	A	3.3	1,4	1,1
	B	4,1	2,2	4,1
	C	1,1	1,4	3,3

Biased Lane Choice

			player2	
		A	B	C
player1	A	3.3	1,5	1,1
	B	5,1	2,2	5,1
	C	1,1	1,5	4,4

Table 1.4: Game Library

1.6.2 Realized Preferences

To translate a game's payoff matrix into one of the six forms we need to consider the relative payoffs achieved by each type paired with each type as both player1 and player2 in the game. What this means is that all connections are assumed to be bi-directional (i.e., symmetric) such that if $agent_i$ is connected to $agent_j$, then $agent_j$ is also connected to $agent_i$.

If the payoff matrix is symmetric along the diagonal then there is no difference between using just the player1 payoffs versus the combined player1 and player2 payoffs. This is so because if the payoff matrix is symmetric then the combined payoffs are simply double the individual payoffs – a positive affine transformation that does not alter the ordinal ranking of outcomes. Of the ten example games used in this research, only two are affected by the combining. The combined payoffs of any 2×2 constant sum game (of which Matching Pennies is an example) will produce equal payoffs for all outcomes via this transformation because the player1 payoff is the winner if and only if the player2 payoff is the loser. When payoffs are combined in the asymmetrical Battle of the Sexes game, this gives the A-types a utility advantage. The ordinal ranking for each player here is still the same, but this change does affect learning and population dynamics since these processes use the combined scores as the fitness values. I present the realized utility values for each pairwise combination in table 1.5 below. The agent type in a row receives the payoff in the column corresponding to the type of the neighbor in such a pairwise interaction.

Though the agents in the models base their behavior on the order of the transformed payoff values, using the games' payoff matrix is actually a very important component of the theory. The point is that a situation faced by some agents really has the multi-dependent outcome structure captured by a game's matrix, but agents cannot change their disposition to act in a certain way – they can't act strategically. Agents can, however, change their interaction partners. When this is the case the realized payoffs presented in table 1.5 include all the information agents need to choose their behavior.

Prisoners' Dilemma

	A	B
A	6	2
B	8	4

Hawk and Dove

	A	B
A	6	4
B	8	2

Battle of the Sexes

	A	B
A	7	4
B	2	7

Coordination Game

	A	B
A	8	4
B	4	8

Lichen

	A	B
A	2	6
B	6	2

Stag Hunt

	A	B
A	8	2
B	4	4

Commensal

	A	B
A	2	8
B	2	2

Matching Pennies

	A	B
A	4	4
B	4	4

Lane Choice

	A	B	C
A	6	2	2
B	8	4	8
C	2	2	6

Biased Lane Choice

	A	B	C
A	6	2	2
B	10	4	10
C	2	2	8

Table 1.5: Realized Pairwise Payoffs: What Row-Type Gets from Column-Type

1.6.3 Game Categorization

By comparing the realized preferences for the pairwise type interactions we can now match each 2×2 game used here to its form and category (presented in table 1.6). The categorization labels are my own description of the social phenomenon modeled by games within that category. The social phenomena are differentiated by what behavior it requires of agents to obtain. So, for example, even though stag hunting is often referred to as the cooperative or coordinated outcome for the Stag Hunt game, the indifference of hare hunters places Stag Hunt in a different category from other cooperative and coordinative games. I call this category contributive because the Stag Hunt models a situation wherein agents risk a loss by contributing to a collective action (vs. operating individually).

Form	Category	Games	
Form1	Cooperative	Prisoners' Dilemma	Hawk and Dove
Form2	Coordinative	Battle of the Sexes	Coordination Game
Form3	Specialized	Lichen	
Form4	Contributive	Stag Hunt	
Form5	Commensal	Commensal	
Form6	Undifferentiated	Matching Pennies	

Table 1.6: Categorized Games

The categories also identify what the prosocial outcome is for all games in it. The details of the prosocial outcome for each game (what it is and why) are described in detail in section 1.8. In table 1.7 I summarize the prosocial outcome by game category considering only the preferential detachment mechanism.

Form	Category	Prosocial Outcome
Form1	Cooperative	A-type agents grouped together
Form2	Coordinative	A-type and B-type agents grouped respectively
Form3	Specialized	A-types and B-types intermixed
Form4	Contributive	A-type agents grouped together
Form5	Commensal	A-types mixed with B-types
Form6	Undifferentiated	Any mix of A-types and B-types

Table 1.7: Prosocial Outcomes by Category

Prosociality is the unifying concept, but each form of situation must achieve it in its own way. In the Cooperative games the agents have to sacrifice utility (defy temptation) in order to achieve the prosocial outcome. In Coordinative games the agents must agree on an outcome, and though both alternatives are acceptable they may not be equally valuable. Specialized games require that opposites attract. Contributive games require agents to risk a guaranteed modest payoff to contribute to a project that would yield better payoffs. In a Commensal game agents attempt to exploit safe players at the risk of meeting another unproductive exploiter. Finally, agents always do the best they can in Undifferentiated games.

If we break down the outcomes by aggregate behavior we see a different comparison of prosociality for these categories. Cooperative, Coordinative, and Contributive games all reach the prosocial outcome when agents are mixed assortatively. Specialized and Commensal games are solved for the social optimum when agents mix disassortatively. And obviously undifferentiated games achieve the same outcome regardless of how the agents are mixed.

Aside from the 2×2 games included here, I also analyze two 3×3 games. These Lane Choice games are combinations of the Prisoners' Dilemma and the Coordination Game: I will refer to this category as Collaborative Games. As such, the prosocial outcome for both is likewise a combination of the Cooperative and Coordinative ones. An exhaustive treatment for 3×3 games could also be provided, but considering that there are 5831 patterns for 3×3 games, and both of the ones analyzed here are of the same pattern, providing an exhaustive categorization would take us far from the point of this project with minor expected benefit. Such a taxonomy is left for future work.

All games matching a form from table 1.3 are indistinguishable with respect to agent behavior under preferential detachment, however differences in the cardinal payoff values will affect the model behavior with respect to learning and population dynamics as described in section 1.7. So, although Prisoners' Dilemma and Hawk and Dove share the same game form and produce identical results with just preferential detachment, differences produced under learning and population dynamics are insightful for later discussions of institutions and the evolution of morality.

1.6.4 Configurations

Agents' preferences over neighbors fix their behavior in each strategic context, but according to preferential detachment agent behavior is contingent on the number and types of *all* its social connections considered together. In fact, agents don't have access to their above preference ranking in the sense that if an agent only has lesser-preferred neighbors then the agent doesn't know they are lesser preferred. They only have access to their own type, the number of their current interaction partners, and the types of their current interaction partners (C1-3 of 1.1.1). So in order to determine an agent's choice we need to specify the whole space of possible types and neighbor types and determine the preference ranking over this agent state-space.

We can capture all of the information used by an agent to determine behavior with what I call the agent *configuration*: the focal agent's type and the number of neighbors of each type present. The schema used here is $type_0[\#type_i]$ in which $type_0$ is the focal agent's type and $[\#type_i]$ is an ordered list of the numbers of neighbors of each type. The items of the list would normally be in brackets and separated by commas, but when each k_i is a single digit value we can omit the brackets and commas for simplicity. For example, if there are two types of agents (A and B) we can represent an A-type agent connected to one A-type agent and two B-type agents as $A[1,2]$ or $A12$. A B-type agent with the same link neighbors would be written $B12$.

Let the maximum degree be K and the degree of an agent its k. The number of possible configurations, $\mathbb{C}$, for each focal type equals the number of possible combinations of T types of neighbors satisfying $\sum_{i=1}^{T} k_i \leq K$. For any $K > 0$ and $T > 0$ we can calculate the number of configurations by

$$\mathbb{C} = \sum_{i=0}^{K} \binom{T-1+i}{T-1} = \binom{T+K}{T} = \frac{(T+K)!}{T!(K!)}. \tag{1.6.1}$$

For example, if $T = 2$ and $K = 4$, then there are 15 different possible configurations for each type θ of agent as follows:

$$\theta 00,\ \theta 01,\ \theta 02,\ \theta 03,\ \theta 04,\ \theta 10,\ \theta 11,\ \theta 12,\ \theta 13,\ \theta 20,\ \theta 21,\ \theta 22,\ \theta 30,\ \theta 31,\ \theta 40.$$

Once we specify a strategic context and a maximum degree the agents can satisfy C1-3 and this induces a preference ranking over configurations that maps to B1-3.

1.6.5 Revealed Preferences

Each game form corresponds to exactly one revealed preference relation over configurations. Each of the six game forms above has a unique *pair* of preference relations; agents of each type in 2×2 games have only three possible relations (prefer A-types, prefer B-types, or indifferent). The preferences are *revealed preferences* because these preference relations are not assumed by the theory, they are produced by the theory once a problem is fully specified. Having the revealed preference ranking in hand allows us to see what action an agent will perform given its current configuration and opportunity set. These revealed preferences will be used to capture agent behavior in the Markov models of chapter II and relate the agent-based model to analytical results.

The preference ranking over the configurations includes all possible agent-states so the ranking does not change based on the current configuration of the agent. However, the *opportunity set* of which configurations can be reached varies with the current configuration. Because some configurations cannot be reached from certain configurations through B1-3 alone, there may be multiple preference rankings compatible with preferential detachment behavior. For example, in the Prisoners' Dilemma A-types prefer two A-type neighbors to just one ($A10 \prec A20$) and prefer fewer B-types when attached to A-types ($A21 \prec A20$), but since it is impossible for an agent to change between $A21$ and $A10$ directly through its own action, no preference between these two states can be revealed. The comparison of $A10$ and $A21$ is underspecified. Both are less preferred to $A20$, but their relation could be anything ($\prec$, $\succ$, or $\approx$). In the revealed preference rankings such gaps are filled in a way that will never be violated by agent behavior (e.g., $A10 \prec A21 \prec A20$ in the case just presented), but is not directly specified by the theory.

I present the three possible revealed preference relations over all configurations for $K = 4$ in table 1.8. So, when $K = 4$ each player in any 2×2 game will behave in a way corresponding to one of these three revealed preference relations, and the realized payoffs for the game being used completely determines which one of the three is used by each player.

$B \prec A:$	$\theta00 \prec \theta01 \prec \theta02 \prec \theta03 \prec \theta04 \prec \theta13 \prec \theta12 \prec \theta11 \prec \theta10 \prec \theta22 \prec \theta21 \prec \theta20 \prec \theta31 \prec \theta30 \prec \theta40$
$A \prec B:$	$\theta00 \prec \theta10 \prec \theta20 \prec \theta30 \prec \theta40 \prec \theta31 \prec \theta21 \prec \theta11 \prec \theta01 \prec \theta22 \prec \theta12 \prec \theta02 \prec \theta13 \prec \theta03 \prec \theta04$
$B \approx A:$	$\theta00 \prec \theta10 \approx \theta01 \prec \theta20 \approx \theta11 \approx \theta02 \prec \theta30 \approx \theta21 \approx \theta12 \approx \theta03 \prec \theta40 \approx \theta31 \approx \theta22 \approx \theta13 \approx \theta04$

Table 1.8: The Three Possible Revealed Preference Relations for K=4

The choice of $K = 4$ was for convenience of presentation – a simple algorithm presented in equation (1.6.2) generates the revealed preference relation for any 2×2 game for any K. A similar algorithm exists for any size game (i.e., any T), and in fact a general approach for any T and K can be devised, but the value of having this in hand is less than the effort required to present it. Only two games analyzed here are not 2×2 games, and their revealed preference rankings are presented with the game details in sections 1.8.9 and 1.8.10.

$$\phi_1 : B \prec A : \begin{cases} \theta(a,b) \prec \theta(a,b+1) & \text{if } a = 0 \\ \theta(a,b) \prec \theta(a,b-1) & \text{if } a > 0 \\ \theta(a,b) \prec \theta(a+1,b) & \forall a,b \\ 0 \leq (a+b) \leq K & \forall a,b \end{cases}$$

$$\phi_2 : A \prec B : \begin{cases} \theta(a,b) \prec \theta(a+1,b) & \text{if } b = 0 \\ \theta(a,b) \prec \theta(a-1,b) & \text{if } b > 0 \\ \theta(a,b) \prec \theta(a,b+1) & \forall a,b \\ 0 \leq a+b \leq K & \forall a,b \end{cases} \tag{1.6.2}$$

$$\phi_3 : B \approx A : \begin{cases} \theta(a,b) \prec \theta(a+1,b) & \forall a,b \\ \theta(a,b) \prec \theta(a,b+1) & \forall a,b \\ \theta(a,b) \approx \theta(c,d) & \text{if } a+b = c+d \\ 0 \leq a+b \leq K & \forall a,b \end{cases}$$

It will be convenient to refer to these three revealed preference relations using the ϕ_i shorthand for each. ϕ_1 generates a lexicographical preference ordering that considers A-types before B-types and favors A-types over B-types unless there are no A-type neighbors. If an agent has no A-type neighbors then it doesn't know they are preferred and won't detach B-types. An agent may acquire more B-type neighbors up to K. The rule-based behavior of preferential detachment acting on only C1-3 induces the revealed preference relation such that $\theta(a,b) \prec \theta(a,b+1)$ if $a = 0$ even though A-types are preferred.

The same reasoning applies to ϕ_2 by exchanging 'a' and 'b' as well as 'A-type' and 'B-type'. ϕ_3 produces a preference relation such that all combinations of A-types and B-types totaling the same number of neighbors are equally preferred, and having more neighbors is preferred to fewer. Thus ϕ_3 represents the behavior of agents that will randomly connect to new neighbors, but detach none.

At this stage we have seen all the exposition necessary to fully understand and model the the preferential detachment mechanism. The base behavior rules, embedded in a strategic context and limited by a maximum number of network neighbors, induces a preference relation over every possible configuration any agent could be in. Any model with agent behavior aligning with these revealed preference relations coheres with the preferential detachment theory. I will now discuss two extensions made to this base framework that are also included in the agent-based model presented in chapter III.

1.7 Learning and Population Dynamics

In this section I present two extensions of the theory that are implemented in the agent-based model of chapter III. The first is an additional agent behavior through which agents imitate the type of their most successful network neighbor. I call this imitative behavior *learning* following [Boyd and Richerson 2005] and others. The second extension is a population-level selection mechanism that duplicates a percentage of the best-performing agents and removes the same percentage of the worst-performing agents. Both of these extensions require an additional cardinality assumption regarding the utility values.

Under pure preferential detachment the agents' preferences are never compared; all actions are performed unilaterally by the agents. Determining the realized (combined player1 and player2) payoffs required adding two payoff values for an agent, and this made an implicit assumption that the payoff values could be treated as cardinal values. This is so even considering that the rankings of the realized payoffs match the ordinal rankings of the original game outcomes. Both of these extensions require comparisons of performance across agents, thus they also require a common currency of agent utility. This common currency is established by using the realized payoff values from table 1.5 as the cardinal utility from the respective pairwise interaction.

The conversion of ordinal ranks into cardinal payoffs is typical in the relevant literature, though not without consequences [Jeffrey 1990, Elster 2007; 2008, Gintis 2009]. The primary limitation is a loss of generality. Any ordinal ranking is compatible with an infinite number of payoff values, a cardinal conversion is just one of those. If some result can be demonstrated with ordinal ranks then it is a more general (and hence more powerful) result. In many analytical models this limitation can be overcome by providing ranges of cardinal values for which the result holds using comparative statics [Rapoport and Chammah 1970, Nowak et al. 2004]. Simulation results are more restricted in this sense because providing ranges often requires running large parameter sweeps to test sensitivity. Generalizability of the results will be discussed in the conclusions section of the agent-based model (see 3.10).

It is important to note that the preferential detachment behavior does **not** utilize these cardinalized utility values. Preferential detachment follows the heuristic presented in section 1.1, which induces the revealed preference relations already discussed. Some of the more preferred configurations yield lower cardinal utility values than lesser preferred configurations. For example, in the Prisoners' Dilemma an agent prefers $A12$ to the $A13$ configuration as revealed by the behavior of detaching the lower-paying B-type agents. But A-types connected to B-types gain two utils (units of utility) from such a connection (as shown in table 1.5). This implies that, according to the converted cardinal utility values, agents are acting to decrease fitness with respect to learning and population dynamics.

One way to interpret this discrepancy is that the preference ranking over configurations reveals what satisfies the agents, what makes them "happy". The cardinal utility values are only relevant in imitations and reproduction, thus we can interpret them as a measure of agent influence, reproductive fitness, or other such factors depending on the situation being modeled. That a person may be more satisfied in a configuration that produces fewer children and/or gives them less influence over his or her peers is quite reasonable. In fact, we will see an interesting result from both the Markov and agent-based model in the Prisoners' Dilemma context that hits on this very point.

1.7.1 Success-Based Imitation

In the game-theoretic literature learning refers to belief updates; with preferential detachment agents are assumed to always have perfect knowledge of their own type and the types of their current network neighbors (so learning in this sense does happen when agents change their configuration, but it is not a research point) [Fudenberg and Levine 1998]. I instead employ a broader meaning of "learning" that is common in the relevant literature: behavior imitation.

Agents learn by changing their type to match the type of one among the network neighbors with the greatest cardinal utility value. This form of success-based imitation is the most widely employed by models in the evolution of prosociality literature [Boyd and Richerson 1987, Bicchieri 1990, Boyd and Richerson 2002, Fehr and Henrich 2003, Hedstrom 2005, Nowak 2006, Bergin and Bernhardt 2009, Fosco and Mengel 2010, Perc and Szolnoki 2010] and the most supported by empirical research in business practices [March et al. 2000, Young 2009], policies [Lemola 2002, Shipan and Volden 2008], hunting techniques of hunter-gatherers [Trivers 1971, Hewlett and Sforza 1986], tool creation by primates [Flack and de Waal 2000], and many others.[2]

[2] Some models use popularity-based imitation such that agent's change their type to copy the most popular type among their neighbors [Boyd and Richerson 1989, Grim et al. 2006]. There are other forms of learning (e.g., signals [Boyd and Richerson 1987, Skyrms 2002]), but these are infrequently incorporated in the relevant literature.

When we include imitation learning we must also consider the order of operations – the timing of imitating, detaching, and random attaching. In some cases the highest utility neighbor will be one from which the focal agent wants to disconnect (e.g., a cooperator with a connected defector). Given the motivation behind the theory of preferential detachment, having agents imitate neighbors they prefer not to interact with fails the face validity test. Nor does it make sense to imitate an agent that has just been randomly attached. So the learning process operates after the detachment phase and before the random connection phase in the agent-based model of chapter III.

1.7.2 Fitness-Based Birth and Death

Population dynamics is a technique that involves selectively eliminating some agents and/or adding new agents. It is often invoked by evolutionary computation (e.g., genetic algorithms) in which solutions are reproduced in proportion to their fitness. Some equation-based models also use this technique in the form of replicator dynamics [Skyrms 2004, Nowak 2006]. A differential equation model may achieve a similar process by the changing the number of agents playing each action as a function of the current number playing each action [Boyd and Richerson 2005]. Markov models may also be of this sort by including transition probabilities between agent types/behaviors; e.g., an SEIR Markov model is of this form. It is also possible to explicitly model the pairing and mating of agents as well fluctuations in morbidity and fertility in agent-based models. All of these are examples of population dynamics as I mean it here.

The approach used here is most similar to that used in evolutionary computation. The fitness of agents is calculated as the sum of the realized payoffs earned from each of their pairwise interactions. Each period the best five percent are replicated and the lowest five percent are removed. The complete details of how this is implemented in the agent-based model are described in section 3.4.4.

Note that, unlike preferential detachment and imitative learning, population dynamics is a global operation. This is not something that agents do, but rather something that happens to them. The selection rate (percent per time step) represents the evolutionary pressure exerted by the environment. The five percent rate chosen for the agent-based model simulations is quite high, and the effects of this rate (versus others) is discussed in the conclusions section of the agent-based model.

1.8 The Game Library

Here I present and describe the relevance and details of each of the ten payoff matrices addressed in this project (presented in table 1.4). Several of these games already exist as named games in the literature with established research histories: Prisoners' Dilemma, Hawk and Dove, Battle of the Sexes, Coordination Game, and Stag Hunt have been studied in numerous studies of the evolution of prosocial behavior. Lichen and Matching Pennies (and possibly Commensal) exist in the game theory literature, but not in the evolution of prosociality literature. The two 3×3 Collaborative games games are novel combinations of common 2×2 games with no precedent.[3]

For each game I provide some hallmarks of its use, focusing on appearances in the evolution of prosociality literature. I display again the payoff matrix and specify the prosocial outcome under pure preferential detachment, learning, population dynamics, and combined learning and population dynamics. All of the details essential for understanding the formal models of chapters II and III are summarized in table 1.10 at the end of this chapter.

The reason that ten different payoff structures are used, and are important, is that collective action and common-pool resource problems take many forms. There is already some debate about which game best represents the "game of life" faced by humans and protohumans, the solving of which best explains the evolution of prosociality [Boyd 1988, Binmore 1998, Skyrms 2004]. The major contenders are the Prisoners' Dilemma, Stag Hunt, and Hawk and Dove games. Some social problems are distinct from these games but clearly also essential to developing and maintaining general prosociality. Specifically, games requiring pure coordination, fairness, specialization, and other behavioral repertoires must be important as well. It is too simplistic to think that a behavioral rule that fosters cooperation in the Prisoners' Dilemma (but no other prosocial behavior) suffices to explain or act as a guide to general prosociality. For example, "Kelley and Thibaut (1978) discuss a large array of mixed-motive games that characterize various social interactions, and Taylor and Ward (1982) argue that the n-person version of the game "chicken" is essential to understanding cooperation" [Boyd and Richerson 2005]. It is for this reason that

[3] The Ultimatum Game also figures prominently in the evolution of prosociality research, but has been left out of the current analysis in order to maintain the focus on the exhaustive coverage of 2×2 games. Preliminary results show that because a discrete version of the Ultimatum Game generates a payoff structure with the action to ask for 50% as an assortatively mixed outcome, preferential detachment succeeds in achieving perfect prosociality.

the current research explores a library of ten strategic contexts, each focused on a different social problem. The apparent inability of other mechanisms to achieve the prosocial outcome in a wide set of problems, and the ability of preferential detachment to do so, makes the current theory a more robust and general explanation of prosocial behavior.

1.8.1 Prisoners' Dilemma

The Prisoners' Dilemma is by far the most examined game in the literature on the evolution of prosociality (including the evolution of norms, morality, language, culture, and other phenomena). It appears in the works of Plato [Plato 2005], Hume [Hume 2000], von Neumann [Morgenstern and Neumann 1980], Rapoport [Rapoport and Chammah 1970], Axelrod [Axelrod 1984], and several hundred researchers since...and continues to dominate research in the field today with new papers coming out every month that analyze the evolution or maintenance of various social features via the Prisoners' Dilemma model. It is also well known and much analyzed in its N-player version (for N greater than two) typically referred to as the Tragedy of the Commons.

Figure 1.1: Prisoners' Dilemma

		player2	
		A	B
player1	A	3,3	1,4
	B	4,1	2,2

The attraction to the Prisoners' Dilemma is clear; the rational action for both players (as identified by rational choice theory) is to play the dominant action of defection. This is the unique Nash equilibrium of the game. Yet empirical studies frequently reveal that people (even people who demonstrate mastery of the game's rules) frequently cooperate [Gintis et al. 2003]. Because this observed cooperative behavior is both helpful and irrational (in this context), and because the formulation is so concise, the Prisoners' Dilemma seems an excellent model for examining various prosocial behaviors (or an absence thereof) in difficult social problems.

The rational action of mutual defection mentioned applies primarily to the original N-player one-shot version of the game. There are, however, a large number of variations and extensions to this version: iteration, spatial structure, continuously variable cooperative investments, and multi-person interactions to name a few

[Doebeli and Hauert 2005]. Each of these variations is then combined with different proposed mechanisms such as punishment, population dynamics, trust, kinship, homophily, reciprocity, and others which each create new sets of assumptions, dynamics, and outcomes [Kuhn 2007, Perc and Szolnoki 2010]. So with respect to the well-researched Prisoner's Dilemma the current project's addition is just a drop in the ocean of extant models. Nevertheless, this particular combination of assumptions and mechanisms is indeed novel in the literature.

Prisoners' Dilemma research focuses on whether the prosocial outcome of mutual cooperation (A,A) can be achieved and maintained. In the indefinitely repeated version cooperation is rationally sustainable as per the Folk Theorem, but if the game ends at some time then through backwards induction the agents will again find that immediate and sustained defection is uniquely rational. If the future is discounted sufficiently then cooperation again becomes sustainable [Alexander 2003, Skyrms 2004]. Another approach is that agents play one of a number of strategies (e.g., Tit-for Tat, Doormat, Grim Trigger) instead of optimizing in each game. Again cooperation is viable under many different model constructions, though possibly open for exploitation. Replicator dynamics have been used to achieve the prosocial outcome under a number of assumptions regarding the likelihood of interactions (signaling, homophily, random, etc.) [Binmore 1998, McElreath et al. 2003, Boyd and Richerson 2005].

There has been an explosion of research over the past few decades in iterated games and computational game theory. This may have been partly fueled by the simplicity and success of Rapoport's Tit-for-Tat strategy in Axelrod's computer tournament [Axelrod 1984], though it has certainly benefited from the availability of powerful desktop computers that make such research accessible to anybody. The Tit-for-Tat strategy has intuitive appeal as a social norm, but it is actually a complicated mechanism. Its success in the iterated Prisoners' Dilemma is attributed to the fact that it is a "combination of being nice, retaliatory, forgiving and clear." Reciprocal altruism is just part of that complicated behavior, punishment is another part, and they have since been examined in detail separately.

Punishment is the most frequently cited mechanism to achieve cooperation in the Prisoners' Dilemma [Bowles and Gintis 2004]. Punishment means different things in the different models because the assumptions and mechanisms are different. Also, punishment is a thick term, so it is not itself a feature of the formal models, but rather an interpretation of the mechanism in the modeled situation. The Prisoners' Dilemma payoff matrix (and associated stories) readily lends itself to this loaded interpretation:

punishment is an action performed by a cooperator or group of cooperators to reduce the utility of defection (possibly at a cost to the punisher).

In this sense preferential detachment in the Prisoners' Dilemma can be interpreted as punishment, but the mechanism itself cannot be considered a real punishment-specific mechanism because the same mechanism is utilized in situations (other games) in which the conceptual requirements to apply punishment are not met. What this suggests is that the mechanism regulating behavior may be simpler than our interpretation of them makes apparent. It is only viewing this behavior in a particular context that makes it punishment, and punishment itself is not specifically important for the evolution of prosocial behavior. These ideas will be explored in detail in chapter IV.

In addition to punishment, there has been a great deal of research on reciprocity's ability to achieve and maintain cooperation in the Prisoners' Dilemma [Trivers 1971, Axelrod and Hamilton 1981, Boyd 1988, Güth and Yarri 1992, Binmore 1998, Lee 1998, Fehr and Henrich 2003, Bowles and Gintis 2004]. Researchers have analyzed reciprocity in combination with a large number of other model assumptions (grids, networks, kinship, signals, exchange, investment, etc.). Reciprocity also gets a credibility boost from observations made of social animals, especially primates, regularly engage in grooming, food sharing, and other behaviors that clearly reflect reciprocal altruistic features[Flack and de Waal 2000, Wilson and Wrangham 2003, Bowles and Gintis 2004, Bekoff and Pierce 2009]. For reciprocation to work agents need to remember who performed a helpful action for them and for whom they have performed helpful actions.

General reciprocity, that an individual will help others (even strangers) as long as that individual has received help from somebody, is simpler because it only requires that agents remember that help was received. Recent studies have shown that even rats exercise general reciprocity in some cases [Bekoff and Pierce 2009] and humans are more likely to help a stranger if they have just found money on the sidewalk. In preferential detachment all behaviors are contingent only upon current social network partners so there is no room for any form of reciprocity.

Cooperating despite temptation to earn a higher payoff through defection is the prosocial behavior we are looking for in the Prisoners' Dilemma. In the base preferential detachment model we will be looking for assortative mixing: cooperators form a group of only cooperators and separated from all defectors. It is possible, and in fact likely, that *some* cooperators will be stuck in stable heterogeneous groups. These arrangements form when the cooperators gain K defectors before attaching to another cooperator. Because behavior is only based on current neighbors these cooperators will not detach their defector neighbors because no neighbor is worse than any other. The defectors attached to these cooperators, however, will detach any defector neighbors they may gain. So these heterogeneous groups will be isolated from both the cooperator and defector groups. Both types of agents are worse-off in this arrangement, but it is stable under preferential detachment. Once we add learning and/or population dynamics the greater fitness of cooperators in a group of cooperators will drive the whole population to A-type dominance – as long as a seed group of cooperators is able to form.

1.8.2 Hawk and Dove

This game (also known as "Chicken") is another symmetric game with no tied outcomes. It appears in the work of von Neumann and Morgenstern [Morgenstern and Neumann 1980] and Anatol Rapoport [Rapoport and Guyer 1966], but was strongly highlighted in the work of J. Maynard Smith [Smith and Price 1973, Smith 1976] where it got this name. Smith's story is that there are two strategies for animals to play when meeting (to defend territory for example). One option is to be a hawk (defect) and fight vigorously at any cost. The other option is to be a dove (cooperate) and acquiesce every time. Like in the Prisoners' Dilemma, the pro-social outcome is for both players to cooperate, yet the two pure-strategy equilibria are those in which one agent plays hawk and the other plays dove. The only evolutionarily stable strategy is a mixed strategy.

Though less examined than Prisoners' Dilemma, Coordination Game, or the Ultimatum Game, Hawk and Dove does appear in the recent iterated games and computational modeling literature. Tomassini's model (which also uses a dynamic social network) has shown that cooperation can be established and maintained through a trust-based mechanism in which potential neighbors are introduced by other agents and reputation percolates through the network [Tomassini et al. 2010]. Other work has shown (or implied) that mechanisms successful in achieving cooperation in the Prisoners' Dilemma would also be successful in the Hawk and Dove game because they both require making a short-term sacrifice for long-term cooperative gains. This similarity is also also shared under preferential detachment since both Hawk and Dove and Prisoners' Dilemma are categorized as cooperative games with the same revealed preference relation.

Though I used the terms "defect" and "cooperate" above, those concepts are not quite appropriate for the story told, or at least the meanings here are different

Figure 1.2: Hawk and Dove

		player2	
		A	B
player1	A	3,3	2,4
	B	4,2	1,1

than in the Prisoners' Dilemma. In Prisoners' Dilemma cooperation protects against mutual ruin, but here cooperation is to keep one agent from exploiting another. This difference is not merely formal, it distinguishes which situations may be appropriately modeled by one game or the other. So though we can see an outcome as being cooperative in the broad sense, and the word is sometimes used in this broad sense, what precisely it means to cooperate changes based on the structure of the payoffs. Insofar as Hawk and Dove shares the features that agents must make a sacrifice to converge on the mutually beneficial outcome, it is consistent to refer to also this game as cooperative; just a slightly different form of cooperation.

I define the prosocial outcome in Hawk and Dove as promoting dove-like (A-type) behavior. This is not an assumption without consequence, especially when comparing to other models. The revealed preferences over configurations here is identical to the Prisoners' Dilemma, so without learning or population dynamics the system behavior will be the same, and the goal here is also assortative mixing. The preference ranks are the same as Prisoners' Dilemma, but the payoffs for action combinations are different, and this difference will produce distinct aggregate behavior when we utilize these as cardinal utility values.

With learning and/or population dynamics A-type domination is the prosocial outcome. Note that this designation is partially at odds with the typical social utility approach of adding all the agent utilities. From a cardinal conversion of the standard payoff matrix a population of all doves is just as social utility satisfying as a population evenly split between exploiting hawks and exploited doves. From this we conclude that it may or may not be more efficient for society to have all doves, it depends on the particular actual cardinal values for that situation. These cardinal values are only accessible in a very narrow range of applications (e.g., when dollars are a reasonable substitute for personal utility). By keeping prosociality in terms of observable social phenomena (groups of doves) we can avoid these complications. Thus preferential detachment theory includes an implicit assumption that efficiency (i.e., achieving the greatest total utility value) is not the criterion for prosociality.

1.8.3 Battle of the Sexes

The Battle of the Sexes is another venerable payoff structure in the study of game theory [Morgenstern and Neumann 1980, Osborne and Rubinstein 1994]. It is within the category of coordination games, but it differs from the pure Coordination Game discussed below because each player has a preferred outcome. This asymmetry is problematic within the formalism of many previous models, so they often use a symmetric version instead. Their "Symmetric Battle of the Sexes" is isomorphic to the Coordination Game below, yet the asymmetry present in the original version **does** affect the trajectory of the system's dynamics (as we will see in the results section 3.9).

Figure 1.3: Battle of the Sexes

		player2	
		A	B
player1	A	4,3	2,1
	B	1,2	3,4

In the Battle of the Sexes context the prosocial outcome is identified as complete assortativity. This is clearly desirable since the highest individual payoffs for both types are achieved in coordinated outcomes. However, this could mean either isolated groups of each type or the dominance of one type. These both count as socially beneficial outcomes, though one of these two outcomes may be preferred in applications of the game to particular social problems. For example, if the problem is two sports teams coordinating on uniforms then clearly maintaining a population of both types is socially better. If the problem is choosing a national outlet voltage, it may not matter if the result is 120V or 240V as long as the whole country uses the same voltage.

In the scenarios involving population dynamics and/or learning, we need to recognize that A-types do better than B-Types in mixed groups because the realized payoffs are greater for A-types in this asymmetric payoff matrix (see 1.5). Thus we expect to see the A-types do better in the beginning, before the two types segregate themselves, but then perform equally well after segregation is complete. The question of whether A-types will dominate the population or have only a slight advantage depends on the time-scales of the learning/population dynamics pressure versus preferential detachment.

Without learning or population dynamics, however, the payoffs work only as ordinal ranks. As you can see from the revealed preference rankings in table 1.8 the configuration order for A-types is the same as in the Prisoners' Dilemma. The order for B-types is simply the reverse, meaning that A-types treat B-types the same way that B-types treat A-types. I will make extensive use of these symmetries and reversals in the Markov model of chapter II. What that means here is that, in distinction from the Prisoners' Dilemma, both types are trying to self-associate. So while stable heterogeneous groups are still possible, they are much less likely (as will also be shown by the Markov model).

Many of the existing mechanisms that achieve prosocial outcomes in the Prisoners' Dilemma also work in the Battle of the Sexes; i.e., anything achieving assortative mixing. So punishment, signaling, mutualism, reciprocity, etc. on fixed networks, grids, and type-probabilistic mixing will also produce assortative mixing in the Battle of the Sexes and hence also produce the prosocial outcome. As with the Hawk and Dove game, the meaning of these mechanisms must be reinterpreted because, considering the stories told for Battle of the Sexes situations, punishment is an inappropriate way to think about behavior in response to mixed-type interaction. Other mechanisms, such as replicator dynamics and those using random mixing, are less suited to the Battle of the Sexes because even if they yield equal populations of both types, there is no notion that they are assortatively grouped.

1.8.4 Coordination Game

As mentioned above, the Coordination Game can be considered a symmetric version of the Battle of the Sexes, and has been studied under both names [Lewis 1969, Schelling 1978, Osborne and Rubinstein 1994]. The Coordination Game represents a scenario in which either action is equally valuable to all participants, but they must agree on which action to use. Like the Battle of Sexes game, assortativity could mean that the society splits into two stable populations of agents each coordinated on a different action, or one could come to dominate the population. The prosocial outcomes in this strategic context are the same as Battle of the Sexes.

Figure 1.4: Coordination Game

		player2	
		A	B
player1	A	4,4	2,2
	B	2,2	4,4

For the Coordination Game we again measure the success of the outcome by the degree to which the two types segregate. An efficient outcome has been reached whenever all interactions are with same-type agents and each agent is using its full potential of interaction partners. There is no requirement on success that the society should be split evenly or that one should dominate. In a further generalization of Coordination games, the payoffs of (A, A) and (B, B) may differ (but still be symmetric) creating a socially preferred outcome. The results of such a situation are covered through the Biased Lane Choice game described below.

The revealed preference ranking here is identical to that of the Battle of the Sexes, thus without learning or population dynamics the system behavior will be identical. However, with learning or population dynamics the behavior of the two strategic contexts will differ. With no bias in the cardinal payoffs, the populations will not separate as in the Coordination Game the way they did in Battle of the Sexes – the two types remain indistinguishable. The comparison of Battle of the Sexes and Coordination Game is particularly relevant to norms of fairness because even though the agents cannot distinguish between the two situations, the difference in actual payoffs produces different social arrangements. One context favors A-types and the other does not. Recognizing such differences is essential for understanding preferential detachment's potential use for institutional design.

1.8.5 Lichen

Lichen is a game of specialization, complementarity, and miscoordination. One might also think of it as the Mating Game or Spin the Bottle because the ideal outcomes are achieved when each type of agent is paired with agents of the other type. My choice of "Lichen" reflects the symbiotic relationship between algae and fungus in lichen. The game is referenced as an example of persistent miscoordination in lectures by Fudenberg and Kreps [Fudenberg and Kreps 1990] and is used to make a point about the generation of norms by Young [Young 1993].

Young uses it as an example of cases such that in fictitious play the distribution of actions converges to a Nash Equilibrium, but the players' behaviors do not. So in Young's model the players have a stochastic adaptive strategy that can lock in coordinated patterns of behavior if they coordinate by chance often enough (what he calls a convention). No player is playing a Nash equilibrium but the aggregated actions match the appropriate percentages. Because in preferential detachment the agents' action is fixed by type, the same system-level behavior is impossible to recreate. However, Young's result is closest in kind to solutions of agent-based models; the solution is a feature of the society of agents rather than an action or strategy of individual agents.

Figure 1.5: Lichen

		player2	
		A	B
player1	A	1,1	3,3
	B	3,3	1,1

As the above discussion already indicates, the prosocial outcome in a Lichen situation is for heterogeneous social structures to dominate. This heterogeneity can be through either static social structures or an unstable, yet reliable, flipping of agent roles. If learning and population dynamics are off, then the prosocial outcomes is a population with a "checkerboard" pattern of alternating A- and B-types (a K-regular bipartite graph). When this alternating arrangement is present, both types of agents perform equally well, so population dynamics acting on this structure would reproduce and eliminate an even number of each type in expectation. The prosocial arrangement is still the same checkerboard pattern under population dynamics, though it will not be as efficiently met because a number of agents comprising the turnover group will not fit the optimal alternating structure.

Behavior under learning is more difficult to anticipate. If the network were fully formed before learning were activated then all agents would be performing equally well and no changes would occur. But while the network is growing through random connections and preferential detachment, stochastic variance in connections will make some agents more successful than others. During this transient period imitation will change some agent types. These changes may balance out and settle to the same alternating pattern as happens without learning. It is also possible that a combination of imitating and connecting will induce a cycle of type-switching as seen in some network models of neuron activation and similar in behavior to Young's adaptive non-static distribution of behaviors.

Lichen is a reversal of the Coordination game and hence the revealed preferences are exactly the reversed preferences. As noted above, there are numerous known ways to achieve reliable assortativity in social structures; however, there is little research in achieving sustained disassortativity. Nevertheless the achievement of effective specialization is key to achieving the best outcome for individuals and groups in many important social situations. The critical importance of finding prosocial outcomes in these specialization situations should therefore reveal itself in moral norms. The success of preferential detachment in this context could also stand as an important result for designing institutional frameworks capable of achieving the appropriate matching of complimentary abilities or resources.

1.8.6 Stag Hunt

The Stag Hunt payoff structure captures a situation in which two individuals can either cooperate to take down a large animal (such as a stag) or go it alone and hunt easy small prey (such as hare). The primary difference between the Stag Hunt and the Prisoners' Dilemma is that in the Stag Hunt both assortative outcomes are Nash equilibria. Rousseau coined the name Stag Hunt for this strategic context [Rousseau 1987], but problems of the same form also occur in Hume as the oarsmen problem [Hume 2000] and in other works before. Examples of behavior exhibiting Stag Hunt relationships include slime molds in forming spore pods and aquatic mammals practicing carousel feeding (collecting fish into a tight area for easier feeding). The salient feature is that the payoff per interaction of B-types is independent of what type each neighbor is, while A-types depend on interactions with other A-types to achieve a good outcome.

Some researchers, most notably Brian Skyrms and Jason Alexander, have highlighted the features of Stag Hunt to argue that it provides a more accurate representation than the Prisoners' Dilemma of the decision conditions facing protohumans – at lease decisions relevant to the development of pro-social behaviors [Skyrms 2004, Alexander 2003]. Specifically, "defecting" is rational and safe. "Cooperation" is also rational, but risky. So this game is easier to solve than the Prisoners' Dilemma since pairs of stag hunters will continue to hunt stag even under conditions of perfectly informed rationality. I used quotes in describing the behaviors because behaviors in this context are distinct from those behaviors in the Prisoners' Dilemma. I refer to the stag hunting behavior as *contributing* and hare hunting as *opting out*.

Figure 1.6: Stag Hunt

		player2	
		A	B
player1	A	4,4	1,2
	B	2,1	2,2

The prosocial outcome for a Stag Hunt situation is a population dominated by contributers (i.e., A-type agents). Without learning or population dynamics this is not possible, so assortative mixing is what we will look for. Because groups of stag hunters do better than hare hunters in terms of fitness, assortativity implies imminent domination under population dynamics. Learning will not lead to complete dominance if the two types segregate too quickly. Though because hare hunters paired with stag hunters don't do better (i.e., no exploitation), it should also be the case that learning and population dynamics drive the population to A-type dominance faster compared to the Prisoners' Dilemma game.

The relative ease of achieving the pro-social outcome in the Stag Hunt compared to the closely related Prisoners' Dilemma has two separate repercussions for the evolution of prosociality. The first result is that certain proposed mechanisms that are not sufficient to achieve (or not robust in achieving) cooperation in the Prisoners' Dilemma are sufficient and robust for contribution in the Stag Hunt game [Skyrms 2004]. The second result is due to Axelrod who showed that applying a discount to future time periods in the iterated Prisoners' Dilemma transforms that game into a Stag Hunt game – where sustained cooperation is rational [Axelrod 1997, Alexander 2003]. This result demonstrates the famous proof by Anatol Rapoport that the only ways to achieve cooperation in the Prisoners' Dilemma with rational agents is to either alter the payoffs or have assortative mixing. The preferential detachment model I am presenting is clearly relying on assortative mixing rather than payoff manipulation, and the mechanism assumes less about the capabilities of those involved.

1.8.7 Commensal

The Commensal game is a reversal of the Stag Hunt game such that the B-types are still acting independently with payoffs insensitive to neighbor type, but now the A-types prefer interactions with B-types instead of other A-types. Commensalism is a symbiotic relationship in which the commensal receives a benefit from the host, but the host is neither benefited nor harmed. I call this game "commensal" with the following story in mind. The human body is host to literally millions of microfauna (such as mites that live in hair follicles) that neither provide benefit nor cause harm to their host. They live largely isolated lives and do not critically depend on interactions with others of their kind (they reproduce asexually by producing offspring from unfertilized eggs, then reproduce sexually with their own clones). The host biomass consumed from feeding makes little to no difference to the tissues surrounding them, but overcrowding results in insufficient nutrients to each individual. Barnacles and cattle egrets are other examples from biology. The biological match is somewhat inapt because hosts and commensals are not comparative interaction partners. Social examples of the same form would be street beggars asking for change and small countries of little wealth or natural resources joining international trade organizations.

Figure 1.7: Commensal

		player2	
		A	B
player1	A	1,1	4,2
	B	2,4	2,2

Depending on the interpretation of the situation we may consider the prosocial outcome as the dominance of B-types or a population of A-types being supported by B-types. Since the B-types are insensitive to the presence of A-types, I favor the fully mixed outcome as the prosocial one. The efficient social outcome is for each B-type to have K A-type neighbors and each A-type to have K B-type neighbors. This is the same "checkerboard" bipartite arrangement as in the Lichen game. However, since the B-types will not disconnect any neighbors we do not expect Commensal to achieve the outcome as well as Lichen. This difference parallels the efficiency gap between Coordination Game and Stag Hunt, the latter of which also has indifferent B-types.

Without learning or population dynamics we may see some A-types isolated as a group without B-types, this happens as a result of B-types attached to other B-types using up their maximum degree allotment. With population dynamics we should see the population distribution converge to the carrying capacity of A-types. The optimal figure is one-half the population, but in the optimal arrangement A-types are doing better than B-types making that arrangement unstable. A-types attached to other A-types are doing the worst, so with the best and worst agents being A-types the population could equilibrate.

Learning in the Commensal game will produce similar system behavior as the Lichen game. B-types supporting a population of A-types will imitate them. This makes that agent undesirable as a neighbor, disconnected, and now another A-type looking for B-types. Such a wandering A-type has few connections and they are often temporary connections with other A-types. Once such an agent finds a B-type, that B-type is likely doing better than itself so the focal agent will change back to B-type completing the cycle. This turbulent behavior will continue for all agents and only generate a periodic attractor or stable arrangement with minuscule probability.

1.8.8 Matching Pennies

Matching Pennies is a competitive constant-sum game. One can only do well in this game at the cost of the other player doing poorly. It is not obvious that prosociality applies to this kind of situation – strategic contexts like this are by their nature antisocial. Perhaps for this reason we do not see Matching Pennies appear in the evolution of prosociality literature, despite its importance in game theory. However, since it is clear that many social situations have the strategic form of competitive games, and it is clear that there are social norms and moral considerations covering competitive situations, Matching Pennies is an important game to study.

Figure 1.8: Matching Pennies

		player2	
		A	B
player1	A	3,1	1,3
	B	1,3	3,1

As just mentioned, it is not prima facie clear what counts as a prosocial outcome in Matching Pennies. Since in each pairwise match there is always a winner and a loser, we may decide that the best outcome is for agents to periodically switch places; sometimes win and sometimes lose and balance time in both positions. Now recall that the implementation of games used here evaluates agent preferences/payoffs as both player1 and player2 across each interaction. What makes the game competitive is that player1 is a winner if and only if player2 is a loser, but since in every case an agent stands as both player1 and player2, every agent is always both a winner and loser. And because the game is constant-sum the two payoffs cancel each other out. So in this implementation the agents of both types are indifferent among all configurations with the same total k.

Because all agents are indifferent to all pairwise interactions, the outcomes will only vary through the stochastic attachment process. Because the the game payoffs themselves are fair and balanced there is no work for detachment to do. Such an outcome meshes well with a fairness consideration for prosociality. So to the limit that competitive situations can be prosocial, preferential detachment achieves this prosocial outcome. Learning will be irrelevant once all agents have reached the full capacity of neighbors, and it will produce only small perturbations to the equal distribution of types during the transient connection period. Population dynamics also has no differences in payoffs to work with, so the only variation will be in the random selection of agents and possibly the lock-in of differences forged during the attachment process.

Examined through one lens, the indifferences in this strategic context seem to reflect a failure of this implementation to represent the competitive nature of the game. A more appropriate interpretation is that as long as a situation satisfies the assumptions of preferential detachment theory, including that agents engage in relationships bilaterally, then the representation is faithful to the situation and *balanced competition* produces the same social behavior as indifference. In future work we can consider non-symmetric interactions, and we can also consider different, non-balanced payoff structures matching the preference rankings.

1.8.9 Lane Choice

I created this game to address a particular social situation and it has no precedent in the literature. It is grounded in common game structures; specifically, it is a combination of the Prisoners' Dilemma and the Coordination Game that I constructed in the following way: Considering just columns and rows A and B, that 2×2 subgame is a Prisoners' Dilemma. Considering just B and C that 2×2 subgame is also a Prisoners' Dilemma. Considering just A and C that 2×2 subgame is a Coordination game. So the unique rational outcome (Nash equilibrium) is (B, B) – defection in both Prisoners' Dilemma subgames (and though both players are playing the same action, the payoff indicates no gains from coordination). In the Lane Choice strategic context not only do agents have to succeed in cooperating despite the payoff benefit of defecting, they also have to coordinate on which cooperative outcome to settle on. This combination of coordinative and cooperative features makes it what I call a Collaborative Game.

The story accompanying the game is that there are three conventions a society can have with respect to choosing lanes to drive in: left-hand side, right-hand side, or free-for-all driving. Free-for-all driving (the defector action) is more convenient when faced with drivers on either side, but does poorly against similarly unpredictable drivers. Driving reliably on either side is equally beneficial as long as everybody agrees on which side. The story is not a particularly great match, but will suffice for our purposes here. Examples of coordinating on a cooperative outcome abound: choosing textbooks for courses, deciding on which public works project to spend funds on, creating a consistent nationwide policy through state-by-state referendums.

Figure 1.9: Lane Choice

		player2		
		A	B	C
player1	A	3.3	1,4	1,1
	B	4,1	2,2	4,1
	C	1,1	1,4	3,3

$Aabc : a+b+c = 0 \prec Aabc : a = 0,\ b+c = 1 \prec Aabc : a = 0,\ b+c = 2 \prec Aabc : a = 0,\ b+c = 3 \prec$
$Aabc : a = 0,\ b+c = 4 \prec Aabc : a = 1,\ b+c = 3 \prec Aabc : a = 1,\ b+c = 2 \prec Aabc : a = 1,\ b+c = 1 \prec$
$Aabc : a = 1,\ b+c = 0 \prec Aabc : a = 2,\ b+c = 2 \prec Aabc : a = 2,\ b+c = 1 \prec Aabc : a = 2,\ b+c = 0 \prec$
$Aabc : a = 3,\ b+c = 1 \prec Aabc : a = 3,\ b+c = 0 \prec Aabc : a = 4,\ b+c = 0$
$Babc : a+b+c = 0 \prec Babc : a+c = 0,\ b = 1 \prec Babc : a+c = 0,\ b = 2 \prec Babc : a+c = 0,\ b = 3 \prec$
$Babc : a+c = 0,\ b = 4 \prec Babc : a+c = 1,\ b = 3 \prec Babc : a+c = 1,\ b = 2 \prec Babc : a+c = 1,\ b = 1 \prec$
$Babc : a+c = 1,\ b = 0 \prec Babc : a+c = 2,\ b = 2 \prec Babc : a+c = 2,\ b = 1 \prec Babc : a+c = 2,\ b = 0 \prec$
$Babc : a+c = 3,\ b = 1 \prec Babc : a+c = 3,\ b = 0 \prec Babc : a+c = 4,\ b = 0$
$Cabc : a+b+c = 0 \prec Cabc : c = 0,\ a+b = 1 \prec Cabc : c = 0,\ a+b = 2 \prec Cabc : c = 0,\ a+b = 3 \prec$
$Cabc : c = 0,\ a+b = 4 \prec Cabc : c = 1,\ a+b = 3 \prec Cabc : c = 1,\ a+b = 2 \prec Cabc : c = 1,\ a+b = 1 \prec$
$Cabc : c = 1,\ a+b = 0 \prec Cabc : c = 2,\ a+b = 2 \prec Cabc : c = 2,\ a+b = 1 \prec Cabc : c = 2,\ a+b = 0 \prec$
$Cabc : c = 3,\ a+b = 1 \prec Cabc : c = 3,\ a+b = 0 \prec Cabc : c = 4,\ a+b = 0$

Table 1.9: Agents' Revealed Preferences in the Lane Choice Game

Each of the three types constitutes an equal share of the initial population. In the experiments without learning or population dynamics the population is constant so the prosocial outcome is assortatively mixed agents. All that is really important is that both A-types and C-types form coordinated, cooperative groups; the B-types will therefore be forced into their own group isolated from others.

When learning and/or population dynamics are involved the prosocial outcome is that A-types and C-types together dominate the population. Because the A-C coordination component is symmetric there should be no significant difference in behavior for these groups except those from stochastic variation. As in the coordination game, for some situations we may wish that one of the two types dominate the population (for example, applied to drivers within a region), but preferential detachment will not produce this outcome in this unbiased version.

Agents get stuck in suboptimal arrangements when all their K neighbors are of the same, less preferred type. In the Lane Choice game A-types are indifferent between

B-types and C-types, B-types are indifferent between A-types and C-types, and C-types are indifferent between A-type and B-types. What all that means is that the potential for stable heterogeneous clusters is much greater in the Lane Choice game compared with either of the games it was built from. There are just more ways that (say) A-types can get stuck with non-A-types. This feature reflects the Lane Choice being a more difficult social situation to "solve" than either of its constituent games.

1.8.10 Biased Lane Choice

In the Lane Choice game both collaborative outcomes were equally valuable. In the Biased Lane Choice game the (C, C) combination is ranked higher than the (A, A) outcome. A situation would have this form if two social conventions were desirable, but homing in on one of the two yields greater benefits. Technology adoption often has this form. We can imagine that (A, A) is VHS technology and (C, C) is Beta video technology. Beta was intrinsically a better standard, but VHS won the race because of network externalities and other social dynamics. As part of forming better institutions to achieve the best social outcome, understanding how to achieve the more beneficial social outcome could have important policy implications.

Figure 1.10: Biased Lane Choice

		player2		
		A	B	C
player1	A	3.3	1,5	1,1
	B	5,1	2,2	5,1
	C	1,1	1,5	4,4

As is easy to determine from the payoff matrices, the revealed preferences rankings here are identical to the unbiased Lane Choice game just presented. Prosociality in the pure preferential detachment case is again defined as assortative mixing for each of the three types; and again we only really care about the A- and C-types.

With learning and/or population dynamics the prosocial outcome is for the higher-payoff outcome of C-type domination occur. Since population dynamics is a global, mechanism achieving assortative mixing suffices for the eventual overtaking of C-types over A-types. Learning, however, is local behavior. So if the agents quickly form type-specific groups through preferential detachment, then the A-type group will remain stable as an isolated group. Before assortativity takes hold, the improved performance of C-types should make them significantly more prevalent in the population even if they don't manage to dominate. With either extension the defectors will perform worse than the other two types, but with learning, they can still persist in isolated groups.

This biased game therefore answers a point brought up in the Coordination and Lane Choice games. If one prefers the domination of one type over an even split of two coordinated outcomes, then one can simply reward the desired outcome a bit more and preferential detachment with population dynamics (but not learning) will converge the population to the superior outcome. When the pressure of preferential detachment is greater than that of population dynamics, agents segregate before significant changes to the numbers of each type occur. Then, because of variations in fitness among groups of A-, B-, and C-types, we will see that first C-types displace B-types and then the A-types until only C-types remain. This (I will argue in section 4.1.3.2) is an example of group selection based solely on processes operating at the agent level.

1.9 Explaining Prosocial Outcomes

Before summarizing the theory of preferential detachment and moving on to the formal modeling chapters, I will now put forth some remarks regarding the theory as an explanation of observed prosocial behavior. Mechanisms such as punishment, mutualism, reciprocity, and altruism are described using terms having familiar meanings; thick terms that carry positive or negative evaluations. The outcomes are also described in evaluatively thick terms: cooperation, defection, coordination, etc. Any formal model must specify a formal meaning for these terms, some of which can be described purely in terms of reproductive fitness [Bekoff and Pierce 2009]. But some formal models are explicitly modeling (say) punishment in the thick evaluative sense of everyday language [Axelrod 1984].

I have made efforts to describe the model in purely behavioristic language, though I also adopted the terminology of other models when comparing them preferential detachment and providing examples. Preferential detachment is presented as a set of behaviors contingent upon a limited set of features possessed by objects as simple as atoms and as sophisticated as humans and nations. The possibility of explaining these evaluatively loaded prosocial phenomena with a purely behavioristic process broaches issues regarding the naturalistic fallacy which will be addressed in chapter IV. Aside from these philosophical issues I address some points regarding the mechanism itself below.

1.9.1 Assortative Mixing

Assortative mixing describes any process that fosters the interaction of like-typed agents (whatever the types and agents are). It correlates with achieving the prosocial outcome in many strategic contexts: Prisoners' Dilemma, Hawk and Dove, Battle of the Sexes, Coordination Game, Stag Hunt, and both Lane Choice games. The success of many proposed mechanisms in achieving prosocial outcomes can be seen as different mechanisms capable of having the assortative mixing pattern emerge. However, their success is often limited to situations in which type-based matching suffices.

But assortative mixing will not work in all scenarios; for example, any problem requiring a diversity of behaviors to solve – including Lichen and Commensal. My motivation for analyzing games from all categories is to canvas the range of possibilities for social problems. We will see that preferential detachment can yield the prosocial behavior over time in each strategic context (with varying success also affected by learning and population dynamics). It is important to notice that (unlike homophily, mutualism, and reciprocity) preferential detachment achieves assortative mixing indirectly, and only in the appropriate strategic contexts. The same mechanism produces disassortative mixing in contexts with heterogeneous prosocial outcomes. The simplicity of the mechanism combined with the breadth of its effectiveness will be made use of in chapter IV and is the main contribution of the project.

1.9.2 Games of Life

As discussed in the introduction of section 1.8, there is already literature discussing the appropriateness of one game versus another to represent the social situation faced by humans and protohumans and leading to the evolution of social norms [Boyd 1988, Binmore 1998, Skyrms 2004]. Preferential detachment theory is presented using only one of the games from the game library at a time. This parallels most previous research (although see [Bednar and Page 2007, Feigel 2003]), and it is a valuable starting point. In future work I will explore the effects of both changing and mixing strategic contexts. This heterogeneity can be used to model fluctuating environments, migration, multi-step problems, and the manifold of situations that favor more complicated prosocial behaviors. Though there is still a good deal of work to be done in understanding one-at-a-time models, I believe that a satisfactory explanation for the evolution of prosociality requires addressing these mixed-context situations. If preferential detachment can also solve these more difficult social problems then this enhances the explanatory power of the mechanism.

1.9.3 Why Preferential Detachment Works

Preferential detachment shares with Tit-for-Tat the "combination of being nice, retaliatory, forgiving and clear" [Axelrod 1984], but in preferential detachment the payoffs of the game determine which behavior counts as nice, retaliatory, and forgiving. And, as mentioned, these are thick terms that conceptually do not apply to behavior in all situations; forgiveness is not meaningful in a coordination context. Details of this conceptual gap will be clarified in chapter IV as part of the discussion considering preferential detachment as an explanation for the evolution of morality.

When games are described using their problem-specific interpretations (e.g., cooperation, coordination, contribution), the different strategic contexts seem to require distinct behaviors to achieve the prosocial outcome. Because behavior through preferential detachment is contingent on an agent's configuration, and which types of neighbors are de facto better for the agents in a situation, it allows agents to be adaptive to whatever situation the agents find themselves in – a meta-rule. The *behavior* isn't different, but the context makes it seem so. The unification of prosocial behaviors under one mechanism stands as a significant contribution to the field, though constrained within the dynamic social network framework.

1.10 Summary of Preferential Detachment

Preferential detachment is a simple mechanism embedded in a theory of how agents (broadly defined) can alter their interaction partners to form aggregates that benefit constituent members: prosociality. We start with a population of agents typed by their action in a game that represents the benefit derived from pairwise agent interactions. Based on an agent's configuration it will choose to either detach an undesirable neighbor, randomly attach a new neighbor, or do nothing (B1-3 of 1.1.1). The agents use only information regarding their relative preferences over their current neighbors to make their behavior decision (C1-3 of 1.1.1). Through multiple iterations of just these processes agents form aggregates that match prosocial arrangements for every situation representable as a 2×2 game.

One implication is that achieving the prosocial outcome may be easier than previous research indicates. When considering the evolution of culture, morality, language, and other social features we may reasonably consider the earliest development to have taken place before advanced cognitive capabilities were available. No such requirements exist for preferential detachment; such processes occur even for very simple creatures and certain molecules. We can therefore interpret results here as general features of self-organizing systems, but our goal of relating to morality will continue to focus the interpretation of the formal models presented in the next two chapters.

A summary of the benefits of preferential detachment as an explanation for the origin and persistence of prosocial behaviors includes:

Benefit1 It requires less of the capacities and capabilities of the agents

Benefit2 It generates the prosocial outcomes.

Benefit3 It does so in a great variety of strategic contexts.

Benefit4 It is robust with respect to population size and social network conditions.

These benefits will be demonstrated through a mathematical model utilizing a Markov model representation of agent configuration dynamics (chapter II) and a dynamic network agent-based model (chapter III). A breakdown of the games, their categories, each type's preferential behavior, and the prosocial outcome is summarized in table 1.10. These include extensions incorporating imitative learning and population dynamics into the agent-based model to make deeper points about the efficacy of preferential detachment and compare with existing models incorporating similar mechanisms.

		A-Type	B-Type	Prosocial Behavior			
Game	Category	ϕ	ϕ	Preferential Detachment	Learning	Population Dynamics	Both
Prisoners' Dilemma	Cooperative	$\phi_1 : B \prec A$	$\phi_1 : B \prec A$	assortative mixing	A-type dominance	A-type dominance	A-type dominance
Hawk and Dove	Cooperative	$\phi_1 : B \prec A$	$\phi_1 : B \prec A$	assortative mixing	A-type dominance	A-type dominance	A-type dominance
Battle of the Sexes	Coordinative	$\phi_1 : B \prec A$	$\phi_2 : A \prec B$	assortative mixing	assortative mixing	assortative mixing	assortative mixing
Coordination Game	Coordinative	$\phi_1 : B \prec A$	$\phi_2 : A \prec B$	assortative mixing	assortative mixing	assortative mixing	assortative mixing
Lichen	Specialized	$\phi_2 : A \prec B$	$\phi_1 : B \prec A$	disassortative mixing	disassortative mixing	disassortative mixing	disassortative mixing
Stag Hunt	Contributive	$\phi_1 : B \prec A$	$\phi_3 : B \approx A$	assortative mixing	A-type dominance	A-type dominance	A-type dominance
Commensal	Commensal	$\phi_2 : A \prec B$	$\phi_3 : B \approx A$	disassortative mixing	disassortative mixing	disassortative mixing	disassortative mixing
Matching Pennies	Undifferentiated	$\phi_3 : B \approx A$	$\phi_3 : B \approx A$	anything	anything	anything	anything
Lane Choice	NA	NA	NA	assortative mixing	assortative A and C	assortative A and C	assortative A and C
Biased Lane Choice	NA	NA	NA	assortative mixing	C-type dominance	C-type dominance	C-type dominance

Table 1.10: Summary of Game Types and Prosocial Outcomes

CHAPTER II

Markov Model of Preferential Detachment

The first approach to formally representing the preferential detachment mechanism is with a time-homogeneous Markov model. This approach assumes that the agent characteristics at time $t+1$ can be determined using information contained in the agent characteristics at time t. Rather than modeling each agent explicitly, I collect the agents into bins by configuration. Recall from section 1.6 that the configuration includes the agent's type and how many of each type of neighbor it has. The configuration thus captures all the information that agent behavior is contingent upon (C1-3 of 1.1.1).

In this exposition I will analyze only 2×2 games with a maximum degree K; the examples set $K = 4$ for calculation purposes. With two types and $K = 4$ there are the following fifteen distinct configurations for each type of agent (see 1.6.4 for how that number is determined).

$$\theta 00,\ \theta 01,\ \theta 02,\ \theta 03,\ \theta 04,\ \theta 10,\ \theta 11,\ \theta 12,\ \theta 13,\ \theta 20,\ \theta 21,\ \theta 22,\ \theta 30,\ \theta 31,\ \theta 40.$$

Specifying a game payoff matrix determines what action(s) an agent of each configuration would take. These action tendencies are captured by the revealed preference ranking over configurations produced by the preferential detachment ruleset. The three possible rankings that an agent may have for $T = 2$ and $K = 4$ are presented in table 2.1.

$\phi_1 : B \prec A :$	$\theta 00 \prec \theta 01 \prec \theta 02 \prec \theta 03 \prec \theta 04 \prec \theta 13 \prec \theta 12 \prec \theta 11 \prec \theta 10 \prec \theta 22 \prec \theta 21 \prec \theta 20 \prec \theta 31 \prec \theta 30 \prec \theta 40$
$\phi_2 : A \prec B :$	$\theta 00 \prec \theta 10 \prec \theta 20 \prec \theta 30 \prec \theta 40 \prec \theta 31 \prec \theta 21 \prec \theta 11 \prec \theta 01 \prec \theta 22 \prec \theta 12 \prec \theta 02 \prec \theta 13 \prec \theta 03 \prec \theta 04$
$\phi_3 : B \approx A :$	$\theta 00 \prec \theta 10 \approx \theta 01 \prec \theta 20 \approx \theta 11 \approx \theta 02 \prec \theta 30 \approx \theta 21 \approx \theta 12 \approx \theta 03 \prec \theta 40 \approx \theta 31 \approx \theta 22 \approx \theta 13 \approx \theta 04$

Table 2.1: The Three Possible Revealed Preference Relations for K=4

Those revealed preferences only indicate which configuration an agent would choose given a set of possibly reachable configurations – its opportunity set. It fails to specify what the opportunity set from each configuration is. It also fails to include what other agents are likely to do to it. For example, in the Prisoners' Dilemma a defector will not disconnect a cooperator, but that cooperator neighbor detaching the focal defector will affect that defector's configuration. Random attachments by other agents also change an agent's configurations independently of that agent's choice of action.

With this in mind I have separated the transitions into three independent behaviors: attachment and detachment ($\mathcal{X}$), being randomly connected to ($\mathcal{R}$), and being detached from ($\mathcal{Y}$). Altogether the resulting transition probabilities include both the action taken by the agents in each configuration and the likely actions taken by other agents with respect to them. Each of these components is described in detail below.

The Markov model is not tracking any system-wide features directly, just the distribution of configurations of the embedded agents. Because of this feature of the model we will have to make some system-level assumptions that are known to be false in general, but are still reasonable approximations. These assumptions are necessary because the Markov model is time-homogeneous, yet the actual probabilities of transitions *do* change in response to the numbers of agents in each configuration. These assumptions reduce the numerical accuracy of the models presented here, but the simplicity and clarity of the mathematical representation (and its qualitative similarity to the agent-based model) has other benefits that make it still useful.

2.1 Random Connections as Bernoulli Trials

This section accounts for agents *receiving* connections from other agents; new connections formed via an agent's own action are covered in section 2.3. Given some focal agent, the probability of having an agent randomly connect to it has the form of a Bernoulli trial. For each of the other agents the probability that it will connect to specific agent is equal to $\frac{1}{N-1}$ (uniformly at random). We can refine this by dividing the total population evenly by type. Let N_A be the number of agents of A-type agents and similarly for N_B. If our focal agent is A-type, then there are N_B B-types and $N_A - 1$ A-types that may try to connect. To represent the feature that agents cannot connect to themselves (and hence must be removed from the pool of possible connectors) I introduce the η function.

$$\eta_A = \begin{cases} N_A - 1 & \text{if focal agent is } A\text{-type} \\ N_A & \text{if focal agent is } B\text{-type} \end{cases}$$
$$\eta_B = \begin{cases} N_B & \text{if focal agent is } A\text{-type} \\ N_B - 1 & \text{if focal agent is } B\text{-type} \end{cases} \tag{2.1.1}$$

So for an A-type focal agent you can think of this as each of the N_B B-type agents and each of the $N_A - 1$ A-type agents rolling a fair die with $N - 1$ sides. An agent connects to the focal agent if the die lands on the number 1. Using this approach we can determine the probability of an agent receiving a connections from A-types and b connections from B-types using the common binomial probability function

$$\rho(a, b) = \binom{\eta_A}{a} \left(\frac{1}{N-1}\right)^{a} \left(1 - \frac{1}{N-1}\right)^{\eta_A - a} \cdot \binom{\eta_B}{b} \left(\frac{1}{N-1}\right)^{b} \left(1 - \frac{1}{N-1}\right)^{\eta_B - b}. \tag{2.1.2}$$

The probability of adding a particular a and b differs for A-types and B-types because of the difference between η_A and η_B, but there is the obvious symmetry that

$$\rho_A(a, b) = \rho_B(b, a). \tag{2.1.3}$$

I present the transition matrix for $K = 4$ in table 2.2. The rows represent the agent configurations at time t and the columns are the states at time $t + 1$. This example is helpful for seeing the pattern produced by the random connection probabilities. Because agents cannot lose neighbors through random connections, it is possible to arrange the matrix (by listing the configurations lexicographically by a then b) to be upper triangular – a feature I will make use of later. Further analysis reveals a perfectly regular pattern to the Bernoulli transition probabilities. Given this configuration order the matrix is composed of smaller upper-diagonal block matrices. The number and size of these block matrices follows a pattern produced by the maximum degree assumption ($a + b \leq K$) that we will make use of in section 2.6.

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	$\rho(0,3)$	$\rho(0,4)$	$\rho(1,0)$	$\rho(1,1)$	$\rho(1,2)$	$\rho(1,3)$	$\rho(2,0)$	$\rho(2,1)$	$\rho(2,2)$	$\rho(3,0)$	$\rho(3,1)$	$\rho(4,0)$
$\theta 01$	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	$\rho(0,3)$	0	$\rho(1,0)$	$\rho(1,1)$	$\rho(1,2)$	0	$\rho(2,0)$	$\rho(2,1)$	0	$\rho(3,0)$	0
$\theta 02$	0	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	0	0	$\rho(1,0)$	$\rho(1,1)$	0	0	$\rho(2,0)$	0	0	0
$\theta 03$	0	0	0	$\rho(0,0)$	$\rho(0,1)$	0	0	0	$\rho(1,0)$	0	0	0	0	0	0
$\theta 04$	0	0	0	0	$\rho(0,0)$	0	0	0	0	0	0	0	0	0	0
$\theta 10$	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	$\rho(0,3)$	$\rho(1,0)$	$\rho(1,1)$	$\rho(1,2)$	$\rho(2,0)$	$\rho(2,1)$	$\rho(3,0)$
$\theta 11$	0	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	0	$\rho(1,0)$	$\rho(1,1)$	0	$\rho(2,0)$	0
$\theta 12$	0	0	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	0	0	$\rho(1,0)$	0	0	0
$\theta 13$	0	0	0	0	0	0	0	0	$\rho(0,0)$	0	0	0	0	0	0
$\theta 20$	0	0	0	0	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(0,2)$	$\rho(1,0)$	$\rho(1,1)$	$\rho(2,0)$
$\theta 21$	0	0	0	0	0	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	0	$\rho(1,0)$	0
$\theta 22$	0	0	0	0	0	0	0	0	0	0	0	$\rho(0,0)$	0	0	0
$\theta 30$	0	0	0	0	0	0	0	0	0	0	0	0	$\rho(0,0)$	$\rho(0,1)$	$\rho(1,0)$
$\theta 31$	0	0	0	0	0	0	0	0	0	0	0	0	0	$\rho(0,0)$	0
$\theta 40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\rho(0,0)$

Table 2.2: Structure of Bernoulli Probabilities of Random Connections for $K = 4$

According to preferential detachment theory an agent will not connect to new agents if that agent already has the maximum number of connections. To make the dynamics more accurate we would need to change the probabilities of receiving random connections based on the numbers of agents in configurations such at $a+b=K$. Doing so in a Markov model is overly cumbersome and accounting for these details is left for the agent-based model. It is important to note that because of the time-homogeneous probabilities the dynamics here systematically overestimate the random connection rate.[1]

Within the confines of the Markov model there is still an adjustment to make for the attempted addition of connections that would yield configurations with $k > K$. Because the binomial probabilities extend to cases in which too many agents attempt to connect there is an "edge effect" at the configurations with $a + b = K$. The probability density for *all* cases with $a + b > K$ is distributed evenly over the cases with $a+b=K$. This reflects the idea that if more than the maximum possible edges would have formed, then the agent would at least reach a configuration in which the maximum is reached, and all ways of doing this are equally probable.

To make this adjustment for the transition probabilities $P(\theta ab \to \theta a'b')$ in the complete random connection matrix $\mathcal{R}$, I utilize the fact that the probabilities across rows of a Markov model always sum to one. First we take one minus the sum of the Bernoulli values $\rho(\cdot,\cdot)$ across a row; this is the leftover probability mass. Then divide this figure by the number of cases in that row where the *column's* $a' + b' = K$. The number of reachable saturated configurations is the size of the set $\{\theta a'b' : a' \geq a, b' \geq b, a' + b' = K\}$ which equals $K - a - b + 1$ for the row with a and b. We call the weighted leftover probability mass $\Gamma(\theta ab)$ and calculate it as

$$\Gamma(\theta ab) = \frac{1 - \sum_{i=a,j=b}^{i+j=K} \rho(i-a, j-b)}{K-a-b+1} \tag{2.1.4}$$

Finally, we add this weighted leftover probability mass to each entry in the row matching columns having maximal capacity. Combining the Bernoulli probabilities with the gamma mass in the appropriate places we can populate the complete random

[1] In the agent-based model of preferential detachment an agent will **not** attempt a random connection if the agent has already detached a neighbor during an iteration. The "one move per turn" restriction is not a core feature of the theory, but I make this note to point out that the Markov model further overestimates the random connection probability compared to the agent-based model because of this.

connection matrix $\mathcal{R}$ for any K, N_A, and N_B with the following calculations:

$$P_{\mathcal{R}}(\theta ab \rightarrow \theta a'b') = \begin{cases} \rho(a'-a, b'-b) & \text{if } a'+b' < K \\ \rho(a'-a, b'-b) + \Gamma(\theta ab) & \text{if } a'+b' = K \end{cases} \tag{2.1.5}$$

2.1.1 Example of Random Connection Probabilities

The probabilities are sensitive to the population of each type and the maximum degree, but not to the strategic context. Our running example continues with the parameters $N_A = 100$, $N_B = 100$, and $K = 4$. The resulting random connection transition matrices for A-types and B-types, referred to as $\mathcal{R}_A$ and $\mathcal{R}_B$, are presented in tables 2.3 and 2.4 respectively.

	$A00$	$A01$	$A02$	$A03$	$A04$	$A10$	$A11$	$A12$	$A13$	$A20$	$A21$	$A22$	$A30$	$A31$	$A40$
$A00$	0.366953	0.18533	0.0463325	0.0076440	0.0016451	0.183477	0.0926649	0.0231662	0.0045310	0.0454058	0.0229322	0.0064420	0.0074147	0.0044538	0.00160774
$A01$	0	0.366953	0.18533	0.0463325	0.012314	0	0.183477	0.0926649	0.0278362	0	0.0454058	0.0276022	0	0.0120847	0
$A02$	0	0	0.366953	0.18533	0.0729448	0	0	0.183477	0.119277	0	0	0.0720182	0	0	0
$A03$	0	0	0	0.366953	0.31745	0	0	0	0.315597	0	0	0	0	0	0
$A04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$A10$	0	0	0	0	0	0.366953	0.18533	0.0463325	0.012314	0.183477	0.0926649	0.0278362	0.0454058	0.0276022	0.0120847
$A11$	0	0	0	0	0	0	0.366953	0.18533	0.0729448	0	0.183477	0.119277	0	0.0720182	0
$A12$	0	0	0	0	0	0	0	0.366953	0.31745	0	0	0.315597	0	0	0
$A13$	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0
$A20$	0	0	0	0	0	0	0	0	0	0.366953	0.18533	0.0729448	0.183477	0.119277	0.0720182
$A21$	0	0	0	0	0	0	0	0	0	0	0.366953	0.31745	0	0.315597	0
$A22$	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0
$A30$	0	0	0	0	0	0	0	0	0	0	0	0	0.366953	0.31745	0.315597
$A31$	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0
$A40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.3: $\mathcal{R}_A$: A-Type Random Connection Transition Probabilities for $N_A = N_B = 100$, $K = 4$

	$B00$	$B01$	$B02$	$B03$	$B04$	$B10$	$B11$	$B12$	$B13$	$B20$	$B21$	$B22$	$B30$	$B31$	$B40$
$B00$	0.366953	0.183477	0.0454058	0.0074147	0.0016077	0.18533	0.0926649	0.0229322	0.0044538	0.0463325	0.0231662	0.0064420	0.0076440	0.0045310	0.00164519
$B01$	0	0.366953	0.183477	0.0454058	0.0120847	0	0.18533	0.0926649	0.0276022	0	0.0463325	0.0278362	0	0.012314	0
$B02$	0	0	0.366953	0.183477	0.0720182	0	0	0.18533	0.119277	0	0	0.0729448	0	0	0
$B03$	0	0	0	0.366953	0.315597	0	0	0	0.31745	0	0	0	0	0	0
$B04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$B10$	0	0	0	0	0	0.366953	0.183477	0.0454058	0.0120847	0.18533	0.0926649	0.0276022	0.0463325	0.0278362	0.012314
$B11$	0	0	0	0	0	0	0.366953	0.183477	0.0720182	0	0.18533	0.119277	0	0.0729448	0
$B12$	0	0	0	0	0	0	0	0.366953	0.315597	0	0	0.31745	0	0	0
$B13$	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0
$B20$	0	0	0	0	0	0	0	0	0	0.366953	0.183477	0.0720182	0.18533	0.119277	0.0729448
$B21$	0	0	0	0	0	0	0	0	0	0	0.366953	0.315597	0	0.31745	0
$B22$	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0
$B30$	0	0	0	0	0	0	0	0	0	0	0	0	0.366953	0.315597	0.31745
$B31$	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0
$B40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.4: $\mathcal{R}_B$: B-Type Random Connection Transition Probabilities for $N_A = N_B = 100$, $K = 4$

2.2 Probabilities of Detachment by Other Agents

Preferential detachment theory states that when an agent has interaction partners of differing types, and one type is preferred less than others, a least preferred agent will be detached. This section derives the transition probabilities resulting from *being detached by* other agents. Like the random connection probabilities, this is something that happens to an agent rather than something that an agent does. Whether agents of a type are detached or not depends on the preferences of *both types* of agent, and the probability depends on an agent's configuration, the number of each type, and indirectly on the maximum degree.

The detachment matrices for each type are called $\mathcal{Y}_A$ and $\mathcal{Y}_B$, and these depend on the ϕ relation of both types. There are six distinct ways these ϕ relations can be arranged corresponding to the six forms of games presented in section 1.3. There are, however, only four distinct $\mathcal{Y}$ structures: 1) both types detach the focal types, 2) only same-type agents detach, 3) only other-type agents detach, 4) or neither type of agent detaches.

Recall again that lesser preferred agents are only detached if an agent currently has a more preferred neighbor. However, in this Markov model we do not have access to who else an agent's neighbors are connected to (i.e., a focal agent's neighbors' neighbors). So, for example, our focal agent may be a defector connected to a cooperator, but we can't know whether that cooperator has any cooperator neighbors. The Markov model is forced to make approximations for these probabilities because it cannot track system-level arrangements of the agent connections. I assume a uniform distribution over the possible configurations a neighbor may be in, though the distribution is unlikely to be uniform (and would be constantly changing).

This uniform approximation is used to determine, given an agent's configuration, the probability that each of its neighbors detaches it. Each neighbors' action is considered independent (a reasonable assumption for networks with low clustering coefficients). Using equation (1.6.1) we know that the total number of configurations, $\mathbb{C}$, for two types of agents is

$$\mathbb{C} = \frac{(2+K)!}{2(K!)} = \frac{(K+2)(K+1)}{2} = \frac{1}{2}(K^2 + 3K + 2).$$

For each type of agent, there are K configurations in which an agent has only that type of neighbor (e.g., $\theta 01, \theta 02, \ldots \theta 0K$) and only one way to have no neighbors. Using this breakdown there are four categories of configurations that we can separate:

Category	Symbol	# of configurations
attached to only A-types	$\mathbb{C}_A$	K
attached to only B-types	$\mathbb{C}_B$	K
attached to both A- and B-types	$\mathbb{C}_M$	$\mathbb{C} - 2K - 1$
attached to none	$\mathbb{C}_0$	1

Table 2.5: Categories of Configurations

Using this information we can determine the likelihood of an agent being detached depending on its configuration and the preferences of both agent types. Note that the appropriate probability to use depends on the likely configuration of each *neighbor*. The neighbors' possible category is further constrained by the focal agent's connection; for an A-type agent and its B-type neighbor we know that the B-type neighbor has at least the one A-type neighbor. If $B : A \prec B$ then the probability of detachment by that B-type is $\frac{\mathbb{C}_M}{\mathbb{C}-K-1} = \frac{\mathbb{C}-2K-1}{\mathbb{C}-K-1} = \frac{1}{2}(K^2 - K)$. The denominator $\mathbb{C} - K - 1$ accounts for the fact that out of the total number of possible configurations $\mathbb{C}$, there are K with no connections to the focal type and 1 configuration with no connections at all. I will use the following shorthand for this denominator: $\mathbb{D} = \mathbb{C} - K - 1 = \frac{1}{2}(K^2 + K)$.

The probabilities filling the $\mathcal{Y}$ matrices can be calculated using the following equations depending on the revealed preferences of agents in game form. For $a' \leq a$ and $b' \leq b$ the detachment probabilities are:

$$P_{\mathcal{Y}_A}(\theta ab \to \theta a'b') =$$
$$\begin{cases} 0^{(a-a')+(b-b')} & \text{if } A:\phi_1 \text{ or } \phi_3 \quad B:\phi_1 \text{ or } \phi_3 \\ \binom{a}{a'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(a-a')}\left(\frac{\mathbb{C}_A}{\mathbb{D}}\right)^{a'} & \text{if } A:\phi_2 \qquad\qquad B:\phi_1 \text{ or } \phi_3 \\ \binom{b}{b'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(b-b')}\left(\frac{\mathbb{C}_A}{\mathbb{D}}\right)^{b'} & \text{if } A:\phi_1 \text{ or } \phi_3 \quad B:\phi_2 \\ \binom{a}{a'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(a-a')}\left(\frac{\mathbb{C}_A}{\mathbb{D}}\right)^{a'}\binom{b}{b'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(b-b')}\left(\frac{\mathbb{C}_A}{\mathbb{D}}\right)^{b'} & \text{if } A:\phi_2 \qquad\qquad B:\phi_2 \end{cases}$$

$$P_{\mathcal{Y}_B}(\theta ab \to \theta a'b') =$$
$$\begin{cases} 0^{(a-a')+(b-b')} & \text{if } A:\phi_2 \text{ or } \phi_3 \quad B:\phi_2 \text{ or } \phi_3 \\ \binom{a}{a'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(a-a')}\left(\frac{\mathbb{C}_B}{\mathbb{D}}\right)^{a'} & \text{if } A:\phi_1 \qquad\qquad B:\phi_2 \text{ or } \phi_3 \\ \binom{b}{b'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(b-b')}\left(\frac{\mathbb{C}_B}{\mathbb{D}}\right)^{b'} & \text{if } A:\phi_2 \text{ or } \phi_3 \quad B:\phi_1 \\ \binom{a}{a'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(a-a')}\left(\frac{\mathbb{C}_B}{\mathbb{D}}\right)^{a'}\binom{b}{b'}\left(\frac{\mathbb{C}_M}{\mathbb{D}}\right)^{(b-b')}\left(\frac{\mathbb{C}_B}{\mathbb{D}}\right)^{b'} & \text{if } A:\phi_1 \qquad\qquad B:\phi_1 \end{cases} \tag{2.2.1}$$

The detachment probability equals 0 if either $a' > a$ or $b' > b$, or the matrix entry represents a loss of a kind of edge not being detached.

If neither type disconnects then $P_{\mathcal{Y}}(\theta ab \to \theta a'b') = 1$ whenever $a' = a$ and $b' = b$, and equals zero otherwise – this is the identity matrix. In the other cases the transition probability is calculated as the multiplication of the probability of losing an edge for each edge lost, times the multiplication of the probability of maintaining an edge for each edge maintained. For ease of calculation, and to foster intuitions regarding the effects of changes in K discussed later, we can translate these symbolic generating functions according to the following scheme:

$$\tfrac{\mathbb{C}_A}{\mathbb{D}} = \tfrac{2}{K+1}, \quad \tfrac{\mathbb{C}_B}{\mathbb{D}} = \tfrac{2}{K+1}, \quad \tfrac{\mathbb{C}_M}{\mathbb{D}} = \tfrac{K-1}{K+1} \tag{2.2.2}$$

Because agents cannot gain connections through detachment, this network has a lower-triangular structure. If both types are detaching from the focal type then $\mathcal{Y}$ also has a block submatrix pattern similar to the random connection matrix – and for the same reason. Every existing edge has an independent probability of being lost just as every non-existing edge had a probability of being gained. The structure of this matrix is therefore the transpose of $\mathcal{R}$, though the probabilities are not related.

As can be seen from the probabilities generated by (2.2.1), there are only four distinct $\mathcal{Y}$ matrices. The entries depend on whether A-types, B-types, both types, or neither type will disconnect from a focal type. We refer to these matrices as

Symbol	Category
$\mathcal{Y}\alpha$	A-types detach
$\mathcal{Y}\beta$	B-types detach
$\mathcal{Y}\delta$	both A- and B-types detach
$\mathcal{Y}0$	neither type detaches

Table 2.6: Categories Detachment Matrices

2.2.1 Example Detachment Probabilities

I will continue with an example calculation for $N_A = 100$, $N_B = 100$, and $K = 4$. From equation (1.6.1) we determine that $\mathbb{C} = 15$ and thus $\mathbb{D} = 10$. Let's consider a focal B-type agent connected to one A-type in the Prisoners' Dilemma. That A-type neighbor must be in one of the following configurations: $A01$, $A02$, $A03$, $A04$, $A11$, $A12$, $A13$, $A21$, $A22$, $A31$. In the first four of the ten configurations the B-type agent will not be disconnected because the A-type has no preferred neighbors. By the uniformity assumption the probability of losing that edge is $\frac{6}{10}$, and of maintaining it is $\frac{4}{10}$. This same reasoning is used for each neighbor to populate each of the $\mathcal{Y}$ matrices in tables 2.7, 2.8, and 2.9. As mentioned, $\mathcal{Y}0$ is the identity matrix for any

N_A, N_B, and K: a square matrix of the appropriate size (15×15 here) with 1 along the diagonal and 0 elsewhere.

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 01$	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 02$	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 03$	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0
$\theta 04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$\theta 10$	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$	0	0	0	0	0	0	0	0	0
$\theta 11$	0	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$	0	0	0	0	0	0	0	0
$\theta 12$	0	0	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$	0	0	0	0	0	0	0
$\theta 13$	0	0	0	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$	0	0	0	0	0	0
$\theta 20$	$\frac{9}{25}$	0	0	0	0	$\frac{12}{25}$	0	0	0	$\frac{4}{25}$	0	0	0	0	0
$\theta 21$	0	$\frac{9}{25}$	0	0	0	0	$\frac{12}{25}$	0	0	0	$\frac{4}{25}$	0	0	0	0
$\theta 22$	0	0	$\frac{9}{25}$	0	0	0	0	$\frac{12}{25}$	0	0	0	$\frac{4}{25}$	0	0	0
$\theta 30$	$\frac{27}{125}$	0	0	0	0	$\frac{54}{125}$	0	0	0	$\frac{36}{125}$	0	0	$\frac{8}{125}$	0	0
$\theta 31$	0	$\frac{27}{125}$	0	0	0	0	$\frac{54}{125}$	0	0	0	$\frac{36}{125}$	0	0	$\frac{8}{125}$	0
$\theta 40$	$\frac{81}{625}$	0	0	0	0	$\frac{216}{625}$	0	0	0	$\frac{216}{625}$	0	0	$\frac{96}{625}$	0	$\frac{16}{625}$

Table 2.7: $\mathcal{Y}\alpha$ Detachment Transition Probabilities for $K = 4$

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 01$	$\frac{3}{5}$	$\frac{2}{5}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 02$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 03$	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$	0	0	0	0	0	0	0	0	0	0	0
$\theta 04$	$\frac{81}{625}$	$\frac{216}{625}$	$\frac{216}{625}$	$\frac{96}{625}$	$\frac{16}{625}$	0	0	0	0	0	0	0	0	0	0
$\theta 10$	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0
$\theta 11$	0	0	0	0	0	$\frac{3}{5}$	$\frac{2}{5}$	0	0	0	0	0	0	0	0
$\theta 12$	0	0	0	0	0	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$	0	0	0	0	0	0	0
$\theta 13$	0	0	0	0	0	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$	0	0	0	0	0	0
$\theta 20$	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0
$\theta 21$	0	0	0	0	0	0	0	0	0	$\frac{3}{5}$	$\frac{2}{5}$	0	0	0	0
$\theta 22$	0	0	0	0	0	0	0	0	0	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$	0	0	0
$\theta 30$	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0
$\theta 31$	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{3}{5}$	$\frac{2}{5}$	0
$\theta 40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.8: $\mathcal{Y}\beta$ Detachment Transition Probabilities for $K = 4$

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 01$	$\frac{3}{5}$	$\frac{2}{5}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 02$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 03$	$\frac{27}{125}$	$\frac{54}{125}$	$\frac{36}{125}$	$\frac{8}{125}$	0	0	0	0	0	0	0	0	0	0	0
$\theta 04$	$\frac{81}{625}$	$\frac{216}{625}$	$\frac{216}{625}$	$\frac{96}{625}$	$\frac{16}{625}$	0	0	0	0	0	0	0	0	0	0
$\theta 10$	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$	0	0	0	0	0	0	0	0	0
$\theta 11$	$\frac{9}{25}$	$\frac{6}{25}$	0	0	0	$\frac{6}{25}$	$\frac{4}{25}$	0	0	0	0	0	0	0	0
$\theta 12$	$\frac{27}{125}$	$\frac{36}{125}$	$\frac{12}{125}$	0	0	$\frac{18}{125}$	$\frac{24}{125}$	$\frac{8}{125}$	0	0	0	0	0	0	0
$\theta 13$	$\frac{81}{625}$	$\frac{162}{625}$	$\frac{108}{625}$	$\frac{24}{625}$	0	$\frac{54}{625}$	$\frac{108}{625}$	$\frac{72}{625}$	$\frac{16}{625}$	0	0	0	0	0	0
$\theta 20$	$\frac{9}{25}$	0	0	0	0	$\frac{12}{25}$	0	0	0	$\frac{4}{25}$	0	0	0	0	0
$\theta 21$	$\frac{27}{125}$	$\frac{18}{125}$	0	0	0	$\frac{36}{125}$	$\frac{24}{125}$	0	0	$\frac{12}{125}$	$\frac{8}{125}$	0	0	0	0
$\theta 22$	$\frac{81}{625}$	$\frac{108}{625}$	$\frac{36}{625}$	0	0	$\frac{108}{625}$	$\frac{144}{625}$	$\frac{48}{625}$	0	$\frac{36}{625}$	$\frac{48}{625}$	$\frac{16}{625}$	0	0	0
$\theta 30$	$\frac{27}{125}$	0	0	0	0	$\frac{54}{125}$	0	0	0	$\frac{36}{125}$	0	0	$\frac{8}{125}$	0	0
$\theta 31$	$\frac{81}{625}$	$\frac{108}{625}$	0	0	0	$\frac{108}{625}$	$\frac{108}{625}$	0	0	$\frac{108}{625}$	$\frac{72}{625}$	0	$\frac{24}{625}$	$\frac{16}{625}$	0
$\theta 40$	$\frac{81}{625}$	0	0	0	0	$\frac{216}{625}$	0	0	0	$\frac{216}{625}$	0	0	$\frac{96}{625}$	0	$\frac{16}{625}$

Table 2.9: $\mathcal{Y}\delta$ Detachment Transition Probabilities for $K = 4$

2.3 Agent Action Probabilities by Strategic Context

So far we have seen the transition probabilities resulting from the population of agents randomly connecting to a focal agent and transitions resulting from the population of agents detaching a focal agent. We have left to account for transitions resulting from the focal agent's own behavior. To recap, the preferential detachment mechanism is that an agent disconnects from any one less preferred neighbor if one exists, and connects to a new agent at random if one does not. There are three possible preference relations an agent may have over agent types. These relations are generated by equations (1.6.2) for any K and presented in table 2.1 for $K = 4$.

Those revealed preferences over configurations do not include the necessary information about which configurations are in the *opportunity set* of agents in each configuration. This information is necessary to populate the entries of the agent action transition matrices $\mathcal{X}_{\phi_1}$, $\mathcal{X}_{\phi_2}$, and $\mathcal{X}_{\phi_3}$. As before ϕ_1 refers to $B \prec A$, ϕ_2 to $A \prec B$, and ϕ_3 to $B \approx A$.

To determine what configurations are accessible through preferential detachment consider that agents can only add or remove one connection per iteration. Only agents with mixed configurations will detach, and they will do so with probability one if either type is preferred. Agents with the maximum number of connections cannot connect to any more. And finally agents that will add a random connection have a uniform probability of doing so over the whole population. Recall from (2.1.1) that η_A is the number of A-types available for connection and equals $N_A - 1$ for A-types and it equals N_A for B-types to reflect that agents cannot connect to themselves. η_B is similarly determined. Using these behavior characteristics we can determine the transition probabilities for $\mathcal{X}$ through the following set of equations:

$$P_{\mathcal{X}_{\phi_1}}(\theta ab \rightarrow \theta a'b') = \begin{cases} \frac{\eta_A}{N-1} & \text{if } a+b<K, a=0 \text{ or } b=0, a'=a+1, b=b' \\ \frac{\eta_B}{N-1} & \text{if } a+b<K, a=0 \text{ or } b=0, b'=b+1, a=a' \\ 1 & \text{if } a>0, b>0, b'=b-1, a=a' \\ 1 & \text{if } a=K \text{ or } b=K, b=b', a=a' \\ 0 & \textit{otherwise} \end{cases}$$

$$P_{\mathcal{X}_{\phi_2}}(\theta ab \rightarrow \theta a'b') = \begin{cases} \frac{\eta_A}{N-1} & \text{if } a+b<K, a=0 \text{ or } b=0, a'=a+1, b=b' \\ \frac{\eta_B}{N-1} & \text{if } a+b<K, a=0 \text{ or } b=0, b'=b+1, a=a' \\ 1 & \text{if } a>0, b>0, a'=a-1, b=b' \\ 1 & \text{if } a=K \text{ or } b=K, b=b', a=a' \\ 0 & \textit{otherwise} \end{cases}$$

$$P_{\mathcal{X}_{\phi_3}}(\theta ab \rightarrow \theta a'b') = \begin{cases} \frac{\eta_A}{N-1} & \text{if } a+b<K, a'=a+1, b=b' \\ \frac{\eta_B}{N-1} & \text{if } a+b<K, b'=b+1, a=a' \\ 1 & \text{if } a+b=K, b=b', a=a' \\ 0 & \textit{otherwise} \end{cases} \tag{2.3.1}$$

2.3.1 Example Action Transition Probabilities

The transition probabilities in tables 2.10, 2.11, and 2.12 reflect the probabilities if $N_A = N_B = 100$ and $K = 4$. The values of η_A and η_B used in the generator functions depend on the type of agent: for A-type agents $\frac{\eta_A}{N-1} = \frac{99}{199} = 0.497487$ and $\frac{\eta_B}{N-1} = \frac{100}{199} = 0.502513$. For B-type agents it is just the reverse: $\frac{\eta_A}{N-1} = \frac{100}{199} = 0.502513$ and $\frac{\eta_B}{N-1} = \frac{99}{199} = 0.497487$. For even population splits, as $N \rightarrow \infty$ these values converge to 0.5 – a feature made use of in generalizing the results of the Markov model.

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0	0
$\theta 01$	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0
$\theta 02$	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0
$\theta 03$	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0
$\theta 04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$\theta 10$	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0
$\theta 11$	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0
$\theta 12$	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0
$\theta 13$	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0
$\theta 20$	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	$\frac{\eta_A}{N-1}$	0	0
$\theta 21$	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0
$\theta 22$	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0
$\theta 30$	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	$\frac{\eta_A}{N-1}$
$\theta 31$	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0
$\theta 40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.10: $\mathcal{X}_{\phi_1}$: Action Transition Probabilities if $B \prec A$

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0	0
$\theta 01$	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0
$\theta 02$	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0
$\theta 03$	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0
$\theta 04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$\theta 10$	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0
$\theta 11$	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 12$	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0
$\theta 13$	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0
$\theta 20$	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	$\frac{\eta_A}{N-1}$	0	0
$\theta 21$	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0
$\theta 22$	0	0	0	0	0	0	0	1.0	0	0	1.0	0	0	0	0
$\theta 30$	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	$\frac{\eta_A}{N-1}$
$\theta 31$	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0
$\theta 40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.11: $\mathcal{X}_{\phi_2}$: Action Transition Probabilities if $A \prec B$

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0	0
$\theta 01$	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0	0
$\theta 02$	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0	0
$\theta 03$	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0	0
$\theta 04$	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0
$\theta 10$	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0	0
$\theta 11$	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	$\frac{\eta_A}{N-1}$	0	0	0	0
$\theta 12$	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	0	$\frac{\eta_A}{N-1}$	0	0	0
$\theta 13$	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0
$\theta 20$	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	$\frac{\eta_A}{N-1}$	0	0
$\theta 21$	0	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	0	$\frac{\eta_A}{N-1}$	0
$\theta 22$	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0
$\theta 30$	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{\eta_B}{N-1}$	$\frac{\eta_A}{N-1}$
$\theta 31$	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0
$\theta 40$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Table 2.12: $\mathcal{X}_{\phi_3}$: Action Transition Probabilities if $B \approx A$

2.4 Combined Probabilities

Finally we combine the three steps above to get the complete transition probability matrices representing all of the dynamics of the preferential detachment mechanism. The behaviors and hence probabilities in the three matrices ($\mathcal{R}$, $\mathcal{Y}$, and $\mathcal{X}$) are considered independent. This is a simplification since the agents' actions captured in $\mathcal{X}$ are exactly those detachments and connections captured in $\mathcal{Y}$ and $\mathcal{R}$. What is true is that an agent's *decision* to detach or connect is independent of the new neighbors it receives or loses from the other agents' decisions, and that is the feature being preserved here.

To produce the combined transition probability matrices for each type $\mathcal{A}$ and $\mathcal{B}$ we take the matrix products $\mathcal{A} = \mathcal{X}_A \mathcal{Y}_A \mathcal{R}_A$ and $\mathcal{B} = \mathcal{X}_B \mathcal{Y}_B \mathcal{R}_B$ for the matching component matrices. The multiplication of the probabilities is appropriate because the three processes are being considered independent. $\mathcal{X}_B$ and $\mathcal{Y}_B$ depend on neighbor preferences, and the six game categories are individuated by the agents' neighbor preferences, so there are distinct matrix triplets for each game category. The matrices corresponding to each game category are presented in table 2.13

Category	Relationship		A-type	B-type
Cooperative	$A : B \prec A$	$B : B \prec A$	$\mathcal{A} = \mathcal{X}_{\phi_1} \mathcal{Y} 0 \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_1} \mathcal{Y} \delta \mathcal{R}_B$
Coordinative	$A : B \prec A$	$B : A \prec B$	$\mathcal{A} = \mathcal{X}_{\phi_1} \mathcal{Y} \beta \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_2} \mathcal{Y} \alpha \mathcal{R}_B$
Specialized	$A : A \prec B$	$B : B \prec A$	$\mathcal{A} = \mathcal{X}_{\phi_2} \mathcal{Y} \alpha \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_1} \mathcal{Y} \beta \mathcal{R}_B$
Contributive	$A : B \prec A$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_1} \mathcal{Y} 0 \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3} \mathcal{Y} \alpha \mathcal{R}_B$
Commensal	$A : A \prec B$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_2} \mathcal{Y} \alpha \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3} \mathcal{Y} 0 \mathcal{R}_B$
Undifferentiated	$A : A \approx B$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_3} \mathcal{Y} 0 \mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3} \mathcal{Y} 0 \mathcal{R}_B$

Table 2.13: Combined Transition Matrices for Each Category

2.4.1 Example Combined Transition Matrices

By encoding the generative functions for each component presented in the previous sections I use Mathematica to produce each of the combined matrices for $N_A = 100$, $N_B = 100$, and $K = 4$. These matrices are iterated on an initial vector of unconnected agents to determine the limiting distributions for each category of strategic context. The complete set of these transition matrices is available in Supplementary Material.

Here I include select matrices that will be useful for demonstrating how the results can be generalized to any N_A, N_B, and K.

The combined transition matrices for A- and B-types in a cooperative game are shown in tables 2.14 and 2.15 respectively. In this case $\mathcal{A} = \mathcal{X}_{\phi_1}\mathcal{Y}0\mathcal{R}_A$; since $\mathcal{Y}0$ is the identity matrix it has no effect on the transitions. The probabilities for the resulting configuration of each agent action are adjusted by the Bernoulli probabilities. So an A-type in configuration $A21$ will, with probability one, detach the B-type neighbor. Yet the probability of transitioning into configuration $A20$ is only $\rho(0,0)$ because it also depends on none of the other agents randomly connecting to it.

The next key point is to notice that while $\mathcal{A}$ is not upper triangular due to detachments it will enact, it is *mostly* upper triangular and is sparse. Of particular importance is that the probability of transitioning into some configurations is low (i.e., the sum of the *column* is low), and no agent can return to the initial configuration of $A00$. By contrast, the combined transition matrix for B-types is complete – every configuration can transition into every other configuration. This is because, as was already noted, the structure of the $\mathcal{Y}\delta$ matrix is the transpose of the $\mathcal{R}$ matrix.

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	0.	0.184399	0.0931306	0.0232826	0.0061879	0.182555	0.184399	0.0696151	0.0201141	0.0912773	0.0689166	0.0277186	0.0225888	0.0198045	0.00601199
$\theta 01$	0.	0.	0.184399	0.0931306	0.0366557	0.	0.182555	0.184399	0.0962275	0.	0.0912773	0.095529	0.	0.0358281	0.
$\theta 02$	0.	0.	0.	0.184399	0.159523	0.	0.	0.182555	0.316519	0.	0.	0.157005	0.	0.	0.
$\theta 03$	0.	0.	0.	0.	0.502513	0.	0.	0.	0.497487	0.	0.	0.	0.	0.	0.
$\theta 04$	0.	0.	0.	0.	1.0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
$\theta 10$	0.	0.	0.	0.	0.	0.	0.184399	0.0931306	0.0366557	0.182555	0.184399	0.0962275	0.0912773	0.095529	0.0358281
$\theta 11$	0.	0.	0.	0.	0.	0.366953	0.18533	0.0463325	0.012314	0.183477	0.0926649	0.0278362	0.0454058	0.0276022	0.0120847
$\theta 12$	0.	0.	0.	0.	0.	0.	0.366953	0.18533	0.0729448	0.	0.183477	0.119277	0.	0.0720182	0.
$\theta 13$	0.	0.	0.	0.	0.	0.	0.	0.366953	0.31745	0.	0.	0.315597	0.	0.	0.
$\theta 20$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.184399	0.159523	0.182555	0.316519	0.157005
$\theta 21$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.366953	0.18533	0.0729448	0.183477	0.119277	0.0720182
$\theta 22$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.366953	0.31745	0.	0.315597	0.
$\theta 30$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.502513	0.497487
$\theta 31$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.366953	0.31745	0.315597
$\theta 40$	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.0

Table 2.14: $\mathcal{A}$: Combined Transition Probabilities for A-Types in Cooperative Games, $N_A = N_B = 100$, $K = 4$

	$\theta 00$	$\theta 01$	$\theta 02$	$\theta 03$	$\theta 04$	$\theta 10$	$\theta 11$	$\theta 12$	$\theta 13$	$\theta 20$	$\theta 21$	$\theta 22$	$\theta 30$	$\theta 31$	$\theta 40$
$\theta 00$	0.220172	0.183108	0.0637544	0.0134844	0.0033694	0.184957	0.129358	0.041326	0.0105941	0.0650517	0.0417458	0.0149527	0.0138995	0.0107643	0.00346229
$\theta 01$	0.132103	0.197933	0.111496	0.0335924	0.0106545	0.110974	0.151598	0.076539	0.0282656	0.039031	0.0510682	0.0310487	0.0083397	0.0152788	0.00207738
$\theta 02$	0.0792619	0.171601	0.146071	0.0647537	0.0285345	0.0665847	0.135349	0.106563	0.0618127	0.0234186	0.0462533	0.0490876	0.0050038	0.0144595	0.00124643
$\theta 03$	0.0475571	0.134666	0.156283	0.0972806	0.0654376	0.0399508	0.107843	0.118077	0.108079	0.0140512	0.0371194	0.0580549	0.0030022	0.011851	0.00074786
$\theta 04$	0.0475571	0.150598	0.196113	0.136427	0.10335	0.0240188	0.0760594	0.099047	0.100099	0.0060046	0.0190148	0.0356648	0.0009906	0.0048429	0.00021322
$\theta 10$	0.132103	0.109865	0.0382526	0.0080906	0.0020216	0.199043	0.150858	0.0502974	0.0149893	0.113014	0.0767908	0.0308808	0.0343604	0.0285356	0.0108977
$\theta 11$	0.220172	0.110086	0.0272435	0.0044488	0.0009646	0.257979	0.12899	0.0319217	0.0075061	0.101931	0.0509657	0.0149061	0.0231194	0.0138531	0.00591272
$\theta 12$	0.132103	0.15412	0.0603805	0.0135667	0.0034791	0.154788	0.180585	0.0707488	0.0226511	0.0611589	0.071352	0.0347087	0.0138717	0.0229384	0.00354763
$\theta 13$	0.0792619	0.145313	0.0978764	0.0322922	0.0107414	0.0928725	0.170266	0.114683	0.056128	0.0366953	0.0672747	0.0636038	0.008323	0.022539	0.00212858
$\theta 20$	0.0792619	0.0659188	0.0229516	0.0048543	0.001213	0.172267	0.134461	0.0454795	0.0141733	0.147426	0.106418	0.0486361	0.0658218	0.0620755	0.0290429
$\theta 21$	0.132103	0.0660516	0.0163461	0.0026693	0.0005787	0.242856	0.121428	0.0300504	0.0074040	0.164351	0.0821753	0.0270911	0.0546442	0.0340769	0.0181742
$\theta 22$	0.0792619	0.0924722	0.0362283	0.0081400	0.0020874	0.145714	0.169999	0.0666015	0.0222446	0.0986103	0.115045	0.0633625	0.0327865	0.0565416	0.0109045
$\theta 30$	0.0475571	0.0395513	0.013771	0.0029126	0.0007277	0.135065	0.107044	0.0364683	0.0116118	0.157362	0.117635	0.0574234	0.0984634	0.108097	0.06631
$\theta 31$	0.0792619	0.0396309	0.0098076	0.0016015	0.0003472	0.198555	0.0992775	0.0245687	0.0061826	0.195753	0.0978764	0.0340569	0.0985268	0.067554	0.0469999
$\theta 40$	0.0475571	0.0237786	0.0058845	0.0009609	0.0002083	0.150838	0.0754189	0.0186643	0.0047536	0.196874	0.0984369	0.0352637	0.137417	0.0999053	0.104039

Table 2.15: $\mathcal{B}$: Combined Transition Probabilities for B-Types in Cooperative Games, $N_A = N_B = 100$, $K = 4$

2.5 Stationary Distributions

To determine the distribution of agents across configurations over time using a Markov model we simply multiply the matrix $\mathcal{A}$ or $\mathcal{B}$ by the vector of agent configuration distributions at time t to produce the distribution at time $t+1$. What we are seeking is a stationary distribution over configurations represented by a vector π_i such that

$$\pi_A \mathcal{A} = \pi_A \text{ and } \pi_B \mathcal{B} = \pi_B;$$

i.e., a vector representing a distribution of the population that is unaffected by further applications of the transition matrix.

Some of the matrices in 2.13 are not positive recurrent (e.g., $\mathcal{A}$ in Prisoners' Dilemma, in which no configurations transition into configuration $A00$) so they are not certain to have a stationary distribution – and if one exists it may not be unique (i.e., it will depend on the initial distribution of agent configurations). Other matrices are ergodic (non-periodic and positive recurrent, e.g., $\mathcal{B}$ in Prisoners' Dilemma) so we know that there exists a unique stationary distribution in those cases. With the ergodic matrices one can use eigenvector analysis to find π_i, but as that technique cannot be used in all cases the iterative approach is used throughout.

For the iterative approach we make use of the property that each multiplication of $\mathcal{A}$ or $\mathcal{B}$ is an iteration of the model's processes on some initial distribution of configurations. We can repeat the multiplication until the distribution differences between the t^{th} and $t+1^{th}$ operation are monotonically decreasing and below some threshold vector τ.

$$\exists T : \forall t > T, \begin{cases} 0 < \pi_0 \mathcal{A}^t - \pi_0 \mathcal{A}^{t+1} < \tau \\ 0 < \pi_0 \mathcal{B}^t - \pi_0 \mathcal{B}^{t+1} < \tau \end{cases} \tag{2.5.1}$$

Then we can approximate π_A and π_B using $\pi_0 \mathcal{A}^T \sim \pi_A$ and $\pi_0 \mathcal{B}^T \sim \pi_B$. The limit as $T \to \infty$ produces the accurate stationary distributions, but we can achieve any desired level of accuracy by setting τ appropriately (e.g., $\tau = 0.0001$ is used in the calculations below).

As just stated, in some cases the stationary distribution will depend on the initial distribution. Any initial distribution in which the entries sum to one is admissible, but here I assume that the whole initial population is in the $\theta 00$ configuration; i.e., initially unconnected. The figures in the distribution vector can be interpreted in two ways: 1) as the probability of any agent being in a given configuration over time or 2) the expected percent of agents in each configuration over time. Since there are 100 agents in all the examples provided below, the probabilities here are also the expected proportions of the agent population in each configuration.[2]

2.5.1 Example Calculation for Cooperative Games

By examining table 2.14 we can clearly see that $A04$ and $A40$ are absorbing states for $K = 4$, so in finite time the probability density will be distributed between only these two states. The number of cooperators stuck in $A04$ compared to $A40$ will depend on the initial distribution, but with all $N_A = 100$ agents starting in configuration $A00$ the figures are: $A04$: 6.5338 and $A40$: 93.4662 (see table 2.16). These results show the probabilities (or proportions) of A-Type agents to settle in each of those configurations starting from $A00$. This limiting distribution is also an equilibrium since A-types will not transition out of these states.

The defectors' transition matrix $\mathcal{B}$ (table 2.15) is complete: there exists a transition probability from every configuration to all configurations. That information alone precludes an equilibrium, but the Perron-Frobenius theorem ensures that there exists at least one π_B satisfying $\pi_B\mathcal{B} = \pi_B$. Since $\mathcal{B}$ is also ergodic we know there exists a unique limiting distribution (i.e., the same for any initial distribution). The limiting distribution for $N_A = N_B = 100$ and $K = 4$ is presented in table 2.16. The interpretation here is that agents are constantly churning among all the configurations, but the probability (or proportion) of B-type agents in each configuration is constant after some transition period.

Note that the number of A-type agents stuck in $A04$ exceeds the number of B-type agents in $B04$ (which is the *least* probable configuration for B-types). This is due to the fact that one of those defectors' defector neighbors may be connected to a cooperator and hence cause it to be disconnected. The more defector neighbors a defector has the greater the probability that the focal defector is disconnected. But this fails to consider that as more cooperators become arranged in $A40$ configurations, the probability that a defector has any cooperative neighbors should decrease.

Upon further examination these outcome distributions reveal an apparent violation of the number of connections between types. The number of cooperators connected to defectors (those in $A04$) is approximately six or seven and each one has four defector neighbors, a total of twenty-four to twenty-eight cooperator-defector links.

[2] Agents are indivisible so the actual proportions would have to be some integer value. Restricting the distribution to integer values *would* affect the outcomes, but exploring this formulation distracts from the main point of the Markov approach.

$A00$	0	$B00$	13.8330
$A01$	0	$B01$	12.4963
$A02$	0	$B02$	5.5655
$A03$	0	$B03$	1.6606
$A04$	6.5338	$B04$	0.6268
$A10$	0	$B10$	17.4611
$A11$	0	$B11$	13.7988
$A12$	0	$B12$	5.3802
$A13$	0	$B13$	2.0060
$A20$	0	$B20$	9.5330
$A21$	0	$B21$	6.8630
$A22$	0	$B22$	3.2347
$A30$	0	$B30$	3.2424
$A31$	0	$B31$	2.9927
A40	**93.4662**	$B40$	1.3058

Table 2.16: Stationary Distributions in Cooperative Games

Yet if we add the number of links from defectors to cooperators on the B column this number clearly exceeds twenty-eight (it is approximately 102).

Both of these anomalies are the result of having used time-homogeneous Markov processes that (as mentioned in the introduction) cannot be contingent on system-level arrangements of the agents. Probabilities of connection and disconnection use approximations for connectability. The discrepancy in connections and the numerically inaccurate representation it produces, are examples of artifacts from this limitation.

However, though the results are not numerically accurate in this respect, the stationary distributions do pass the face validity test and do represent outcomes that we would expect from such a process as preferential detachment. The result that a cooperator has more than a ninety-three percent chance of forming a group with other cooperators means that the prosocial outcome was over ninety-three percent successfully achieved. That defectors would continuously change configurations, constantly unhappy with defector neighbors yet unsuccessful at maintaining cooperative ones, also matches the interpretation and understanding of the theory provided in the first chapter.

2.5.1.1 Markov Dynamics of Cooperative Games

Determining the limiting distribution suffices to answer the question of whether preferential detachment succeeds in achieving the prosocial outcome, but we can gain more insight into the mechanism of preferential detachment by examining the dynamics of the process that led there. Doing so also fosters comparison to the agent based model of chapter III and to other models.

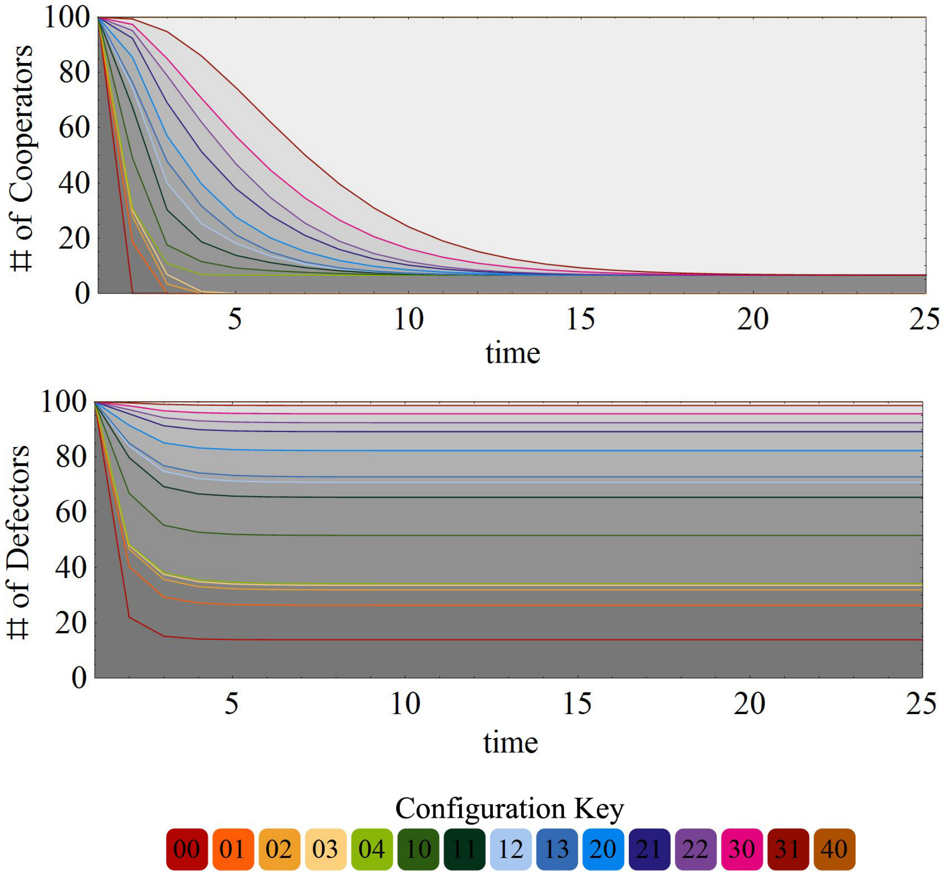


Figure 2.1: Configuration Dynamics in Cooperative Games

We can get a clearer picture of the comparative dynamics by looking at the distribution of configurations for the two types as presented in figure 2.1. From this we can see that the defectors settle near to their limiting distribution more quickly than cooperators. Both agent types' distribution dynamics are monotonic which is prima facie surprising for the defectors since the whole matrix $\mathcal{B}$ is positive recurrent. This results partly from having started all the defectors in the $B00$ configuration; initially the probability of being in any of the other configurations can only increase. Furthermore for each row the transitions across the columns are all between 0.001 and 0.216 and most are less than 0.1. Compare this to a uniform distribution of 0.667 per matrix entry. This transition structure puts slightly more force on a few transitions and little overall change. Because each row has similar figures it doesn't take many iterations to get close to the limiting distribution; each iteration is essentially smoothing out already similar distributions.

2.5.2 Stationary Distributions for Each Game Form

Following the same approach described above for the cooperative games I determine the stationary distributions in each of the strategic context categories (table 2.17). The prosocial outcomes are presented in boldface for ease of reference. These results demonstrate the consistently high level of prosociality that the preferential detachment mechanism achieves across the full range of strategic contexts. They also reveal the degree to which system-level approximations of agent arrangements affect the interpretations of the results.

I have already discussed the cooperative game outcome and the mismatch between the number of cross-type connections. This same anomaly appears in the Contributive and Commensal games and is caused by 1) using time-homogeneous Markov models and 2) treating the A-types and B-types independently. The numerical discrepancy does not interfere with our interpretation of the results with respect to prosociality because its magnitude is small compared to the overall problem scale.

The outcome for A-types in the Cooperative and Contributive games is the same because the A-types have the same matrix combination; so both are over 93% successful in achieving the prosocial outcome. The interpretations differ because of the distribution of B-types. In Contributive games (e.g., Stag Hunt) the B-types are indifferent, but are detached by A-types in mixed configurations, so they end up grouped together. In both cases the number of stable heterogeneous groups is determined by the number of A-types who receive K consecutive B-type neighbors. The difference is that Cooperative games limit the other attachments defectors will allow (i.e., only more isolated cooperators, which are in short supply), whereas in the Contributive games the hare hunters can fill their network neighbors with other hare-hunters. The difference reflects that in Cooperative games the B-types would rather not have the neighbors they get stuck with, but in Contributive games the B-types are equally satisfied any anybody. These differences parallel the description in the literature that the Stag Hunt is an “easier” game than the Prisoners’ Dilemma in which to sustain cooperation.

The results from all four of the other game categories demonstrate a complete achievement of the prosocial outcome. For the Coordinative games this means that assortative mixing is achieved, while for the Specialized games it means that disassortative mixing was achieved. Though it is possible to have stable heterogeneous groups of the unsocial pattern arise in these games, it would require a group of A-types getting K consecutive less-preferred neighbor connections *and* each of those neighbors *also* getting K consecutive less-preferred neighbors. This will be a very rare event, though it will be possible to detect it in the agent-based model. The results here reliably converge to 100% because the probabilities are time-homogeneous, and hence there is *always* a probability of detachment by types that prefer other neighbors.

The results of the Undifferentiated game category, as well as the B-types of the Commensal games, demonstrate the steady attachment of connections until the full allotment is reached. The distribution reflects the binomial coefficients for how many ways that combination of neighbors can be achieved through an add-one-per-turn process. Though it is unclear what prosociality may mean in these contexts, the outcome reflects a fair (e.g., probabilistically unbiased) distribution over configurations. Insofar as no agent type is given preferred treatment lacking a motivation for that special treatment, and since a violation of this feature could also be interpreted as a violation of fairness norms, we can consider this unbiased result as achieving prosociality in this way.

Cooperative Games

$A00$	0	$B00$	13.8330
$A01$	0	$B01$	12.4963
$A02$	0	$B02$	5.5655
$A03$	0	$B03$	1.6606
$A04$	6.5338	$B04$	0.6268
$A10$	0	$B10$	17.4611
$A11$	0	$B11$	13.7988
$A12$	0	$B12$	5.3802
$A13$	0	$B13$	2.0060
$A20$	0	$B20$	9.5330
$A21$	0	$B21$	6.8630
$A22$	0	$B22$	3.2347
$A30$	0	$B30$	3.2424
$A31$	0	$B31$	2.9927
A40	**93.4662**	$B40$	1.3058

Coordinative Games

$A00$	0	$B00$	0
$A01$	0	$B01$	0
$A02$	0	$B02$	0
$A03$	0	$B03$	0
$A04$	0	**B04**	**100**
$A10$	0	$B10$	0
$A11$	0	$B11$	0
$A12$	0	$B12$	0
$A13$	0	$B13$	0
$A20$	0	$B20$	0
$A21$	0	$B21$	0
$A22$	0	$B22$	0
$A30$	0	$B30$	0
$A31$	0	$B31$	0
A40	**100**	$B40$	0

Specialized Games

$A00$	0	$B00$	0
$A01$	0	$B01$	0
$A02$	0	$B02$	0
$A03$	0	$B03$	0
A04	**100**	$B04$	0
$A10$	0	$B10$	0
$A11$	0	$B11$	0
$A12$	0	$B12$	0
$A13$	0	$B13$	0
$A20$	0	$B20$	0
$A21$	0	$B21$	0
$A22$	0	$B22$	0
$A30$	0	$B30$	0
$A31$	0	$B31$	0
$A40$	0	**B40**	**100**

Contributive Games

$A00$	0	$B00$	0
$A01$	0	$B01$	0
$A02$	0	$B02$	0
$A03$	0	$B03$	0
$A04$	6.5338	$B04$	100
$A10$	0	$B10$	0
$A11$	0	$B11$	0
$A12$	0	$B12$	0
$A13$	0	$B13$	0
$A20$	0	$B20$	0
$A21$	0	$B21$	0
$A22$	0	$B22$	0
$A30$	0	$B30$	0
$A31$	0	$B31$	0
A40	**93.4662**	$B40$	0

Commensal Games

$A00$	0	$B00$	0
$A01$	0	$B01$	0
$A02$	0	$B02$	0
$A03$	0	$B03$	0
A04	**100**	$B04$	6.2960
$A10$	0	$B10$	0
$A11$	0	$B11$	0
$A12$	0	$B12$	0
$A13$	0	$B13$	24.7664
$A20$	0	$B20$	0
$A21$	0	$B21$	0
$A22$	0	$B22$	37.1703
$A30$	0	$B30$	0
$A31$	0	$B31$	25.2336
$A40$	0	$B40$	6.5338

Undifferentiated Games

$A00$	0	$B00$	0
$A01$	0	$B01$	0
$A02$	0	$B02$	0
$A03$	0	$B03$	0
$A04$	6.5338	$B04$	6.2960
$A10$	0	$B10$	0
$A11$	0	$B11$	0
$A12$	0	$B12$	0
$A13$	25.2336	$B13$	24.7664
$A20$	0	$B20$	0
$A21$	0	$B21$	0
$A22$	37.1703	$B22$	37.1703
$A30$	0	$B30$	0
$A31$	24.7664	$B31$	25.2336
$A40$	6.2960	$B40$	6.5338

Table 2.17: Stationary Distributions for All Game Categories

In addition to the stand-alone conclusions we can draw from these results, they also offer a validation set for the agent-based model presented in the next chapter. Since both models are implementations of the same mechanism they should produce a high degree of agreement in outcome distributions. And because the ABM can manage the system-level network arrangement information, we can further validate the claim made here that the probabilistic approximations are close enough to be useful.

2.6 Generalizing the Results

The results above are encouraging, but are specific to $N_A = 100$, $N_B = 100$, and $K = 4$. Changes in the population sizes alter the transition probabilities in $\mathcal{R}$ and $\mathcal{X}$ in straightforward ways shown below. The value of K determines the number of configurations, so changing it generates a set of transition matrices of a different size. The generating functions provided for each component can be directly applied to calculate the outcomes for any N_A, N_B, and K, but it is more insightful to demonstrate more general features of results that hold across all parameter values. To generalize the results just presented I reveal the salient patterns in the component matrices and demonstrate to what degree they persist through changes in N_A, N_B, and K. I then relate these patterns to the outcomes they produce in each of the strategic contexts.

2.6.1 Generalizing over the Maximum Degree

Increasing the maximum degree would require a complete recalculation of all the transition probability matrices because it increases the number of configurations agents can be in, but it does this in a systematic way. Restricting the current analysis to 2×2 games, the number of possible configurations can be found using equation (1.6.1) with $T = 2$:

$$\mathbb{C} = \frac{(2+K)!}{2(K!)} = \frac{(K+2)(K+1)}{2} = \frac{1}{2}K^2 + \frac{3}{2}K + 1.$$

To determine the affect on $\mathbb{C}$ of an increase in K we find $\mathbb{C}(K+1) - \mathbb{C}(K)$:

$$\frac{\Delta\mathbb{C}}{\Delta K} = \frac{(K+3)(K+2)}{2} - \frac{(K+2)(K+1)}{2} = K + 2 \tag{2.6.1}$$

What this means is that for every in increase K by one the *number* of configurations, and hence each dimension of the matrix, increases by $K + 2$ entries. For example, if we increase $K = 4$ to $K = 5$, the number of configurations increases by six; from fifteen to twenty-one. The *rate* of increase is just one; as K increases we add one additional configuration compared to the previous increase (because K is increasing by one).

2.6.1.1 Effect of K on $\mathcal{R}$

Recall table 2.2 showing the structure of the binomial random connection probabilities through the Bernoulli trial process for $K = 4$. The effect of this process is obviously to move the agents into configurations with larger numbers of total connections. $\mathcal{R}$ is an upper triangular matrix with a regularly repeating pattern of upper triangular submatrices. Each of these submatrices corresponds to a particular value of a (or symmetrically using b); so the visual pattern depends on the order of the configurations but the mathematical properties do not. The overall upper-triangular structure is created by the fact agents cannot lose edges through random connection. The submatrices have their structure because the total edges cannot sum to more than K. Thus for any K, $\mathcal{R}$ will maintain the same tiered upper triangular structure, but on a larger scale.

Given the chosen ordering of the configurations (i.e., lexicographically by a, then b), another way to gain an intuition regarding the effect of different K values is to visualize $\mathcal{R}$ for some large value of K. Then for each *decrease* in K we eliminate the rows and columns with $a + b > K$ (using the lower K). These columns will all be the right-most column in each submatrix; the rows will be the bottom rows. If one removes columns from the right, and rows from the bottom, of an upper-triangular matrix the resulting matrix will also be upper-triangular. It is clear that the *pattern* is maintained, so now I address the effect K has on the *values* of $\mathcal{R}$.

There are two aspects of $\mathcal{R}$ to consider: the Bernoulli probabilities $\rho(a' - a, b' - b)$ and the edge effect $\Gamma(\theta ab)$. We can see from equation (2.1.2) that $\rho(a' - a, b' - b)$ is a function of N_A, N_B, Δa, Δb, and the focal agent's type (recall that these are probabilities of the number of connections *added*). So these values are independent of the value of K except through the expansion of the range of a and b for which they must be computed. As K increases the probability mass assigned to the edge configurations through $\Gamma(\theta ab)$ decreases for any fixed a and b. This is because the edge cases (where $a' + b' = K$) are pushed to higher values of a' and b'; hence there is less leftover probability mass. But $\Gamma(\theta ab)$ is constant for any $\theta(K - a, K - b)$ since the same number of possible configurations are excluded.

Because adding connections is modeled as Bernoulli trials subject to the constraint of $a + b <= K$, as K increases the *rate* at which agents gain new random edges is unaffected. This is partly a byproduct of using a time-homogeneous Markov model. The implication for relaxing the maximum degree assumption is that, with respect to random connections, both the magnitude and direction of the force exerted is constant for any K.

2.6.1.2 Effect of K on $\mathcal{Y}$

We next consider the effect that K has on the detachment transition matrices $\mathcal{Y}\alpha$, $\mathcal{Y}\beta$, $\mathcal{Y}\delta$, and $\mathcal{Y}0$. Recall that these matrices correspond to cases in which A-types, B-types, both types, or no types will disconnect. For the less preferred type, a detachment only occurs if a neighbor has both types of agents. The probability of a detachment is thus proportional to the number of configurations in which neighbors' neighbors are expected to be in a mixed configuration.

Using the four categories of configurations in 2.5 and the generating functions for detachment probabilities in (2.2.1) we can identify a clear pattern as K increases. The number of configurations in which an agent has only same-type neighbors equals K for each type of agent. The number of unconnected configurations is always one. The number of mixed configurations $\mathbb{C}_M$ equals $\mathbb{C} - 2K - 1 = \frac{1}{2}(K^2 - K)$. Thus as we increase K the change in $\mathbb{C}_M$ is

$$\frac{\Delta \mathbb{C}_M}{\Delta K} = \frac{1}{2}\left((K+1)^2 - (K+1)\right) - \frac{1}{2}\left(K^2 - K\right) = K \tag{2.6.2}$$

This rate is also clear since we already saw that $\mathbb{C}$ increases by $K + 2$ and each $\mathbb{C}_A = K$ and $\mathbb{C}_B = K$ increase by one for each increase in K.

Recall now the functions for the proportions of configurations that are mixed or unmixed from (2.2.2) and recreated here:

$$\tfrac{\mathbb{C}_A}{\mathbb{D}} = \tfrac{2}{K+1}, \quad \tfrac{\mathbb{C}_B}{\mathbb{D}} = \tfrac{2}{K+1}, \quad \tfrac{\mathbb{C}_M}{\mathbb{D}} = \tfrac{K-1}{K+1}$$

From these proportions we can clearly see the effect of an increase in K on the probabilities of keeping or losing a connection.

$$\begin{aligned} \tfrac{\Delta \frac{\mathbb{C}_A}{\mathbb{D}}}{\Delta K} = \tfrac{\Delta \frac{\mathbb{C}_B}{\mathbb{D}}}{\Delta K} &= \tfrac{2}{K+2} - \tfrac{2}{K+1} = & -\tfrac{2}{K^2+3K+2} &= -\tfrac{1}{\mathbb{C}}, \\ \tfrac{\Delta \frac{\mathbb{C}_M}{\mathbb{D}}}{\Delta K} &= \tfrac{K}{K+2} - \tfrac{K-1}{K+1} = & \tfrac{2}{K^2+3K+2} &= \tfrac{1}{\mathbb{C}} \end{aligned} \tag{2.6.3}$$

The result of this analysis is that for a lesser-preferred type of agent, the probability of keeping each edge to the types not preferring them decreases, and the probability of losing each such edge increases as K increases. That is, the force exerted by preferential detachment on removing less-preferred neighbors increases with increasing K.

2.6.1.3 Effect of K on $\mathcal{X}$

He have left to examine the agent action transition probabilities $\mathcal{X}_{\phi_1}$, $\mathcal{X}_{\phi_2}$, and $\mathcal{X}_{\phi_3}$ presented in tables 2.10, 2.11, and 2.12 respectively. Here we assume that an agent may only perform one action per iteration: a detachment or a random connection, though it is also possible that an agent performs no action. If a configuration indicates that a detachment is appropriate, then that detachment happens with probability one. Otherwise, if an agent's $k = K$ then it does nothing, and if $k < K$ one random connection is made. The probability of connecting to an agent of each type depends only on the focal agent's type, N_A, and N_B.

The overall force of $\mathcal{X}_{\phi_1}$ is to increase the number of connections to A-types over time, and this force strengthens as K increases. The reason is again related to the number of mixed versus pure B-type configurations. When the configuration at time t is mixed the configuration at time $t+1$ has one fewer B-types. When the configuration is $\theta 0k$ (and $k < K$) there is an $\frac{\eta_A}{N-1}$ probability that it becomes mixed and an $\frac{\eta_B}{N-1}$ probability that it stays pure. Once an agent gains an A-type connection it can never become pure B-type again (through this process). Thus the probability of achieving the stable configuration $\theta 0K$ equals $\left(\frac{\eta_B}{N-1}\right)^K$, which is obviously decreasing with increasing K. The only other stable configuration is $\theta K0$, and this is eventually achieved unless $\theta 0K$ is. So the probability of achieving a pure A-configuration increases quickly with K.

The dynamics for $\mathcal{X}_{\phi_2}$ follow the same pattern as $\mathcal{X}_{\phi_1}$, but switching the A-types and B-types. This process therefore takes the agents into the $\theta 0K$ configuration with greater probability as K increases.

Finally, $\mathcal{X}_{\phi_3}$ (the transitions for indifferent agents) contains no detachment probabilities and all θab such that $a + b = K$ are stable. The force exerted by this process increases the number of connections steadily toward one of these stable configurations. As K increases by one there is one more stable configuration; the probability of ending up in any one of them is $\binom{K}{b}\left(\frac{\eta_A}{N-1}\right)^a\left(\frac{\eta_B}{N-1}\right)^b$. So it is not the case that all stable configurations are equally probable because there are more paths to some (e.g., evenly mixed) than others (e.g., pure).

2.6.2 Generalizing over the Population Size

The numbers of each type of agent (as well as the total population) affect the transition probabilities through the random connections. Detachment operates on an agent's current actual neighbors, so it depends only on K. Changes in the population sizes do not affect the structure or patterns of the transition matrices as long as N_A and N_B are greater than or equal to K. However, for the Markov models presented here to be appropriate, N_A and N_B should be appreciably larger than K. This is because the approximations made to compensate for the lack of system-wide connection information are better approximates the larger the population is.[3]

2.6.2.1 Effect of N_A and N_B on $\mathcal{R}$

The values in the random attachment matrices are determined by the binomial probability function presented in (2.1.2). By utilizing the symmetry between A-types and B-types presented in (2.1.3) I can demonstrate the effect that the population has on A-types knowing that the results hold mutatis mutandis for B-types. For A-type agents, the probability of *gaining* exactly a connections with A-types and b connections with B types equals

$$\binom{N_A-1}{a}\left(\frac{1}{N-1}\right)^a\left(1-\frac{1}{N-1}\right)^{N_A-1-a}\cdot\binom{N_B}{b}\left(\frac{1}{N-1}\right)^b\left(1-\frac{1}{N-1}\right)^{N_B-b}.$$

As N_A or N_B increases, the number of possible connections increases through the binomial coefficients and through the exponents of success or failure. But the individual probability of being connected to $(\frac{1}{N-1})$ shrinks. Every new agent added that can attempt more connections is also a new agent to receive connections from other agents. If N_A and N_B stay equal then these two differences balance out and the binomial probabilities stay the same. You can see from plot 2.2 of binomial probabilities for each possible addition with $a+b \leq 4$ that (except for extremely low values for which the distribution is ill-defined) the probabilities are constant across population changes.

[3] The reason for the increasing accuracy can be seen clearly through the difference between transitions as proportions or probabilities. There is no such thing as part of an agent insofar as agents represent people or animals. The larger N_i is, the smaller the discrepancy between this real-valued Markov model and an interpretation requiring integer values. Also, in cases for which a distribution of agents over configurations is assumed (as in the detachment matrices) the larger the population the greater the confidence that the distribution fits the population.

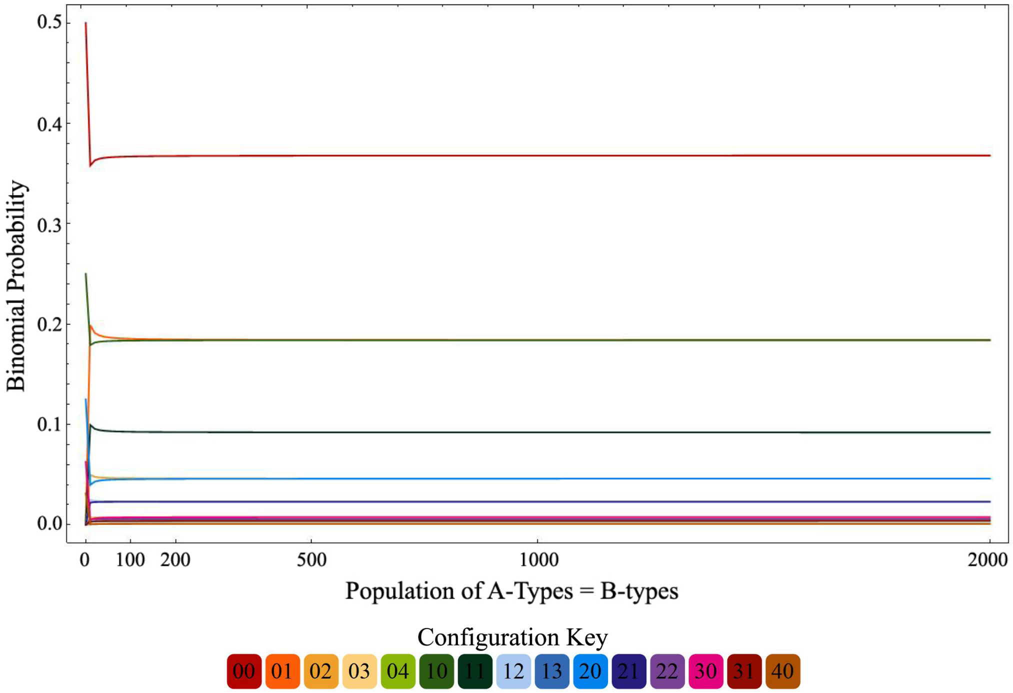


Figure 2.2: Probabilities of random connections for $N_A = N_B = 0$ to 2000.

When we increase N_A relative to N_B this has an obvious affect on the binomial probabilities of getting connected to by an A-type agent: they increase. As is clear from plot 2.3, as the ratio of A-types to B-types increases, the probabilities of getting multiple attachments from A-types (olive, cyan, and magenta representing $\rho(1,0)$, $\rho(2,0)$, and $\rho(3,0)$ respectively) increase rapidly at first and then level off. This is also true of $\rho(4,0)$ (in brown), but the probability is still so low that it is difficult to discern from the plot. Obviously the probabilities of receiving one, two, three, or four connections to B-types must drop simultaneously and proportionally.

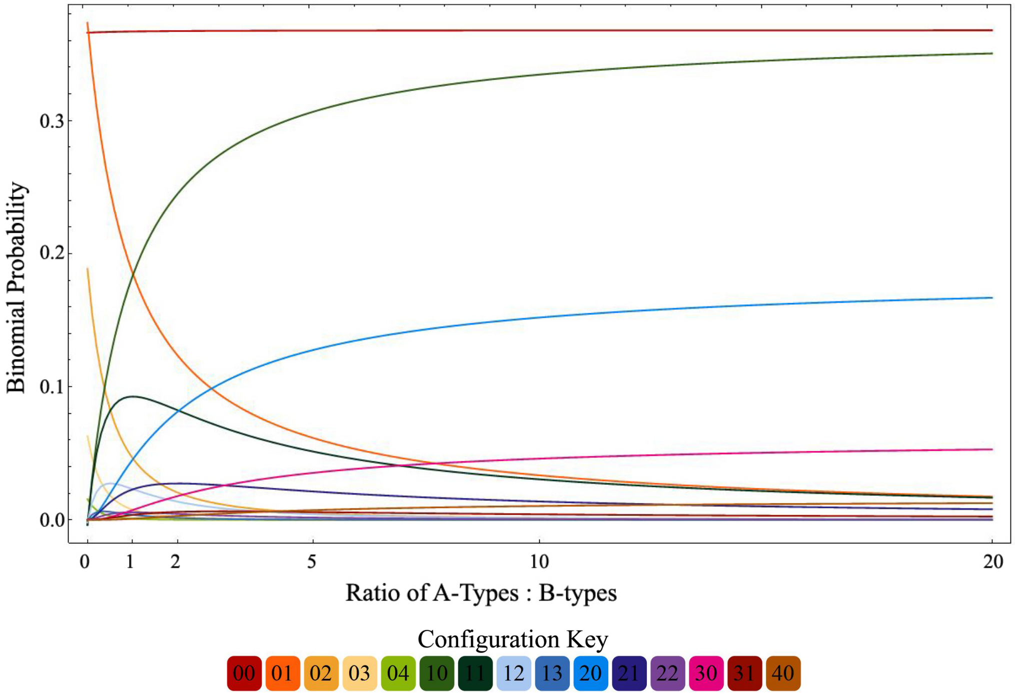


Figure 2.3: Probabilities of random connections for N_A/N_B.

Recall that the "left-over" probability density (having more agents try to connect than can actually connect) is accumulated at the $a' + b' = K$ configurations and included in $\mathcal{R}$ via $\Gamma(\theta ab)$. If we increase both N_A and N_B equally, the probability of adding a too large number of agents stays unchanged because the binomial probabilities do. When changing one population relative to the other the shape of the distribution changes, and these changes can alter the amount of probability mass not accounted for. However, the probability of getting four A-type connections in the same iteration when there are twenty times more A-types is already less than 0.025. Receiving four connections from mixed types is even lower. The values accounted for with Γ are low enough that changes in it have only minor effects on the impact of $\mathcal{R}$.

As noted already, because the values in $\mathcal{R}$ correspond to the addition of a number of connections, they are unaffected by K except through K's effect on the size of the matrix and hence the range of a and b they need to be calculated for. Therefore it is clear that there are no combination or interference effects between the maximum degree and population sizes. The final result is that the qualitative effects of $\mathcal{R}$ on preferential detachment's overall force are robust against changes in N_A, N_B, and K if $N_A = N_B$; and when populations change are unequally this has the obvious effect of increasing the number of connections to the larger population.

2.6.2.2 Effect of N_A and N_B on $\mathcal{Y}$

The probability of being detached by other agents depends on the actual current neighbors an agent has and *the neighbors'* expected configuration. The expected configuration of neighbors is based on a uniform distribution over their possible configurations, given that they are attached to the focal agent. From this description (and a fortiori by examining the generating functions in equations (2.2.1)) it is clear that changes in N_A or N_B do not affect the detachment matrices $\mathcal{Y}$.

2.6.2.3 Effect of N_A and N_B on $\mathcal{X}$

Transition probabilities corresponding to agents following the preferential detachment mechanism are captured in $\mathcal{X}$. This means detachment when agents have both preferred and less preferred neighbors, and (if not full) a new connection uniformly chosen at random from the whole population otherwise. Detachment is unaffected by N_A and N_B because it happens with probability one in the appropriate configurations.

Because the probability to connect to an agent of a certain type is proportional to the existence of those types in the population, the entries in $\mathcal{X}$ corresponding to random connection events change with population changes. The generating function and matrix structure are identical, but the values change in an obvious way. If we change $N_A = 100$, $N_B = 100$ to $N_A = 50$, $N_B = 150$, for example, an A-types' probability of connecting to another A-type changes from $\frac{99}{200}$ to $\frac{49}{200}$ and the probability of connecting to a B-type changes from $\frac{100}{200}$ to $\frac{150}{200}$. Hence *unequal* changes in the population levels can greatly affect the dynamics and outcome of the preferential detachment process.

From this analysis we can conclude that the preferential detachment mechanism is highly robust against changes in scale; as long as $N_A = N_B$ the particular values have little effect on any of the transition matrices. This feature is in contradistinction to other proposed mechanism for the evolution of prosociality (esp. punishment, mutualism, and kinship) which breakdown with increasing population levels. Uneven changes in the population levels affect the outcome differently among the strategic contexts, and these changes (along with changes due to varying K) are the subject of the next section.

2.6.3 Effects on the Strategic Contexts

Now that we have seen the effects of changes in N_A, N_B, and K on each transition matrix, these analyses are combined to demonstrate the changes in the outcomes of each category of strategic context. This will be done by exploiting the structural and numerical patterns identified in the previous two sections, and how they affect the stationary distributions presented in table 2.17 for $N_A = N_B = 100$, and $K = 4$. Because we are primarily concerned with the degree to which prosociality is achieved, the discussion will be focused on this aspect of the outcomes.

To facilitate the discussion I again show the matrix composition for each game category from table 2.13. These will make clear the translation of the independent effects on $\mathcal{R}$, $\mathcal{Y}$, and $\mathcal{X}$ into the effects on each game category.

Category	Relationship		A-type	B-type
Cooperative	$A : B \prec A$	$B : B \prec A$	$\mathcal{A} = \mathcal{X}_{\phi_1}\mathcal{Y}0\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_1}\mathcal{Y}\delta\mathcal{R}_B$
Coordinative	$A : B \prec A$	$B : A \prec B$	$\mathcal{A} = \mathcal{X}_{\phi_1}\mathcal{Y}\beta\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_2}\mathcal{Y}\alpha\mathcal{R}_B$
Specialized	$A : A \prec B$	$B : B \prec A$	$\mathcal{A} = \mathcal{X}_{\phi_2}\mathcal{Y}\alpha\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_1}\mathcal{Y}\beta\mathcal{R}_B$
Contributive	$A : B \prec A$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_1}\mathcal{Y}0\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3}\mathcal{Y}\alpha\mathcal{R}_B$
Commensal	$A : A \prec B$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_2}\mathcal{Y}\alpha\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3}\mathcal{Y}0\mathcal{R}_B$
Undifferentiated	$A : A \approx B$	$B : A \approx B$	$\mathcal{A} = \mathcal{X}_{\phi_3}\mathcal{Y}0\mathcal{R}_A$	$\mathcal{B} = \mathcal{X}_{\phi_3}\mathcal{Y}0\mathcal{R}_B$

Table 2.18: Combined Transition Matrices for Each Category

2.6.3.1 Cooperative

Cooperative games (e.g., Prisoners' Dilemma and Hawk and Dove) are such that both types prefer A-type neighbors. For A-types (cooperators) the only stable configurations are $AK0$ and $A0K$, so for any K the population will be divided into these two configurations. As K increases the probability of getting K consecutive connections to B-types decreases, hence the probability of ending up in $A0K$ decreases. With increasing K preferential detachment *increases* the degree of prosociality.

Changing the population levels in equal numbers has a negligible effect on the outcome of cooperative games. As we increase the proportion of A-types in the population we quickly increase the probability of reaching $AK0$ because A-types are more likely to be connect with A-types through $\mathcal{R}$ and $\mathcal{X}$. Scenarios with many fewer cooperators than defectors will rarely produce the prosocial outcome; a result that is realistic and expected. So scaling up either K or $N_A = N_B$ increases the prosociality of these situations beyond the stationary distribution level presented in table 2.17 of 93.4662%.

2.6.3.2 Coordinative

To generalize the results for Coordination games recall that both types of agents detach each other; the behavior of the two types are mirror images (replace $A \leftrightarrow$

B) and both types prefer assortative arrangements. Thus the prosocial outcome is probability density focused on $AK0$ and $B0K$. As K increases, the $\mathcal{R}$ matrices will push fewer agents into pure configurations of the other type. This means more mixed configurations through which $\mathcal{X}_{\phi_1}$ and $\mathcal{Y}\beta$ can push A-types into configurations with only A-types. The dynamics for B-types is, naturally, the mirror of those for A-types so they increasingly form arrangements containing only B-types.

Unlike the Cooperative games, when the population levels change unevenly in Coordinative games they will both still achieve a high degree of prosociality. The reason is that both are moving toward assortative arrangements; if the proportion of A-types drops then it has a greater probability to get stuck with all B-type neighbors, but the B-types will have a lower probability of getting stuck with all A-type neighbors. The resulting imbalance underscores the limitation of this modeling technique, but the intuition is clear. As the population and/or K increases, the probability of getting stuck in mixed-type groups decreases and the degree of prosociality increases. Within the constraints of the Markov model prosociality is already achieved with 100% certainty for $K = 4$ and $N_A = N_B = 100$.

2.6.3.3 Specialized

The example game in this category is Lichen, a miscoordination game in which the payoffs and preferences are a reversal of the Coordination game. Both types of agents detach same-type agents and the prosocial outcome is probability density focused on $A0K$ and $BK0$. A-types here use the same transition matrices as B-types in Coordinative games, and vice versa, with the only numerical difference arising from the difference between η_A and η_B.

When $N_A = N_B$ the difference is small and decreasing with increasing values. When N_A increases relative to N_B, the A-types are more likely to get stuck in $AK0$. Though this was beneficial in the Coordinative games, the opposite is true here. But again the decrease in prosociality of one type balances the increase in prosociality of the other type since with more As, the B-types are less likely to get stuck with Bs. And just like in the Coordinative games, within the Markov model's constraints the dynamics produce perfect prosociality for both types at the current parameter values.

2.6.3.4 Contributive

Stag Hunt dynamics are similar to dynamics in the Prisoners' Dilemma with one key difference: hare hunters disconnect from nobody. So the matrices for A-types are identical with Cooperative games, but only the A-types will detach B-types here. The result is that for any set of parameters, B-types have more B-type neighbors in comparison to Cooperative games (and only B-type neighbors in the Markov model results). Since prosociality depends only on the success of A-types to form groups with other A-types, this difference has no effect on the qualitative results. After all, B-types are indifferent between neighbor types.

The degree of prosociality (as measured by the assortativity of A-types) matches the result of the Cooperative games because it is also driven by the number of A-types that do not get K consecutive connections with B-type agents. We have already seen that assortativity increases with increases in K and N_A, and decreases with increases in N_B.

2.6.3.5 Commensal

The Commensal game has the same relationship with the Specialization games (e.g., Lichen) that Stag Hunt had with the Cooperative Games: A-types have the same dynamics and B-types are now indifferent. Since we are again only concerned with the success of A-types in associating with B-types, and since in this Markov model the types are considered independently, the level of prosociality increases with K and with increasing N_B. Thus increasing the availability of hosts to the commensals improves the commensals' success –an intuitive and desirable result.

In Lichen the lower prosociality of A-types from higher relative N_A levels was offset by increased B-type prosociality. Because B-types in this game are indifferent, there is no such offset. So while the indifference B-types in Stag Hunt makes that game "easier to solve" than the Prisoners' Dilemma, the indifference of B-types here should reduce the success of agents in achieving prosociality. However, within the constraints of the Markov model we will see perfect disassortativity by A-types and a roughly binomial distribution of B-types for any values of K, N_A, and N_B.

2.6.3.6 Undifferentiated

The outcome for games in this category is clear. Both agent types will gain connections until some configuration in which $a + b = K$ is reached. The distribution of agents over these final configurations will approximate the binomial distribution of getting a connections from K trials. Only approximately binomial because of the edge effects in $\mathcal{R}$ and the additional connections made through $\mathcal{Y}$ (which will also create a binomial distribution over final configurations, but adding on top of those from $\mathcal{R}$). This outcome distribution will obtain for any N_A, N_B and K, and as these parameters increase the realized distribution becomes closer to the ideal binomial distribution.

2.6.3.7 Lane Choice and Biased Lane Choice

The Markov analysis has considered only 2 games, but the results obtained can also be brought to bear on the two 3 × 3 included in the current project. We can combine the generalized results of the Cooperative and Coordinative games and extend these insights to the Lane Choice and Biased Lane Choice games. Since these games combine aspects of both game forms, the forces of the transition matrices in these games would have a key feature in common: assortativity.

Recall that each types' dynamics are considered independent for purposes of calculation. The A-types in this 3×3 game have 1) an $\frac{N_A - 1}{N-1}$ probability of connecting to A-types, 2) will be detached by C-types in configurations with C-types (but not by B-types), and 3) will detach B- and C-types only if connected to other A-types. So with respect to achieving prosociality we can compare the processes to existing ones by considering types bundled together by behavior.

For A-types here, the random attachment process is similar to $\mathcal{R}_A$ since the probability of gaining an A-type neighbor is the same when we consider B- and C- types together as the "other" type. The detachment by C-types is similar to $\mathcal{Y}\beta$ with A- and B-types both considered as As and the C-types taking the place of the B-types in that matrix. Finally the agent actions are similar to $\mathcal{X}_{\phi_1}$ if we again consider B- and C- types together as the "other" type.

The matrices are similar, but not the same. For one, the detachment force decreases because the increased number of configurations available with three agent types means that there are more configurations (in number and in proportion) in which A-types will be attached to non-preferred types (because this now includes all configurations with combinations of B- and C-types). Instead of just two stable configurations for A-types, there is now the configurations $Aabc$ with $a = K$ as well as all the configurations in which $b + c = K$. The probability of reaching one of these latter configurations is still small compared to reaching the prosocial one, but larger than in the 2×2 games. The A-types are, of course, completely symmetrical to the A-types, even in the Biased Lane Choice, because the preference rankings are the same for all agent types. And finally, the defectors in these games will still have a positive recurrent combined transition matrix and hence have a population distributed among all the configurations.

2.7 Summary of the Markov Model

The Markov model reveals trends in agent connections resulting from the three separate parts of the preferential detachment process: being random attached to, being detached, and individual-based behavior to connect, detach, or do nothing. This is done by tracking the proportion of agents in each configuration (or the probability for any agent that it will be in that configuration) across iterations of multiplying the current distribution by each behavior's transition matrix. The multiplication process is repeated until a stationary distribution of configurations is reached (which may be an equilibrium or sustained distribution of churning agents).

The Markov model captures all the individual-based features of the preferential detachment theory, but it fails to keep track of the particulars of the social arrangement. Because of this, being detached and detaching are accounted for separately, even though being detached is just some other agent's detaching. Furthermore, an inability to accommodate the full social structure means that some transitions are probabilistic because the change depends on features (such as a neighbors' neighbors) that are not known. Uniform distributions over possible social arrangements provide a reasonable approximation, but the transition probabilities are static through configuration distributions that render those probabilities impossible. These inadequacies, however, exert only a small effect in comparison to the behaviors that are appropriately captured by the Markov modeling technique.

The results reveal that the prosocial outcomes described in chapter I are achieved completely in Coordinative, Specialized, Commensal, and Undifferentiated games and to a very high degree in Cooperative and Contributive games. With a maximum degree of four and one-hundred agents of each type, we find that six and a half percent of the A-types in Cooperative and Contributive get stuck in stable heterogeneous groups; or equivalently that there is a 6.5% chance for A-types to collect four B-type neighbors before becoming attached to any other A-types. The precise quantitative values are not the focus. What is more elucidating is the pattern in outcomes generated by the preferential detachment process; i.e., how the same rule applied in each context produces a distinct and appropriate outcome distribution.

The two questions remaining are: How sensitive is the analysis to changes in the parameters used in the above examples? and To what degree do the system-level assumptions affect the outcome distributions? The latter questions is left to the agent based model of the next chapter. The former question can be answered by solving for outcomes using the generating functions for the transition probabilities, but it is more insightful to reveal the salient patterns.

The random connections follow a binomial probability distribution for the number of connections which may be added. These probabilities are a function of the population levels of each type: for equal populations the probabilities are unchanged, and for unequal populations there is an increasing probability of connecting to the relatively more populous type. These increasing probabilities converge to adding 1,2,. . . ,K agents distributed binomially of the more prevalent type. Increasing the maximum degree expands the range of possible configurations, but has no effect on the rate at which agents gain new neighbors.

As K increases the probability of detaching less-preferred neighbors increases because there is a greater probability of having preferred-type neighbors as well. The same reasoning applies to being detached by agents who prefer the focal type less. Detachment depends only on an agent's configuration, so it is insensitive to population changes. Individual attachment depends on the proportion of each type available, so equal increases changes nothing and unequal changes produce linear transition probability changes.

2.8 Next Steps

The three major limitations of this representation of the preferential detachment mechanism are 1) the static transition probabilities, 2) the inability to access specific network structures, and 3) the lack of support for the information needed for population dynamics or imitation-based learning (see below). Because the likelihood of certain transitions should change depending on the number of agents in each configuration, the magnitudes of the probabilities here are approximations to the theorized underlying process. If including information about the number of agents in each category only changes the magnitude of the probabilities slightly (without altering the structure or patterns of the transition matrix) then including this detail will only affect the trajectory of the dynamics and not the final stationary distributions. As already indicated in the discussion of the model results and in the generalizations of strategic contexts, there are outcomes excluding mixed configurations that we expect to persist, and these are due to a combination of (1) and (2).

A partial solution limitation (3) is to include type-changing behavior (e.g., mutation, learning, population dynamics) within this modeling technique by specifying a transition matrix

$$M = \begin{pmatrix} \mathcal{A} & \mathcal{L}_A \\ \mathcal{L}_B & \mathcal{B} \end{pmatrix}$$

in which $\mathcal{L}_i$ is a block matrix of transition probabilities from type i to the other type. It would also be possible to make the transition probabilities a function of the numbers in each category. But doing either of these is putting a round peg in a square hole. The processes would still be approximations, albeit closer approximations, and the non-ergodicity of the matrices along with the lack of structural information would still limit the value of such a model.

Adaptive transition rates and population dynamics could be addressed using a difference or differential equation model that has a detail level in between this Markov model and an agent-based model. Specifically, the agents in each configuration are the variables, and there is an equation specifying the rate of change in the number of agents in each configuration as a function of the number in each configuration. This creates a system with $\mathbb{C}$ equations and $\mathbb{C}$ unknowns. This system has the same matrix form as the Markov model, but the transition probabilities (the coefficients) are now functions of the values of the number of agents in each configuration so they can change appropriately to reflect changes in the system arrangement. Solving these systems of equations would also foster comparisons to some of the other research in the field, and make explicit what the basins of attraction are for different outcomes.

If the details of network structure significantly impact the preferential detachment dynamics then neither this Markov model nor a difference equation model will be able to capture those effects. This could be the case for many reasons including 1) the presence of certain low-probability community configurations being persistent, 2) variations in structural characteristics (e.g., clustering, betweenness, component size) affecting agent action opportunities, and/or 3) subtle path-dependencies generated through the random connection mechanism. Furthermore, including the full network structure facilitates the comparison of neighbor satisfaction levels and hence allows for imitation-based learning. The agent-based model presented in the next chapter includes all of these features. Even with the results of the agent-based models in hand, this Markov model provides useful insight by identifying an underlying causal tendency in the dynamics produced by the preferential detachment mechanism. Comparing the results of multiple models will indicate the benefits of a more complete model as well as highlight the usefulness of an approximate mathematical formulation.

CHAPTER III

Agent-Based Model of Preferential Detachment

We now turn to the description and analysis of an agent-based model of preferential detachment. As described in chapter I the theory of preferential detachment is that a particular mechanism (along with other environmental and social conditions) drives the behavior of agents into social structures that promote the interest of the agents in them. This chapter demonstrates that an agent-based model of the preferential detachment theory produces the prosocial outcome in all the strategic contexts presented in chapter I. It also demonstrates the robustness of the mechanism against variations in the two main parameters (maximum degree and total population) of the model. The effects of adding imitative learning and population dynamics are then examined in detail – elucidating these effects is especially important for comparing preferential detachment with previous work in the field as will be done at the end of this chapter.

Included below are descriptions of 1) the parameters used, 2) the agent properties and their behaviors, 3) the measures taken of system dynamics and outcomes, 4) the results from two batteries of experiments, and 5) an analysis of the model itself with respect to those results. The purpose of this description is to facilitate an understanding of exactly what has been simulated and analyzed insofar as this bears on the interpretive questions raised in chapter IV. The details are described sufficiently to allow an independent researcher to be able to recreate the computer model. Though the model is simple by design, and operationalizes the theory of preferential detachment presented in chapter I with as much fidelity as possible, it is necessary to make some additional assumptions in the construction process and these are explained and justified.

3.1 Model Procedure Overview

Before digging into the fine details of the model I will now describe the workflow of the process. This treatment acts as an outline that should help the reader understand the big picture while I detail individual components below.

Throughout each run of the model the game structure, total population size, maximum degree, and all other parameters are held constant. The number of each agent type can change via learning and population dynamics; and obviously the agents' utility, current degree, and other measures can change throughout a simulation. All of the experiments discussed here utilize a simultaneous updating rule, so all agents' behavior in period t is based on the same state of the world (as it was at the end of period $t-1$) with the exception of the bookkeeping necessary to ensure agents do not exceed the maximum degree value. What this means is that a pair of connected agents can both "decide" to disconnect the same connection (even though the relation can be unilaterally broken) because they are making their decision of what behavior to enact based on the same arrangement.

The model is initialized with an unconnected population of each relevant type of agent (e.g., two types for 2×2 games). Each iteration, agents assess the payoffs received from each network neighbor. Any agent earning differing payoffs from its neighbors will detach one least preferred agent.

If learning is activated then the agents compare the cardinal utilities of their network neighbors to their own. These utilities equal the sum of the realized payoffs of table 1.5 over the set of interaction partners. If any neighbor has a higher total utility then the focal agent changes its type to match the type of one of its neighbors with the highest total payoff.

If an agent is indifferent among all its connections (equally preferred or it is unconnected), and its degree is less than the maximum degree, then it will randomly connect to a new agent from the pool of agents that have also not reached the maximum degree (if any exist).

If population dynamics are activated then at the end of each iteration a pool of agents numbering five percent of the total population are chosen from the highest payoff agents for replication. The lowest five percent are eliminated. New agents are introduced with types matching the distribution of types in the top agents, and they start with no connections.

After all the behaviors are completed the model collects various metrics of the agents and their social structure. These procedures repeat until no agent can act to improve its configuration, or until system behavior is locked in a cycle.

3.2 Parameters

Here I describe the parameters of the model and where they leverage into the model. I will leave the parameters for the visualization out of this discussion and leave the parameters of the metrics for the discussion in section 3.4.6. A summary of the values typically used for experiments is included in table 3.1, and deviations from these values will be described with the experiments in section 3.5 below.

parameter	values
num-typeA-E	200/#*types*
max-degree	5
population-dynamics?	true & false
learning?	true & false
turnover-rate	5%
game	10 payoff matrices

Table 3.1: Summary of agent-based model parameters.

Parameter 3.2.1. *num-typeA* ... *num-typeE*: The initial number of agents for each type. The number of distinct types equals the number of rows/columns in the payoff matrix used in the given experiment (e.g., two types for 2×2 games). The proportions of agent types can change through learning or population dynamics, but the total population is fixed. See section 1.3 for more details and motivations. Except for the experiments exploring the effects of different population sizes, a value of 200/#*types* is used (rounded to the nearest integer).

Parameter 3.2.2. *max-degree*: The maximum number of interaction partners that an agent may have at a time; also referred to as *max-edges* or K. The degree of any given agent will always be between 0 and *max-degree*. See section 1.4 for more details and motivations. Except for the experiments testing sensitivity to the maximum degree, this value is set to *five*.

Parameter 3.2.3. *population-dynamics?*: A Boolean variable for whether the population dynamics procedure is activated or not. Every parameter combination is experimented with using both *true* and *false*.

Parameter 3.2.4. *learning?*: A Boolean variable for whether the learning procedure is activated or not. See section 1.7 for more details on learning and population dynamics. Every parameter combination is experimented with using both *true* and *false*.

Parameter 3.2.5. *turnover-rate*: The percentage of the total population that is replicated and removed each iteration when population dynamics is activated. In all the experiments with population dynamics *five percent* of the total population is replicated and removed each iteration.

Parameter 3.2.6. *game*: The payoff structure that determines the utility gains for each agent for each possible combination of action types. Agents receive payoffs from each connected agent according to the current types of both agents. Agents receive both the player-one payoff and the player-two payoff for each interaction, which only has an effect for non-symmetric games as discussed in section 1.6.

3.3 Model Setup and Time 0

The model is initialized with *num-typeA* ... *num-typeE* agents for each type in *game*. No agents have any connections to other agents (which gives them all utility zero). Since no agents have any neighbors to disconnect, all agents will connect to a uniformly randomly chosen other agent from the whole pool of agents.

With true simultaneous updating random connections could result in more than *max degree* agents trying to connect to some one agent. This constraint violation is avoided through the use of simulated simultaneity. In simulated simultaneity the agents use the same reference system state (the beginning of the iteration) to make decisions, but each behavior is computed sequentially. This allows the bookkeeping for available connections to be done during each agent's action and ensure the maximum degree constraint is not violated. This is less problematic than deleting, and more parsimonious than rewiring, any extra edges. Detachment, learning, and population dynamics are all effectively simultaneous, but random connections done in the way just described are effectively done with agents chosen randomly without replacement.

3.4 Model Iterations

In this section I detail the steps to the preferential detachment process that occur during each iteration (aka *tick* or *time-step* or *t*) of the model.

3.4.1 Detachment Step

The first step is that all the agents reset their utility values and update their list of connected agents. Then each agent with network neighbors assesses the payoffs received from each neighbor. Agents do this by collecting a list of the payoffs they receive as Player1 in the game and a separate list of the payoffs they receive as Player2. These payoff lists are then summed element-by-element making a combined list of payoffs from each neighbor (the elements of the combined list are the realized payoffs of 1.5). If the maximum entry in the combined list is greater than the minimum entry then an agent will be detached. The detached connection is chosen as the neighbor providing utility equal to the combined list's minimum-value entry and with the lowest agent ID number.[1]

Edges are not actually removed until each agent has had an opportunity to evaluate its current (end of iteration $t-1$) neighbors. So it is possible for both ends of a link to choose that link for disconnection, but at the end of this step a link is eliminated if either agent has chosen to detach it. Every agent that has chosen to detach a neighbor is marked as having acted this iteration. Then each agent's total utility is calculated as the sum of the remaining elements of the combined payoff list for use in learning and population dynamics.

3.4.2 Learning Step

The current model only explores one learning rule: success-based imitation (recall section 1.7 covering other learning techniques). Each agent compares their own total utility to the total utility levels of all its neighbors. If any of an agent's neighbors is earning greater utility than itself this iteration (after detachment), then the focal agent collects the set of neighbors with the greatest utility value. If multiple neighbors share the greatest utility level (higher than itself) then the imitated type is selected uniformly at random among the set of greatest-utility neighboring agents. Agents that have changed their type through this process are marked as having acted; thus imitating a higher-payoff agent of the same type does not count as acting.

[1] Unique ID numbers are assigned to each agent when they are created and, since connections and the agent activation order are random, are essentially arbitrary with respect to an agent's list of neighbors. This only has a biasing effect when there are three types of agents, and the topic is addressed when the Lane Choice game is analyzed in section 3.9.7 below.

3.4.3 Random Connection Step

Agents that have not yet acted (neither detached nor changed type through imitation) will form a new connection to an agent picked randomly from the set of agents whose degree is less than the maximum allowed degree. As described above in the model setup (3.3), the behavior rule for connections effectively works as random without replacement to ensure no more than the allowed number of connections are ever formed. For each iteration a different random draw is performed for connection-making – shuffling the connection order minimizes any "first mover" bias.

3.4.4 Population Dynamics Step

The final step within each iteration is to have the agents undergo population dynamics. To do this the model identifies a number of agents equaling five percent of the total population (rounded to the nearest integer) among those with the highest and lowest total utility values. If the five percent mark cuts through a set of agents with equal utility, then the appropriate number to make up a total of five percent of the whole population is picked randomly from that marginal set. The worst five percent agents and their connections are removed from the model. Each of the best five percent creates a new agent with the same type of itself and no initial connections. Note that the agents chosen for removal and replication are not affected by learning or random connections made during this iteration because the agents' utility values are not recalculated during those steps.

3.4.5 Visualization Update Step

At this point the network visualization for the graphical user interface is updated with the current agent types and current edges. The layout of the nodes and edges are also updated using Netlogo's built-in spring-force/repulsion algorithm. Utility values are represented by node size – greater utility produces larger nodes. Colors are updated to match agent types according to the scheme in table 3.5. We will later see the results of these visualization rules in the screenshots presented in the results section. No aspect of the model – neither agent behavior nor measure calculation – is effected by the visualization, so during batch runs the visualization procedures are switched off to save processor effort.

3.4.6 Measures

The final operation is that all the measures of agent behavior, agent characteristics, and collective properties are evaluated. More aspects of agent behavior and properties are tracked than are presented here; I will focus on the measures specifically relevant to the topic of the evolution of prosocial behavior. These include properties for tracking group formation, the make-up of these groups, and their fitness performance. All measures are collected *by type*, and whole-system properties will be inferred from cross-comparisons in the results section.

Measure 3.4.1. Population of Each Type: Recall that under learning and population dynamics the number of agents of each type can change, though the total population is fixed. Achieving the prosocial outcome in specific strategic contexts implies specific dynamics for the population levels of each agent type (e.g., cooperators dominating or coordinators staying balanced). Population figures are plotted as percentages to ease comparison even through the battery of experiments exploring changes in the population levels in Prisoners' Dilemma.

Measure 3.4.2. Mean Percent Similar of Neighbors: Each agent records the percent of its neighbors which have the same type as itself, and the model plots the mean of these values over time. This is a measure of assortativity which is key to many of the predictions and explanations of prosocial behavior.

Measure 3.4.3. Mean Percent Similar of Max Neighbors: While the measure just described above records how assortative agents of each type are, this measure captures utilization. Each agent counts the number of same-type neighbors, these individual measures are averaged across the agents, and the result is divided by the maximum degree. When this figure is less than one it means that at least some agents of this type cannot maintain connections to agents of the same type and/or are maintaining connections to other-type agents.

Measure 3.4.4. Mean Utility: This is the straightforward summation of the total utility values by type divided by the number of agents of that type. The dynamics of this measure are interesting for the analysis of imitation and replicating vs detachment discussed in section 1.7. Mean utility is also helpful in determining whether agents are stuck in suboptimal groups.

Measure 3.4.5. Mean Clustering Coefficient: The clustering coefficient measures what percentage of one's network neighbor pairs are also connected to each other. It has been observed that social networks are hierarchal when subgroups are defined by the degree of clustering among nodes. Measuring clustering allows us to determine if groups of same-types agents are amalgamations of tightly connected subgroups, or a linear accumulation of agents. The aggregated measures for each type fosters comparison of group dynamics between behavior roles. Note that in preferential detachment agents disconnect without any information regarding their neighbors' neighbors and connect randomly. Thus the level of clustering among agents is an indirect effect of the behavior rule.

Measure 3.4.6. Number of Components: A component of a network is set of nodes such that there exists a path along edges connecting them. Here I individuate components by type, so each set of contiguous agents of the same type counts as a component. These components are what I take as type-specific *groups* for analysis and interpretation. Although I will also talk about heterogeneous groups, these are not measured directly.

Measure 3.4.7. Largest Component Size: Because a component can be a single agent or all the agents of a type, the number of components is by itself not enough to understand group dynamics. So in addition to the number of components the model tracks the number of agents in the largest component as a percentage of the agents of that type in the population. Together these figures allow us to understand whether and when a core group of agents develops, whether and how many agents are trapped in marginal groups, and whether and when groups tend to fuse or split.

Measure 3.4.8. Mixed Agents: This variable holds the number of agents with neighbors of a different type than itself. It is a direct derivative of the percent similar measure and mainly used for the halting condition discussed below.

Measure 3.4.9. Mixed Agent Moving Average: A moving average of the number of mixed agents recorded over 300 ticks is also recorded. The volatility of this moving average is an indicator for how settled the model's dynamics have become and is also mainly used for the halting conditions immediately to follow.

3.4.7 Halting Conditions

When model behaviors or properties match any of the halting conditions the simulation stops running. The tick at which this happens can (and often does) vary from run to run. The idea behind halting conditions is that collecting further data will not enhance our understanding of the model. The durations for which conditions must hold are to ensure that a premature satisfaction of the condition is highly improbable.

Condition 3.4.10. Stasis: If the set of edges is constant and no agent acts (detaches, connects, or imitates) for ten consecutive iterations then the system behavior has reached an equilibrium and the run stops.

Condition 3.4.11. Dominance: If the entire population is comprised of agents of a single type then the run stops. The dominance of one type of agent entails impending stasis for the games considered.

Condition 3.4.12. Learning Cycle: If learning is activated and no agents have changed types for twenty consecutive iterations then the run stops. This condition handles cycles that occur under learning behavior when there are insufficient openings within a same-type group. An agent in this situation makes a connection to an opposite-type agent, disconnects it, then randomly makes another connection, but only other-type agents are available. Since agents are not changing types at this point, this cycle would continue indefinitely without effecting the general network structure or prosociality.

Condition 3.4.13. Population Convergence: If population dynamics are activated and the number of mixed agents is fewer than five percent of the population for five consecutive iterations then the model stops. The birth/death process ensures that there are always new agents appearing, connecting, and making available connections to them; thus stasis is impossible. The population convergence condition obtains when the agents have settled into separated self-maintaining groups. At this point only the newly introduced agents could be connecting to agents of a different type. Since the number of new agents per iteration is five percent of the population, if only five percent or fewer agents are mixed for multiple consecutive iterations then (except for these agents) the population is very likely assortatively mixed.

Condition 3.4.14. Mixed Agent Convergence: If neither learning nor population dynamics are activated then it is possible for agents to get trapped into groups of heterogeneously typed agents. They are "trapped" because all the other agents have filled their network neighborhood and reached stable configurations. In these mixed group arrangements it is typically the case that random edges are made among agents within the group, and then quickly eliminated. This process repeats indefinitely because there will always be agents of both types with available connections, thus stasis will never be achieved. But the set of agents undergoing this connection-detachment process is quite stable. As mentioned above, the model records the moving average of the number of mixed agents over the preceding 300 iterations. If the difference between a) the moving average and b) the current number of mixed agents is less than three agents for twenty consecutive iterations, then the model stops.

Condition 3.4.15. Timed Out: If the model runs for 2000 iterations the model stops. This is the catch-all halting condition in case no measurable features of the model indicate that the prosocially-relevant dynamics have run their course. As it happens, this condition is only utilized for the Lichen game which we expect to undergo persistent system-wide cyclic behavior.

Even with all these halting conditions, it sometimes happens that a simulation continues past the time for which it is informative. For example, simulations often require several hundred iterations before mixed-agent convergence is satisfied, yet this is also sometimes several hundred iterations beyond the time when the last important structural modification happens. For this reason, some plots presented in the results section are truncated in order to improve the clarity within the temporal range of scientific value. Plots are only truncated when the remaining data exhibits no further systematic variation

3.5 Experiments

We now have all the details on the procedures of the model, enough to understand exactly what the model does. An experiment is identified by a specific parameter combination. All experiments are run through 100 simulations using unique random seeds.

Because the Prisoners' Dilemma is so popular in the literature, and because it contains many of the interesting features of prosocial strategic situations, I perform a battery of experiments including a sweep of the two major parameters: population size and maximum degree. These experiments parallel the generalizations made in the Markov model, and the results here will be compared to the mathematical results. These parameters are not swept for the other strategic contexts, but the results from the Markov model, combined with the results in this battery of experiments, support arguments that such sweeps are unnecessary.

3.5.1 Branch One: Sweeping Population Size

I perform twenty-eight experiments with the Prisoners' Dilemma. This includes varying the population level evenly for both agent types at seven levels from 10 to 1000 (see table 3.2). Each of the seven population levels is tested with and without learning and population dynamics. As per table 3.1 all of these experiments use a maximum degree of five and a turnover rate of five percent for population dynamics.

num-typeA	*num-typeB*
10	10
25	25
50	50
100	100
250	250
500	500
1000	1000

Table 3.2: Population size sensitivity experiments.

Among other things, these experiments seek to determine whether the success of the mechanism scales well with population size, and whether there is a minimum population size required for the preferential detachment mechanism to achieve the prosocial outcome (cf. [Helbing and Yu 2009]). The analysis of experiments with uneven populations is left for future work.

3.5.2 Branch Two: Sweeping Maximum Degree

Six values for maximum degree (see table 3.3) are tested with learning and population dynamics both on and off – creating twenty-four experiments. These experiments use 100 agents of each type and the other default parameters. Because a maximum degree of five is low on the theoretically proposed range of values for many species, this battery of experiments tests whether achieving the prosocial outcome is sensitive to this parameter.

max-degree
3
4
5
10
15
20

Table 3.3: Maximum degree sensitivity experiments.

The results of the Markov model indicate that achieving the prosocial outcomes should be easier (faster convergence and a greater percentage of benefiting agents) for higher maximum degree values. These experiments will investigate a validation of that result, as well as the affect on other measures of social structure unavailable with the Markov model.

3.5.3 Branch Three: Sweeping Strategic Contexts

The third branch of experiments uses two hundred total agents divided equally among the types involved and a maximum degree of five. Eight of the ten games are 2×2 games so initially there are 100 of each. The two Lane Choice games use 67 agents of each of the three types. For this branch of experiments the parameter being swept is the payoff matrix itself. For ease of references I present those payoff matrices again here. Other parameters are set to the defaults already described. There are forty experiments in this branch if we including the ones redundant with experiments from branches one and two. As always, each experiment is run 100 times with a unique random seed.

Without learning or population dynamics the results should be identical within the six game categories of table 1.6, and getting such a result helps to verify the model's operation. With learning and/or population dynamics we still expect to see the prosocial outcome reliably obtain, but the dynamics and social measures may differ in significant and informative ways. This applies to comparisons of Prisoners' Dilemma to Hawk and Dove, and Battle of the Sexes to Coordination Game. In all cases we are looking for the level of success in achieving prosociality and to the particular details of the system-level arrangements.

Prisoners' Dilemma

		player2	
		A	B
player1	A	3,3	1,4
	B	4,1	2,2

Hawk and Dove

		player2	
		A	B
player1	A	3,3	2,4
	B	4,2	1,1

Battle of the Sexes

		player2	
		A	B
player1	A	4,3	2,1
	B	1,2	3,4

Coordination Game

		player2	
		A	B
player1	A	4,4	2,2
	B	2,2	4,4

Lichen

		player2	
		A	B
player1	A	1,1	3,3
	B	3,3	1,1

Stag Hunt

		player2	
		A	B
player1	A	4,4	1,2
	B	2,1	2,2

Commensal

		player2	
		A	B
player1	A	1,1	4,2
	B	2,4	2,2

Matching Pennies

		player2	
		A	B
player1	A	3,1	1,3
	B	1,3	3,1

Lane Choice

			player2	
		A	B	C
player1	A	3.3	1,4	1,1
	B	4,1	2,2	4,1
	C	1,1	1,4	3,3

Biased Lane Choice

			player2	
		A	B	C
player1	A	3.3	1,5	1,1
	B	5,1	2,2	5,1
	C	1,1	1,5	4,4

Table 3.4: Game Library

3.6 Analysis Techniques

Before presenting the results of the experiments, I will first explain the analysis and visualization techniques used to describe those results. Once these are clear I move to summarize and show highlights of the experimental results. Because seven measures are tracked over 100 runs for each of 84 unique experiments there is too much output to present it all (the complete set of output plots is available as supplementary material). Furthermore, much of it is uninteresting and/or irrelevant to the questions regarding the evolution of prosocial behavior.

Methodology for analyzing agent-based models is still in its infancy. Statistical treatments typically used to aggregate and categorize datasets are often inappropriate for understanding and explaining the behavior of agent-based models because they smooth out exactly those features we wish to explore with this modeling approach. These features include path sensitivities, "black swan" events, multi-modal outcomes, threshold phenomena, and the gamut of behaviors falling under the name "complexity". The agent based model for preferential detachment is quite simple in terms of parameters and measurable features, but even this model poses formidable challenges in capturing, analyzing, and comparing the results in insightful ways. Now I describe the techniques used to describe behavior, compare dynamics within and between experiments, and uncover complex features of preferential detachment.

3.6.1 Measures of System Properties

Seven measures of tick-by-tick agent and system properties were presented in section 3.4.6. These measures produce time-series data that are plotted to reveal changes over time for each run (as well as the mean over runs) in each experiment. Such plots are useful for determining whether some variable converges over time, fluctuates periodically or sporadically, and other straightforward features reflecting to what degree prosociality has resulted. For example, in the Prisoners' Dilemma with learning and/or population dynamics the prosocial outcome of the model is for cooperators to dominate the population. So if the percent of A-type agents rises to one then we have a clear answer as to whether the prosocial outcome occurred. With learning or population dynamics we can often evaluate the prosociality of the outcome by simply reading off the proportions of each type of agent.

Given an interpretation of the model appropriate for the evolution of social norms, we desire to avoid outcomes in which (for example) defectors are exploiting cooperators (in stable heterogeneous configurations). And for any of the games for which the prosocial outcome was identified as assortatively mixed agents we can measure the success of the experiment as having a %-similar-neighbors as high as possible for both types of agents. By identifying the target outcome this way (instead of through aggregate utility measures), I avoid the problem of defining a social utility measure. I also collect the mean utility values for each type. These are helpful for understanding learning and selection dynamics, but they also foster a comparison to other models. The prosocial outcomes defined in my behavior-based approach (see table 1.10) coincide with higher social utility defined as the sum of the realized payoffs, but my identifications require more of the social configurations. (as described in section 1.9).

Some complex behaviors cannot be practically measured directly, but can be inferred from a combination of the measures collected. We can infer aspects of the overall network structure by looking at the number of components and the percent of agents in each component. If there are few components, and most agents of a type are in the largest component, then that implies that the remaining agents of that type are separated from the dominant group. This could mean that they are unconnected, or connected to other types, or connected to similar types in a minor tightly knit group. The % of max degree, % similar, and clustering coefficient measures can then be analyzed to distinguish which is the case.

These sorts of detections would be very difficult to automate, and doing so would significantly impede the model's performance. There are too many different ways to achieve all the possible configurations and phenomena we may be interested in. By examining all the model output with methods such as those just described I am able to categorize, summarize, and explain model behavior in appreciable detail. But there is too much to cover it all, so I provide summary tables, expose similarities, and highlight interesting cases. All of this is done with prosociality in mind. The complete collection of output plots is available as supplementary material for confirmation and deeper analysis.

3.6.2 Statistical Treatment

Much of the value from this research comes in the comparison of agents of different types within each game. Furthermore, we will also want to compare and contrast the dynamics of agents across games to ascertain what effects the game has on social features. The plots (presented below) facilitate easy ocular analysis of many features of system dynamics (i.e., what happened in the experiment) such as whether the means of two data streams are roughly equivalent or which is greater. But variation among the runs of an experiment can be highly variable. Due to this variability two collections of data streams may have dissimilar behavior and yet have similar means. Furthermore, the differences between the two data collections may appear similar in the plot, but be insufficient to claim the processes are distinct. We need to be able to test whether the differences between two data samples are significant.

The standard approach to perform this task relies on assumptions regarding the shape of the data's underlying distribution – typically that it is normally distributed. But we know that for some of my measures, in some situations, the distribution will certainly not be normal. And, for example, when the run data has a fat tailed and/or one-sided distribution, the mean and standard deviation yield poor and inappropriate significance tests. Not only is it a priori unclear what the distribution of a measure will be across runs, it is likely that the shape of the distribution will change through the course of a run. For all these reasons a non-distributional, nonparametric technique is better suited to evaluate the differences among agent type dynamics and among the various experiments.

One statistical test that is particularly useful for comparing two data samples of this variety is the Kolmogorov-Smirnov two-sample test (K-S test). This technique determines the probability that two observed samples could have come from the same distribution, but without specifying what that generating distribution might be. The first step of the test is to generate the empirical cumulative distribution functions (CDF) for both samples. The y-value of the function at $x = X$ describes what proportion of the data have x-values lower than X. This is essentially a histogram of the data where the bars are stacked as we move to greater x-values (see figure 3.1).

The test statistic for the K-S test is the greatest difference between the two empirical distribution functions (D in the figure). The null hypothesis is that the two samples are from the same continuous distribution, and so the alternative hypothesis is that they are from different continuous distributions. From this test statistic, and the sizes of the samples used, we calculate the probability of observing that distance. The end result is that we can determine the likelihood that our two data samples are distinct [Kaner et al. 1980, Wilcox 1997].[2]

In practice what this means is that I can determine whether and when the measure values from A-Type and B-Type agents are significantly different. The K-S p-value (the probability that the two data sets are the same) is calculated for every iteration and displayed on the plots along with the data. This does not report whether the two

[2] I would like to thank Ray Koopman of Simon Fraser University in Burnaby, British Columbia for concise, procedural Mathematica code to implement the Kolgomorov-Smirnov two-sample test. Koopman's code is an implementation of Thomas Waterhouse's original code written in R.

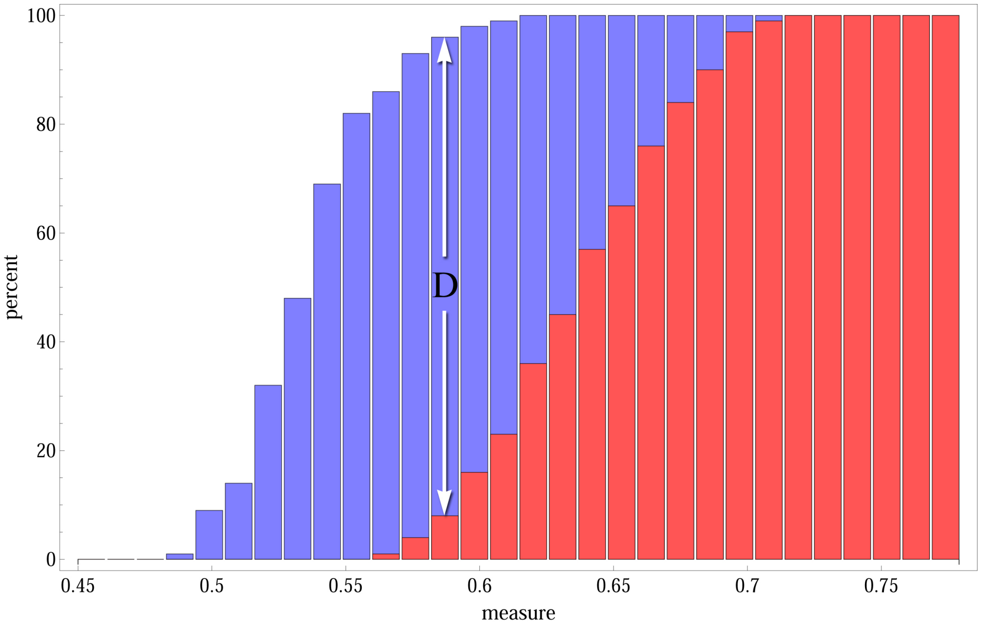


Figure 3.1: Empirical CDFs of a measure at one time slice.

data *streams* were generated by the same process, but by combing these tick-by-tick values we can determine the temporal range at which the differences between the types is revealed as differences in measures.

In some situations (e.g., the Coordination Game) we will test the hypothesis from preferential detachment theory that the types' behaviors are **not** significantly distinct. Other nonparametric techniques (such as the Wilcoxon signed-rank test, MannWhitney U test, and the Siegel-Tukey variance comparison) could also be applied to the data, but the K-S test suffices to demonstrate all the system features we are immediately interested in.

There is one final point worth highlighting for clarity. In many plots you will see the K-S p-value rises sharply to insignificant values near the end of an experiment's time frame. This (typically) results from the fact that runs take different numbers of iterations to complete. As the number of data points per tick decreases, the confidence level of whether two data sets are distinct decreases. This is not because the distributions are getting closer, but because the significance levels are sensitive to the samples sizes. In some comparisons we will find it helpful to truncate the data sets for ease of comparison, and the K-S statistic is helpful in determining appropriate cut points.

3.6.3 Visualization

Each measure collected during simulation produces its own output plot. The measures are recorded at the end of every iteration for each of the one hundred runs. This unaggregated data streams per run are plotted as thin lines with time on the x-axis and the value of the measure on the y-axis. The color of a line represents the agent type to which it corresponds according to the color scheme in table 3.5.

A-Type	Blue
B-Type	Red
C-Type	Green

Table 3.5: Agent type color encoding.

The mean of the data from individual runs is plotted as a thick, partially transparent line of the corresponding color. In addition, the K-S p-values are plotted on the same diagrams with the data. They appear as a thin line that is black in regions where it is significant at the 95% confidence level, and orange otherwise (see the plot reading guide in figure 3.2).

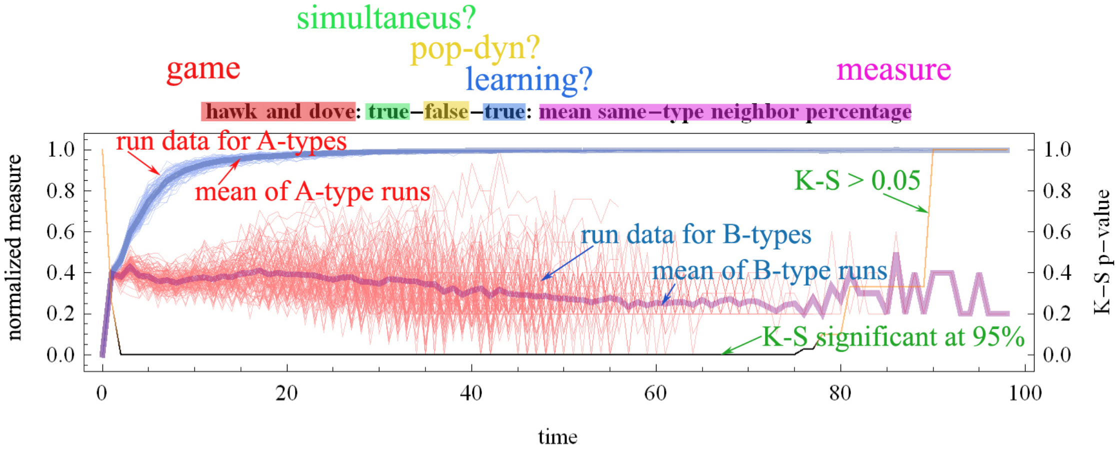
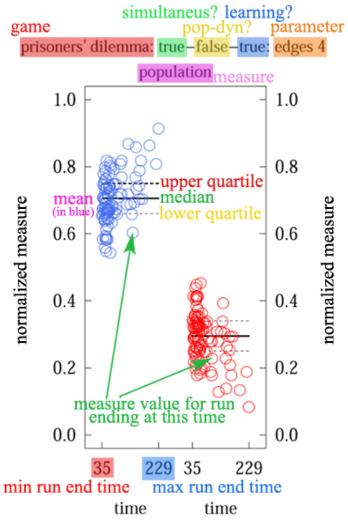


Figure 3.2: Behavior Plot Reading Guide

This plotting technique fosters the quick intuitive grasp of 1) typical behavior by situation, 2) variance in behaviors for each variable, 3) comparisons among agent types, and 4) uncommon fluctuations in measures. Overlaying the K-S p-value allows us to also see immediately when differences between the types indicate genuinely distinct behaviors. The full set of these plots is produced for all experiments and used to determine the outcomes with respect to prosociality, and to detect occurrences worthy of more in-depth analysis. Some of these analyses are performed using a similar plotting/analysis technique (adapted to a different comparison); some analyses require different approaches that are discussed below where appropriate.

In addition to the plots of simulation behavior across time, I present results as the measure variables' value at the final iteration of each run (see 3.3 for an example and legend of plot features). The outcome results are presented in a scatter plot with rings indicating the final value on the y-axis and the time at which the run ended on the x-axis. These plots therefore summarize information about the run time of the simulation and whether that correlates with the value of the measure presented. In addition to the scatter plot, standard box-and-whisker information is provided via horizontal lines at the appropriate values. Precise values for the summary statistics are presented in a box under the plot for careful evaluation and comparison.



Mean:	0.706	0.294
Maximum:	0.915	0.455
Upper Quartile:	0.750	0.340
Median:	0.705	0.295
Lower Quartile:	0.660	0.250
Minimum:	0.545	0.085

Figure 3.3: Summary Plot Reading Guide

Finally, there are occasions when a screenshot of the simulation environment is the simplest way to describe a phenomenon under study. The node colors also match agent types according to the scheme in table 3.5. Typically, the size of a node represents the utility value of that agent; however, there are occasions when it is more illuminating to represent another measures using the node size.

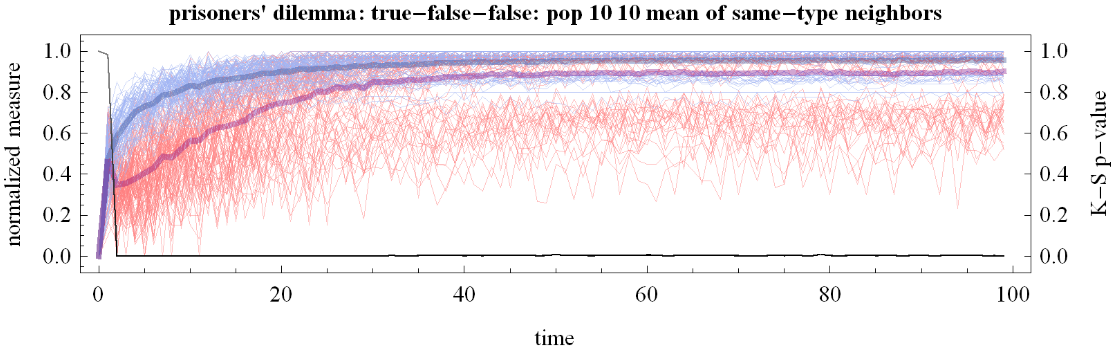
3.7 Branch One Results: Sweeping Population

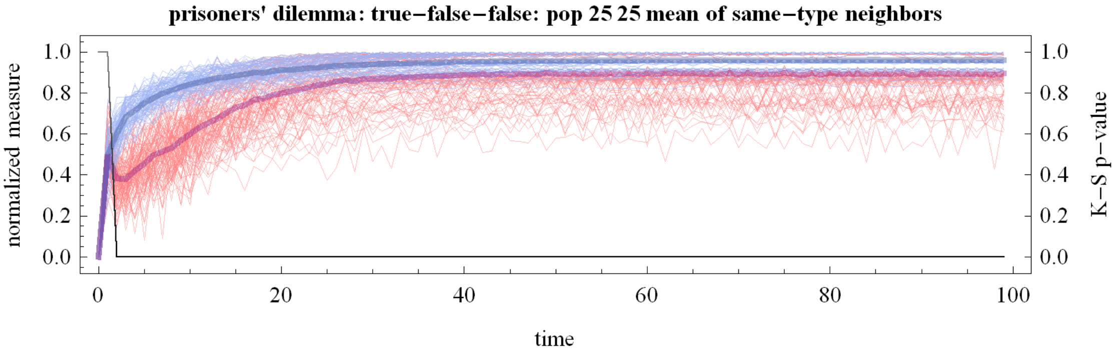
This section describes what happened in the simulations, but leaves the interpretation and impact of those results for section 3.10. Separating the tasks in this way allows me to systematically cover the system behavior for each situation here, and then later discuss the more substantive issues requiring the comparison of particular results. As mentioned earlier, the complete output is too vast (588 plots) to display all of it, so the presentation here includes only the results that bear on the questions of the evolution of prosociality. Some results are described without the corresponding plots embedded in the text; you may refer to the Supplementary Material containing the complete collection of output plots for more detail.

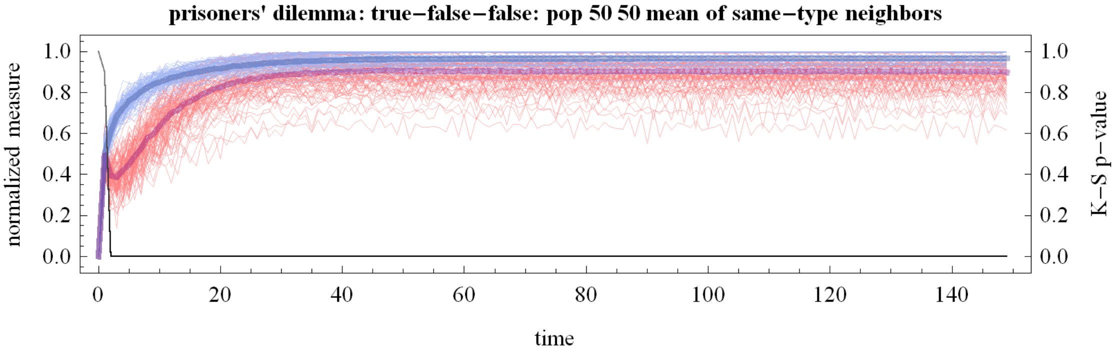
The first branch of experiments tests the sensitivity of behavior in the Prisoners' Dilemma to changes in population size. Since Prisoners' Dilemma also appears in section 3.9 (results of experiments sweeping across the strategic contexts), the discussion here focuses on those features that are specific to population changes.

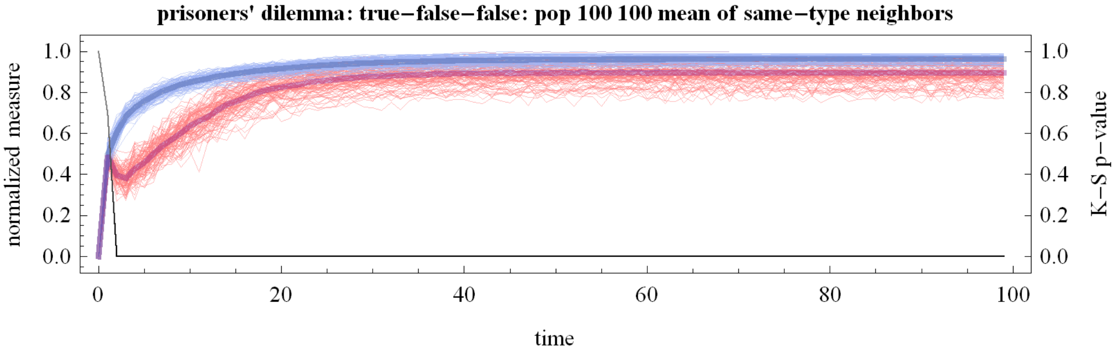
3.7.1 Prosociality of Preferential Detachment Alone

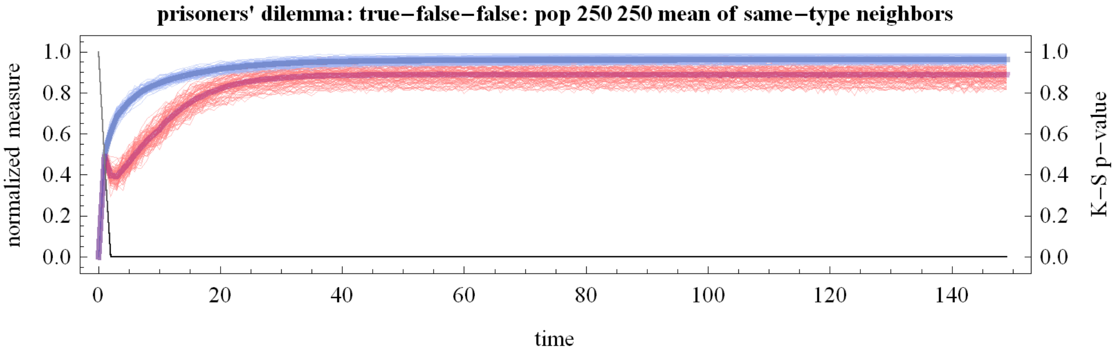
To demonstrate the effect of population size on prosociality in the absence of learning or population dynamics it suffices to present the level of assortativity of A-types. As a measure of assortativity I present the plots of the mean percentage of neighbors having the same type. This is the mean value of the agents separated by type *within a run*; both individual run data and the average over runs are presented in each plot (as described in section 3.6.3 above).

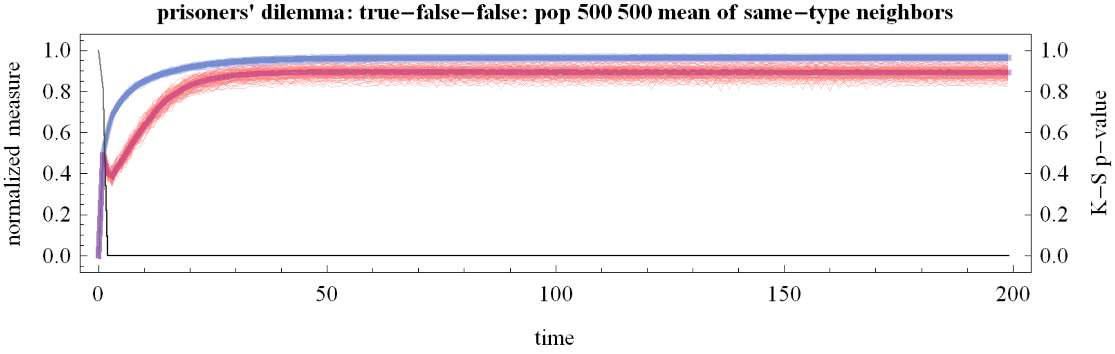


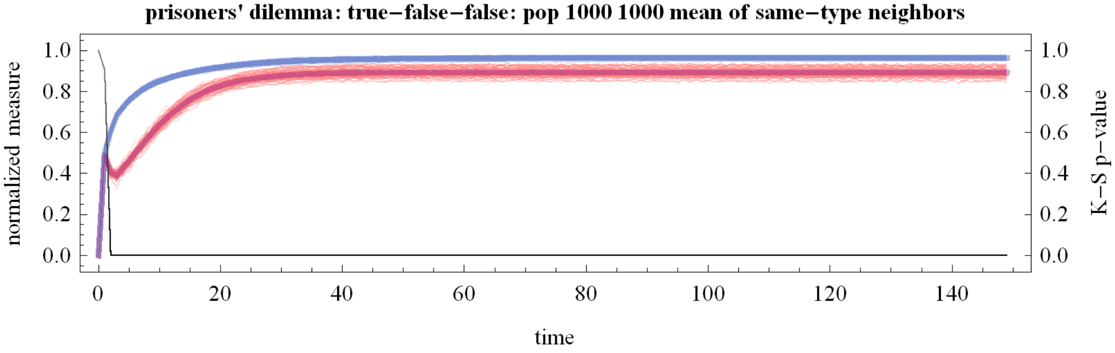

prisoners' dilemma: true−false−false: pop 25 25 mean of same−type neighbors
normalized measure
K−S p−value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
time


prisoners' dilemma: true−false−false: pop 50 50 mean of same−type neighbors
normalized measure
K−S p−value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
time


prisoners' dilemma: true−false−false: pop 100 100 mean of same−type neighbors
normalized measure
K−S p−value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
time

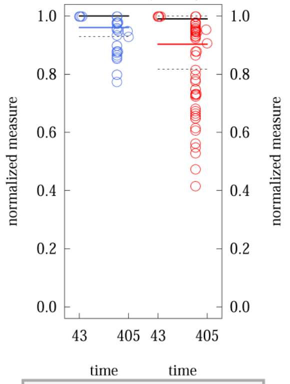

prisoners' dilemma: true−false−false: pop 250 250 mean of same−type neighbors
normalized measure
K−S p−value
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
time





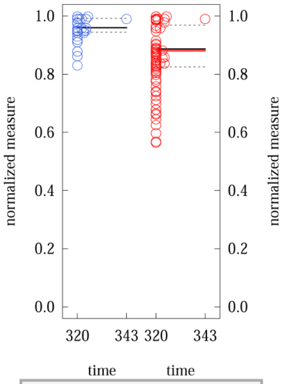
As the population increases the volatility decreases, so that each run converges to a similar trajectory. This can be explained by two considerations: 1) the number of agents at the perimeter of a homogeneous group increases slower than the total population, and 2) with more agents the run-by-run values are averages over a greater number of agents, which tends to smooth variations. From the outcome plots below we can see that for all population levels the degree of assortativity among A-types, and hence prosociality, is approximately 96% across population levels. The B-types have a significantly lower level of assortativity (significant as indicated by the Kolmogorov-Smirnov two-sample test's p-value), and their behavior shows more run-by-run variance at each population level compared to A-types.

prisoners' dilemma: true−false−false: pop 10 10
mean of same−type neighbors



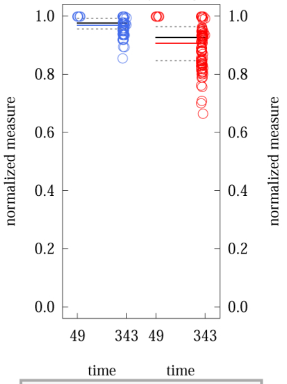
Mean:	0.960	0.903
Maximum:	1.000	1.000
Upper Quartile:	1.000	1.000
Median:	1.000	0.990
Lower Quartile:	0.930	0.817
Minimum:	0.775	0.417
K−S p−value:	0.006	

prisoners' dilemma: true−false−false: pop 25 25
mean of same−type neighbors



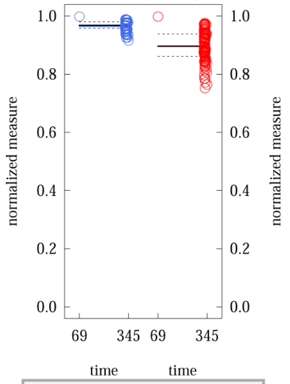
Mean:	0.960	0.881
Maximum:	1.000	1.000
Upper Quartile:	0.992	0.969
Median:	0.960	0.887
Lower Quartile:	0.944	0.825
Minimum:	0.832	0.567
K−S p−value:	0.000	

prisoners' dilemma: true−false−false: pop 50 50
mean of same−type neighbors



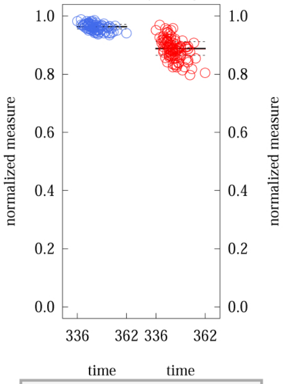
Mean:	0.968	0.906
Maximum:	1.000	1.000
Upper Quartile:	0.992	0.963
Median:	0.976	0.926
Lower Quartile:	0.955	0.847
Minimum:	0.856	0.666
K−S p−value:	0.000	

prisoners' dilemma: true−false−false: pop 100 100
mean of same−type neighbors



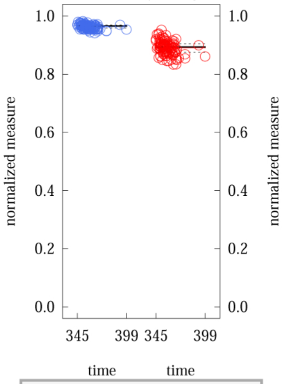
Mean:	0.965	0.896
Maximum:	1.000	1.000
Upper Quartile:	0.980	0.938
Median:	0.968	0.896
Lower Quartile:	0.958	0.862
Minimum:	0.918	0.754
K−S p−value:	0.000	

prisoners' dilemma: true−false−false: pop 250 250
mean of same−type neighbors

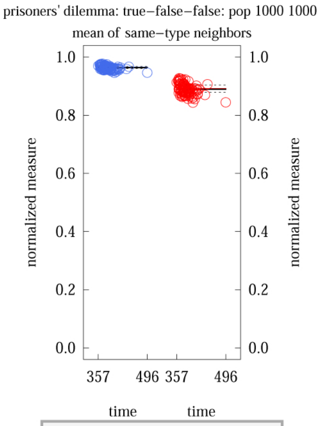


Mean:	0.963	0.888
Maximum:	0.990	0.972
Upper Quartile:	0.971	0.912
Median:	0.963	0.888
Lower Quartile:	0.955	0.865
Minimum:	0.936	0.799
K−S p−value:	0.000	

prisoners' dilemma: true−false−false: pop 500 500
mean of same−type neighbors



Mean:	0.965	0.892
Maximum:	0.984	0.954
Upper Quartile:	0.970	0.905
Median:	0.966	0.894
Lower Quartile:	0.959	0.875
Minimum:	0.943	0.835
K−S p−value:	0.000	

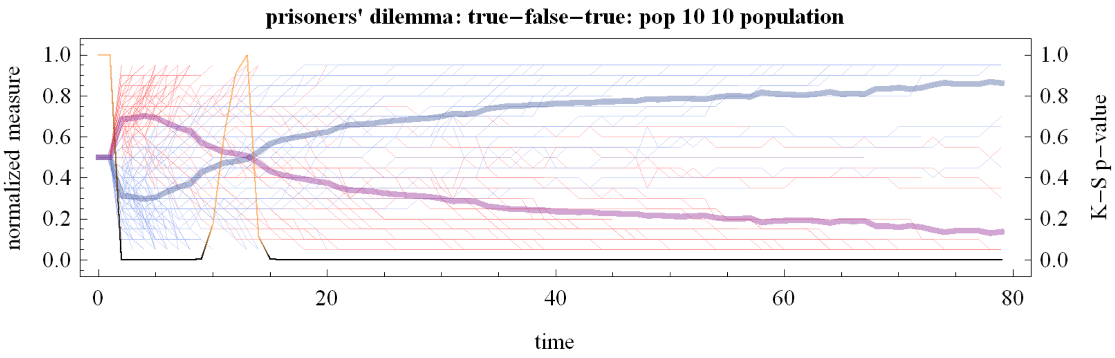


Mean:	0.965	0.891
Maximum:	0.979	0.929
Upper Quartile:	0.968	0.904
Median:	0.964	0.889
Lower Quartile:	0.961	0.879
Minimum:	0.948	0.845
K–S p–value:	0.000	

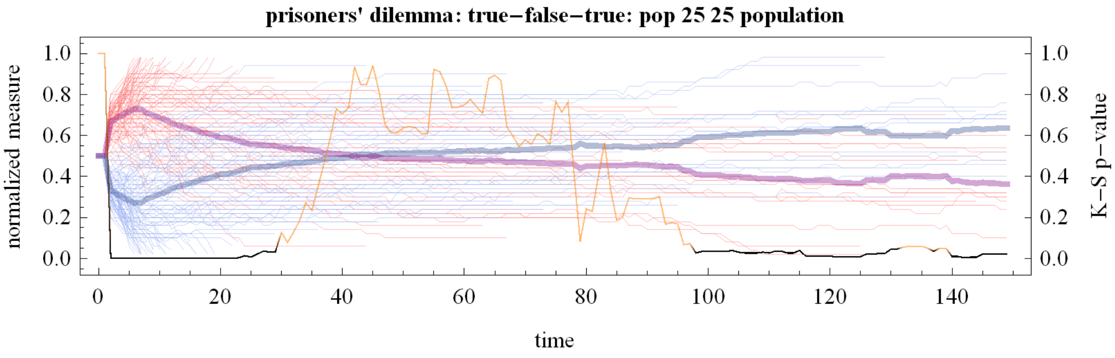
This result echoes the effects of changes in population levels in the Markov model: even for low population levels, achieving high levels of prosociality is expected; and the accuracy of that expectation increases with increasing populations because the results are more precise around a shared mean value.

3.7.2 Prosociality with Imitative Learning

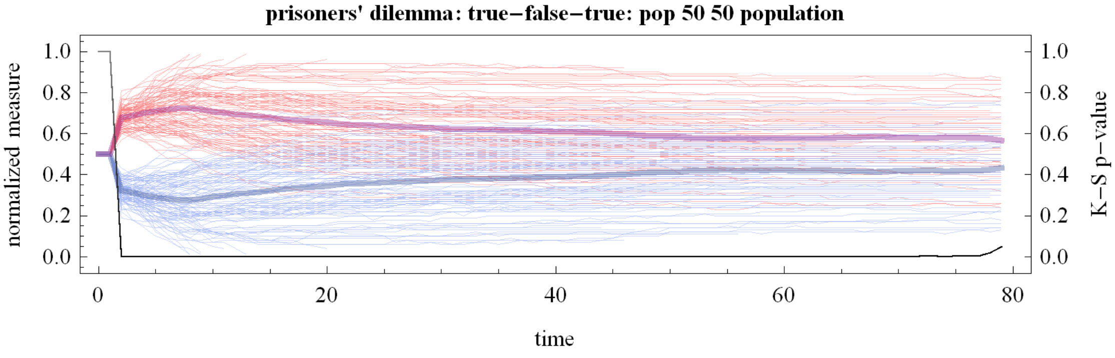
Recall that prosociality in learning experiments means A-type dominance, which we can measure as the percent of A-types in the population. Because defectors receive higher payoffs in mixed arrangements, they are expected to grow in number early in the simulation. We are looking to uncover when and how much defector success early in the simulation shapes the later trajectory and outcome. And because learning is implemented as a local behavior, if the types segregate themselves through preferential detachment, then populations of both types can survive in separated groups. So another outcome we need to investigate is the effect of population size on the formation of stable segregated homogeneous groups. I now present plots of population percentages of each agent type across the population sweep to examine population trajectories and judge outcome success.



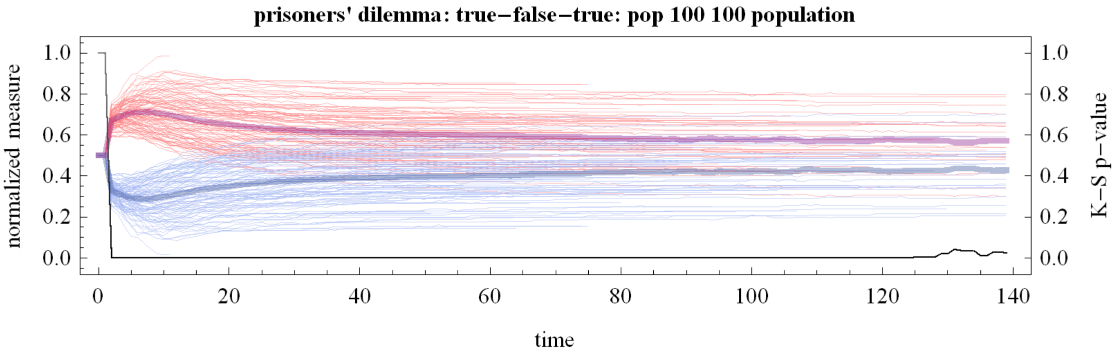
prisoners' dilemma: true−false−true: pop 10 10 population



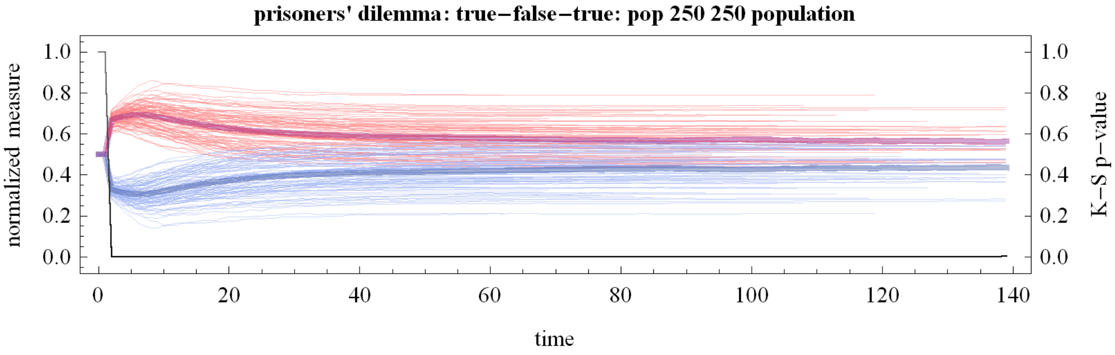
prisoners' dilemma: true−false−true: pop 25 25 population

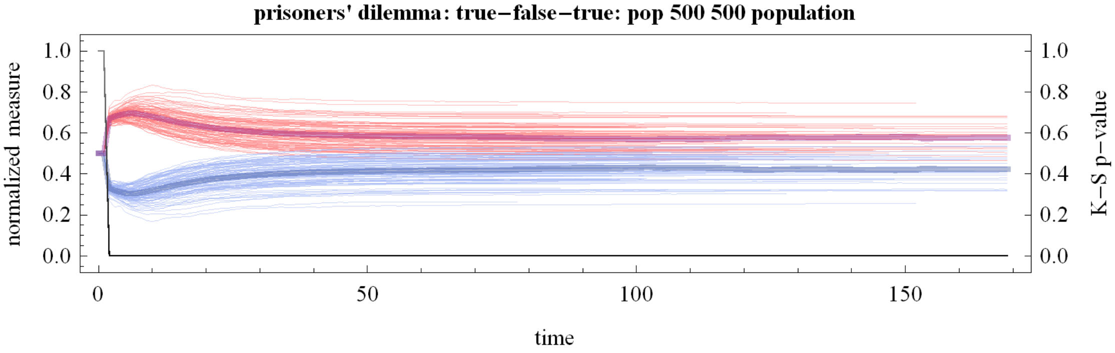


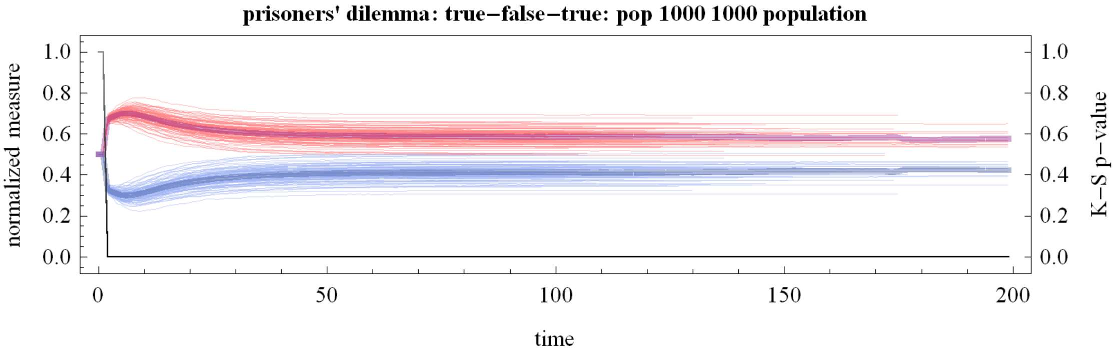
prisoners' dilemma: true−false−true: pop 50 50 population



prisoners' dilemma: true−false−true: pop 100 100 population

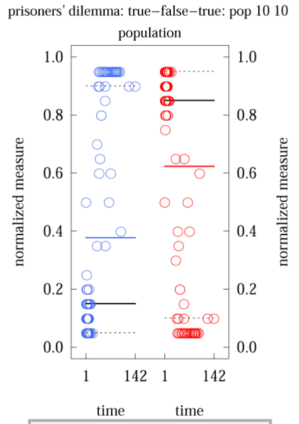

prisoners' dilemma: true−false−true: pop 250 250 population


prisoners' dilemma: true−false−true: pop 500 500 population

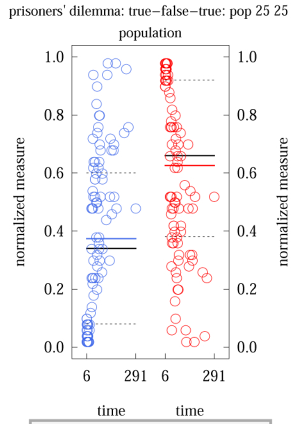

prisoners' dilemma: true−false−true: pop 1000 1000 population

As before, the shape of the average trajectories is similar at all population levels, though this result is clouded by the fact that the runs end when one type reaches dominance, and more runs end early when the population is low. It is also again true that the deviation of the run data from the average behavior *decreases* with increasing total population.

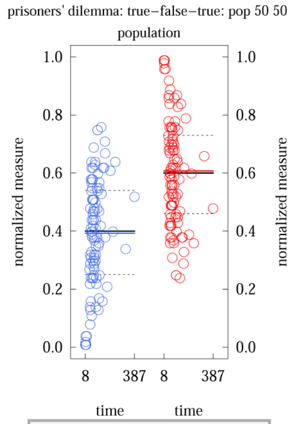
The summary plots below show the distribution of population data across the $N_i =$ 10 to $N_i = 1000$ range for the last time step of each run. This is particularly helpful to understand the lower population cases because the length of the runs is highly variable (making it difficult to read the outcome values from the above plots). The runs end whenever a halting condition is met, and in this situation that often happens a little prematurely (i.e., a condition is satisfied a few periods before the arrangement it is intended to detect). The stasis values for the cooperator population are typically greater than those presented as output here, but only by a small quantity (e.g., ten more agent convert before stasis is actually reached in the separated populations), and the qualitative result of 1) two separated homogeneous populations, and 2) a larger one for defectors is still accurate.


prisoners' dilemma: true–false–true: pop 10 10

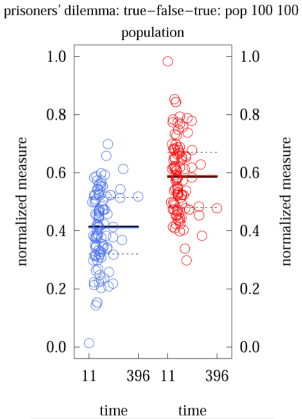
Mean:	0.377	0.623
Maximum:	0.950	0.950
Upper Quartile:	0.900	0.950
Median:	0.150	0.850
Lower Quartile:	0.050	0.100
Minimum:	0.050	0.050
K–S p–value:	0.000	


prisoners' dilemma: true–false–true: pop 25 25

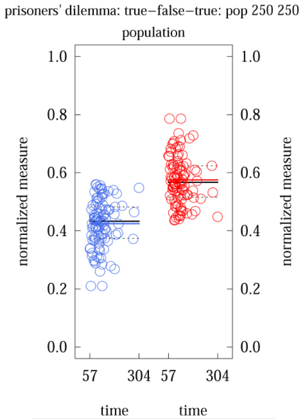
Mean:	0.374	0.626
Maximum:	0.980	0.980
Upper Quartile:	0.600	0.920
Median:	0.340	0.660
Lower Quartile:	0.080	0.380
Minimum:	0.020	0.020
K–S p–value:	0.000	


prisoners' dilemma: true–false–true: pop 50 50

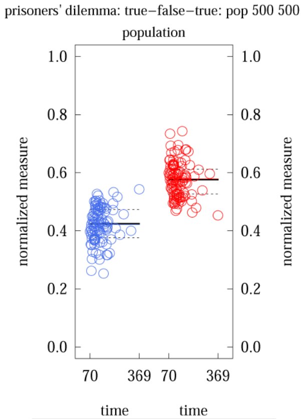
Mean:	0.393	0.607
Maximum:	0.760	0.990
Upper Quartile:	0.540	0.730
Median:	0.400	0.600
Lower Quartile:	0.250	0.460
Minimum:	0.010	0.240
K–S p–value:	0.000	


prisoners' dilemma: true–false–true: pop 100 100

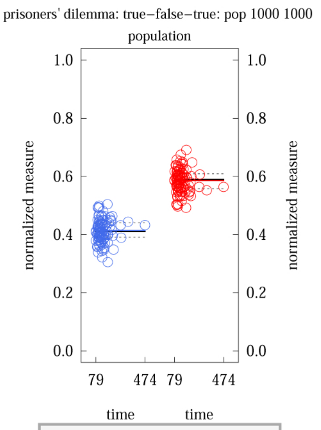
Mean:	0.411	0.589
Maximum:	0.700	0.985
Upper Quartile:	0.515	0.670
Median:	0.415	0.585
Lower Quartile:	0.320	0.480
Minimum:	0.015	0.300
K–S p–value:	0.000	


prisoners' dilemma: true–false–true: pop 250 250

Mean:	0.425	0.575
Maximum:	0.562	0.788
Upper Quartile:	0.482	0.624
Median:	0.433	0.567
Lower Quartile:	0.374	0.516
Minimum:	0.212	0.438
K–S p–value:	0.000	


prisoners' dilemma: true–false–true: pop 500 500

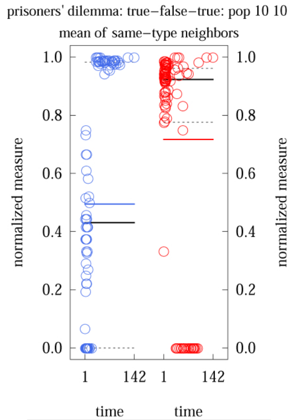
Mean:	0.423	0.577
Maximum:	0.545	0.745
Upper Quartile:	0.473	0.612
Median:	0.424	0.575
Lower Quartile:	0.376	0.527
Minimum:	0.255	0.455
K–S p–value:	0.000	



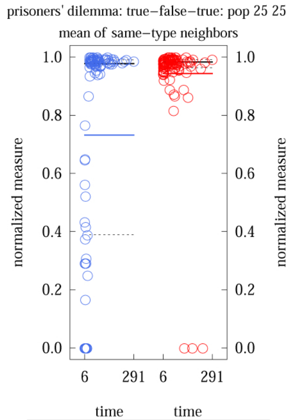
Mean:	0.414	0.586
Maximum:	0.506	0.693
Upper Quartile:	0.440	0.609
Median:	0.410	0.590
Lower Quartile:	0.390	0.558
Minimum:	0.307	0.494
K−S p−value:	0.000	

The outcome plots confirm the result that when defectors dominate it is only when they do so very quickly. The duration of the run and the success of cooperators are positively correlated. The endemic cooperator population level appears to be in the narrow 0.410 to 0.425 range; extending the population above 2000 total agents would likely continue to produce outcomes populations near these values. That there is such consistency in the mean population levels is surprising. Explaining why that particular value obtains would require analyzing 1) the connection probabilities, 2) the timing of neighbor saturation, 3) the relative utility levels of each type over time, and 4) the timing of detachment and learning. For our purposes here, the important result is that the expected result is consistent above a population of 50 agents of each type, and that the value of the result favors defectors.

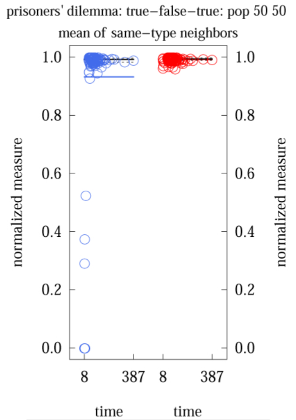
In light of the result that cooperators fail to dominate the population, let us take a closer look at their ability to segregate themselves. Assortativity is a second-best outcome, especially as a precursor to the results below that combine both type-changing behaviors. The outcome plots below show that at population levels of $N_i = 50$ both types are highly segregated, and as the population increases this percent increases (because there are just a few "odd men out" in each case regardless of the total population). High assortativity is exactly what we expect imitative learning to produce, and the level attained is higher here than in the base model because stable heterogeneous communities are impossible (due to payoff differences among the agents).



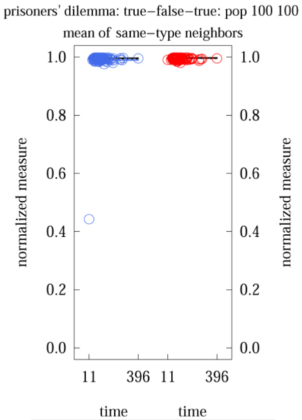
Mean:	0.494	0.717
Maximum:	1.000	1.000
Upper Quartile:	0.982	0.961
Median:	0.431	0.923
Lower Quartile:	0.000	0.776
Minimum:	0.000	0.000
K−S p−value:	0.000	



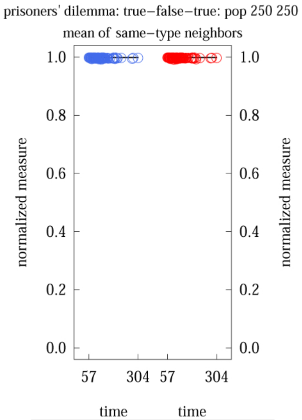
Mean:	0.732	0.943
Maximum:	1.000	1.000
Upper Quartile:	0.990	0.992
Median:	0.978	0.983
Lower Quartile:	0.389	0.964
Minimum:	0.000	0.000
K−S p−value:	0.000	



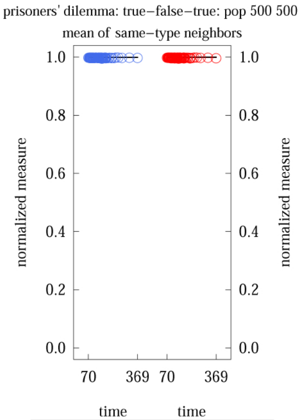
Mean:	0.932	0.992
Maximum:	1.000	1.000
Upper Quartile:	0.996	0.997
Median:	0.992	0.993
Lower Quartile:	0.982	0.989
Minimum:	0.000	0.960
K−S p−value:	0.023	



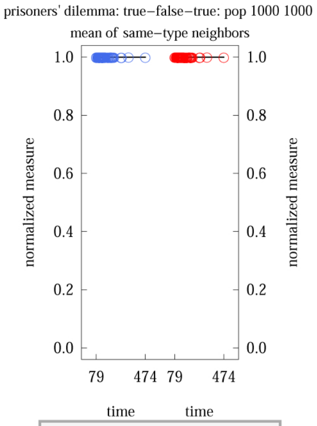
Mean:	0.990	0.996
Maximum:	1.000	1.000
Upper Quartile:	0.998	0.998
Median:	0.996	0.997
Lower Quartile:	0.993	0.994
Minimum:	0.444	0.982
K−S p−value:	0.149	



Mean:	0.998	0.999
Maximum:	1.000	1.000
Upper Quartile:	0.999	0.999
Median:	0.998	0.999
Lower Quartile:	0.998	0.998
Minimum:	0.993	0.993
K−S p−value:	0.023	



Mean:	0.999	0.999
Maximum:	1.000	1.000
Upper Quartile:	1.000	1.000
Median:	0.999	0.999
Lower Quartile:	0.999	0.999
Minimum:	0.996	0.996
K−S p−value:	0.015	

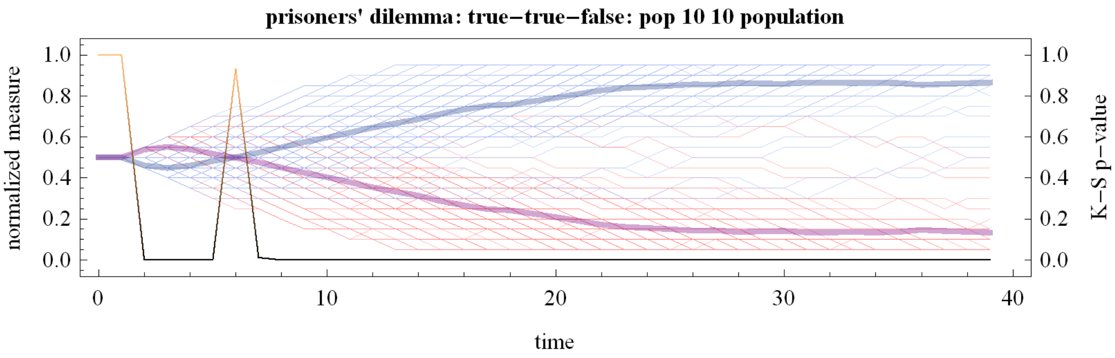


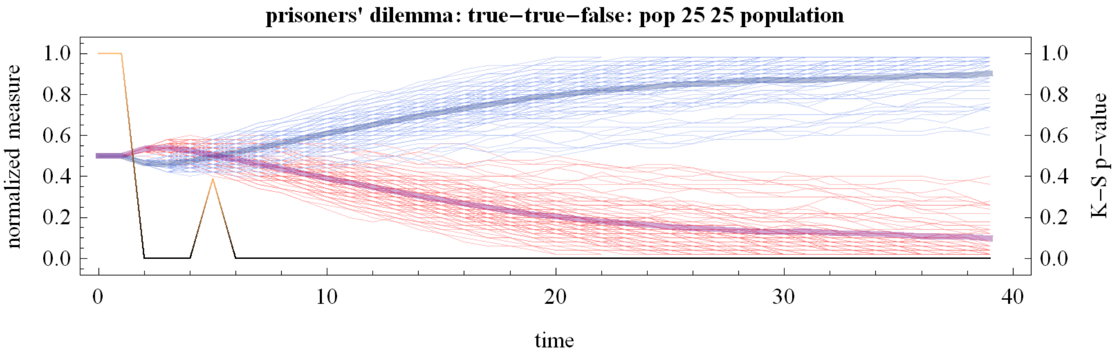
Mean:	1.000	1.000
Maximum:	1.000	1.000
Upper Quartile:	1.000	1.000
Median:	1.000	1.000
Lower Quartile:	0.999	0.999
Minimum:	0.998	0.998
K−S p−value:	0.149	

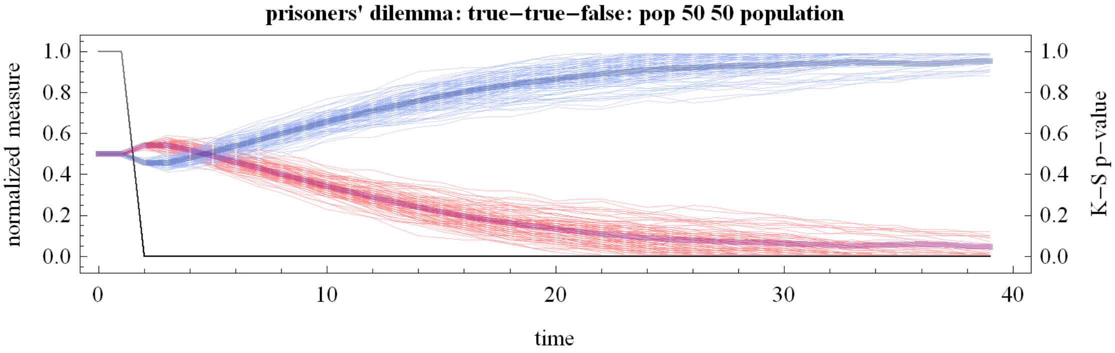
An outcome of stable, homogeneous, separated populations is certainly better than defector dominance, but it falls short of the cooperator dominance we are looking for. Because the social learning rule is local, and because of the particular arrangement required for a defector to imitate a cooperator, preferential detachment makes learning an ineffective means to spread cooperator behavior across the dynamic network arrangements. However, from these results it is clear that even a small amount of mixing (e.g., random rewiring) would take the population to an all-cooperator outcome because once their community is formed they perform better. This point is revisited in the conclusions section (3.10).

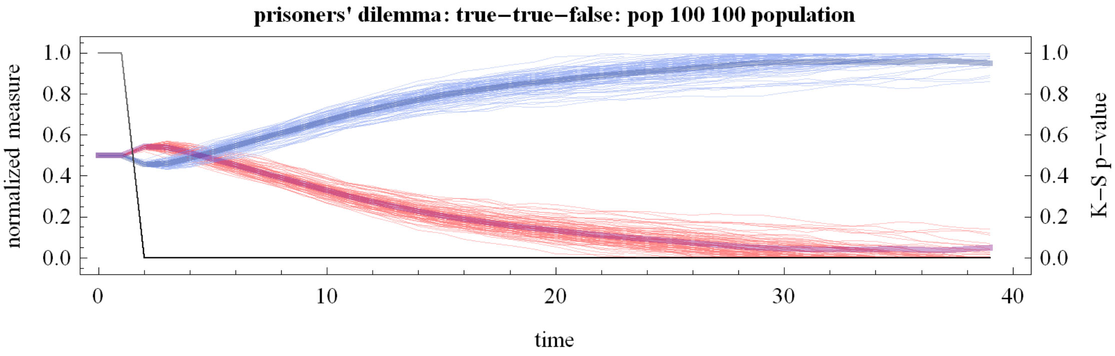
3.7.3 Prosociality with Population Dynamics

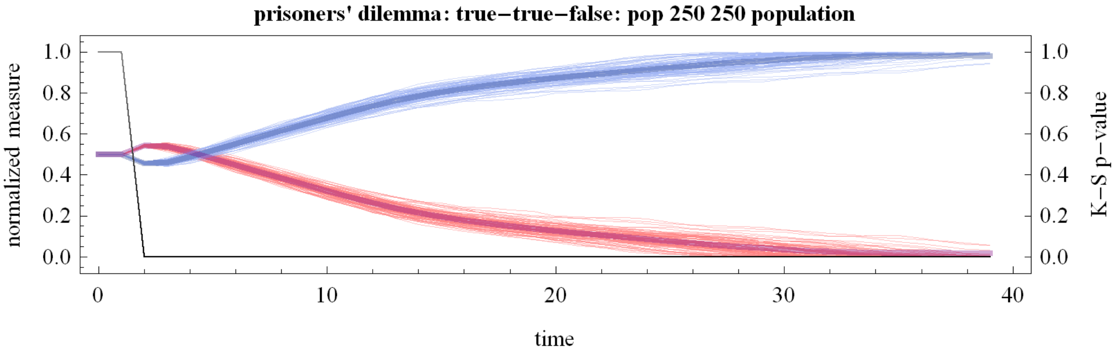
With population dynamics added, the agents achieve high levels of prosociality (which is again defined as A-type domination). The following plots across population levels show that the average system behavior is again consistent with increasing populations, and the run-by-run variation is again monotonically decreasing with increasing populations.

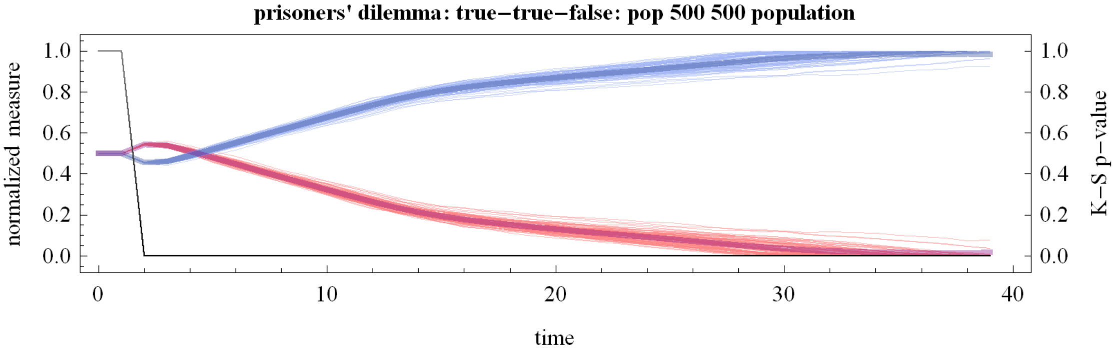

prisoners' dilemma: true−true−false: pop 10 10 population
normalized measure
K−S p−value
time
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40

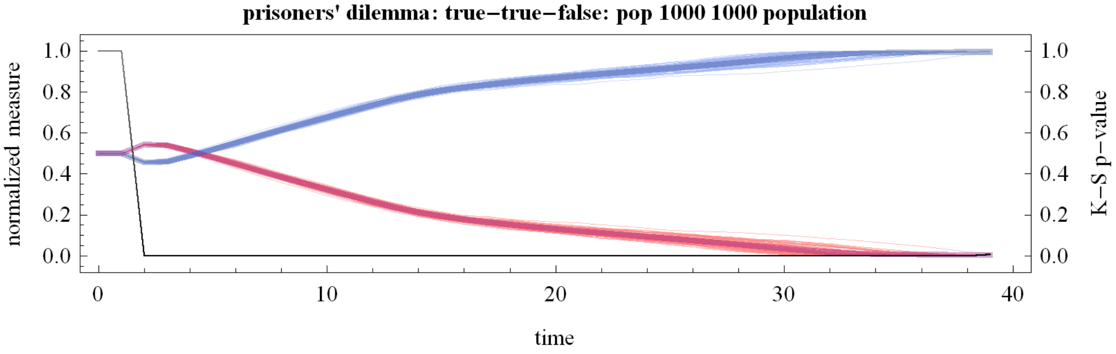

prisoners' dilemma: true−true−false: pop 25 25 population
normalized measure
K−S p−value
time
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40


prisoners' dilemma: true−true−false: pop 50 50 population
normalized measure
K−S p−value
time
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40


prisoners' dilemma: true−true−false: pop 100 100 population
normalized measure
K−S p−value
time
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40

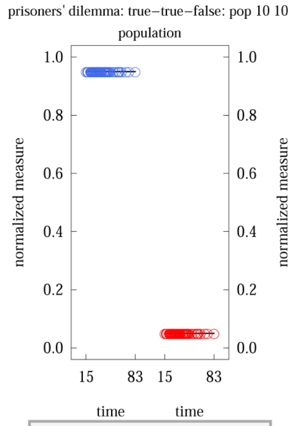




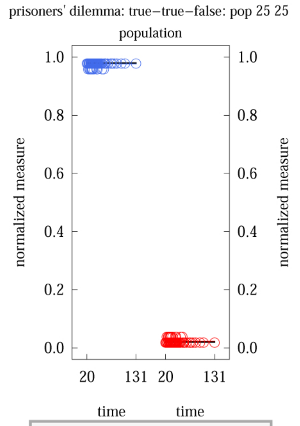


The behavior in these experiments converges more quickly with population dynamics than with learning or neither type-changing capability. The faster convergence is partially a function of the high 5% turnover rate, but it also depends on the structure of the prisoners' dilemma. We will see in branch three that the game structure plays a role in determining the effectiveness of learning vs. population dynamics in time needed to achieve population convergence. Another behavior of note is that the jump in defector success in early iterations is more prominent with population dynamics compared to imitative learning.

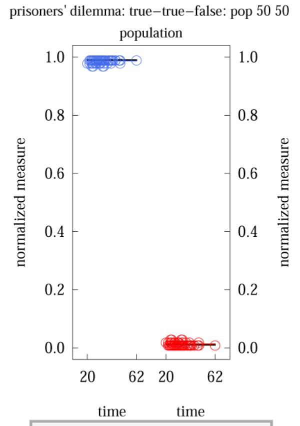
These plots are truncated to a time before every run achieves a halting condition; this makes it seem as though some runs end before full convergence. But by examining the outcomes across population levels at each run's final iteration we can compare the prosociality levels across population levels – and to the results with and without learning seen above.



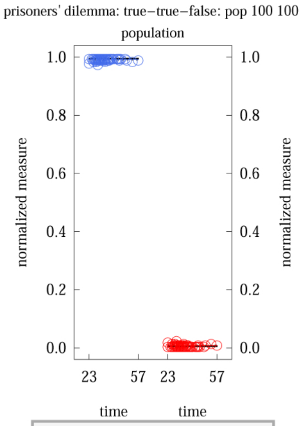
Mean:	0.950	0.050
Maximum:	0.950	0.050
Upper Quartile:	0.950	0.050
Median:	0.950	0.050
Lower Quartile:	0.950	0.050
Minimum:	0.950	0.050
K−S p−value:	0.000	



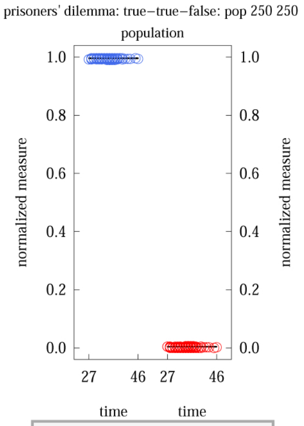
Mean:	0.978	0.022
Maximum:	0.980	0.040
Upper Quartile:	0.980	0.020
Median:	0.980	0.020
Lower Quartile:	0.980	0.020
Minimum:	0.960	0.020
K−S p−value:	0.000	



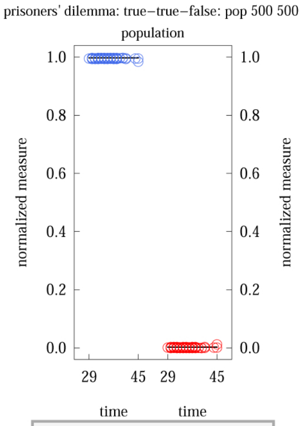
Mean:	0.987	0.013
Maximum:	0.990	0.030
Upper Quartile:	0.990	0.010
Median:	0.990	0.010
Lower Quartile:	0.990	0.010
Minimum:	0.970	0.010
K−S p−value:	0.000	



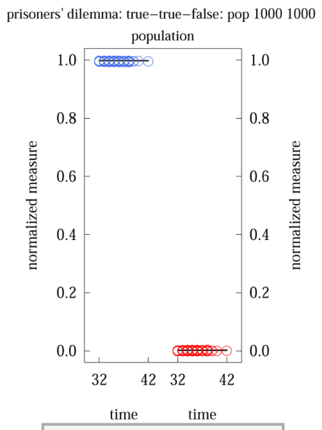
Mean:	0.993	0.007
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K−S p−value:	0.000	



Mean:	0.996	0.004
Maximum:	0.998	0.008
Upper Quartile:	0.998	0.006
Median:	0.996	0.004
Lower Quartile:	0.994	0.002
Minimum:	0.992	0.002
K−S p−value:	0.000	

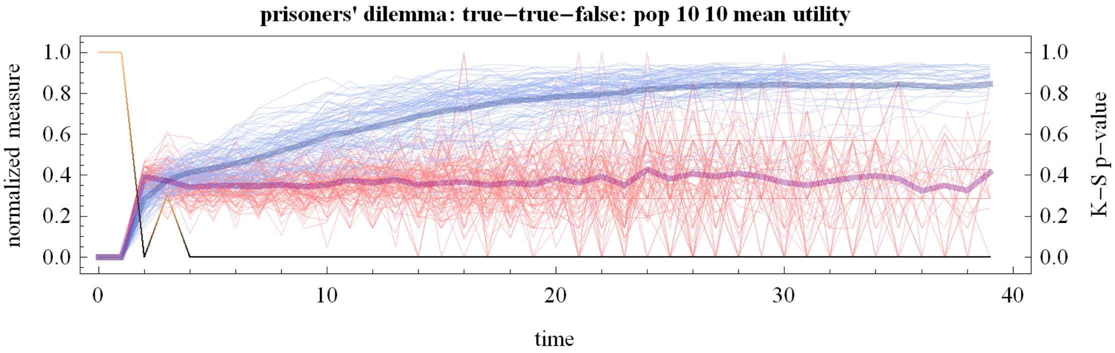


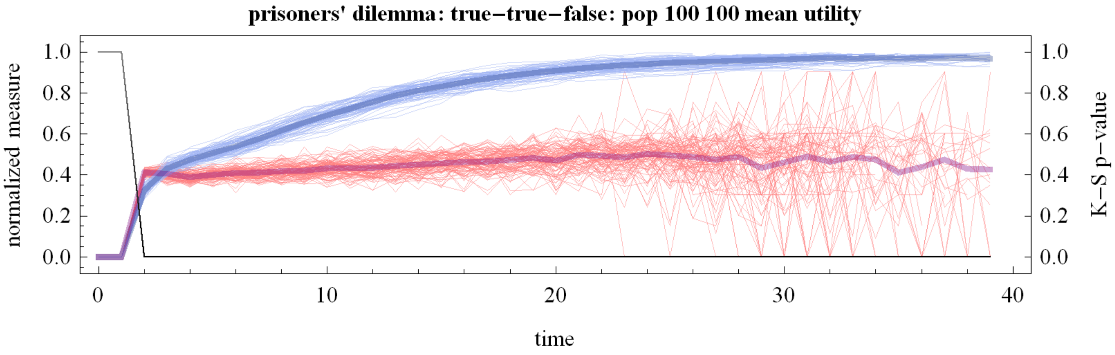
Mean:	0.997	0.003
Maximum:	0.999	0.014
Upper Quartile:	0.998	0.003
Median:	0.998	0.002
Lower Quartile:	0.997	0.002
Minimum:	0.986	0.001
K−S p−value:	0.000	

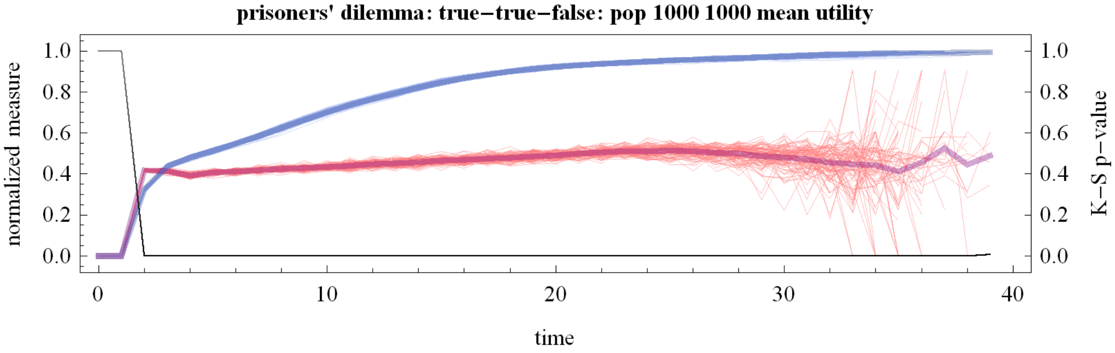


Mean:	0.998	0.002
Maximum:	1.000	0.006
Upper Quartile:	0.999	0.003
Median:	0.998	0.002
Lower Quartile:	0.997	0.001
Minimum:	0.994	0.000
K–S p–value:	0.000	

This set of summary plots shows us that population dynamics ensures cooperator success at all population levels. Once enough cooperators are contained in groups with other cooperators through preferential detachment their utility will fill the slots for the top five percent. More cooperators enter the simulation, but these new cooperators have a larger probability of connecting to defectors because the preexisting cooperators have mostly saturated their neighbor allotment. Through the random connection process some defectors will attach to multiple cooperators and receive a spike in utility sufficient to bump it into the top 5%. Obviously the possibility of this happening shrinks with larger populations because a lucky defector needs to outperform more individual cooperators to breach the upper echelon. This phenomenon can be seen in the plots of mean agent utility for $N_i = 10$, $N_i = 100$, and $N_i = 1000$.



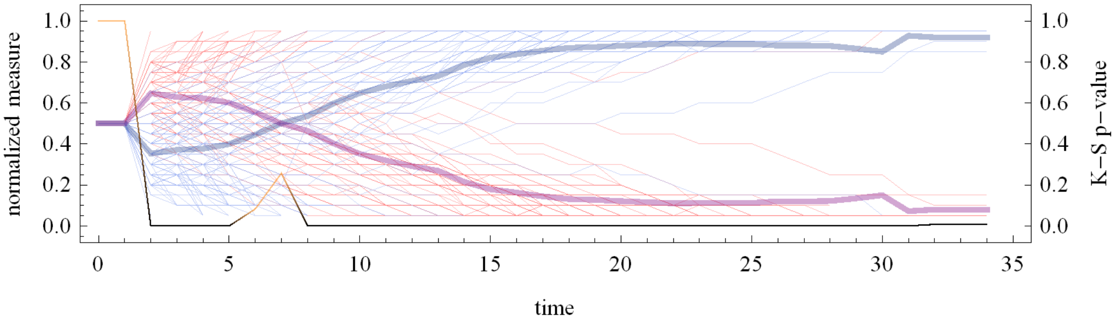


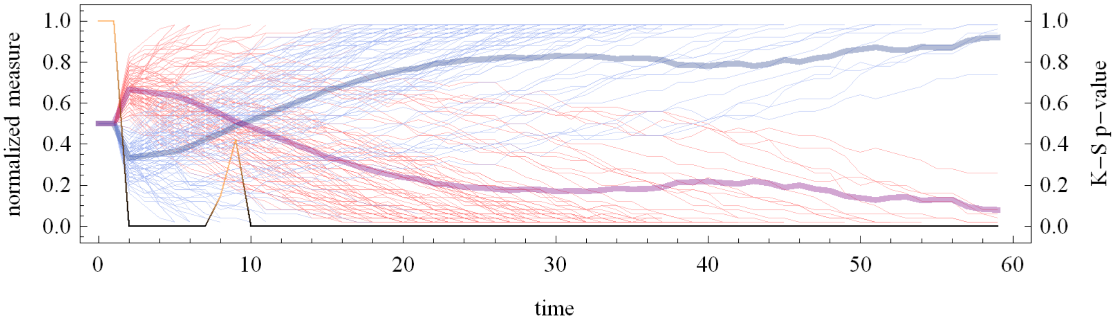


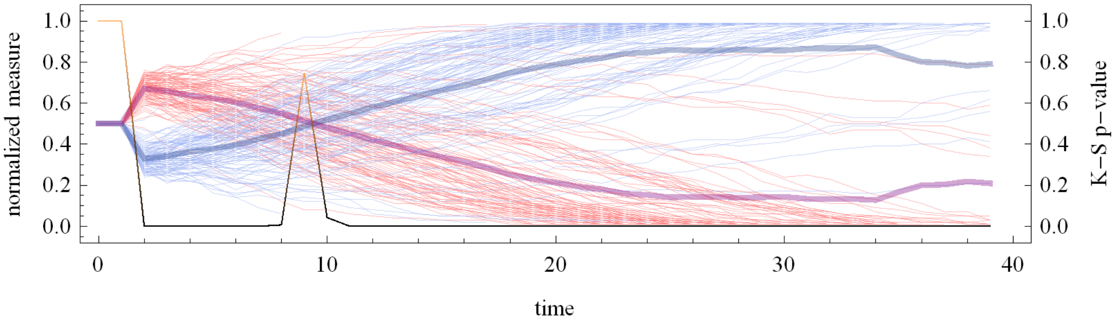
Because these are mean values for the whole population within that run, the spikes are smoothed out at larger populations. It also means that if the mean utility shows a large spike then either some individual utility has increased by a very large amount or several individuals have experienced a more moderate increase. Runs in which defectors are more successful take longer to complete, a result that is revealed by the increasing volatility in the data toward larger run times (because runs in which cooperators quickly dominate have already ended). When utilizing population dynamics, the success of cooperators is guaranteed, but the time until the prosocial outcome is dependent on the population (in a systematic way) and on the contingent behavior within each run of the simulation.

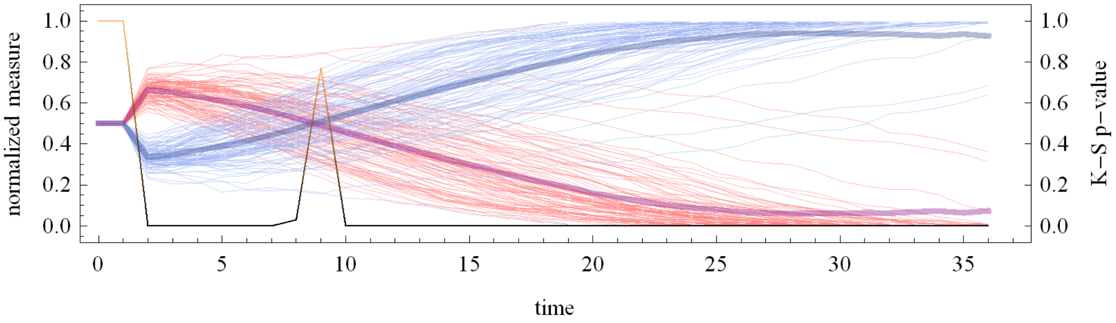
3.7.4 Prosociality with Learning and Population Dynamics

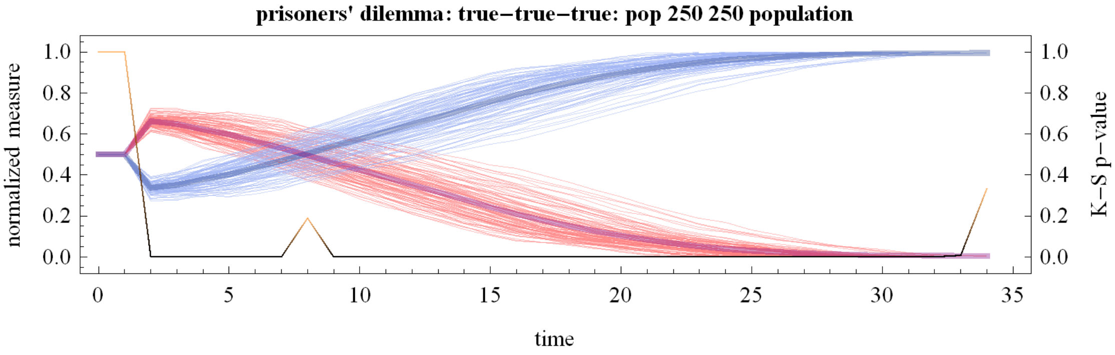
When we combine the mechanisms of learning and population dynamics the results share features of each mechanisms alone. As before, the run-by-run variability decreases with increasing populations, and the results converge to signature system behavior pattern. The shape of the trajectory resembles the population dynamics experiments, but with a greater increase in defector numbers early on from the learning rule. And it is again the case that defectors can achieve population dominance with smaller populations, but only if they do so before a dense group of cooperators forms; i.e., typically during the early periods of a run.

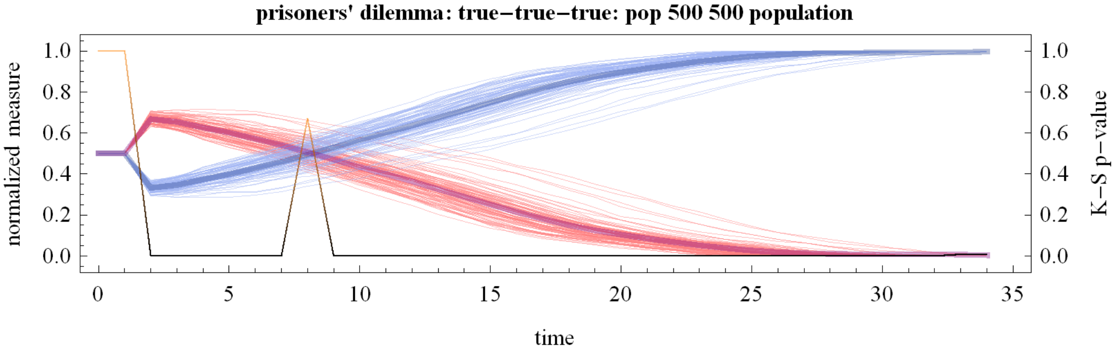

prisoners' dilemma: true−true−true: pop 10 10 population

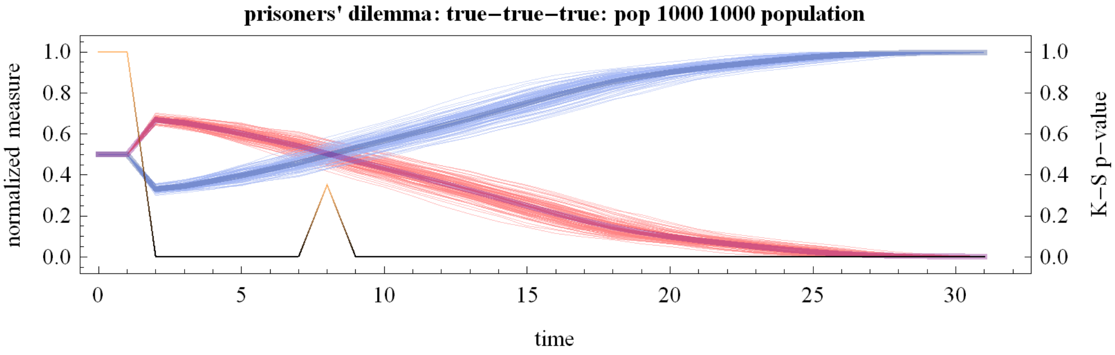

prisoners' dilemma: true−true−true: pop 25 25 population


prisoners' dilemma: true−true−true: pop 50 50 population


prisoners' dilemma: true−true−true: pop 100 100 population

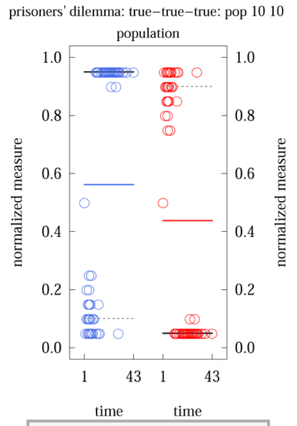




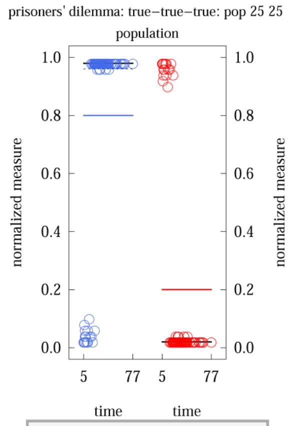


The effect of the combination is that each type-changing behavior exacerbates the effects of the other. During the early iterations, population dynamics increases the number of defectors, and new defectors are more likely to make and receive connections with cooperators. These connections then set the stage for imitative learning which again favors defectors in mixed arrangements. Defector success, however, depends on there being cooperators with open connections; as the population swells with defectors those defectors become less successful. Once some cooperators do form a cohesive and exclusive group, that group can exclude defector neighbors. The amplified combined behavior then very quickly takes the system to cooperator dominance.

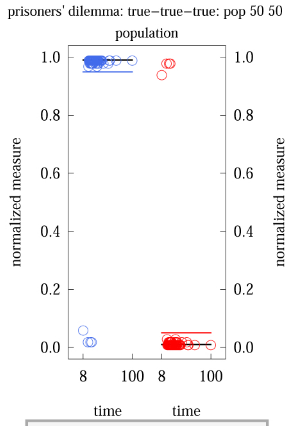
As before, and for the same reasons, the behavior of the model's runs become increasingly more consistent with larger population sizes. In the combined mechanism experiments stable heterogeneous groups are impossible because of learning, and stable separated populations are impossible because of population dynamics and differing payoffs. Domination of one type is the only possible outcome (though some other halting condition may be satisfied before domination occurs).



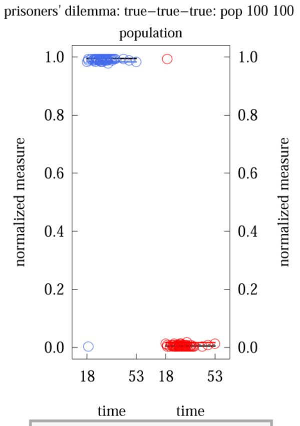
Mean:	0.562	0.438
Maximum:	0.950	0.950
Upper Quartile:	0.950	0.900
Median:	0.950	0.050
Lower Quartile:	0.100	0.050
Minimum:	0.050	0.050
K−S p−value:	0.000	



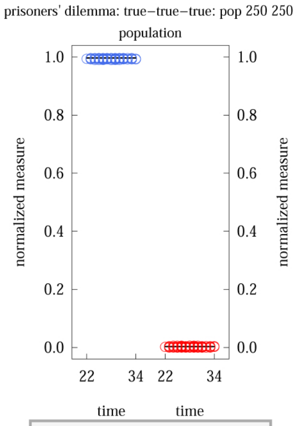
Mean:	0.800	0.200
Maximum:	0.980	0.980
Upper Quartile:	0.980	0.020
Median:	0.980	0.020
Lower Quartile:	0.960	0.020
Minimum:	0.020	0.020
K−S p−value:	0.000	



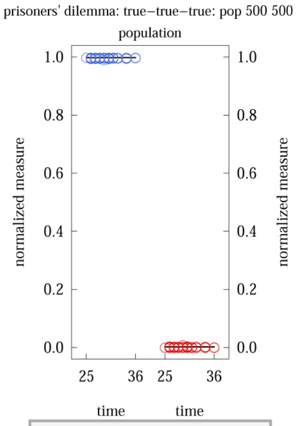
Mean:	0.950	0.050
Maximum:	0.990	0.980
Upper Quartile:	0.990	0.010
Median:	0.990	0.010
Lower Quartile:	0.990	0.010
Minimum:	0.020	0.010
K−S p−value:	0.000	



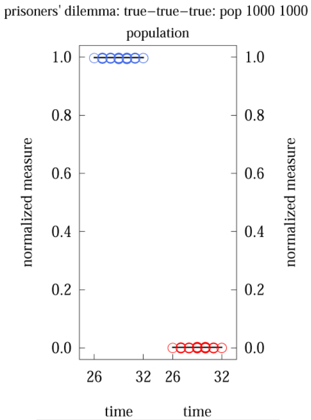
Mean:	0.983	0.017
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.005	0.005
K−S p−value:	0.000	

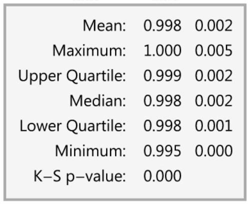


Mean:	0.997	0.003
Maximum:	0.998	0.008
Upper Quartile:	0.998	0.004
Median:	0.997	0.003
Lower Quartile:	0.996	0.002
Minimum:	0.992	0.002
K−S p−value:	0.000	



Mean:	0.998	0.002
Maximum:	0.999	0.009
Upper Quartile:	0.999	0.003
Median:	0.998	0.002
Lower Quartile:	0.997	0.001
Minimum:	0.991	0.001
K−S p−value:	0.000	





Mean:	0.998	0.002
Maximum:	1.000	0.005
Upper Quartile:	0.999	0.002
Median:	0.998	0.002
Lower Quartile:	0.998	0.001
Minimum:	0.995	0.000
K−S p−value:	0.000	

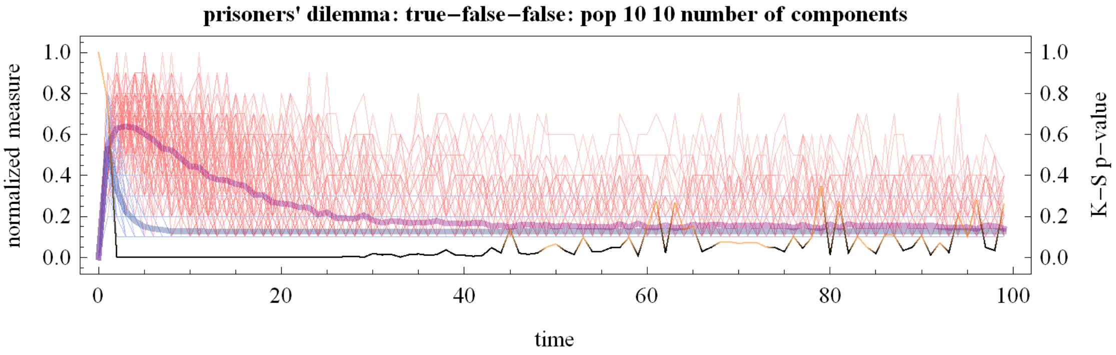
The prosocial outcome is again cooperator dominance, and this result occurs with increasing frequency as the population increases. The number steadily increases with increasing population size. At populations of 250 agents of each type or above, the outcome is reliably cooperator dominance. In all but a few runs the cooperators are successful with as few as fifty agents, and only twenty runs were unsuccessful at achieving the prosocial outcome when $N_i = 25$. The majority of runs end with cooperator dominance even when $N_i = 10$. We can also see the iteration at which the run ended, which reveals two features of the model. For all cases in which defectors dominate the population they do so very early in the run. Furthermore the runs complete more quickly compared to the three previous cases.

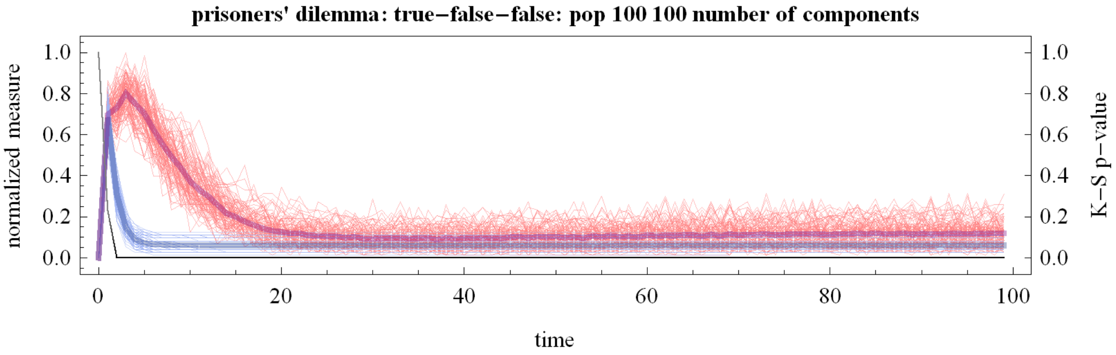
3.7.5 Stable Heterogeneous Groups

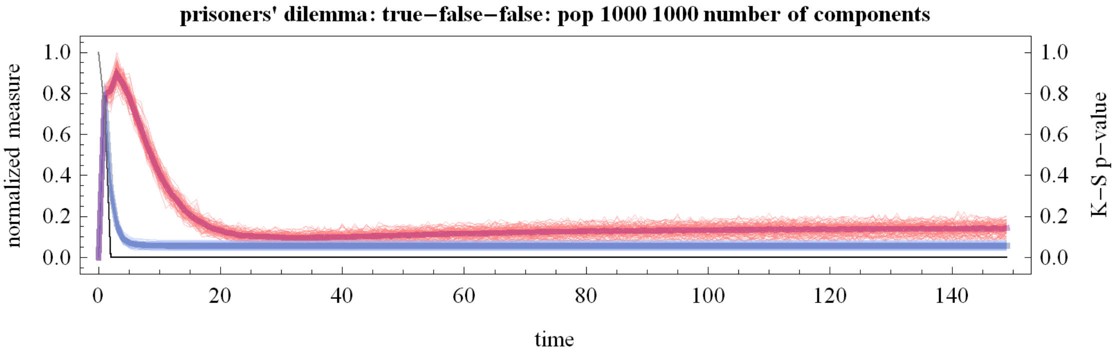
Stable heterogeneous clusters in the Prisoners' Dilemma must be composed of a single cooperator with K defector neighbors...possibly linked in chains via a shared defector. Thus the persistence of lower-than-one levels of same-type neighbors, and a much lower value for B-types compared to A-types, reveals that when the population is lower, a greater percentage of agents are attached to agents of the opposite type. Though even with $N_i = 10$ fewer than half of the runs produce any stable mixed groups.

Furthermore, by examining the number of components for $N_i = 10$, $N_i = 100$, and $N_i = 1000$ in the plots below we can see that there are significantly fewer groups of

A-types compared to B-types in large populations, but the difference is insignificant for small populations. In intermediate populations, the outcome is variable among the runs for both types, but the volatility is greatly reduced for A-types and the difference is significant. Because for 100 and 1000 the number of A-type agents in the largest component is very nearly 50% (i.e., most of the A-types), the outcome is not multiple communities, but one large community and isolated agents with other-type neighbors.







Thus as the population increases a smaller percentage of cooperators become stuck in stable groups with defectors. Screenshots from simulations with $N_i = 10$, $N_i = 100$, and $N_i = 1000$ exemplify this result. As described in the visualization section (3.6.3) the blue nodes represent A-types (cooperators) and the red nodes are the B-types (defectors).

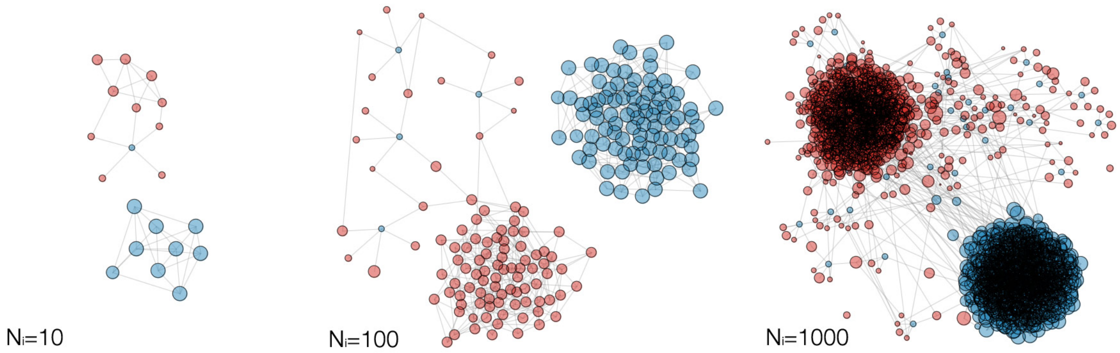


Figure 3.4: Stable Heterogeneous Groups in the Prisoners' Dilemma

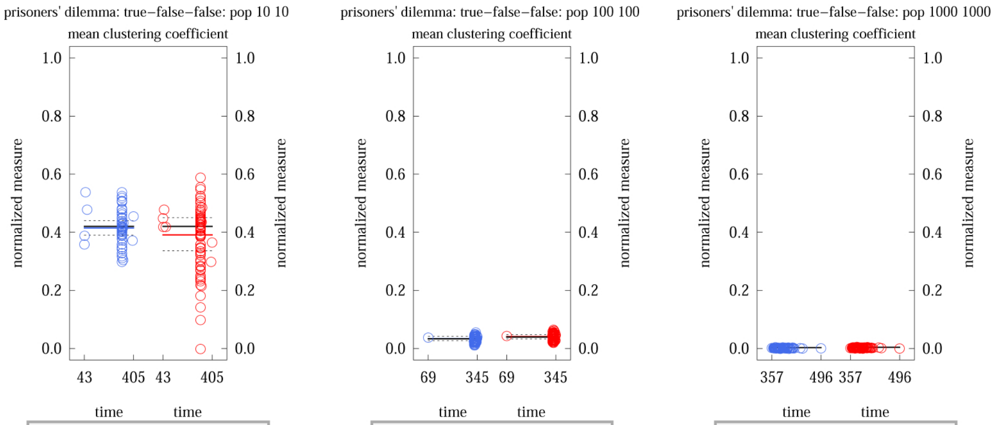
The result is that the expected *percent* of cooperators stuck in heterogeneous groups is consistent across population levels, so the expected *number* of such agents increases. Having a large number of agents creates more opportunities for the rare series of attachment events necessary to form these heterogeneous groups; however, the larger number also means that the likelihoods of connecting to an agent of each type stay similar longer (these feature is partly responsible for the smoothing out of the dynamics as larger population levels).

The measure of percent similar neighbors at each run's final step reported above shows values reliably above 80% of A-types and increasing to nearly 100% at large populations. Because agents in heterogeneous groups have a value of zero for this measure, the aggregated value actually reports how many agents have values of zero vs one-hundred as a percentage of the population. As mentioned above, when $N_i = 10$ fewer than half the runs have *any* mixed groups, though some runs have as many as one-fifth of the cooperators (i.e., two of ten cooperator agents) stuck with defectors. When $N_i = 1000$ the worst-case run is 94.8% similar (approximately fifty cooperators in mixed groups), but there are *no runs* with zero heterogeneous groups.

3.7.6 Clustering and Group Formation

Much of the literature on group formation on social networks refers to clustering coefficients ([Watts and Strogatz 1998, Ravasz and Barabási 2003]). High clustering has also been shown to promote cooperation in simulations using learning behavior over static social network structures [Abramson and Kuperman 2001, Assenza et al. 2008]. Yet other models have noted that scale-free networks having very low clustering coefficients also promote cooperation through the imitation of hubs [Yang et al. 2009]. Clustering is less prominent in the literature involving dynamic network models (though see [Shi et al. 2007]), but investigating the clustering properties of final agent arrangements here may provide a bridge (and hence a possible explanation) for why the observed clustering in static network structures came into being.

Each component of the final networks produced by preferential detachment is nearly a regular K-graph; i.e., a graph in which each node has degree K. This by itself does not indicate any particular level of clustering (e.g., both fully connected and ring graphs are regular K-graphs), but in context it does preclude certain graph structures (e.g., power law distributions, wagon wheels, and a fully connected graph since $K \ll N$). There is a great deal of network theory research on random graphs and their properties (Newman [2009]), but the graphs here are randomly *generated* regular graphs, and the existing results do not apply. We can examine the value of the clustering coefficient to evaluate its role and to determine the relationship, if any, between clustering and the success of cooperators in the Prisoners' Dilemma.



Mean:	0.415	0.391
Maximum:	0.540	0.590
Upper Quartile:	0.440	0.450
Median:	0.420	0.420
Lower Quartile:	0.390	0.337
Minimum:	0.300	0.000
K–S p–value:	0.023	

Mean:	0.034	0.041
Maximum:	0.057	0.066
Upper Quartile:	0.042	0.048
Median:	0.033	0.039
Lower Quartile:	0.027	0.033
Minimum:	0.012	0.021
K–S p–value:	0.000	

Mean:	0.003	0.004
Maximum:	0.006	0.007
Upper Quartile:	0.004	0.005
Median:	0.003	0.004
Lower Quartile:	0.003	0.003
Minimum:	0.001	0.001
K–S p–value:	0.000	

First we examine the results from simulations with $N_i = 10$, $N_i = 100$, and $N_i = 1000$ and neither learning nor population dynamics. The first observation is that the difference between the clustering coefficients of the two types of agents is small, but the slightly greater defector clustering is (barely) statistically significant at all population levels. There are long portions of the runs in which the difference among the runs are *not* significant, but it is significant in the final structures that form.

The other result is the rather obvious connection between clustering and population size. Considering that final arrangements are characterized by self-similar groups of agents with $K = 5$ neighbors each, when population levels are small, a connection among neighbors is likely. In fact the values of clustering seen correspond to what one expects from random connections under the given constraints ([Newman 2009]). For example, note that as the population goes up by a factor of ten, the mean clustering coefficient goes down by approximately a factor or ten.

Simulations with learning and/or population dynamics end with all agents of the same type, usually cooperators, so the clustering of defectors at the final step is meaningless. For all these experiments the mean clustering coefficients for cooperators at the different population levels mirrors the values presented above. Before the outcome is reached, the results with learning and/or population dynamics produce differences in the clustering coefficient that are significant between the two agent types through much of the simulation time, and cooperators have systemically higher clustering coefficients.

Since we know that a group of cooperators must achieve a high utility level to survive the learning and/or population dynamics process, high clustering may be expected among cooperators during this critical stage of social development in cooperator dominating scenarios. This does not seem to be the case except when the population is very small; and in that case there is a more plausible alternative explanation for the high clustering (i.e., agent availability). Clustering in these networks seems to play a negligible role in the dynamics or outcome for all experiments. Though clustering is low in moderate or larger populations, there are no hubs in these networks, so previous results do not explain the success of imitation in achieving the cooperative outcome here. The ability of preferential detachment to efficiently separate agents into sparsely connected assortative groups (which yield greater payoffs to cooperators) still stands as the dominant force behind cooperator success in all of the experiments presented in this section.

3.7.7 Summary of Population Sweep Results

A common theme in the population sweeps is that for defectors to gain the upper hand they must prevent cooperators from forming an exclusive group in the early iterations. Preferential detachment exerts a pressure toward assortativity in the Prisoners' Dilemma, but defectors can preempt cooperator success if they happen to connect to cooperators rather defectors. In the base model this will allow defectors to create stable heterogeneous groups. In the other cases it may allow defectors to take over while learning and/or population dynamics are exerting a greater force than preferential detachment in the early stages of an experiment.

Because the system behaviors just described require essentially luck (a confluence of low-probability events) on the part of defectors to form the correct arrangements early on, they are relatively rare phenomena. Population dynamics never actually produces an antisocial result, but learning does favor defectors enough in early periods to give them a lasting edge on population levels, or achieve dominance when populations are small. We saw in section 3.7.5 that stable heterogeneous groups play a large role in only a few cases when populations are small, and they play a negligible (but non-zero) role in all runs with large populations. Either way the *expected* impact on prosociality is minor, no matter how intriguing their appearance.

The summary result of experiments sweeping a range of equal population levels is that preferential detachment is highly successful at achieving assortativity (the prosocial outcome) in the Prisoners' Dilemma in populations as small as ten agents of each type. Occasional defector dominance has a devastating impact on prosociality when learning is activated, but it is only possible when the pool of cooperators is small enough to be exhausted within the first few iterations. Preferential detachment is increasingly successful at segregating the populations as the number of agents increases, and hence defector dominance through neighbor imitation becomes impossible at higher population levels, the the separated populations do prevent cooperators from dominating. Since groups of cooperators have greater fitness than groups of defectors, for populations with more than a couple dozen cooperators, cooperator dominance is a nearly certain outcome with population dynamics. When learning and population dynamics are combined, the success of cooperator dominance is primarily driven by the success of preferential detachment to segregate the agents which promoted cooperator replication, hence it performs increasingly well as the population grows and only allows defector dominance in small populations.

Preferential Detachment

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
Pop 10 10	0.500	0.960	0.500	0.903
Pop 25 25	0.500	0.960	0.500	0.881
Pop 50 50	0.500	0.968	0.500	0.906
Pop 100 100	0.500	0.965	0.500	0.896
Pop 250 250	0.500	0.963	0.500	0.888
Pop 500 500	0.500	0.965	0.500	0.892
Pop 1000 1000	0.500	0.965	0.500	0.891

Learning

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
Pop 10 10	0.377	0.494	0.623	0.717
Pop 25 25	0.374	0.732	0.626	0.943
Pop 50 50	0.393	0.932	0.607	0.992
Pop 100 100	0.411	0.990	0.589	0.996
Pop 250 250	0.425	0.998	0.575	0.999
Pop 500 500	0.423	0.999	0.577	0.999
Pop 1000 1000	0.414	1.000	0.586	1.000

Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
Pop 10 10	0.950	0.912	0.050	0.000
Pop 25 25	0.978	0.941	0.022	0.050
Pop 50 50	0.987	0.933	0.013	0.027
Pop 100 100	0.993	0.936	0.007	0.004
Pop 250 250	0.996	0.937	0.004	0.010
Pop 500 500	0.997	0.938	0.003	0.008
Pop 1000 1000	0.998	0.938	0.002	0.006

Learning and Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
Pop 10 10	0.562	0.598	0.438	0.392
Pop 25 25	0.800	0.803	0.200	0.188
Pop 50 50	0.950	0.913	0.050	0.052
Pop 100 100	0.983	0.927	0.017	0.039
Pop 250 250	0.997	0.938	0.003	0.033
Pop 500 500	0.998	0.939	0.002	0.010
Pop 1000 1000	0.998	0.939	0.002	0.006

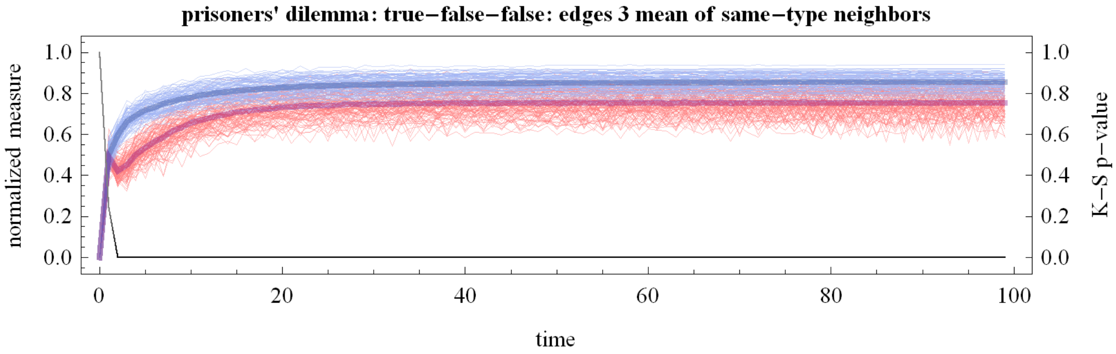
Table 3.6: Results from Prisoners' Dilemma Across Population Size

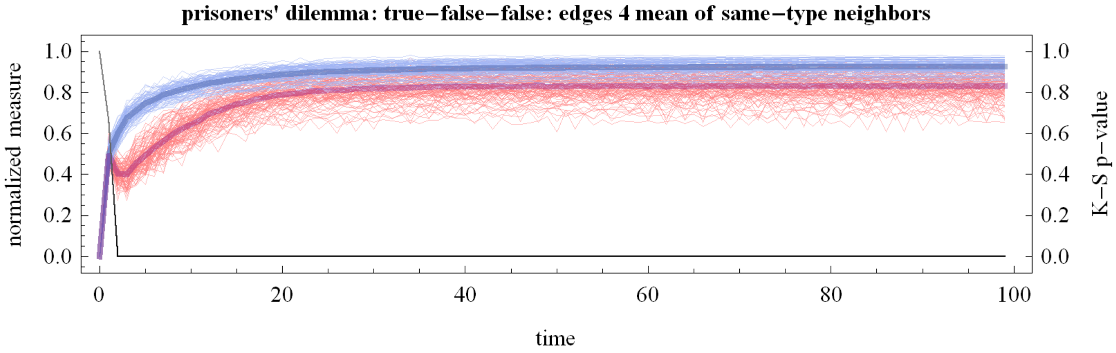
3.8 Branch Two Results: Sweeping Maximum Degree

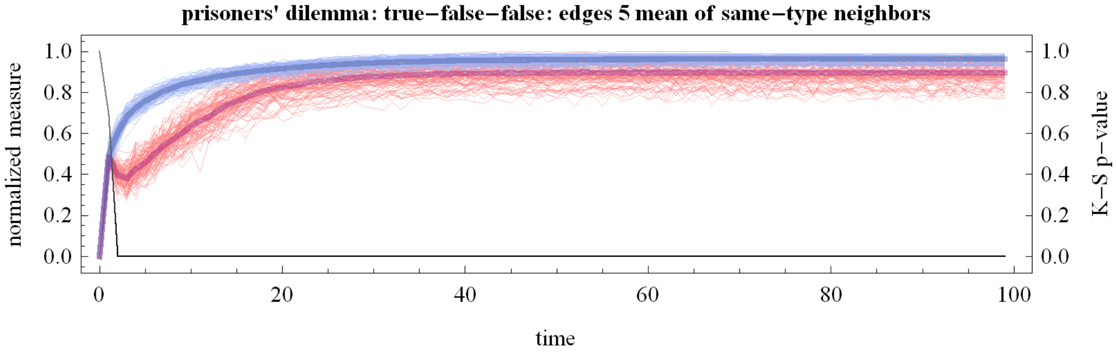
We now turn to a set of experiments that sweep the number of social connections that agents are allowed to maintain. Recall from 1.4 that the default value for maximum degree ($K = 5$) was based on neurological and anthropological research regarding primates' abilities to keep track of social relationships [Dunbar 1993]. To accommodate the uncertainty produced by extrapolating these results to modern humans, and to expand the range of applicability of the model, I now present the results of running the simulations with maximum degree ranging from three to twenty. All simulations in this section use one-hundred agents of each type.

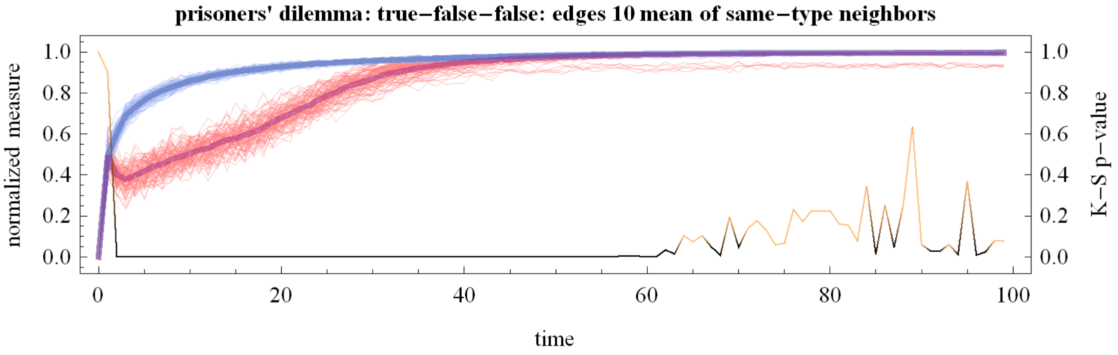
3.8.1 Prosociality of Preferential Detachment Alone

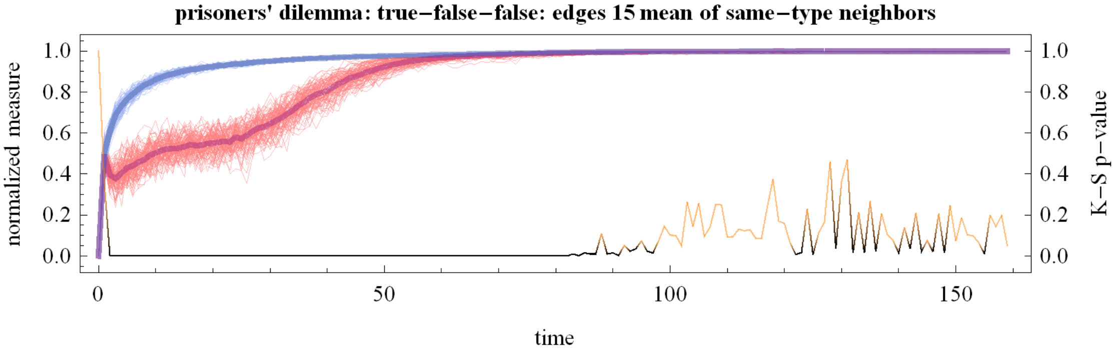
We first examine the effects of changing the maximum degree on the base preferential detachment model. Recall that in these scenarios, because agents cannot change type, the prosocial outcome is the formation of an exclusively A-type group. As a measure of success we again present the percent of same-type neighbors averaged over the agents in a run. The plots below reveal the change in this measure over time for each $K = 3, 4, 5, 10, 15, 20$.

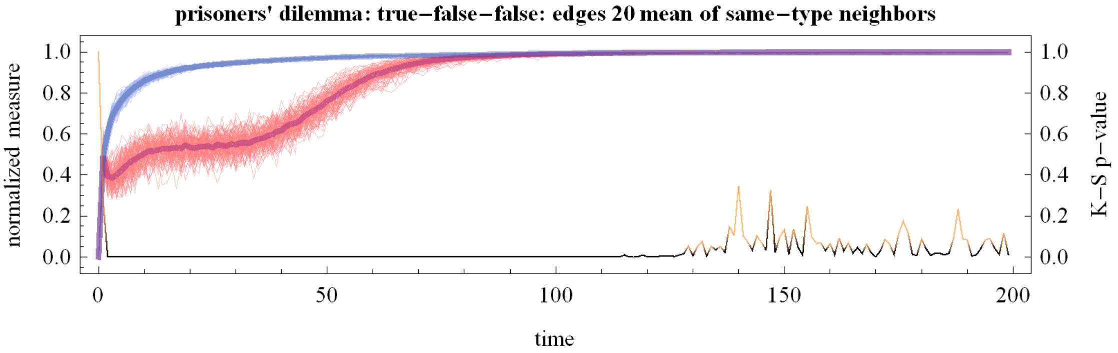




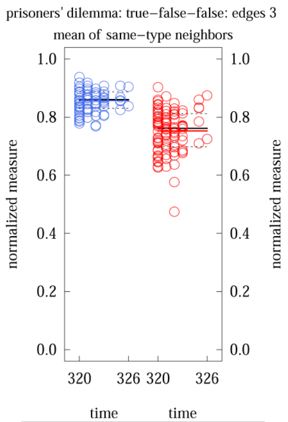



prisoners' dilemma: true−false−false: edges 10 mean of same−type neighbors

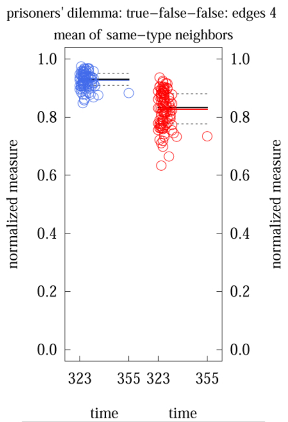

prisoners' dilemma: true−false−false: edges 15 mean of same−type neighbors


prisoners' dilemma: true−false−false: edges 20 mean of same−type neighbors

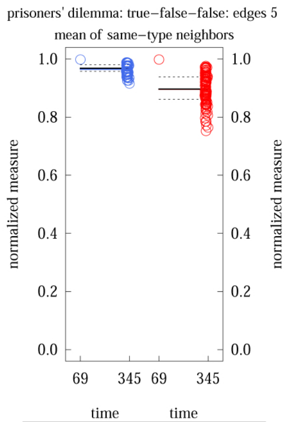
During the transition period, increasing values of K decrease variability of the cooperators, but not the defectors. The trajectory of the cooperators is similar at all levels; they consistently increase the percent of cooperator neighbors, but at a decreasing rate as more available cooperator connections are used up. Defectors, on the other hand, experience longer transition periods with increasing K through which they have lower overall degree and a few connections to cooperators. Recall that in the current implementation agents can only detach one agent per iteration. Since defectors will also detach defectors if they have a cooperator neighbor, defectors form sparse arrangements because they shed defector neighbors while trying to maintain cooperator neighbors. Eventually the cooperators form an exclusive group, after which time defectors settle on keeping connections with other defectors and saturate their configurations. This dynamic does not effect the trajectory of cooperators in the plots because even though when K is larger they have more defector neighbors for longer, the total number of neighbors is also larger, and so the percentage is nearly the same.



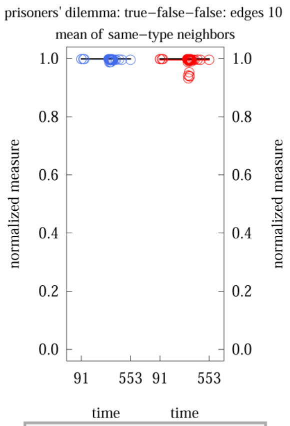
Mean:	0.857	0.753
Maximum:	0.940	0.905
Upper Quartile:	0.887	0.812
Median:	0.860	0.762
Lower Quartile:	0.830	0.698
Minimum:	0.770	0.477
K–S p–value:	0.000	



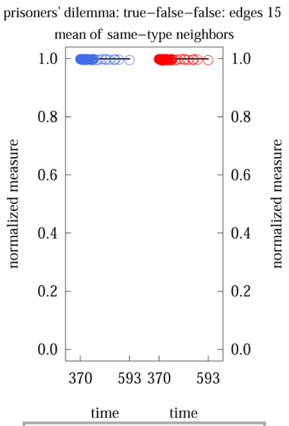
Mean:	0.927	0.827
Maximum:	0.977	0.938
Upper Quartile:	0.950	0.880
Median:	0.930	0.833
Lower Quartile:	0.910	0.777
Minimum:	0.850	0.635
K–S p–value:	0.000	



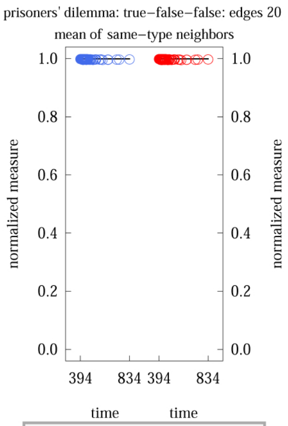
Mean:	0.965	0.896
Maximum:	1.000	1.000
Upper Quartile:	0.980	0.938
Median:	0.968	0.896
Lower Quartile:	0.958	0.862
Minimum:	0.918	0.754
K–S p–value:	0.000	



Mean:	0.999	0.996
Maximum:	1.000	1.000
Upper Quartile:	1.000	1.000
Median:	0.999	0.999
Lower Quartile:	0.998	0.998
Minimum:	0.989	0.934
K–S p–value:	0.033	



Mean:	0.999	0.999
Maximum:	1.000	1.000
Upper Quartile:	1.000	1.000
Median:	0.999	0.999
Lower Quartile:	0.999	0.998
Minimum:	0.997	0.995
K–S p–value:	0.521	



Mean:	0.999	0.999
Maximum:	1.000	1.000
Upper Quartile:	1.000	0.999
Median:	0.999	0.999
Lower Quartile:	0.998	0.998
Minimum:	0.997	0.997
K–S p–value:	0.062	

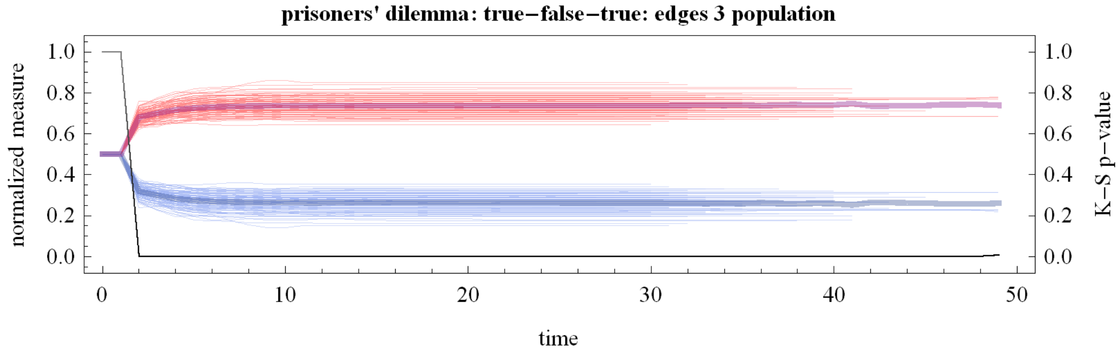
The behavior plots show that for values of $K = 10$ or larger the outcomes of both types are statistically indistinguishable: near perfect assortativity. But note that when $K = 10$ there are five (of 100) runs that settle to an outcome of noticeably less than 100% self-similarity. These reveal the presence of stable heterogeneous groups. Other runs also have outcome levels of mean same-type neighbors less than 100%, but these minor deviations result from *unstable* connections between cooperators and defectors. These unstable connections form when the arrangement of the giant components of each type leave a cooperator with open connections that cannot be satisfied by another cooperator. Such a cooperator constantly attaches and detaches defectors: and this is why the end times are so large ($t > 300$) even though the agents have clearly settled by $t = 100$ in most cases.

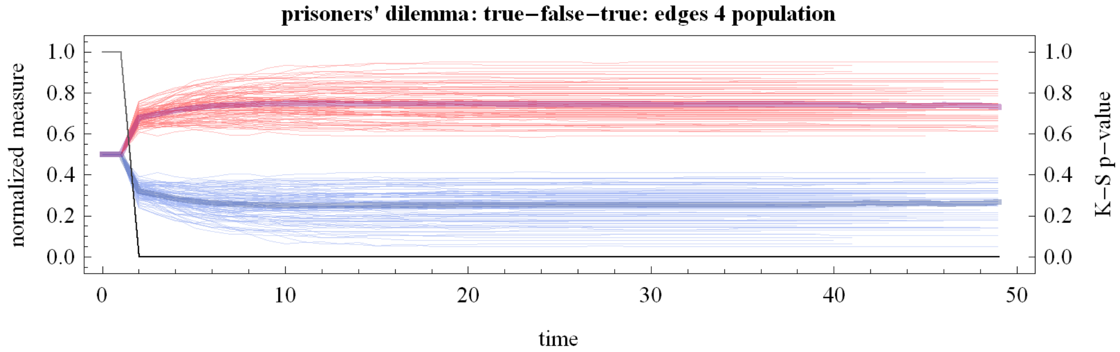
Stable heterogeneous groups are common at smaller values of K, but disappear before $K = 15$. In order for a cooperator to get stuck in a heterogeneous group it must have maxed out its number of allowable connections with defectors through consecutive random attachments (recall that this can happen in one iteration if several defectors happen to randomly attach to a focal cooperator). The larger the maximum degree, the lower the probability that a cooperator will find itself with all defector neighbors. And as long as a cooperator has at least one cooperator neighbor it will eventually shed off all its defector neighbors through preferential detachment. Thus $K = 10$ is likely near the threshold value for the persistence of heterogeneous groups. Future work will investigate more deeply this threshold effect and its sensitivity to other parameters.

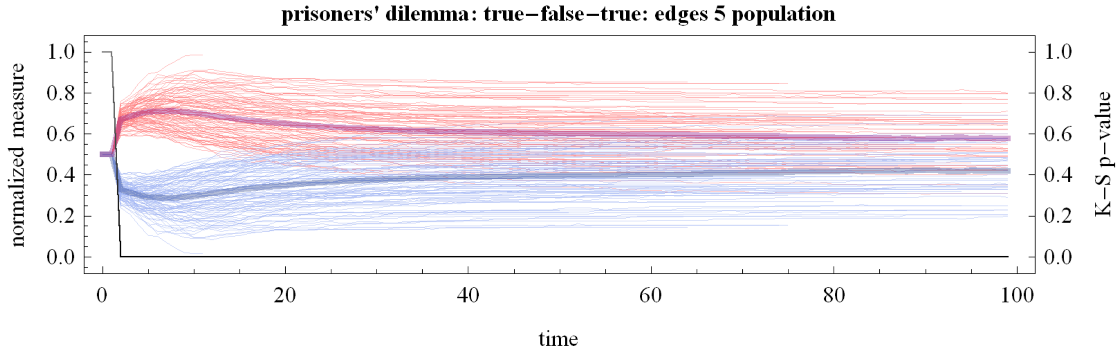
The success of preferential detachment in achieving prosociality across ranges of maximum degree is clear from these results. The worse-case run was 77% successful in terms of self-similarity, and the expected success level is above 85%. Those both occurred at the smallest sensible maximum degree value of three. As K increases the expected success level increases *and* the outcomes become more consistent. Whatever the most appropriate value for K may be for a given species, preferential detachment would largely succeed in forming prosocial arrangements in the Prisoners' Dilemma contexts with one-hundred agents of each type.

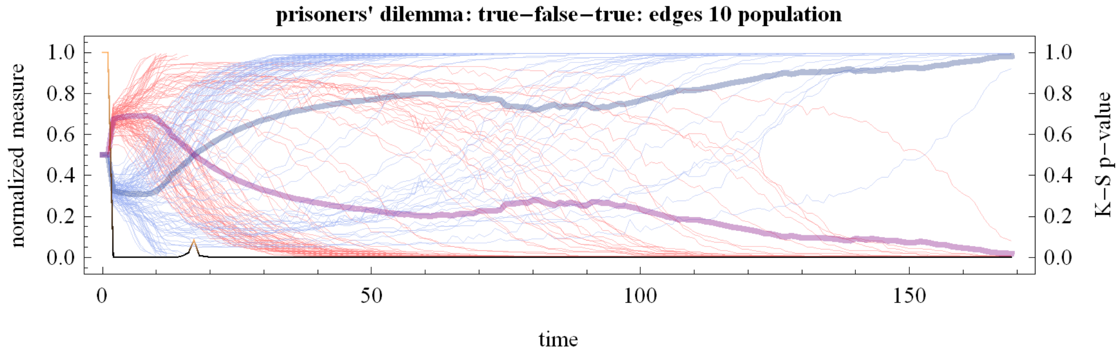
3.8.2 Prosociality with Imitative Learning

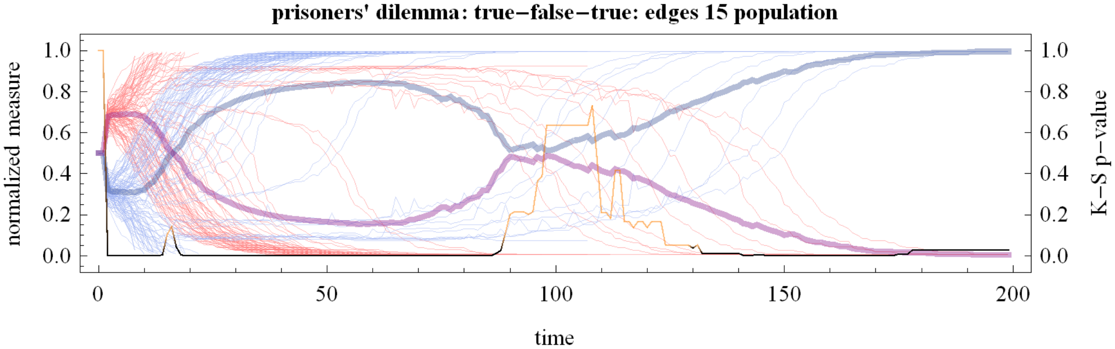
With imitative learning now added to preferential detachment we are seeking to discover what effect the number of edges has on cooperator dominance. Because learning is a local behavior, agent connectability is expected to play a large role in the success of cooperators. In mixed configurations the defectors are more successful; it is only after a cooperator is itself attached to multiple cooperators that a defector attached to that subgroup would become a cooperator. Furthermore, because defectors are detached before the learning step, a cooperator must have two defector neighbors for one (the one not detached) to be able to copy the cooperator. So if there are too few connections then cooperation will not be able to spread, but if there are too many then it may be difficult for cooperators to exclude defectors.

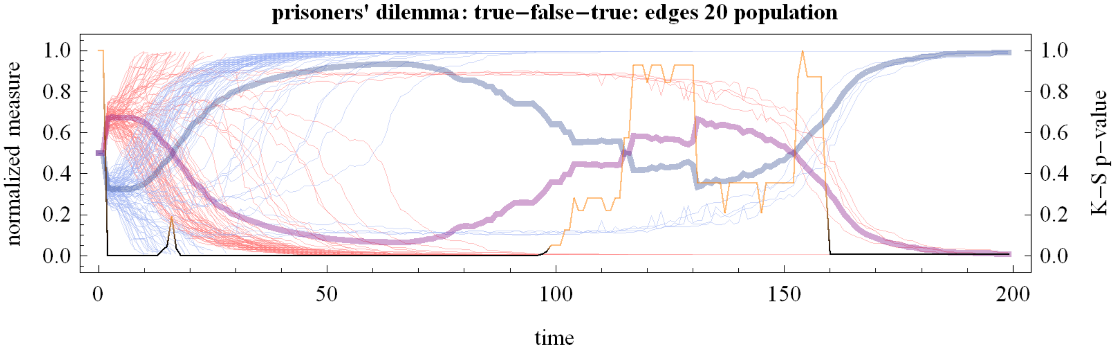








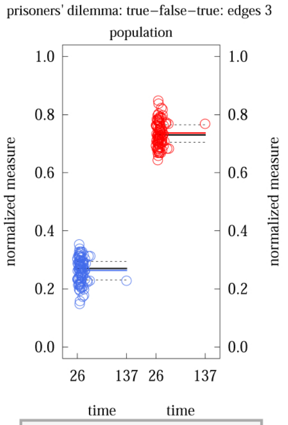




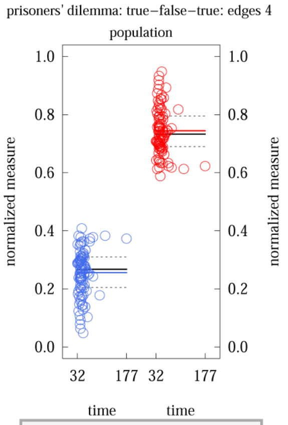
When the maximum degree is low, agents quickly fill their maximum degree. Because cooperators need two attached defectors for cooperation to spread (one is detached before it can learn), and this is less likely with fewer edges, low degree means that defectors are less likely to imitate the cooperative type. Initially the defectors are doing better, so the imitation goes the other way, and then preferential detachment produces segregated neighborhoods.

As we increase K, the model exhibits a dramatic qualitative shift in behavior. When $K = 3$ each run ends with segregated groups in which the cooperators make up approximately 26% of the population. As K goes to four, then five, the outcome population level of cooperators shifts upward, though defectors still get the same benefit in early periods. Clearly the increased connectivity is allowing cooperators to influence defectors after they've formed their community, and hence convert more of them. The variability of the runs also *increases* with greater maximum degree; so as the average behavior shows greater cooperation, the runs take more extreme trajectories. This trend of increasing cooperation continues so that by the time $K = 10$ cooperators can rise to dominance, but the increased variation trend also increases so that more runs end with a quick rise of defection. Depending on what happens in the first five or ten iterations, the outcome will move to the domination of one or the other type.

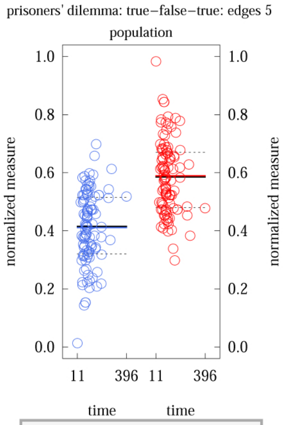
A closer look at the behavior plots for $K = 10, 15, 20$ shows that in a few runs there is a small population of cooperators that persists, separated from the large defector component, for a hundred or more iterations, and then explodes to dominate. These sleeper cells indicate that the cooperator community is highly stable, but that there are a few connections to agents outside the core group. It takes a particular combination of micro-arrangements at the edges of the two communities to foster the cooperative recovery and the increasing time until this happens reveals that higher maximum degree complicates the series of maneuvers necessary for cooperator dominance. This complication may also explain why more runs end with cooperator dominance when $K = 10$ than when $K = 15$ or 20.



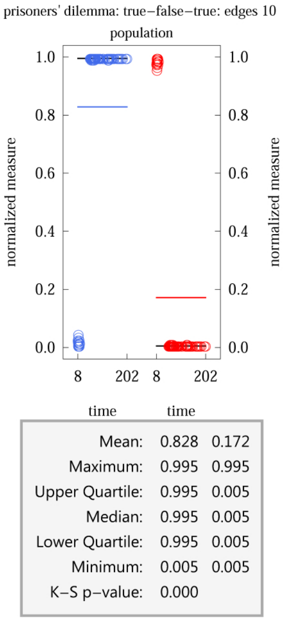
Mean:	0.264	0.736
Maximum:	0.355	0.850
Upper Quartile:	0.295	0.765
Median:	0.270	0.730
Lower Quartile:	0.230	0.705
Minimum:	0.150	0.645
K−S p−value:	0.000	



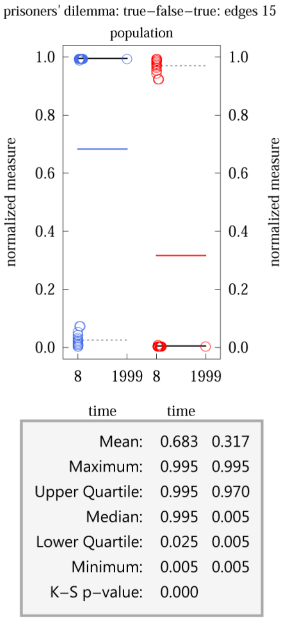
Mean:	0.256	0.744
Maximum:	0.410	0.950
Upper Quartile:	0.310	0.795
Median:	0.268	0.732
Lower Quartile:	0.205	0.690
Minimum:	0.050	0.590
K−S p−value:	0.000	



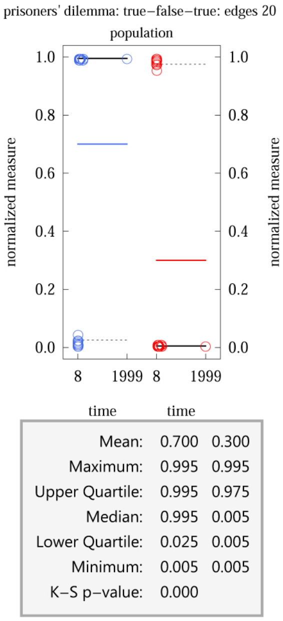
Mean:	0.411	0.589
Maximum:	0.700	0.985
Upper Quartile:	0.515	0.670
Median:	0.415	0.585
Lower Quartile:	0.320	0.480
Minimum:	0.015	0.300
K−S p−value:	0.000	



Mean:	0.828	0.172
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.005
Median:	0.995	0.005
Lower Quartile:	0.995	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	



Mean:	0.683	0.317
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.970
Median:	0.995	0.005
Lower Quartile:	0.025	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	



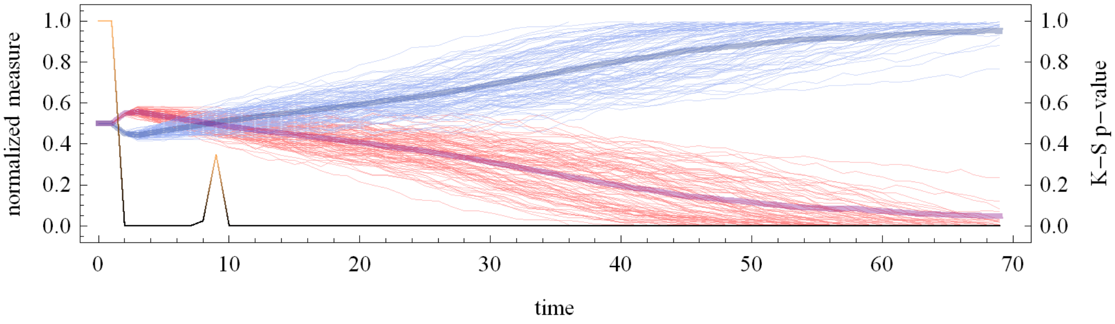
Mean:	0.700	0.300
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.975
Median:	0.995	0.005
Lower Quartile:	0.025	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	

The outcome plots show that for larger values of the maximum degree, stable and separated homogeneous groups do not form. The behavior plots show that such arrangements almost form in a few runs, but over enough time the necessary arrangement is reached for the spread of cooperation. The results present a very interesting story: low connectivity guarantees separated populations with low (less than 30%) cooperators, and high connectivity produces total cooperation, but only in approximately 70% or the runs. So the expected level of prosociality rises consistently with increasing K, but there is a trade off of reliable low levels or high levels with a moderate risk defector domination. That such a qualitative shift results from this parameter sweep reveals the effect that maximum degree has on the agent configurations and network arrangements, and the agent behavior with respect to those configurations and arrangements. I present more implications of this result for morality and institutions in the conclusions (3.10) and in the next chapter.

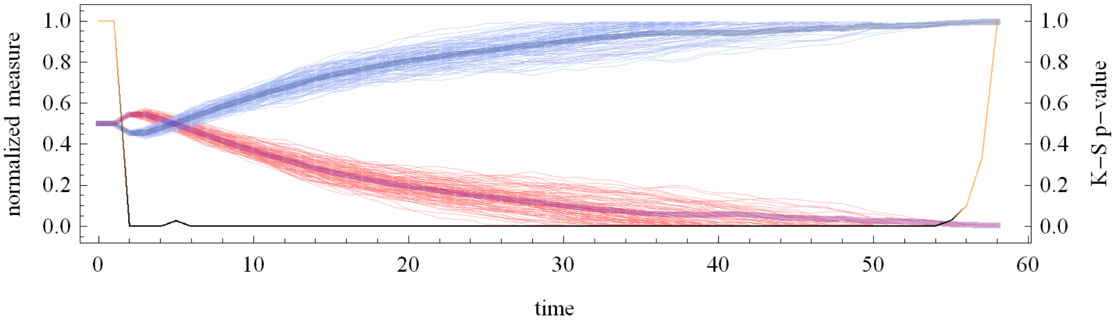
3.8.3 Prosociality with Population Dynamics

The global mechanism of population dynamics in the Prisoners' Dilemma can only end in complete dominance of either cooperators defectors. We will see below that in every run of the experiments here the cooperators completely succeed in achieving the prosocial outcome. The effects of changing the maximum degree are, therefore, revealed in the path taken to reach the outcome. This can manifest in the time needed to reach dominance, or in the specific trajectory that the population levels take.

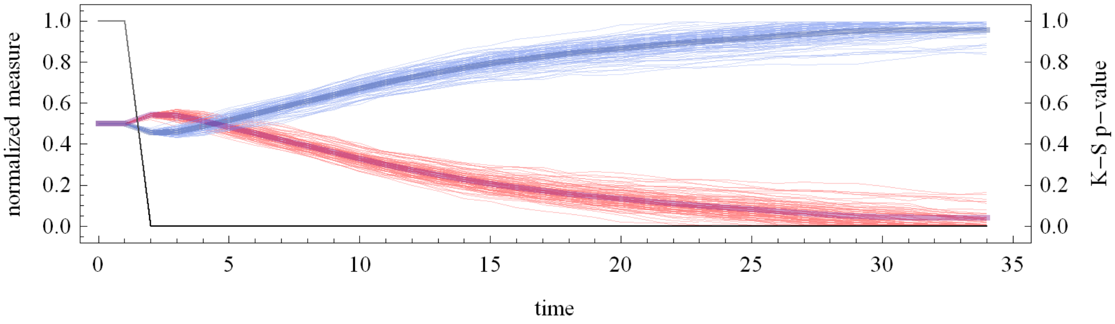
prisoners' dilemma: true−true−false: edges 3 population



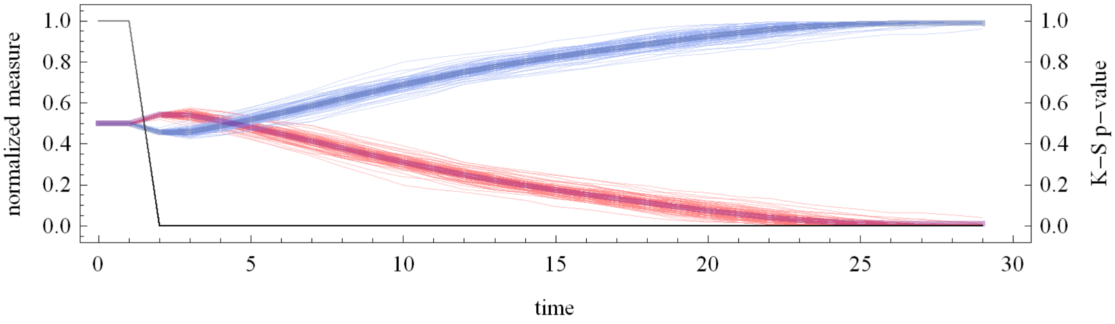
prisoners' dilemma: true−true−false: edges 4 population

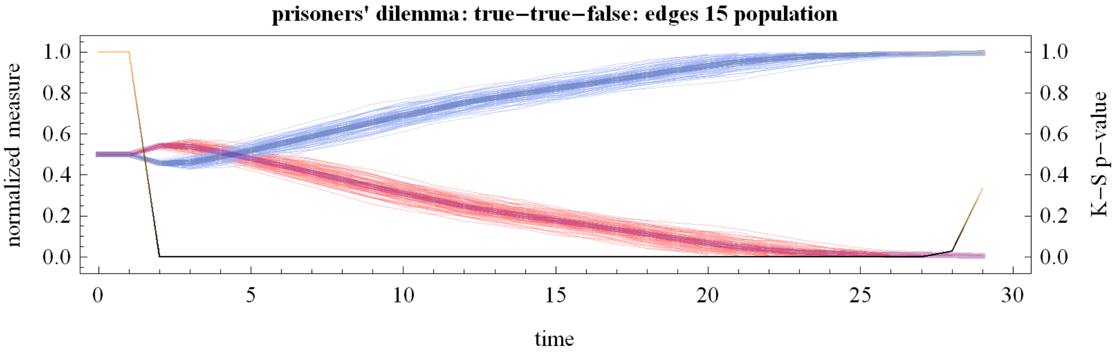


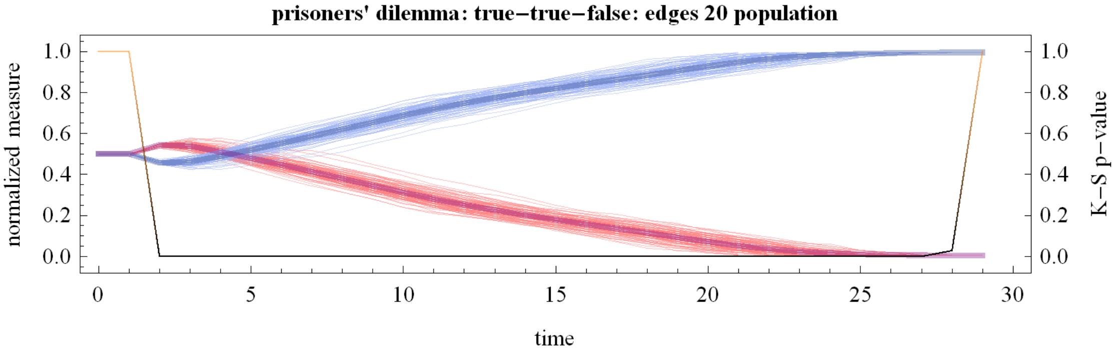
prisoners' dilemma: true−true−false: edges 5 population



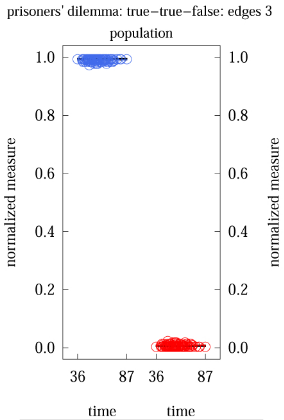
prisoners' dilemma: true−true−false: edges 10 population



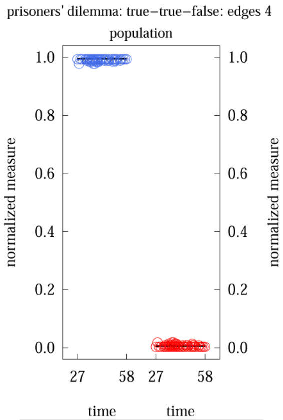




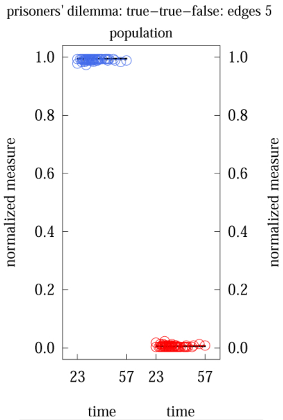
As we have by now come to expect, increases in the maximum degree reduce the variability among the runs. The reason is again the fact that with greater degree the probabilities faced by agents become more similar. Particular events, such as a cooperator receiving three or four defector connections in one early turn, can produce large and long-lasting effects on the dynamics when that many additions saturates an agent's network neighborhood. The probability of receiving k connections in one iteration is nearly unchanged with greater K because each agent only acts once per iteration (recall 2.6.1.1 from the Markov model). Since $K - k$ is increasing with K, the opportunity to bounce back from low probability events becomes greater.



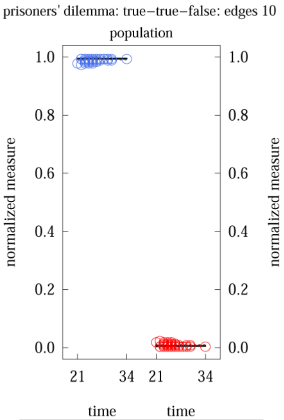
Mean:	0.992	0.008
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	



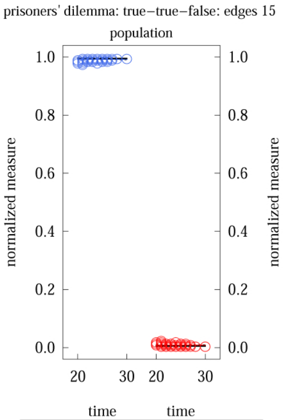
Mean:	0.993	0.007
Maximum:	0.995	0.020
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.980	0.005
K–S p–value:	0.000	



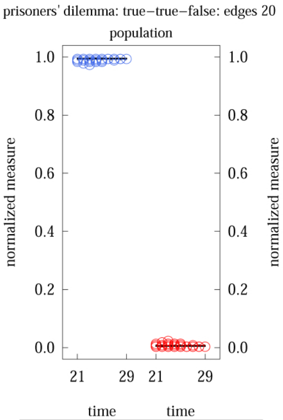
Mean:	0.993	0.007
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	



Mean:	0.992	0.008
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	

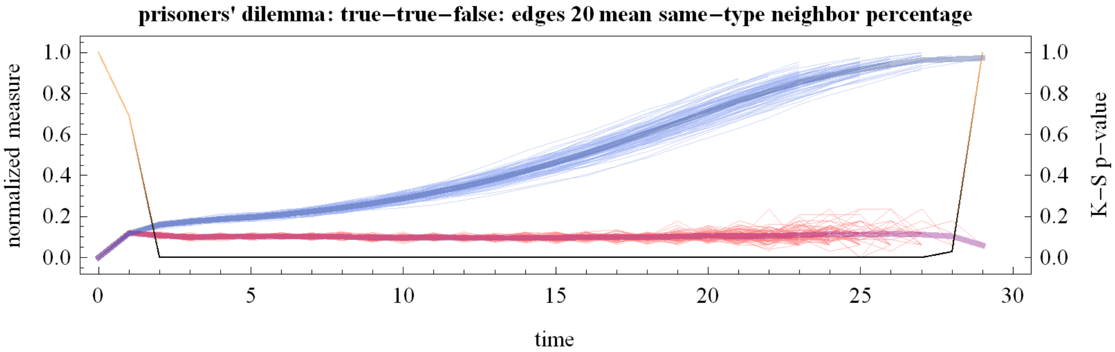


Mean:	0.992	0.008
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	



Mean:	0.993	0.007
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	

The outcome plots confirm the result that cooperator dominance is the concluding state in every run of the simulation at all parameter values. The minor deviations seen in the plot result from the satisfaction of another halting condition just before complete dominance is achieved. We can see from these plots that there is a trend toward decreasing run times and less variation among run times. In the case of $K = 20$ the population converges on cooperators in only 21 to 29 steps. We can see from the plot of mean same-type neighbor percentage (the utilization measure) below that in most runs the cooperators dominate the population before all the agents saturate their neighborhood. This feature of underutilization will be important in the discussion of the evolution of morality and for the general scaling properties of preferential detachment discussed in the conclusions section of this chapter (3.10).

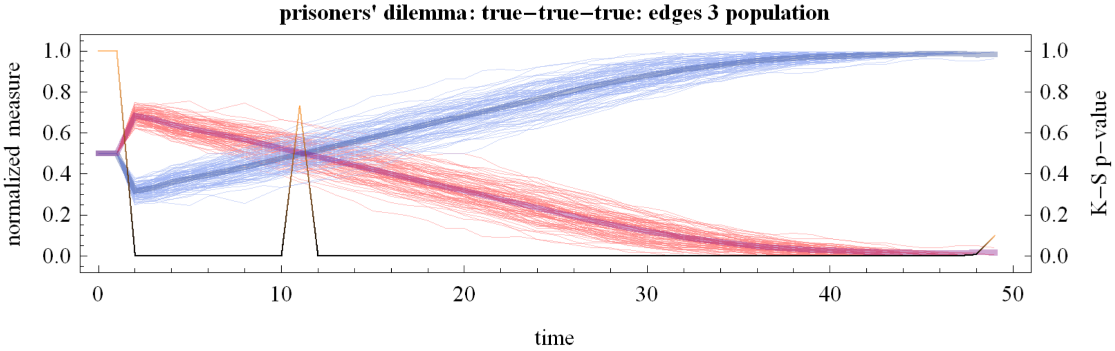


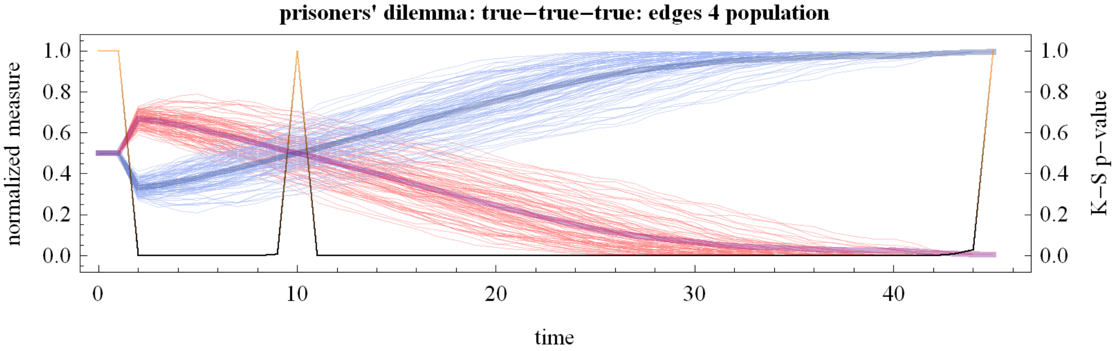
What these results tell us is that even for cases in which preferential detachment alone is less than completely successful in segregating the population (i.e., $K = 3, 4, 5$), it *is* consistently successful in producing a group of cooperators. Because cooperators in a group of cooperators receive greater payoffs than defectors in defector groups (and even defectors in typical stable heterogeneous groups), population dynamics ensures that the number of cooperators grows. Once the highest earning five percent of agents (ten agents) are cooperators through preferential detachment's effects, it would only take ten iterations for cooperators to dominate from an evenly split total population. And this happens regardless of whatever other structural or behavior properties the system exhibits. Longer run times indicate that some of the top performing agents are defectors for longer; a property of arrangements in which defectors maintain connections with cooperators.

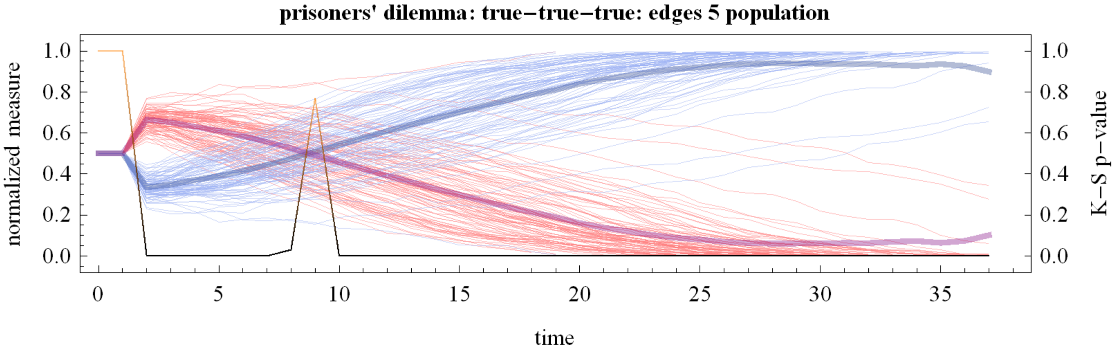
Since all runs end in cooperator dominance it is clear that achievement of the prosocial outcome in the Prisoners' Dilemma with population dynamics is completely insensitive to the maximum degree parameter. The transition behavior is sensitive to K up to a point, and in ways that are completely expected from an analysis of the probabilities of connections and the process of detachment.

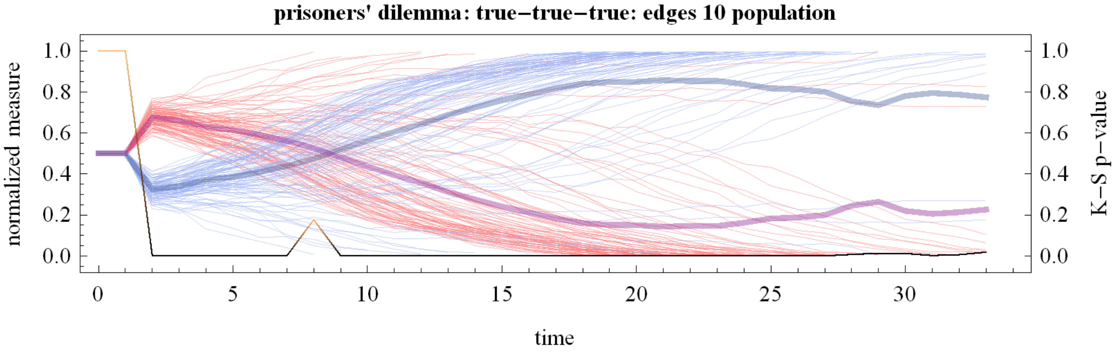
3.8.4 Prosociality with Learning and Population Dynamics

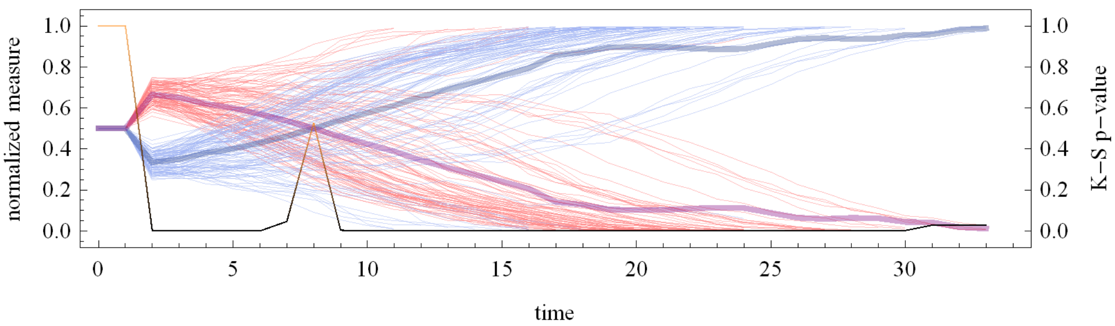
The results below show that the system behavior of the combined mechanisms does indeed combine aspects of the three mechanisms presented above. The overall behavior most closely resembles the plots of population dynamics, but with increasing (rather than decreasing) run variability with increasing K as seen with learning.

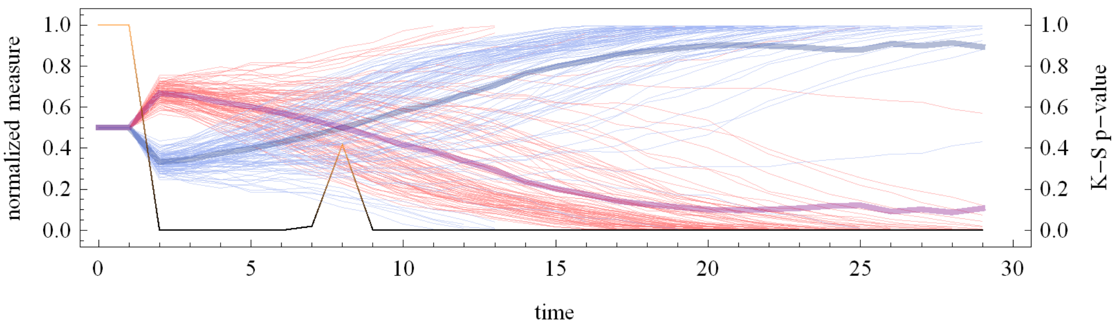

prisoners' dilemma: true−true−true: edges 3 population


prisoners' dilemma: true−true−true: edges 4 population


prisoners' dilemma: true−true−true: edges 5 population

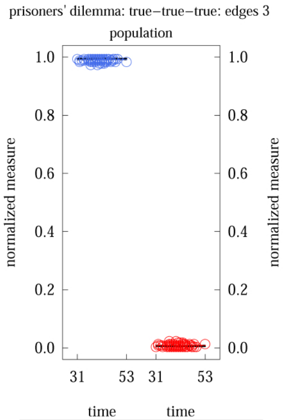

prisoners' dilemma: true−true−true: edges 10 population


prisoners' dilemma: true−true−true: edges 15 population

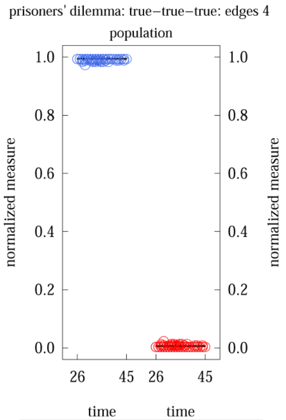

prisoners' dilemma: true−true−true: edges 20 population

When the maximum degree is low, the agents separate into segregated groups; and then group of cooperators performs better so they take over the population through replication. As we turn up the connectivity of the network, learning becomes a greater force in shaping the distribution of types and their social arrangements. This allows for more defectors to attach to more cooperators early on, and hence become more likely to be imitated. With $K = 5$ one of 100 runs ends with defector dominance, and this number increases at higher K. The combined mechanisms generate greater prosociality than learning alone, but less than population dynamics alone.

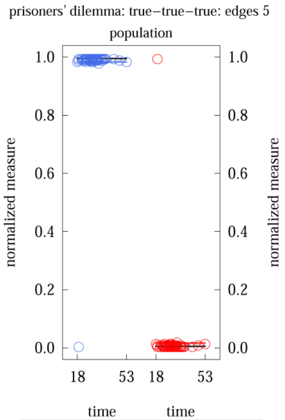
The results reveal the role of luck in defector success. Their higher initial success must be sufficiently established to disrupt cooperator group formation. This is not just a matter of the number of defectors increases by a threshold amount. It is critical that two defectors attach to each cooperator for defection to spread. So even though defectors get an early population bump in every run, they only succeed in generating the necessary arrangement in a moderately small number of runs (fewer than twelve in the worst case).


prisoners' dilemma: true–true–true: edges 3

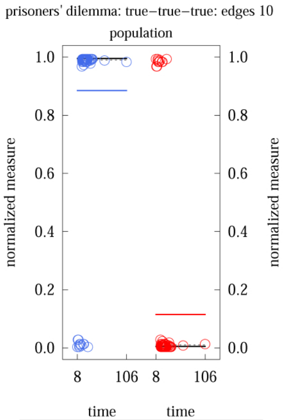
Mean:	0.991	0.009
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	


prisoners' dilemma: true–true–true: edges 4

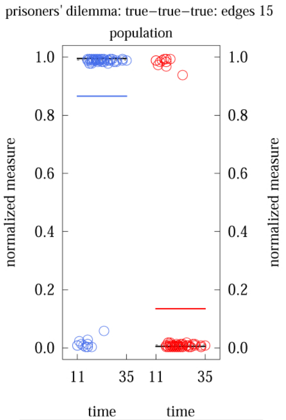
Mean:	0.993	0.007
Maximum:	0.995	0.025
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.975	0.005
K–S p–value:	0.000	


prisoners' dilemma: true–true–true: edges 5

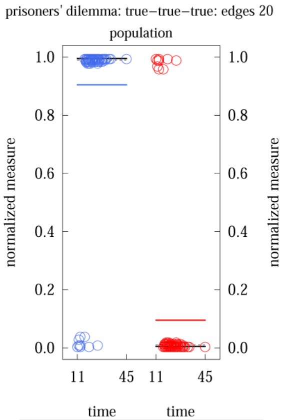
Mean:	0.983	0.017
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	


prisoners' dilemma: true–true–true: edges 10

Mean:	0.885	0.115
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	


prisoners' dilemma: true–true–true: edges 15

Mean:	0.865	0.135
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	


prisoners' dilemma: true–true–true: edges 20

Mean:	0.905	0.095
Maximum:	0.995	0.995
Upper Quartile:	0.995	0.010
Median:	0.995	0.005
Lower Quartile:	0.990	0.005
Minimum:	0.005	0.005
K–S p–value:	0.000	

With the combined mechanisms every outcome is an extreme-valued outcome. For low K this is always the prosocial outcome, but the number of defector-dominated runs increases between $K = 5$ and 10. Still, the level of prosociality is quite high and this result is important for the evolution of morality in human populations. If we consider each run as an independent protohuman group competition, then what these results show is that in some cases the defectors will succeed in taking over the population, but usually they won't. The implications of this are discussed below.

3.8.5 Summary of Maximum Degree Sweep

In each case presented above, a greater maximum degree either improved the prosociality of the simulation outcome, or there was no room for improvement. For the base model, increasing the maximum degree drastically reduces the probability of cooperators gaining K consecutive defector neighbors and getting stuck in stable heterogeneous groups – the only arrangement for which preferential detachment is insufficient to bring to assortative outcomes. To gain dominance with population dynamics and/or learning, cooperators need to earn higher payoffs than defectors. To have higher payoffs, the cooperators need to form a community sufficiently large so that they are connected to mostly other cooperators.

The only barrier to cooperator dominance with learning is the fact imitative learning is a local behavior. If agents segregate too quickly, then populations of both types will endure regardless of their relative payoffs. However, if there is insufficient separation, the defectors can take over before cooperators form a community. Increasing the maximum degree fosters more agent interaction, which has two effects: it helps cooperators eventually take over the whole population if they succeed in forming the core community, but it also helps defectors dominate if they get sufficient momentum early on. We see a trade-off in outcomes between a guarantee of a moderately low percentage of cooperators with low K, or a moderately high probability of cooperator dominance tempered by an alternative of defector dominance for higher K.

With population dynamics added, every run ends with complete cooperator dominance; and it is clear why increasing the maximum degree increases the speed at which cooperators can dominate the society. A high maximum degree reduces the number of heterogeneous groups, which is the only arrangement in which defectors can outperform cooperators. Because preferential detachment alone ensures mostly assortative groups, and the agents in cooperator groups perform better than one in defector groups, this cooperator dominated outcome may have appeared ex ante obvious. But keep in mind that during the early stages of most runs the defectors outperform the cooperators before stable groups have formed. Thus it is possible for defectors to win out in some scenarios – future work will explore the boundaries of the parameter space at which cooperators vs defectors dominate the population with population dynamics.

The combination of preferential detachment, imitative learning, and population dynamics is quite successful in supporting cooperation. Interestingly, the effects of learning are overwhelmed by the other two mechanisms when K is low. However, when K increases to ten or higher, the increased connectivity plus a constant supply of new, unconnected agents, provides the opportunity for defectors to take over in some scenarios. This is relatively rare, and the prosocial outcome of cooperator dominance is observed in 86.5% the worst-performing experiment ($K = 15$).

The collective results of this branch of experiments is that for a variety of behavioral variations the establishment of successful cooperative communities is highly probable, and it becomes more probable with increasing numbers of immediate social interaction partners. These results do not depend on any particular arrangement of the social network except those wrought through the preferential detachment mechanism. The changes in system behaviors for larger maximum degree are produced through the altered connection probabilities, which corresponds to the effect captured in the Markov model of chapter II. The result that prosociality *increases* with increasing network connections without learning, but decreases with learning, is important for topics discussed in the conclusions section.

Preferential Detachment

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
$K = 3$	0.500	0.857	0.500	0.857
$K = 4$	0.500	0.927	0.500	0.827
$K = 5$	0.500	0.965	0.500	0.896
$K = 10$	0.500	0.999	0.500	0.996
$K = 15$	0.500	0.999	0.500	0.999
$K = 20$	0.500	0.999	0.500	0.999

Learning

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
$K = 3$	0.264	0.995	0.736	0.997
$K = 4$	0.256	0.994	0.744	0.997
$K = 5$	0.411	0.990	0.586	0.996
$K = 10$	0.828	0.849	0.172	0.175
$K = 15$	0.683	0.737	0.317	0.318
$K = 20$	0.700	0.733	0.300	0.311

Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
$K = 3$	0.992	0.935	0.008	0.037
$K = 4$	0.993	0.935	0.007	0.008
$K = 5$	0.993	0.936	0.007	0.004
$K = 10$	0.992	0.941	0.008	0.003
$K = 15$	0.992	0.945	0.008	0.012
$K = 20$	0.993	0.946	0.007	0.000

Learning and Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim
$K = 3$	0.991	0.935	0.009	0.009
$K = 4$	0.993	0.937	0.007	0.014
$K = 5$	0.983	0.927	0.017	0.039
$K = 10$	0.885	0.864	0.115	0.135
$K = 15$	0.865	0.845	0.135	0.122
$K = 20$	0.905	0.877	0.095	0.084

Table 3.7: Results from Prisoners' Dilemma Across Maximum Degree

3.9 Branch Three Results: Sweeping Strategic Contexts

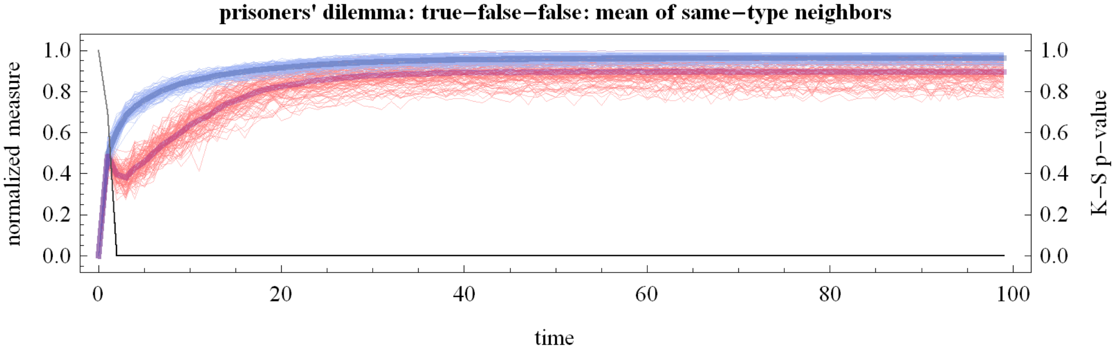
Now that we have seen how sensitive the results of preferential detachment are to changes in the two most important numerical parameters, we now explore the results for each strategic context in the game library (1.8). One hundred simulations for each payoff matrix for each of the four sets of behavior rules are performed with an initial population of two hundred total agents, and a maximum degree of five (as explained in detail in section 3.5). Though some comparisons are made throughout the results presentation, the next section is used to highlight differences and similarities of results among the strategic contexts, parameter values, and results of the Markov Model. The focus of this section is to describe what happened in each experiment, and to explain that system behavior in terms of the mechanisms of the agent-based model.

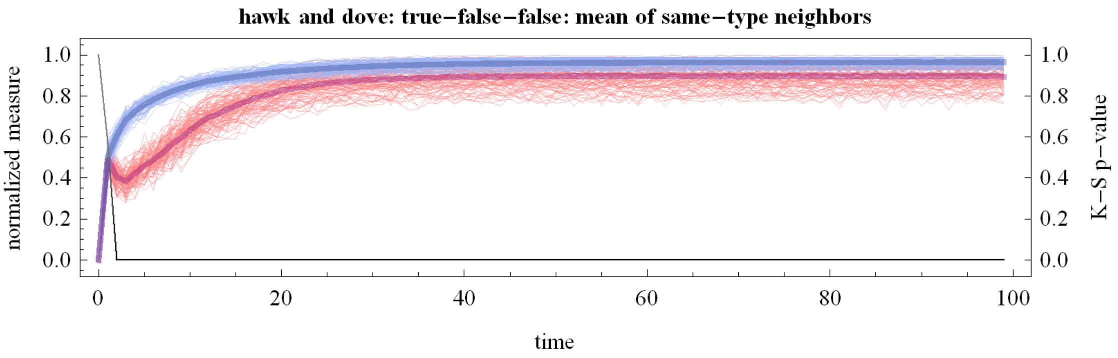
3.9.1 Prisoners' Dilemma and Hawk and Dove

The results of the Prisoners' Dilemma in all the scenarios covered here have already been revealed and discussed in the previous two parameter sweeps and do not need to be given detailed covered again. However we will see the relevant output plots immediately below because of their relevance to the Hawk and Dove game, and to points made about the categorization of the games in chapter I. Here it suffices to recap the high, but not perfect, level of achievement of prosociality in the base model and with learning, and the perfectly prosocial outcome whenever population dynamics are involved.

The Hawk and Dove game shares the cooperative category with the Prisoners' Dilemma. What this means is that the payoff structures are indistinguishable to agents who only have access to the relative utility of the game's payoffs. In this category agents must resist a temptation to defect for a greater payoff, and rather cooperate which, if mutual, earns a good, but suboptimal, payoff. But the cardinal payoffs for the two contexts are distinct, and thus produce distinct environments for learning and population dynamics.

As expected from the description of the theory, the experiments without learning or population dynamics involving games from the same category produce qualitatively identical results. Within each plot it is clear that the variation among the runs is small, and that each run follows the mean value closely. Comparing the run values from the two experiments reveals that they are insignificantly different – the minimum K-S p-value is 0.204, which means that we cannot be 95% confident that at any time step the two sets of run data came from distinct continuous distributions. Presenting these plots and statistics serves as a check that the agent-based model correctly represents the preferential detachment theory.





For what follows it is helpful to recall how the payoff matrices of the two games differ:

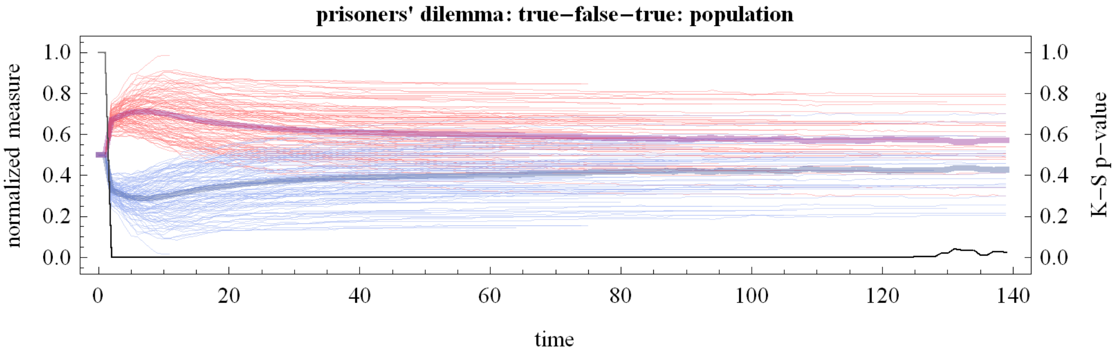
Prisoners’ Dilemma

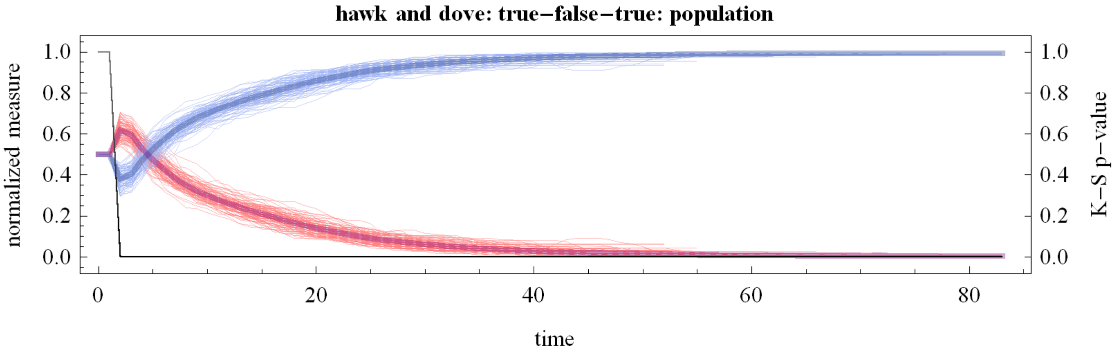
		player2	
		A	B
player1	A	3,3	1,4
	B	4,1	2,2

Hawk and Dove

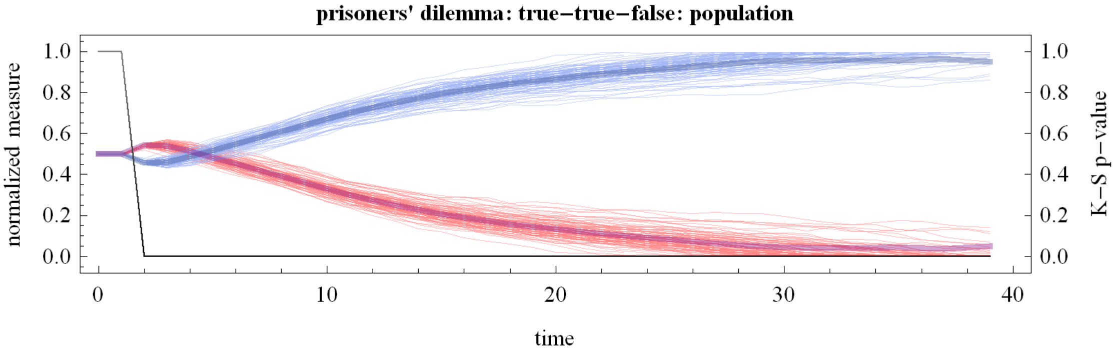
		player2	
		A	B
player1	A	3,3	2,4
	B	4,2	1,1

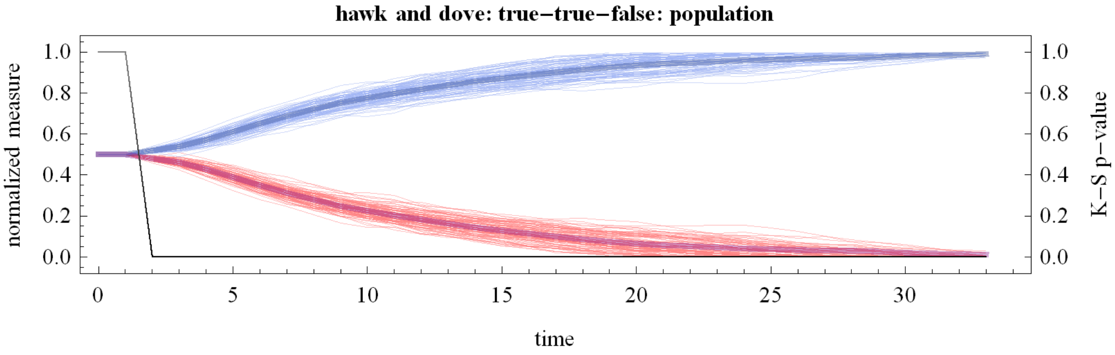
In game theoretic terms the difference is that in the Prisoners’ Dilemma, defection is a dominant action and hence (B, B) is the only Nash equilibrium; whereas in Hawk and Dove there are two the Nash equilibria of (A, B) and (B, A) and a mixed strategy equilibrium. In preferential detachment theory the difference materializes as a greater payoff to A-types connected to B-types (4 vs 2) and lower B-to-B payoffs (2 vs 4) in Hawk and Dove compared to Prisoners’ Dilemma situations. This difference does not alter the payoff ranking of neighbors, but it does affect the utility scores used by learning and population dynamics. The effects of learning on these differing payoffs are revealed by the differences in system behavior in the following plots.



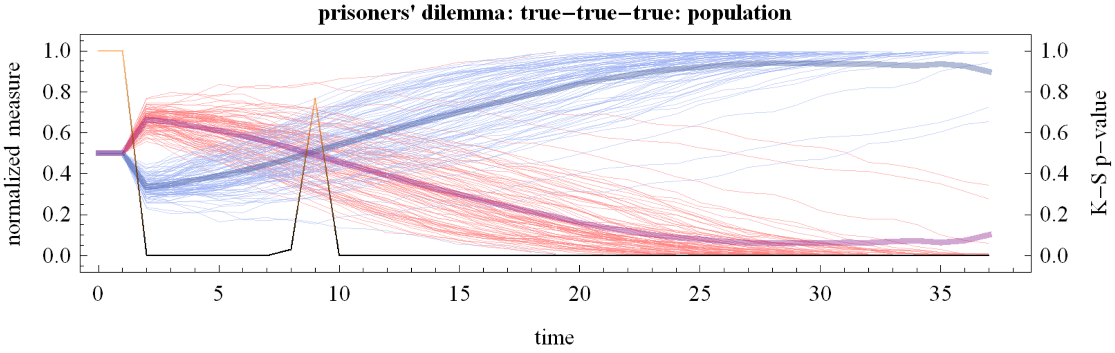


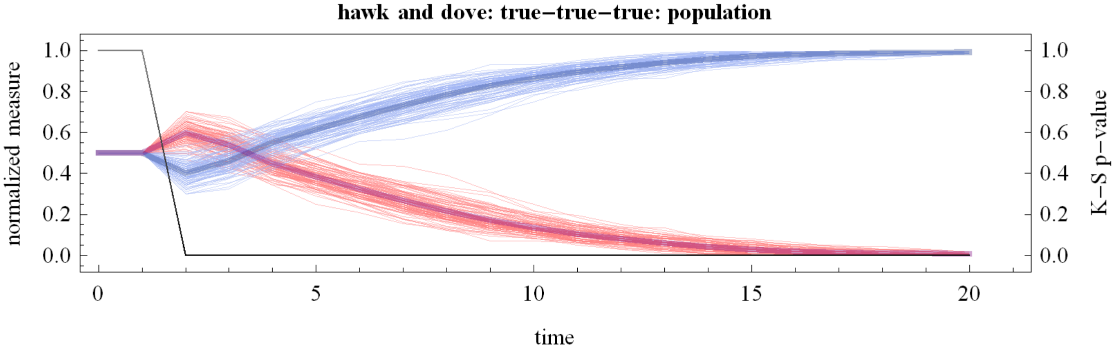
In Hawk and Dove there is greater cohesiveness and convergence to a greater degree of prosociality (i.e., higher A-type dominance). The difference in A-type performance between the two games reveals that the details of the payoff values *do* make a difference when a local imitation rule is added to the agent actions. The process through which this happens is clear: B-types do worse in all interactions, and hence there are fewer arrangements in which they will be imitated, and more arrangements in which A-types will be imitated. Though it is still the case that preferential detachment segregates the types into separate communities before learning can promote complete A-type domination, an initial population of $N_A = N_B = 100$ is sufficient for an average final value of 99.1% A-type agents in Hawk and Dove (with a minimum run value of 93.5%). Comparing that to 41.1% in Prisoners' Dilemma tells us that when applying preferential detachment to situations in which imitative learning is operating it is crucially important to represent the cardinal payoffs accurately.





When population dynamics are added instead of learning, the population trajectories have slightly greater cohesiveness and quicker convergence in the Hawk and Dove situation. The degree of prosociality is the same since A-types dominate in every run. In Hawk and Dove, due to the lower payoffs received by B-types in $B-B$ interactions, and the greater payoffs received by A-types connected to B-types, the B-types no longer enjoy the population boost in early periods (when mixed arrangements are common). We will see in further down that this "defector boost" occurs in several strategic contexts and can lead to defector domination in some cases. The size and duration of the initial benefit to anti-social agents is sensitive to the parameters in ways that are explored in the conclusions section (3.10).





As was already pointed out, the combination of preferential detachment, imitative learning, and population dynamics is effective and robust in establishing prosocial communities in cooperative games. Hawk and Dove behavior has the same-shaped trajectories as Prisoners' Dilemma behavior, but the increased cohesiveness makes an important difference here. Cooperators need to form a sufficiently large homogeneous group in order to reach high enough payoffs to outperform perimeter defectors. A lucky series of connections by defectors in the Prisoners' Dilemma can bring them to dominance; e.g., by getting just the right connections to cooperators early on. Due to the lower payoffs received by defectors in Hawk and Dove, such an occurrence is nearly impossible.

The major implication of these results is that within a game category some social problems are easier to solve than others. Even though preferential detachment does well in achieving the prosocial outcome in the Prisoners' Dilemma, it does better with the Hawk and Dove payoffs (especially when the type-changing behaviors are involved). These effects can be separated from the effectiveness of preferential detachment itself because the models perform identically in the base model. One point is an obvious one: if one desires numerically accurate results from an application of the theory to a particular problem, then one needs accurate representations of the cardinal interaction payoffs.

A more nuanced research question that arises from these results is a need to establish the relationship between the payoffs and the prosociality of an outcome within a game form. Future work can explore payoff sensitivity by sweeping across values for the individual interaction payoffs within the bounds of the cooperative game form. Recall that in Prisoners' Dilemma a greater maximum degree increased the probability that cooperators would achieve the necessary connections for achieving domination; future work can explore whether and when a similar relationship exists in Hawk and Dove. Another way to spin this question is to uncover what the payoffs need to be to motivate actors to cooperate in games of this form.

3.9.2 Battle of the Sexes and Coordination Game

The next category of games, the coordinative games, includes games in which both of the assortative outcomes are Nash equilibria. Furthermore, both assortative outcomes produce equally total utility, but in the Battle of the Sexes each player prefers a different outcome. This distinction is removed in this fixed-type preferential detachment model because agents receive payoffs as both Player1 and Player2, and the realized payoffs for both types is seven. Using the standard payoff matrix in preferential detachment results in a situation in which both types do equally well when assortatively mixed, but A-types do better in heterogeneous arrangements.

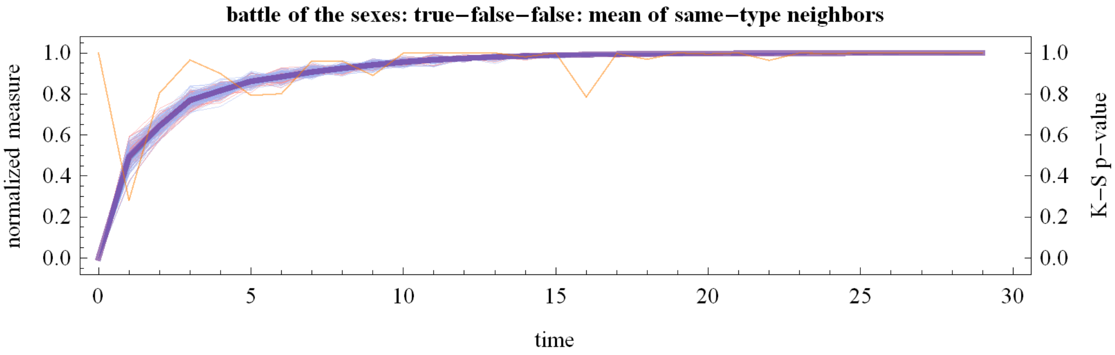
Battle of the Sexes

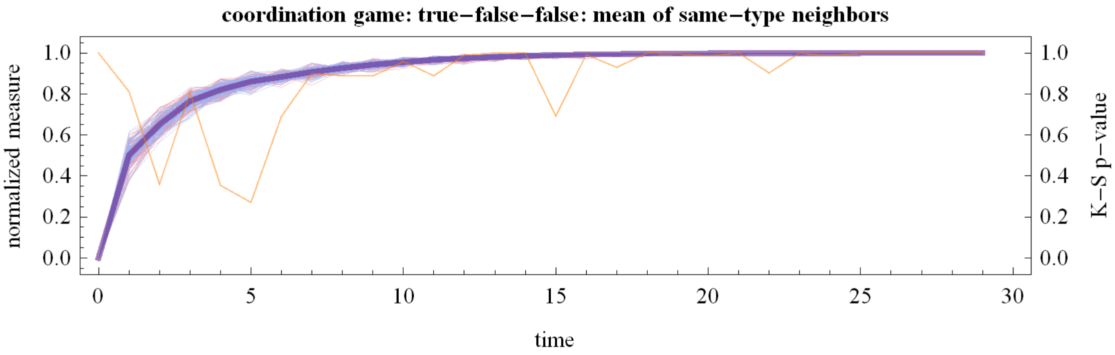
		player2	
		A	B
player1	A	4,3	2,1
	B	1,2	3,4

Coordination Game

		player2	
		A	B
player1	A	4,4	2,2
	B	2,2	4,4

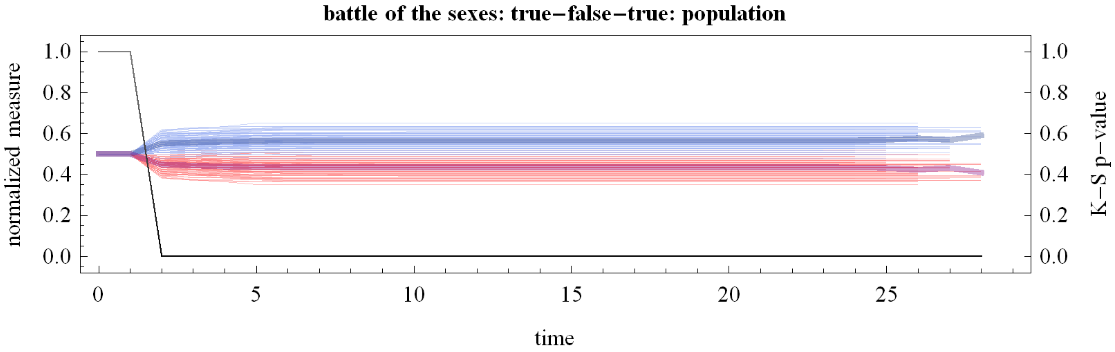
In this model agents are not choosing over types/actions, so the asymmetry in the game's payoff matrix is reflected in preferential detachment theory as a greater realized payoff for A-types than B-types in mixed interactions (4 vs 2) in the Coordination Game.

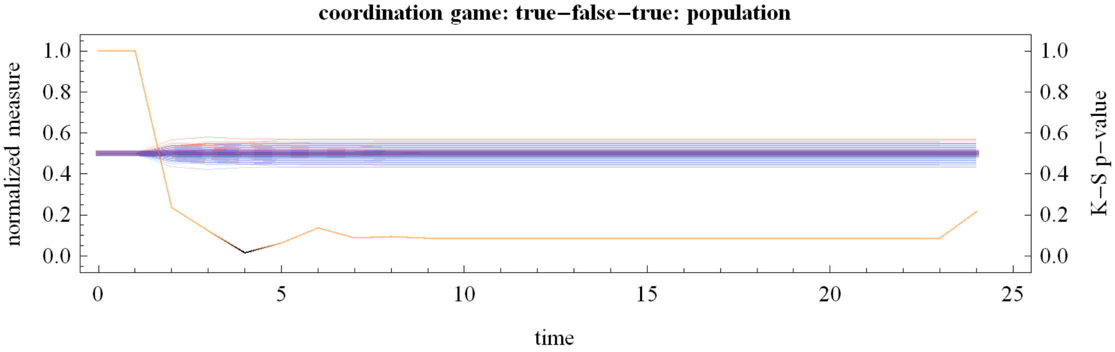

battle of the sexes: true–false–false: mean of same–type neighbors


coordination game: true–false–false: mean of same–type neighbors

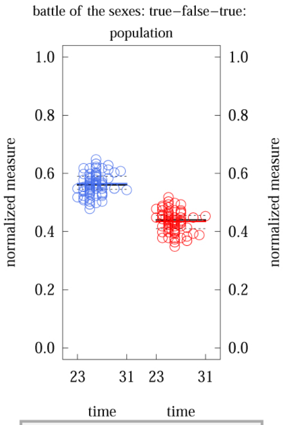
The plots for the base model again demonstrates that within-category model behavior is indistinguishable. Producing this result is a check on the model implementation and verifies that the simulations operate in accordance with the theory they are meant to encode. We can see from the plots that the values for same-type neighbors follow similar trajectories. Though there are three time steps at which the differences are significant at the 95% confidence level, the mean Kolmogorov-Smirnov p-value across the time-steps is 0.409 – the two experiments are *not* significantly different throughout the simulations.

The behavior itself reveals exactly the features we expect from games in this category. Without consideration of the exact payoff values, the A-types and B-types are inverses of each other: they act identically when types are considered as "same-type" and "other-type". We can also see that both types form completely segregated groups, thus perfectly satisfying the prosociality condition.

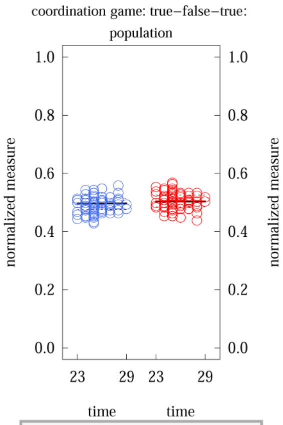




Though the trajectory plots reveal that, on average, the A-types settle into a group with a slightly, but statistically significant, greater number of agents than B-types in the battle of the sexes, and that in the coordination game the figures are insignificantly different, the behavior plots obscure the final outcomes and how they differ. The outcome plots below better reflect the the similarities and differences we care about.

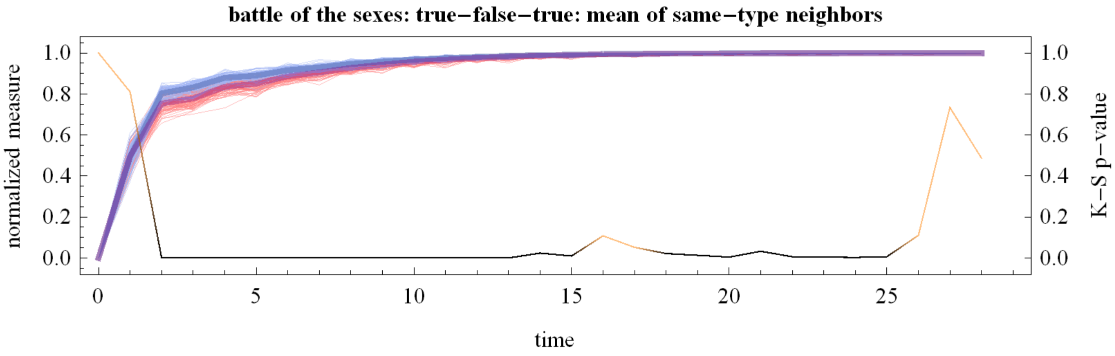

battle of the sexes: true−false−true: population

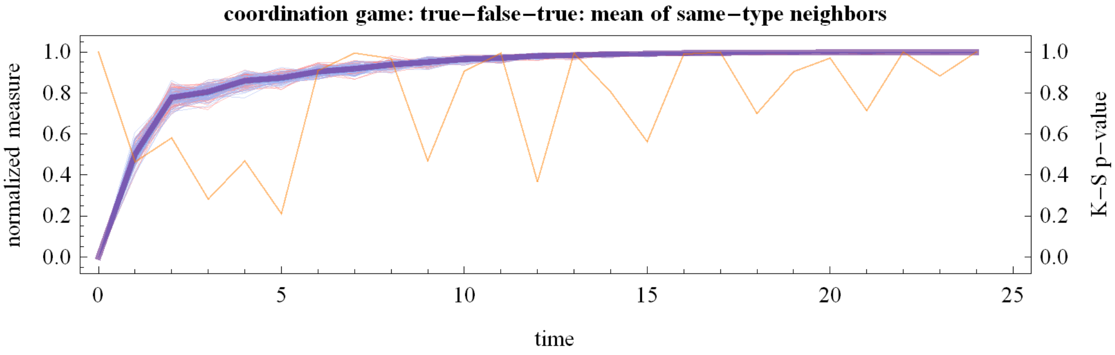
Mean:	0.565	0.435
Maximum:	0.650	0.520
Upper Quartile:	0.590	0.455
Median:	0.560	0.440
Lower Quartile:	0.545	0.410
Minimum:	0.480	0.350
K−S p−value:	0.000	


coordination game: true−false−true: population

Mean:	0.495	0.505
Maximum:	0.560	0.570
Upper Quartile:	0.515	0.525
Median:	0.498	0.502
Lower Quartile:	0.475	0.485
Minimum:	0.430	0.440
K−S p−value:	0.087	

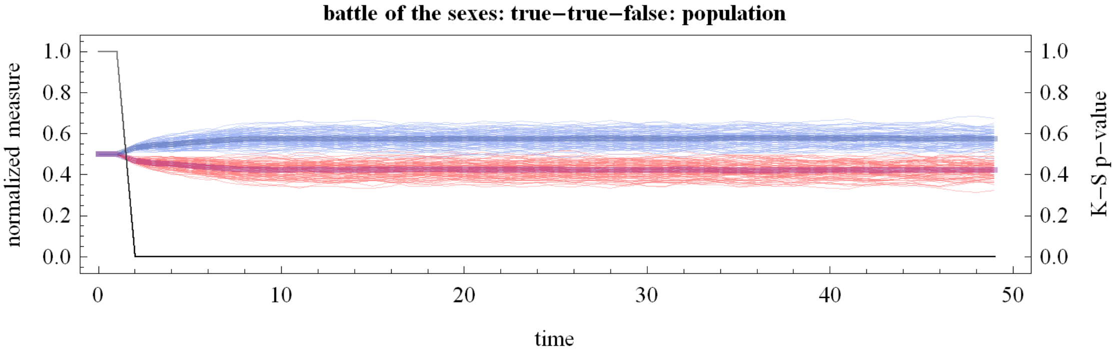
We can safely conclude that the Battle of the Sexes outcomes for *A*-types and *B*-types are produced from distinct processes, and hence that the asymmetry in the game *does* matter for the outcome. The outcomes, however, are not reliably distinct: there is approximately 20% overlap in the outcome populations. The most lopsided outcome is a 65/35 split, but the mean value for the *A*-type population is 56.6%. Even though the difference in statistically significant, the populations are similarly sized. Though these population trajectories and outcomes are interesting, and essential for understanding the model's operation, the prosocial outcome across all the coordinative games is identified as assortativity for both types, regardless of population levels.

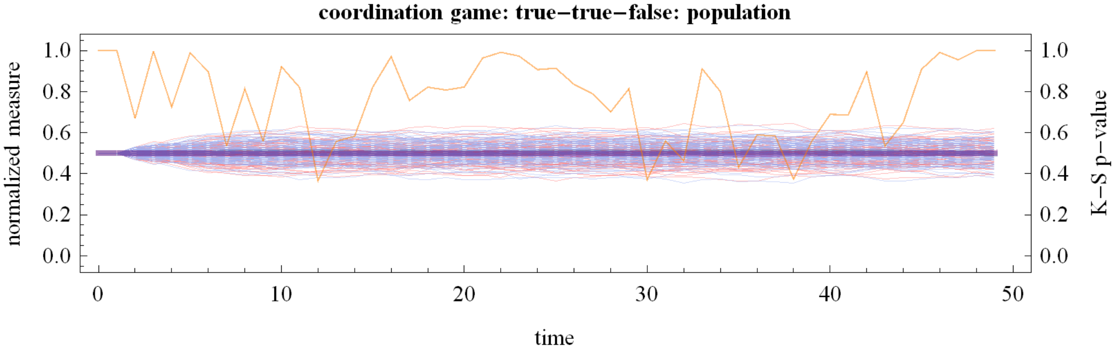

battle of the sexes: true−false−true: mean of same−type neighbors



The values of same-type neighbors presented in the above plots show that both types achieve indistinguishably high levels of assortativity in both strategic contexts. The behavior of Battle of the Sexes shows more variation in the early iterations as the A-types are being imitated by attached B-types; but preferential detachment quickly severs the population, and both segregated groups gain connections until their maximum degree is saturated. The social structure reaches the identified prosocial outcome in every run.

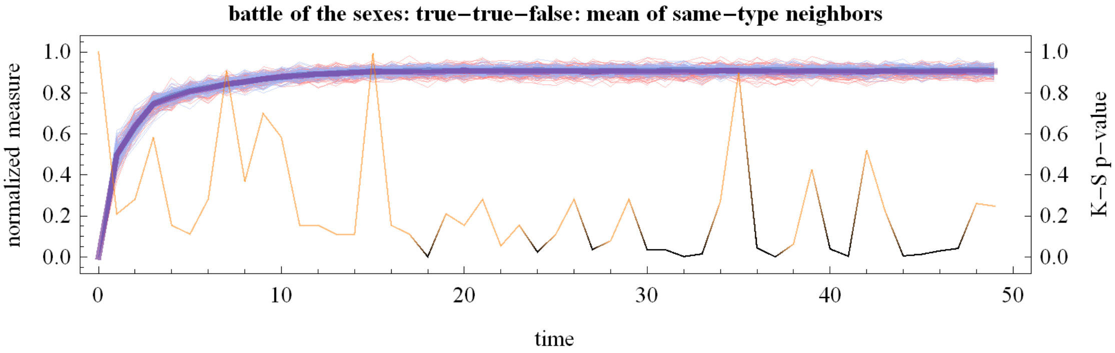
Now we examine the case with population dynamics instead of learning. In the Battle of the Sexes the A-types do better than B-types in mixed arrangements, but they do equally well after the two types have segregated. We can see in the plot below that during the early iterations the A-types increase to reliably higher population levels in every run, and then the populations level off.

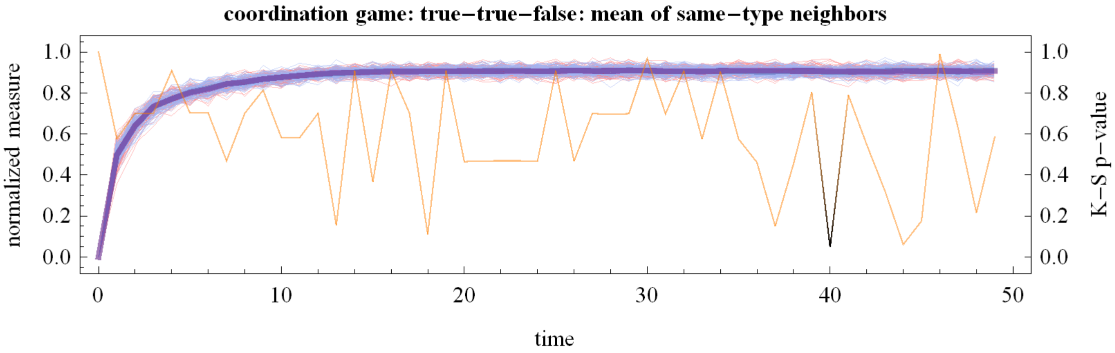




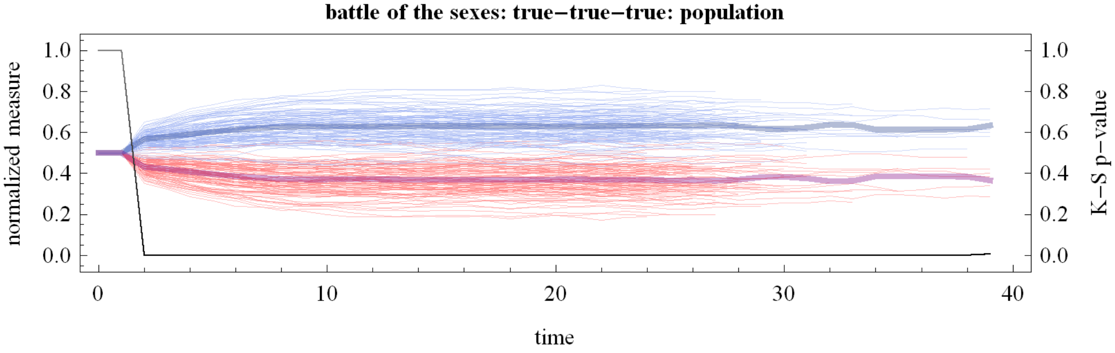
This trajectory is absent in the Coordination Game. Though there is run-by-run variations in the population values, neither type has a systematic benefit. The variation is entirely explained by random events in early iterations producing population differences that are locked into the social arrangement after the types split into separate groups. The fact that population dynamics produce a significant and reliable increase in the A-type population in Battle of the Sexes is an important result for discussions of equality and fairness of social norms and institutions in the final chapter.

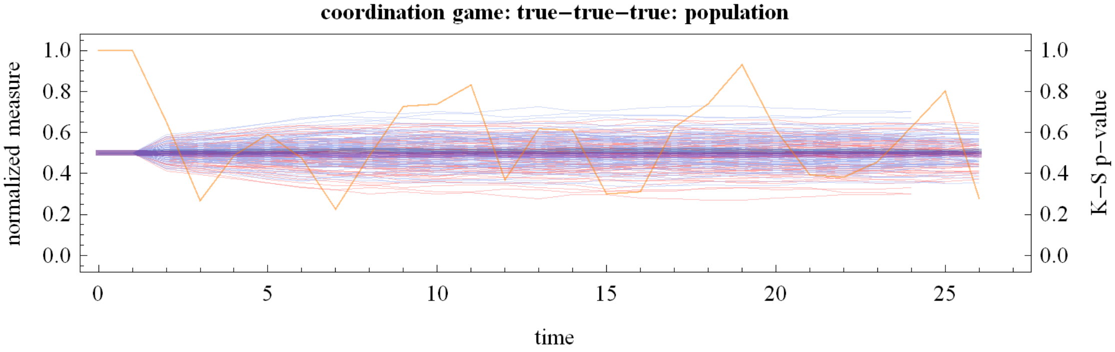
The differences in population levels, which are larger than the differences in the previous two cases, have an effect on the assortativity as well. The plots reveal that the B-types in the Battle of the Sexes have greater volatility in their same-type neighbor measure in comparison to the Coordination Game. This volatility yields differences in the two types which are occasionally significant, but not reliably so.





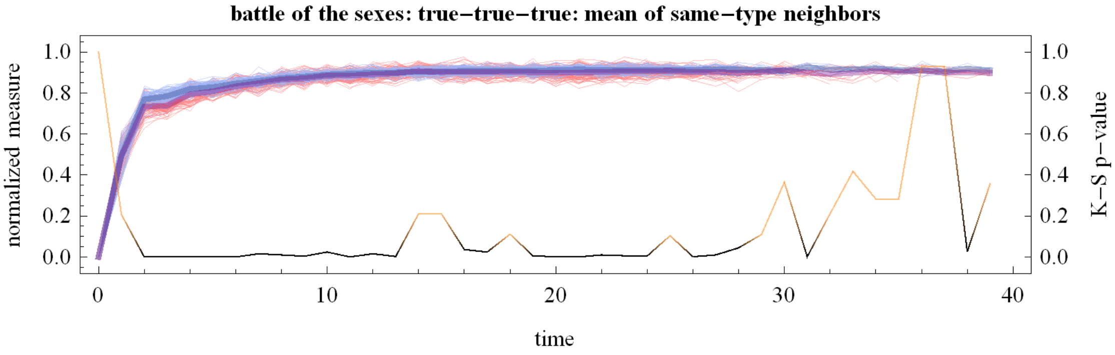
The results of these experiments show that, again, both types of agents achieve high levels of prosociality. The final levels are mitigated by the fact that, since in segregated arrangements both types are performing equally well, both types are replicated in proportion to their presence in the population. New agents of both types are introduced each iteration and are forming connections among themselves. The new agents have fewer connections than the incumbent agents, and hence comprise the lowest five percent utility agents. Population dynamics continuously eliminates and creates nearly the same pool of fringe agents, which have one iteration to form random connections and impose this effect on the measure of prosociality.

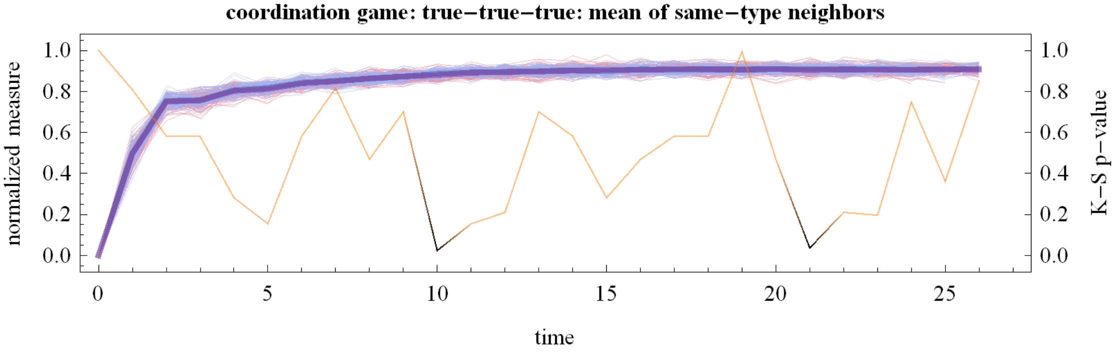




In the combined mechanism case we again expect that the Battle of the Sexes would produce a difference in behavior (favoring the A-types) that would not show up in the Coordination Game, and this is exactly what we observe. The run-by-run variation is greatest with the combined mechanisms in both games, and the difference between the two types in Battle of the Sexes is also greatest here. Learning is a local behavior that stops influencing model behavior once social arrangements have solidified. Population dynamics has a constant effect, but again, once the social arrangements have solidified the same set of agents is being replaced each iterations. By combining these mechanisms we create a situation with a constant new supply of agents with no connections. These new agents are both a vehicle for imitative learning across newly formed connections, and also a source of utility differentiation. Both aspects make the system behavior more volatile, and over time keep outcomes more volatile.

The relevant result in these experiments is the level of assortativity, which is presented in the following plots.



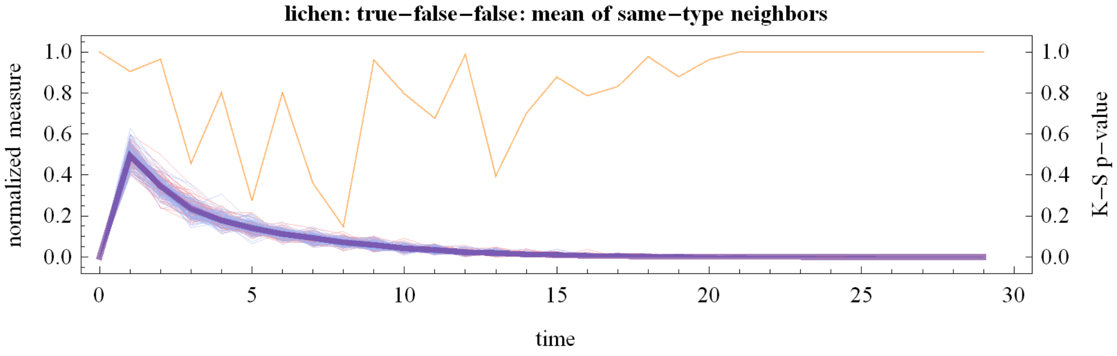


From these we can see that the assortativity results with combined learning and population dynamics are qualitatively similar to the results of all the other cases: a sharp increase followed by increasing, but slowing, percent of same-type neighbors. Just as with population levels, population dynamics prevents agents from ever reaching perfect assortativity because there is a constant churning of agents of both types.

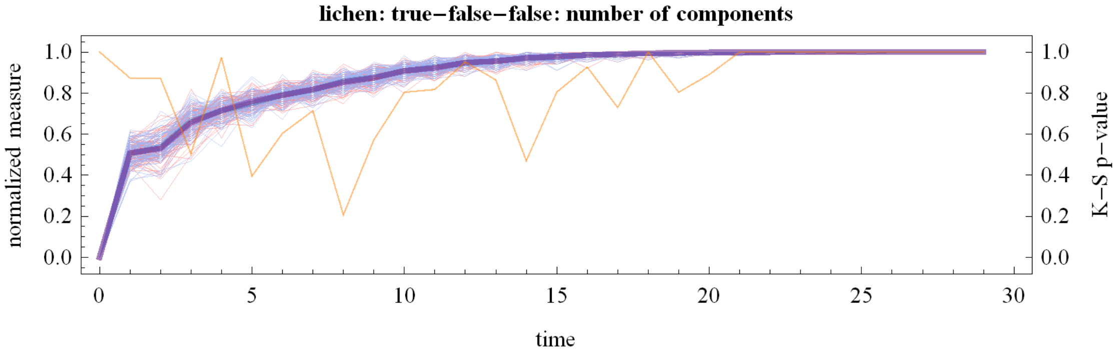
Despite the measured decrease to prosociality, this feature of the model is highly desirable because it is a realistic behavior for natural systems, and because it demonstrates equilibrium of process without stasis. The split of the population, and the behavior of agents, become dynamically stable over time while also exhibiting the properties indicative of the evolution of a social convention of the type often cited in the literature.

3.9.3 Lichen

Our next strategic context models situations in which miscoordination, specialization, codependence, etc., are essential to agents' interests. In all four experiments on this game a social arrangement is identified as prosocial to the degree that *different* types interact. This will be measured with the mean percent similar neighbors, just as assortativity was for the coordinative games above, but here we are looking for minimal values. We will also be looking for other results of this rarely examined situation, such as population stability, connectivity, and utility.



The disassortativity produced by preferential detachment can be seen from the decreasing value of same-type neighbors. In fact, this value reaches zero in every run. The resulting social arrangement exhibits a "checkerboard" pattern in which each agents has $K = 5$ connections to agents of the other type (a degree-5 regular bipartite graph). This pattern is also revealed in the number of same-type components produced; shown in the following plot.

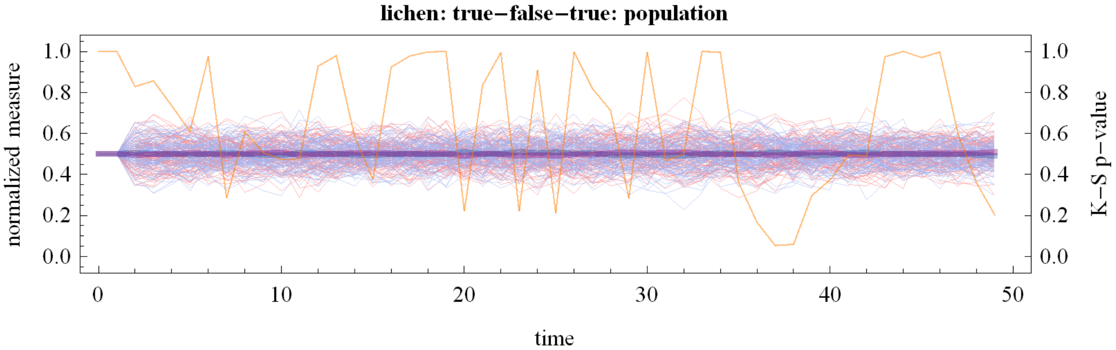


The plot of number of components over time shows the trajectory of the plot toward the limiting level of 100% (of the largest observed number), but what it doesn't show is that the largest observed number of components is 100. What this means is that at the end of almost every run the number of components of each type equals the number of agents of each type; i.e., no agent is connected to one of the same type. It is also the case that the two types' behaviors are indistinguishable with respect to assortativity and the number of components – there is symmetric behavior between types. The results demonstrate that the base model produces the ideal social arrangement for the Lichen strategic context.

The addition of imitative learning complicates the results of the Lichen game. In the ideal checkerboard arrangement every agent is receiving the same payoff, so learning initialized at that stage would not disrupt anything. But because connections are made randomly from an initially unconnected set of agents, and because the payoffs are different when the number of connections to each type of neighbor differ, learning produces erratic changes in the agents' types. Agent types determine detachment behavior and also the payoffs received by both a focal agent *and* all its neighbors. So a change in one agent's type sparks a cascade of type changing behavior across the network.

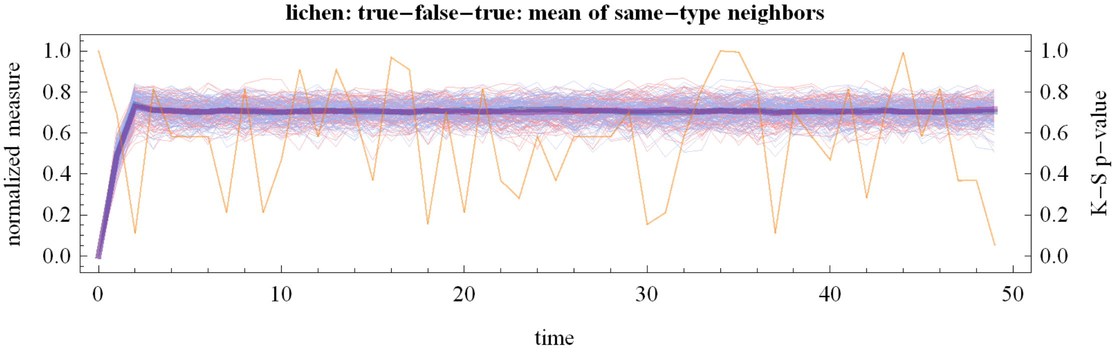
An example may help elucidate how this progresses. Imagine two connected focal agents – one of each type. The A-type agent has four other B-type neighbors, while the B-type focal agent has only three other A-type neighbors. In this arrangement the focal A-type is earning a higher payoff than the focal B-type because of the extra connection, so the B-type agent will imitate its neighbor and become an A-type. But now that B-turned-A-type agent has four A-type neighbors and a very low utility. Furthermore, the original A-type agent has one fewer B-type neighbor, thus reducing its payoff. If any of *its* other B-type neighbors have four other A-type neighbors, then the focal A-type will imitate one of them and become a B-type. This series of events switches, but stabilizes, the relationship between the focal agents, but it has pushed the same situation to all of those agents' neighbors. The example demonstrates how things would progress if only one pair were unbalanced, but naturally the situation being faced by our focal agents in this example is widespread among the agents in the formative iterations of each simulation.

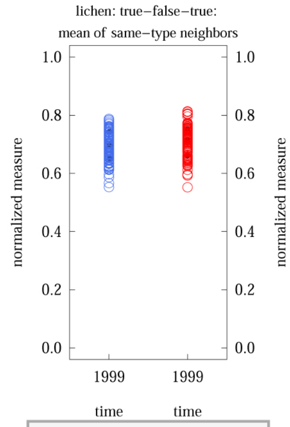
Considering the success of preferential detachment in the base model, one might reasonably expect that some arrangement could be reached that would stabilize the network structure and types. However, we will see from the following plots that the system behavior exhibits a constrained randomness that is not conducive to prosociality.



The population of each type of agent fluctuates throughout each run, though clearly there is a damping effect because the success of each type of agent depends on the presence of other-type agents. The trajectories of the number of each type of agent are statistically insignificant. Though the plots are truncated to fifty periods, each run actually continues until the "time-out" halting condition at 2000 iterations. Recalling the set of halting conditions from 3.4.7, timing out implies that the system never reaches stasis, never enters a edge formation/elimination cycle, and never converges to a fixed-size set of mixed agents. A constant, near random churning of agent types precludes any of the established halting conditions from obtaining. However this process does not by itself eliminate the possibility that the levels of disassortativity (and hence prosociality) would be high. For example, it could be the case that the cascade only affects agents at a moving edge of a network community, the inside of which is well organized. It could also be the case that this process eventually finds a stable arrangement given enough time. We may be especially hopeful since the population levels do, in fact, stay close to the 50/50 split that prosociality requires.

For our measure of prosociality we examine the percent of same-type neighbors across time and at the final iteration of each run. We can see that after a very short formative period, the system behavior develops into a pattern with moderately high, but constantly changing neighbor similarity. The stability the average of this measure over time, despite the fluctuations in step-by-step and run-by-run values, indicates that a systematic and reliable feature of the model creates the observed levels. What this means is that the result is not produced via a locked-in feature from early stages or pure stochasticity. The result is also not through stable equal populations and network connections, but rather through dramatically fluctuating populations and network structures that, because of feedback within the model, sustain a coherent range of properties. As interesting as this result is, the coherent structure that develops is not conducive to prosociality.

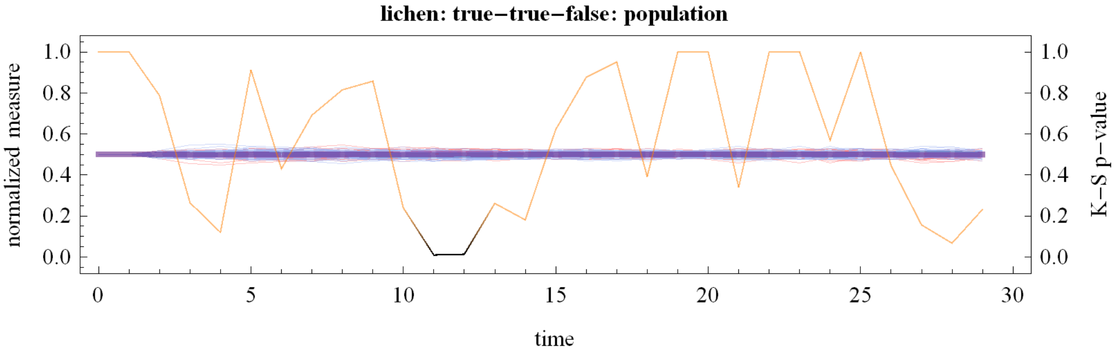


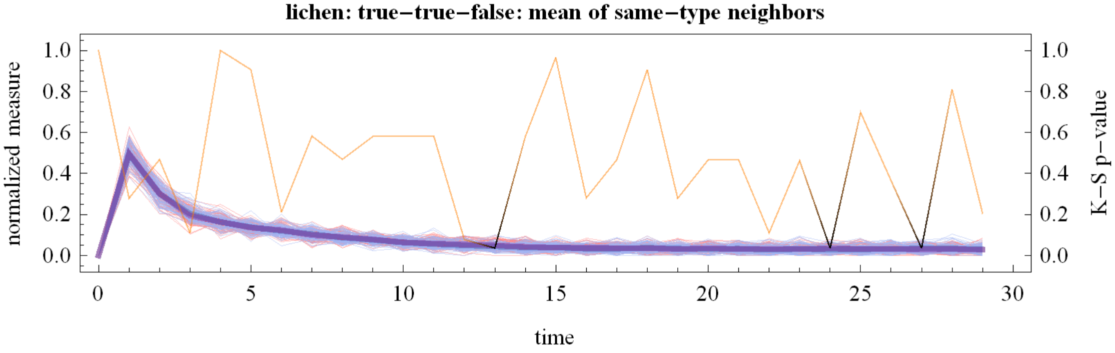


Mean:	0.694	0.709
Maximum:	0.789	0.815
Upper Quartile:	0.742	0.751
Median:	0.695	0.720
Lower Quartile:	0.655	0.661
Minimum:	0.555	0.554
K−S p−value:	0.154	

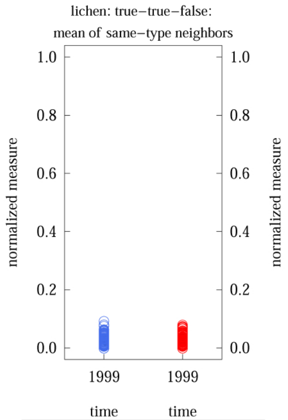
Though we have seen that preferential detachment itself can facilitate perfectly assortative or perfectly disassortative arrangements depending on the payoff structure, learning can only promote greater assortativity because imitation decreases local variation in types. When preferential detachment itself moves the arrangement toward greater assortativity, learning will magnify the effects of preferential detachment, but here we see a conflict of mechanism. Agents with dissimilar neighbors perform better and soon find themselves with self-similar neighbors. Because imitative learning counters the separating effects of preferential detachment in this strategic context, the combination is only 30% successful on average; i.e., it can be considered *unsuccessful* in achieving prosociality in this situation.

Unlike learning, population dynamics does not directly impede prosociality in the Lichen game, though it is unclear that population dynamics plays any role to help either. The base model was successful because in an even split of the agents the two types could remove same-type connections and find stable partners of the other type. Even in the final ideal checkerboard state, population dynamics would still remove five percent of the population and force them to endure the random attachment phase again. But it would do so equally for both types; and this symmetry balances the disruptive effect between both types. This symmetry ensures that each type continues to have a roughly equal share of the total population. And roughly equal shares of the population implies roughly equal opportunities for dissimilar agents to connect.





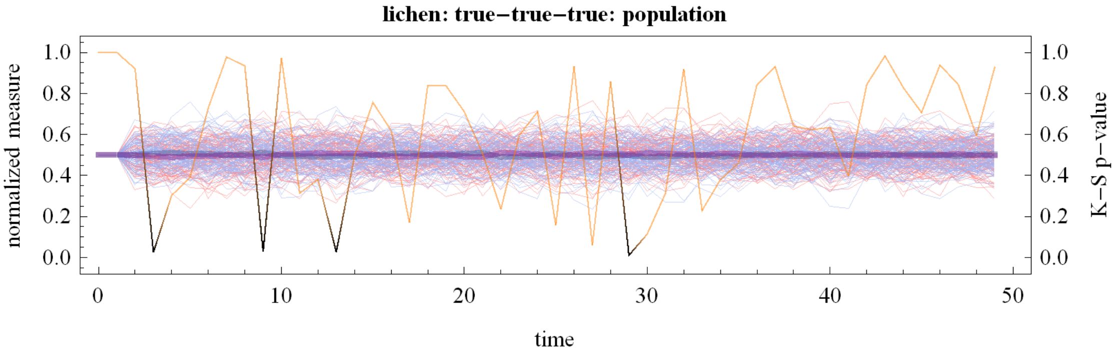
Once the simulation matures, the lowest five percent utility agents are the agents created in the previous iteration. The bulk of the population remains stable, fully connected, and disassortatively mixed à la the base model. Agents that formed recently have fewer connections, hence lower utility, and are therefore removed. The replacement agents are copies of the top five percent utility agents. Since all fully other-connected agents, regardless of type, receive the same payoff, this top five percent is chosen uniformly at random from this set of agents. Hence the distribution of agents replaced approximately equals the distribution of new agents and the population profile remains largely stable. This stability is enhanced by feedback via the feature that agents of each type perform better under the presence of more other-type agents.

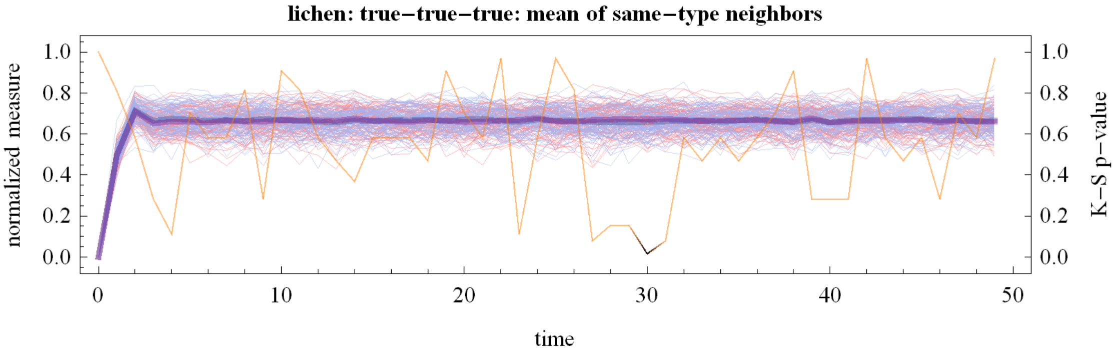


Mean:	0.028	0.031
Maximum:	0.094	0.081
Upper Quartile:	0.040	0.046
Median:	0.026	0.027
Lower Quartile:	0.014	0.015
Minimum:	0.000	0.000
K−S p−value:	0.452	

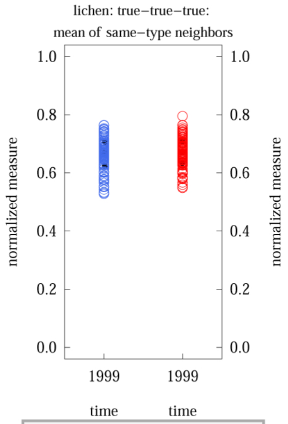
The end result is that, though the population is constantly churning, it is doing so in a systematic way that leaves the bulk of the population alone to maintain a near-ideal regular bipartite graph. This arrangement constitutes a prosocial outcome for situations in which mixed-type interactions are beneficial. Every run reproduces this high level of disassortativity, reduced only slightly by the random connections of the constant supply of nascent agents and short-term path sensitivities in the number of agents left unsaturated.

Finally, we reach the analysis of the combined learning and population dynamics scenario. As you can see from the plots below, the behavior of the system is dominated by the effects of learning. The volatility and variation are both enhanced by population dynamics, so the results are less reliable even though average assortativity has declined slightly.





As before, the erratic nature of the model output effects both types equally across the set of runs, so the two sets of run data are not significantly different. The feedback between the types keeps the agent population and assortativity within regular bounds, and produces a dynamically stable distribution of behaviors across time. This sort of system behavior is paradigmatic of complex systems, and the Lichen game's behavior under a broader set of circumstances deserves much further study.

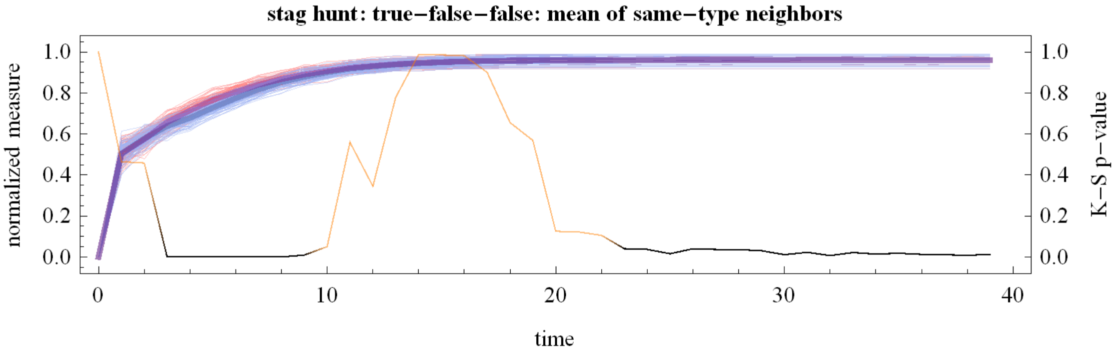


Mean:	0.663	0.670
Maximum:	0.766	0.797
Upper Quartile:	0.704	0.704
Median:	0.668	0.667
Lower Quartile:	0.626	0.637
Minimum:	0.530	0.550
K−S p−value:	0.701	

For our purposes here it suffices to acknowledge that imitative learning has a disastrous effect on prosociality in this strategic context, and this despite preferential detachment (with or without population dynamics) being extremely successful in reaching the desired outcome. Also of consequence is that despite the fact that the Lichen game requires a drastically different outcome for procosiality than is seen in commonly analyzed games (disassortativity rather than assortativity), the contingent rule of preferential detachment succeeds in both contexts. We will discuss the implications of the success of preferential detachment across a breadth of situations in the conclusions section (3.10).

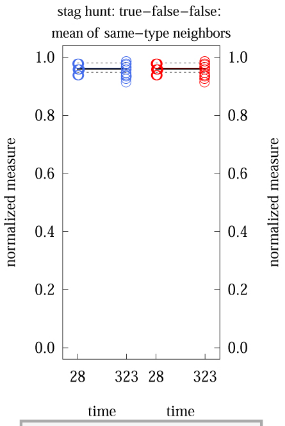
3.9.4 Stag Hunt

The Stag Hunt game represents a situation in which the A-types benefit from interacting with other A-types, but the B-types are indifferent to their neighbors' type. Because B-types have no preferences regarding their partner types (though they do prefer to have as many neighbors as possible), the prosocial outcome of the base model is defined as assortativity; specifically that A-types are able to find and maintain interactions with other A-types.



The trajectories here match closely the system behavior in the Battle of Sexes and Coordination Game, though there is more between-type variation here. The reason for this is because, just like in those games, the A-types are detaching any B-type agents they encounter unless they have only encountered B-types. Because the probability of randomly attaching to and/or receiving attachments from five consecutive agents of the same type in an evenly distributed population is quite small, having just one type initiating detachment suffices to produce high levels of segregation between the types.

We can see from the output plot below that the final levels of between-type connectivity for both types is very low. Hence the agents are highly assortatively mixed and the outcome achieves a high degree of prosociality.



stag hunt: true−false−false:
mean of same−type neighbors

Mean:	0.961	0.961
Maximum:	0.988	0.988
Upper Quartile:	0.980	0.980
Median:	0.960	0.960
Lower Quartile:	0.948	0.948
Minimum:	0.916	0.916
K−S p−value:	0.006	

The imperfect achievement of segregated populations are the result of stable heterogeneous groups, but ones of a different structure than those we saw in the Prisoners' Dilemma. In the cooperative games the B-types also preferred A-types, and so when they maintained an A-type neighbor, they eschewed other B-type neighbors. This left them with few connections because the pool of A-type agents that can be stably exploited is small. In the Stag Hunt game the B-types have no preferences over type, so even when an A-type neighbor is present they continue to acquire additional A- and B-type neighbors. It is for this reason that the B-types have the same value of same-type neighbor as the A-types in this game, rather than the much lower value seen in the Prisoners' Dilemma.

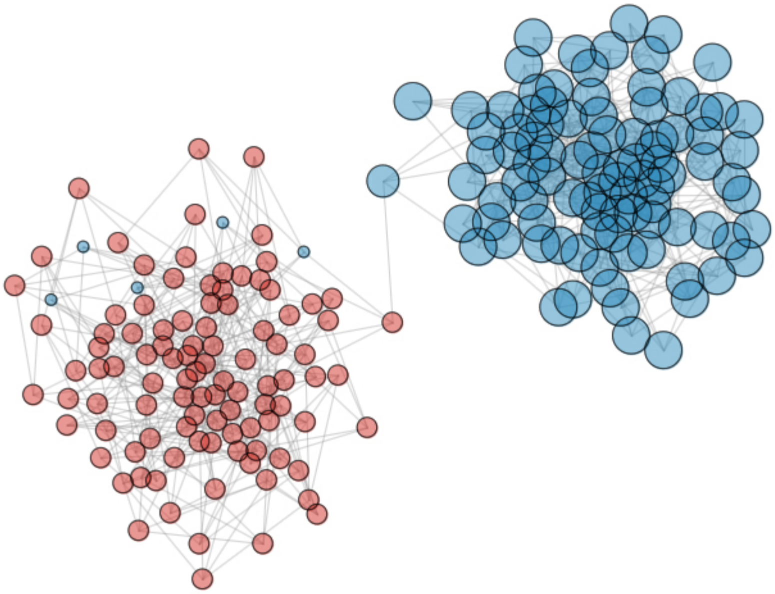
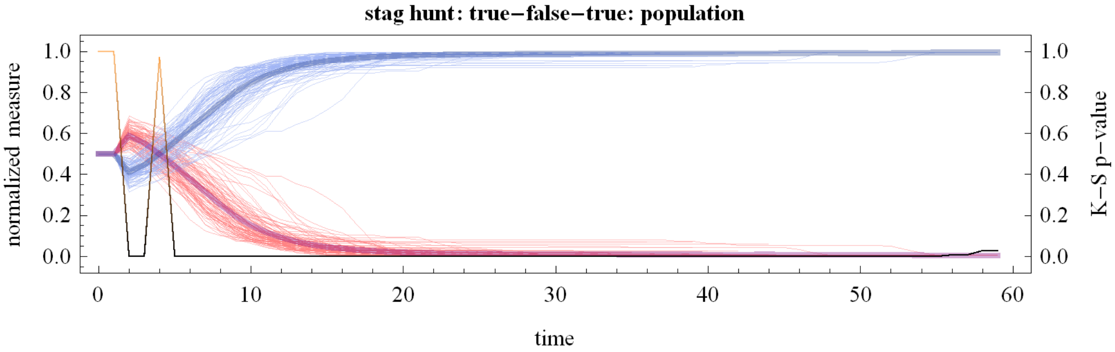
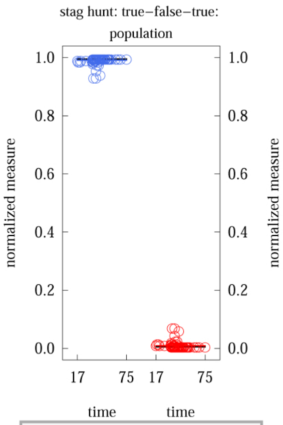


Figure 3.5: Stable Heterogeneity in Stag Hunt

When learning is incorporated into the agent behavior we need to consider the payoff benefit that A-types receive when interacting with other A-types vs the penalty for interacting with B-types. A-types only receive one util from a connection with a B-type, but four utils from A-type interaction. B-types receive two utils from any connection.



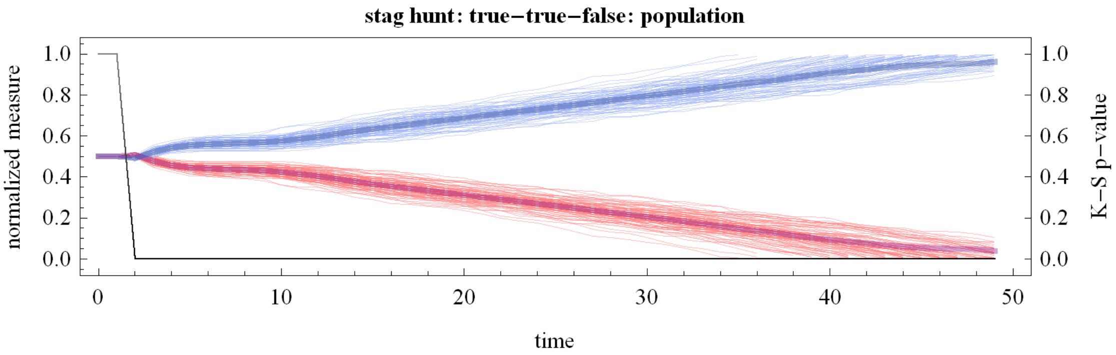
In the early stages of the simulation, when mixing is widespread, it is typical for B-types to perform better and be imitated. However, once preferential detachment has produced a moderate amount of separation, the A-types quickly gain the upper hand. Because B-types in mixed groups are *not* disconnecting B-types in this situation (as they are in the Prisoners' Dilemma), the network stays more connected. As a result, when A-type imitation becomes the utility-enhancing behavior, it spreads very quickly through the population. However, note that because detachment can still occur before imitating A-types pays better, some groups of stable, isolated B-types can form.



Mean:	0.991	0.009
Maximum:	0.995	0.070
Upper Quartile:	0.995	0.005
Median:	0.995	0.005
Lower Quartile:	0.995	0.005
Minimum:	0.930	0.005
K–S p–value:	0.000	

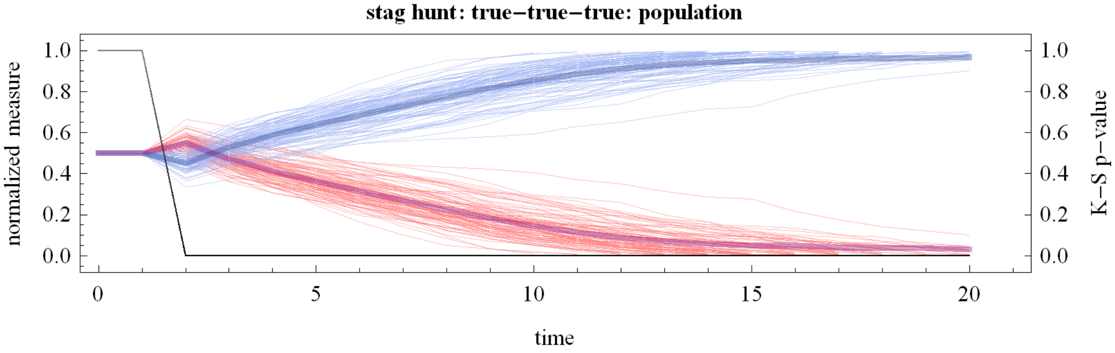
The outcome plot shows that, indeed, some level of B-type isolation and perpetuation is the norm. But they constitute only a small part of the population. The prosocial outcome of A-type dominance is achieved to a high degree across all runs.

We already know that, because population dynamics is a global operation, if the intra-type payoffs of one type are higher than another, then assortativity ensures that the higher-payoff type eventually dominates (as long as the the temporary enhancement yielded from inter-type connections doesn't produce premature convergence).



Unlike learning, population dynamics does not interfere much with the force of preferential detachment. In early mixed stages the B-types are doing better on average, and some get reproduced, but through random connections $A-A$ connections are expected to comprise one-quarter of the connections made, and so many of these will make it into the top five percent. After a few periods of sorting into separate groups (see the bend near $t = 10$) the A-types will grow in number by 5% every turn until they dominate.

Just as we saw in the Prisoners' Dilemma and Hawk and Dove games, the combined mechanisms increase the run-by-run variation, but also compliment each other to ensure the prosocial outcome (segregation then dominance).



Local heterogeneity is not sustainable because learning ensures that A-types in an otherwise B-dominated group will become Bs, and vice versa. And population dynamics ensures that any isolated groups of B-types is removed and replaced by A-types. This happens very quickly and the prosocial outcome is perfectly attained in every run. The similarity of the trajectory paths here and with Prisoners' Dilemma hints that if the maximum degree were increased enough, then the variation of connections may produce some runs that converge to B-type domination, but the value of K would have to be greater than was used for Prisoners' Dilemma.

3.9.5 Commensal

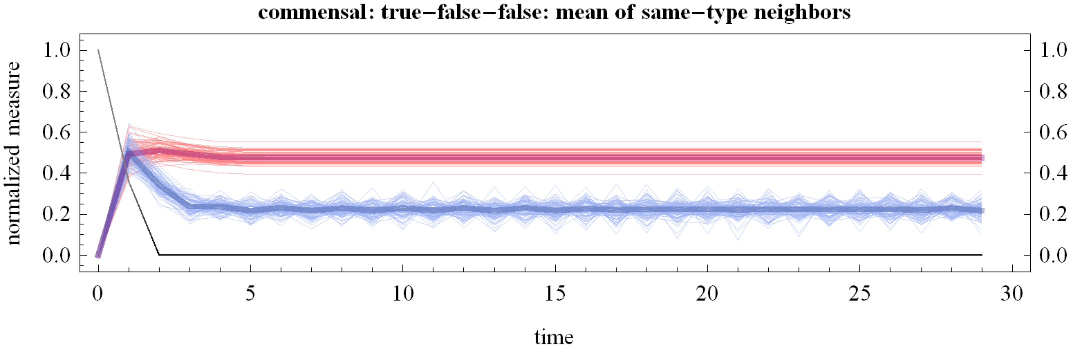
The Commensal game models a situation that flips the role of A-types in the Stag Hunt. Instead of doing better when associating with other A-type agents, they now prefer interactions with B-types. The B-type agents are still indifferent between neighbor types.

Commensal

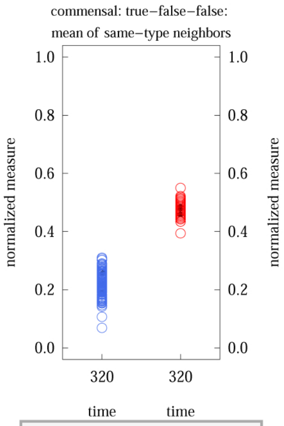
player2

		A	B
player1	A	1,1	4,2
	B	2,4	2,2

This situation poses a novel difficulty for prosociality: disassortativity is desired, but only by A-types. Since A-types connected to B-types receive the greatest utility, the greatest social utility comes from a balance of agents such that each A-type has a full compliment of B-types and every B-type is saturated with A-types. This is again the regular bipartite graph seen in the Lichen game. Though for different reasons than those present in the Lichen game, perfect disassortativity is the prosocial outcome in each of the four cases presented below.



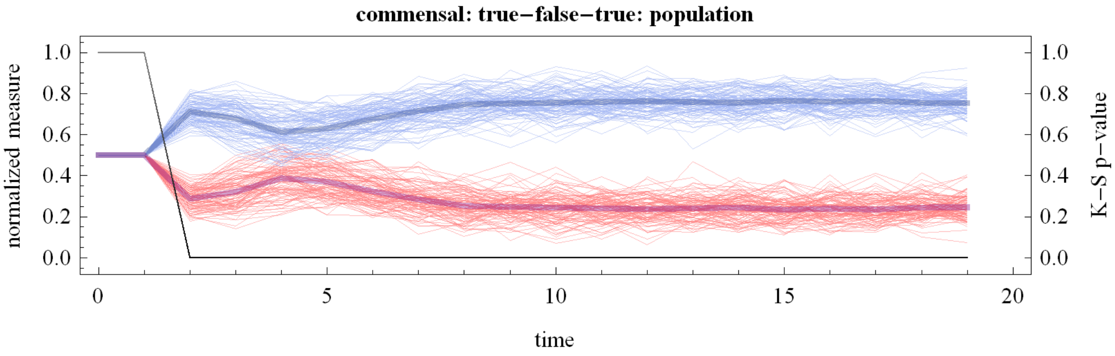
In the Lichen game we saw perfect disassortativity in the base model, but here the agents are much less successful in achieving the prosocial result. The reason is that in this situation the B-types simply collect neighbors until they fill their maximum degree, detaching nobody. Thus we see that the values for B-types level off at $t = 5$ – at which time roughly fifty percent of the connections to B-types have A-types at the other end. A-types that manage to gain at least one connection to a B-type will detach any A-type neighbors they receive. Though A-types can form stable isolated groups, few A-type agents will have zero connections to B-types (the same number as those stuck in stable heterogeneous groups in the Prisoners' Dilemma). The rest of the A-types' available connections will endlessly attach and detach to A-types because there are no more B-type connections to be had. Thus the lower level of assortativity comes from having fewer connections.



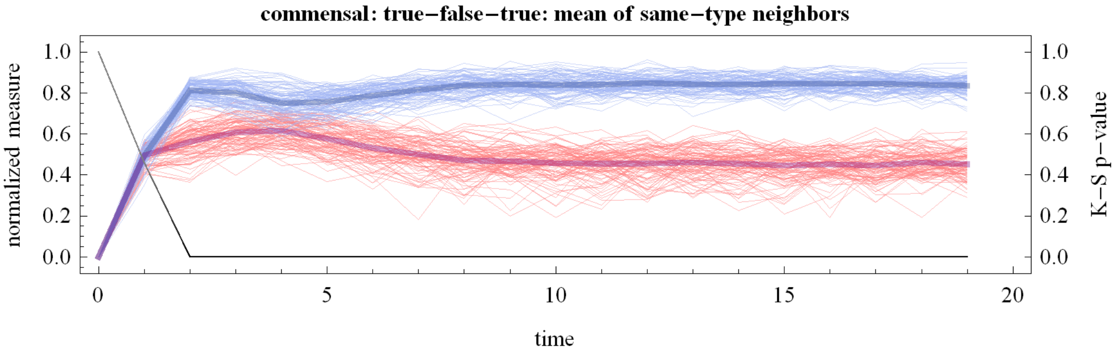
Mean:	0.220	0.475
Maximum:	0.312	0.552
Upper Quartile:	0.254	0.488
Median:	0.216	0.476
Lower Quartile:	0.194	0.456
Minimum:	0.070	0.396
K−S p−value:	0.000	

We can see this cycle from the output plot in the fact that *every* run ends at $t = 320$: this means that every run stops due to the Mixed Agent Convergence halting condition. Considering this analysis, the results are exactly as expected, but they are not prosocial to a high degree. The indifference of B-types in this context greatly reduces the success of agents to form efficient specialized arrangements.

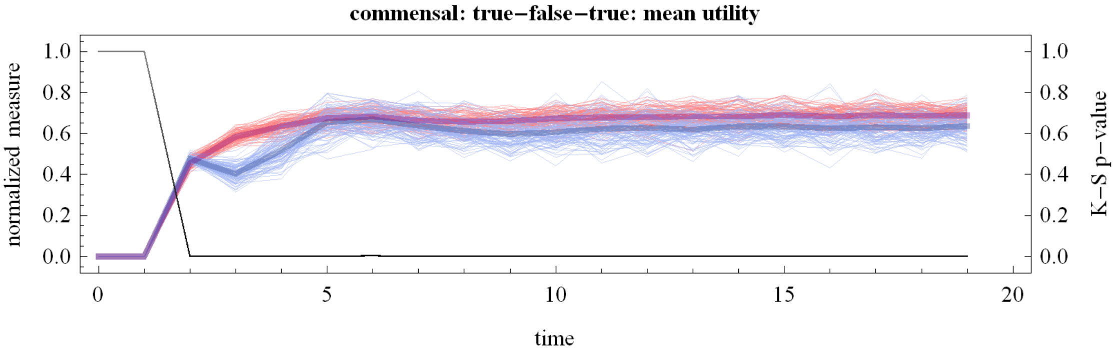
The system behavior with learning produces a distinctive behavior due to the negative feedback between A-types doing better when attached to B-types, and B-types imitating A-types because they are doing better. Unlike the Lichen game, if the agents in this situation were in the ideal checkerboard configuration it would not persist. B-types always earn a payoff of two, and A-types attached to them earn four. So imitation produces a trend towards a greater number of A-types, but only up to a point, because the higher payoff received by the A-types depends on the presence of B-types. We can see the resulting population levels in the plot below.



Because A-types are imitated, they end up with more A-type neighbors than B-types...even though the B-types are indifferent to neighbors and A-types prefer B-types. This dynamic, therefore, frustrates prosociality in the sense that A-types are not connecting to as many B-types as would benefit the members of society the most.

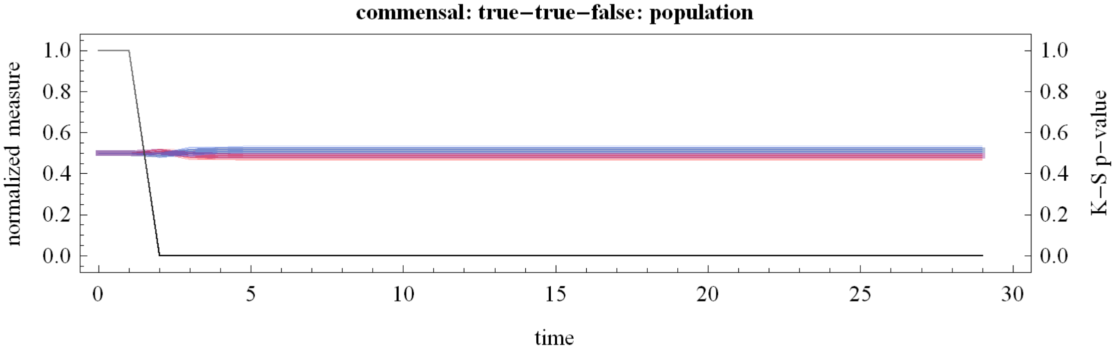


The levels of population and assortativity are constantly fluctuating, but within limited ranges that are consistent from run to run. After an early transition period, the feedback between the types maintains a social structure with constantly churning agent types and connections. The driver of this consistency is the payoff ratio of the interactions. The number of A-types and B-types, as well as the number of connections among them, are such that each type maintains a similar utility level – as seen from the following plot.

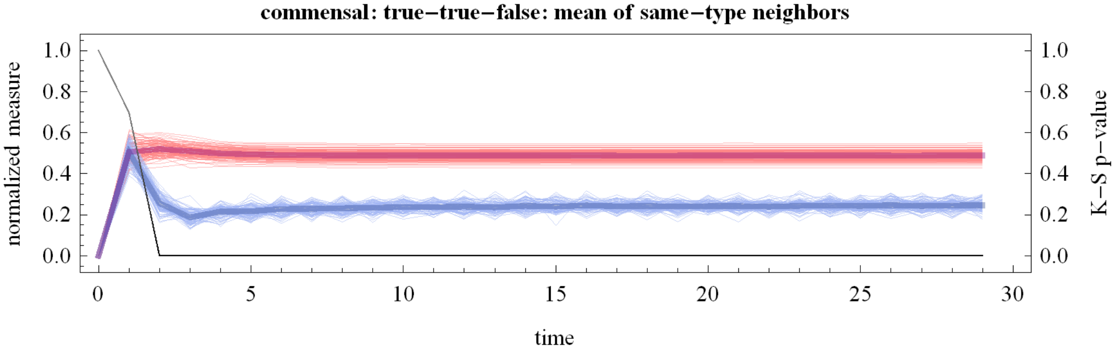


The indifferent B-types do better on average, because the A-types quickly find themselves attached to many other A-types if they are doing better. This can happen because *all* the neighboring B-types can convert in one turn, but the rate of connection/detachment is only one per iteration. So the A-types suffer a short-term payoff penalty greater than the utility benefit. The numbers of each type, and the numbers of connections, reach a dynamically stable set of values that maintains a social arrangement that doesn't benefit the commensals at all. The prosocial outcome is not achieved; and worse, the payoffs are lower than a population filled with indifferent agents. The results from learning and dissortativity are clear (an intuitive): the imitation mechanism is incompatible with specialization-requiring situations, even in combination with preferential detachment.

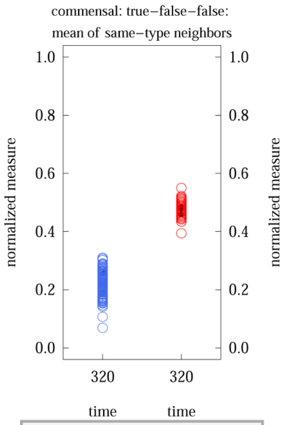
We now turn to analyzing the scenarios with population dynamics. In the outcome arrangements of the base model, the A-types are earning a higher payoff than the B-types, but that does not affect the behavior in the base model. Despite a lower connectivity, the cardinal value assignment chosen here still provides them with a small fitness advantage.



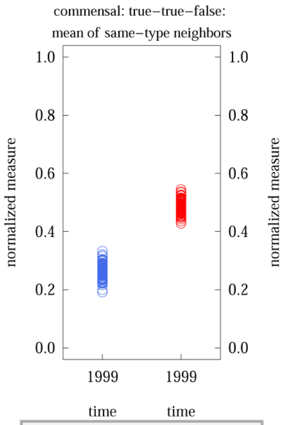
When we add population dynamics, that fitness advantage translates into greater share of the population. But because the fitness advantage was small, and because the fitness of A-types depends on connecting to B-types, the increase in A-type population is also small. The population quickly converges to a stable near-equal level in which all incumbent agents are earning an equal payoff. For this to be true, the A-types must be underutilizing their maximum degree, which is the case. In fact the system behavior is qualitatively identical to what we saw in the base model.



Furthermore, the outcome values of percent similar neighbors with population dynamics (below right) are close to the base model values (below left).



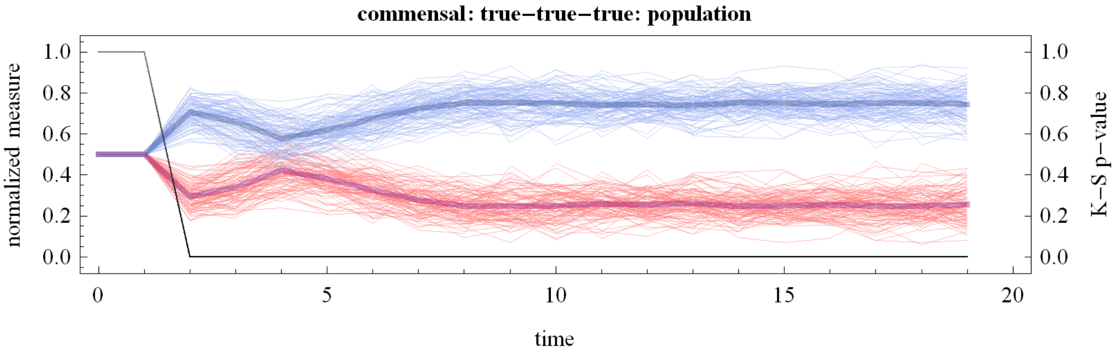
Mean:	0.220	0.475
Maximum:	0.312	0.552
Upper Quartile:	0.254	0.488
Median:	0.216	0.476
Lower Quartile:	0.194	0.456
Minimum:	0.070	0.396
K−S p−value:	0.000	



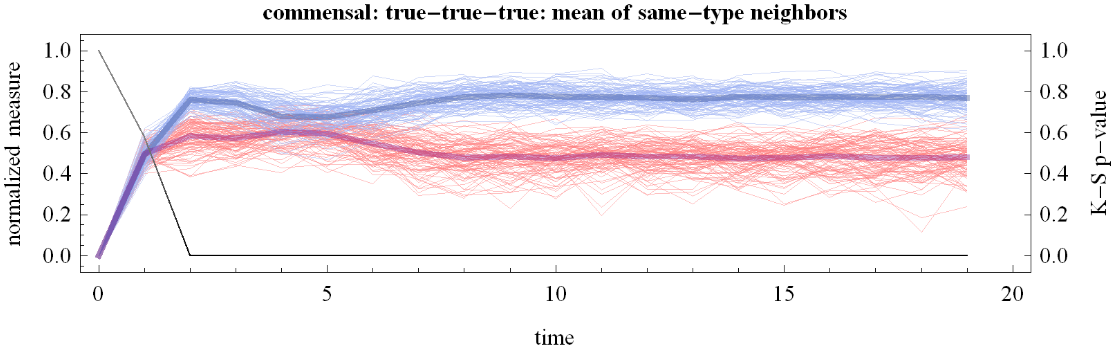
Mean:	0.267	0.488
Maximum:	0.335	0.547
Upper Quartile:	0.283	0.504
Median:	0.266	0.488
Lower Quartile:	0.253	0.473
Minimum:	0.194	0.430
K−S p−value:	0.000	

Though achieving a balanced, unstable, population of the two types based on payoff and connection feedback is interesting and deserves greater study, that is not the purpose of the current study. For our purposes we point out that the level of prosociality is low. A very small amount better than the base model, and much better than the learning case; but still very far from the ideal condition of zero assortativity for both types. Although, because of the indifference of B-types, a fully disassortative outcome was never expected, or even thought possible. In situations like Commensal, in which only one agent type benefits from the prosocial outcome, it is reasonable to find that attaining the prosocial outcome is more difficult.

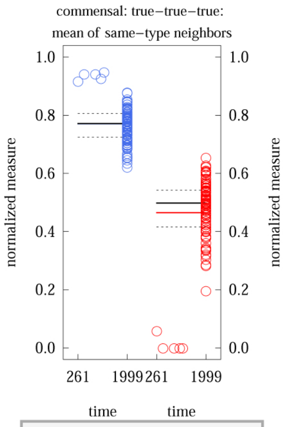
We finally turn to the case with learning and population dynamics. Because learning is involved, lessons learned from just above, and from the Lichen game, tell us to expect erratic behavior within bounds that is also not conducive to prosociality. The plots below confirm this expectation.



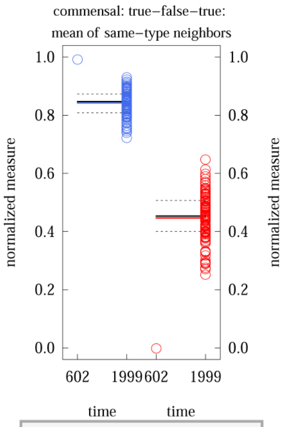
The population levels are very close to the trajectories seen in the learning experiment, with variation in the early stages intensified by population dynamics. Once the populations and connectivity reach levels at which both types are earning near-equal utility, the population dynamics mechanisms just roughly sustains those population levels. The population levels are highly volatile from period to period, nearly twenty percent difference between the minimal and maximal values with in a run. But this variation acts as noise against a steady expected level for all parameters when taken as the mean value across the one hundred runs. We can see this phenomenon again in the measure of assortativity below.



The trajectories of both agent populations and same-type neighbors indicate that learning is the dominant force in determining the outcome values of these variables. We already saw that learning creates populations such that each type earns a similar utility level. And the numbers of each settle on values such that the feedback maintains those levels. We can see from the outcome plots below how the results here (left) differ from those of the learning experiment (right). The final results are numerically distinct as we expect from the effects of population dynamics, but qualitatively similar enough to identify learning as playing a larger role than population dynamics in determining the outcomes.



Mean:	0.771	0.465
Maximum:	0.949	0.655
Upper Quartile:	0.806	0.542
Median:	0.771	0.498
Lower Quartile:	0.724	0.416
Minimum:	0.622	0.000
K−S p−value:	0.000	



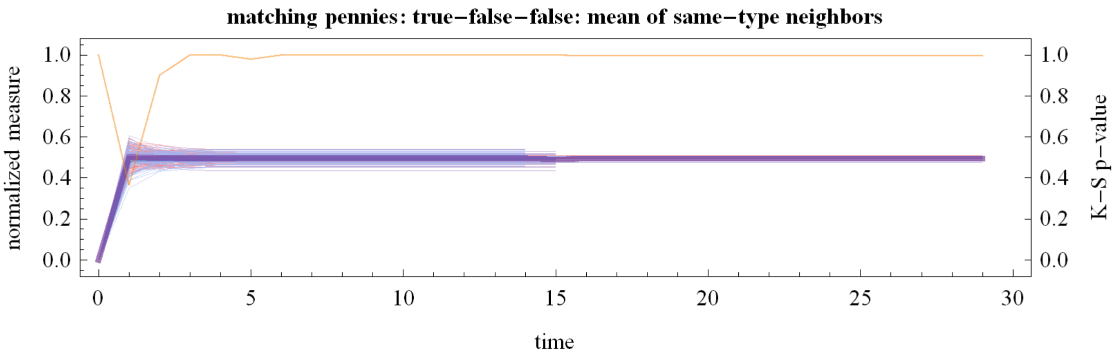
Mean:	0.842	0.446
Maximum:	0.994	0.650
Upper Quartile:	0.873	0.507
Median:	0.847	0.453
Lower Quartile:	0.808	0.400
Minimum:	0.724	0.000
K−S p−value:	0.000	

I have already noted that miscoordination games are absent from the evolution of prosociality literature. For this reason it is difficult to make any comparison of the success of preferential detachment to other mechanisms...other mechanisms haven't been tried. Though the series of micro-events that produces these results is well understood, it is a complicated mix of agent detachment, imitation, and replication. Preferential detachment, and the additional mechanisms, create fascinating dynamically stable patterns of agent behavior in these contexts, but they do not do well in achieving the desired social outcome. We will revisit this result in the conclusions section (3.10) to connect to the issue of social problem difficulty and the higher bar necessary to solve these problems.

3.9.6 Matching Pennies

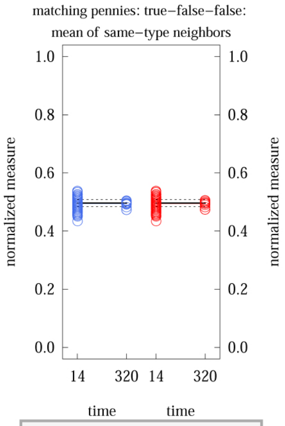
This strategic context is not strategic at all; all agents are indifferent to all connection types. The game is still important to analyze for points made about the general applicability of preferential detachment to "solve" social problems of all kinds. This includes not producing anomalous and undesired results in neutral environments.

For the base model of Matching Pennies I report the mean percent similar neighbors. In fact, it doesn't matter which variable's trajectory I present because they are all statistically indistinguishable.



Agents start unconnected and simply randomly acquire interaction partners of any type until the maximum degree is reached. No agents detach any agents. Because the population levels are equal, there is a roughly equal probability of attaching to each type (only roughly due to the no-self-edges effect that $\frac{99}{199}$ differs from $\frac{100}{199}$ – see 2.3.1).

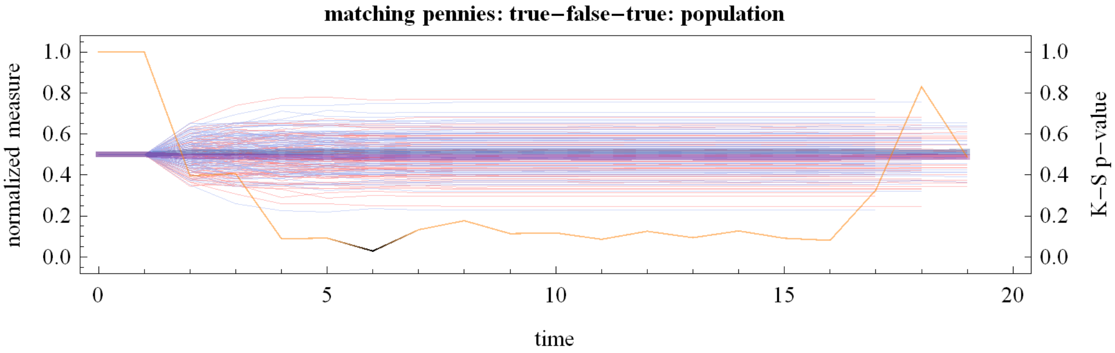
The spread of outcome values seen in the output plot reflects natural stochastic variation in the connection process. The summary values captured over the one hundred runs are identical.



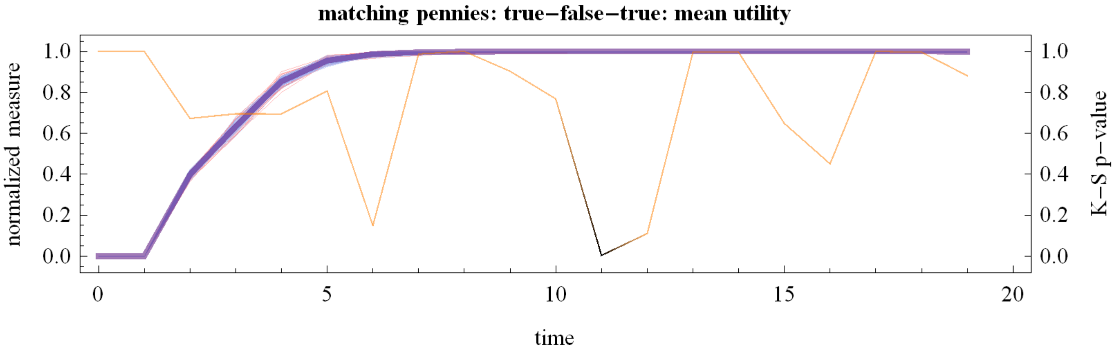
Mean:	0.496	0.496
Maximum:	0.540	0.540
Upper Quartile:	0.508	0.508
Median:	0.496	0.496
Lower Quartile:	0.484	0.484
Minimum:	0.436	0.436
K−S p−value:	1.000	

Also of methodological interest is the two outcome time spans. Most runs end at fourteen or fifteen iterations from the stasis condition; however, in a few runs there is an agent left with open connectability and nobody else open to interact with. These runs halt from mixed agent convergence. Thus even in an undifferentiated environment it is possibly to have an “odd man out” receiving lower utility than the norm from purely historical accident in how the arrangements formed over time.

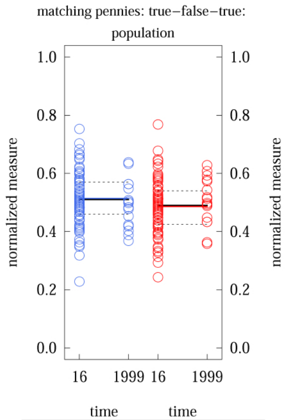
The addition of the imitative learning rule creates greater variation in the outcomes only through the impact of utility variation in early iterations. During the random connection period some agents will, by chance, collect neighbors more quickly than others. These agents will therefore have a greater utility than some of their neighbors, and therefore be imitated. Agents of each type have an equal probability of having a greater actual degree, so averaged over 100 runs the effects of imitation are similarly distributed. The result is high run-by-run variation, but stochastically indistinguishable mean population values for the experiment.



Because all agents are indifferent to the configurations, all agents receive equal utility once they all reach $k = K$ as revealed by the following plot.



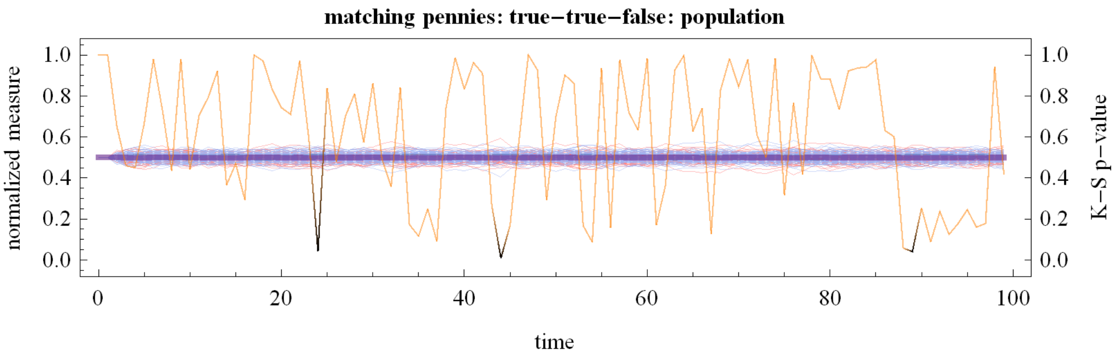
With equal utility for all agents, the variation in the observed population differences and early utility differences is not indicative of a systemic bias in the types. The variation in outcome populations and completion time can be clearly seen in the outcome plot below.



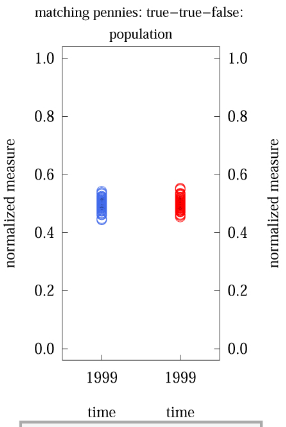
Mean:	0.514	0.486
Maximum:	0.755	0.770
Upper Quartile:	0.570	0.540
Median:	0.510	0.490
Lower Quartile:	0.460	0.425
Minimum:	0.230	0.245
K–S p–value:	0.127	

The result that as much as 75% of the population can become one type or the other in a particular run despite there being no distinguishing characteristic between the types has important implications for the morality and institutional applications. Recall that Matching Pennies is a purely competitive environment by nature. It is only the mechanics of preferential detachment in which all interactions are considered bi-directional that translates the payoffs into uniform indifference. An uninformed observer of such a situation – one who knew the game and the result, but not the agent heuristics – may conclude that one agent type had an unfair advantage because of a resulting population difference. Yet we have just seen that the population differences are the result of locked-in random events, and ones that furthermore have no utility effects for any of the agents.

Population dynamics brings us back to a more balanced population and less variable trajectories. The reason is that unlike imitative learning, the population increase resulting from a few extra random connections only has a small effect on the population change. No cascades can happen in earlier periods and variations cannot lock in because after all agents are saturated (at around $t = 15$) and earning equal payoffs, the ones chosen for replication are chosen randomly. So the population levels in every run are essentially a random walk around the initially even split.



The variations among runs are not consistent, biased, or maintained. Each run ends via the time-out condition, hence the distribution of outcome population values seen below is just one sample – an essentially randomly chosen slice of the population trajectories. The mean value is exactly 50% of each type, and the difference between the distributions is not significant.

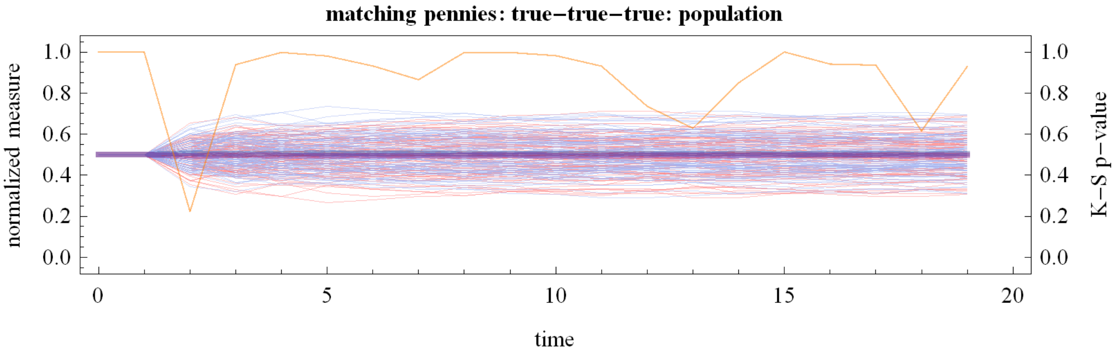


Mean:	0.500	0.500
Maximum:	0.545	0.555
Upper Quartile:	0.513	0.510
Median:	0.501	0.499
Lower Quartile:	0.490	0.485
Minimum:	0.445	0.455
K–S p–value:	0.895	

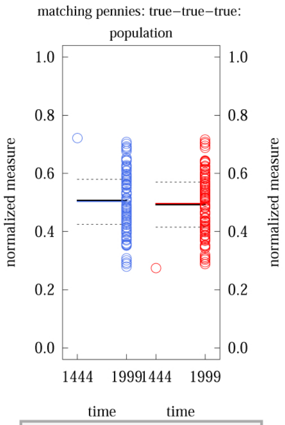
The result is that, unlike learning, population dynamics produces results that look like a fair outcome in the Matching Pennies situation. There is minor variation within a run, but it is randomly perturbed over time such that early increases do not translate to sustained higher populations. There is no reliable difference between the types across the runs.

The combined mechanisms of learning and population dynamics produces a straightforward combination of the results. Like the learning experiment, early variations through random connections crystallize in the population values so that the population values take a broad range (between roughly a quarter and three quarters). And like the population dynamics experiment the values fluctuate in a random walk fashion after all the agents have saturated their connections. The difference is that the random walk is now anchored at the population value produced by learning's early-period effect.

Additionally, nodes recently created through the population dynamics process have comparatively fewer connections, and hence are likely to copy each other and existing agents. But whatever the overall distribution of types in the population, all the incumbent (i.e., persisting) agents receive the same utility and are therefore copied with probability equal to their percentage of the total population. The lowest utility agents are the new ones (due to having fewer edges), so these are the ones selected for elimination and these too have a distribution mirroring the total population's distribution.



The final outcomes have a distribution similar to the learning experiment. This is not surprising because it is again an essentially randomly chosen time slice taken at the time-out halting condition, and population dynamics is just adding noise to the learning outcomes. One run ends earlier than all the others (at $t = 1444$), and this means that some other halting condition was satisfied – most likely the learning cycle 3.4.7.



Mean:	0.504	0.496
Maximum:	0.724	0.719
Upper Quartile:	0.580	0.570
Median:	0.508	0.492
Lower Quartile:	0.425	0.415
Minimum:	0.281	0.276
K–S p–value:	0.999	

The Matching Pennies strategic context, despite describing a situation in which all agents are indifferent to their neighbors, produces system behavior that is ex ante unexpected and important. Because each set of rules produces different dynamics and outcomes, and because these sometimes belie the indifference of the agents, identifying the structure of the game and the preferences of the agents from the observed phenomena is complicated. Below we will see a comparison of the results here to other games, and highlight the interest of including this situation in the library of strategic contexts.

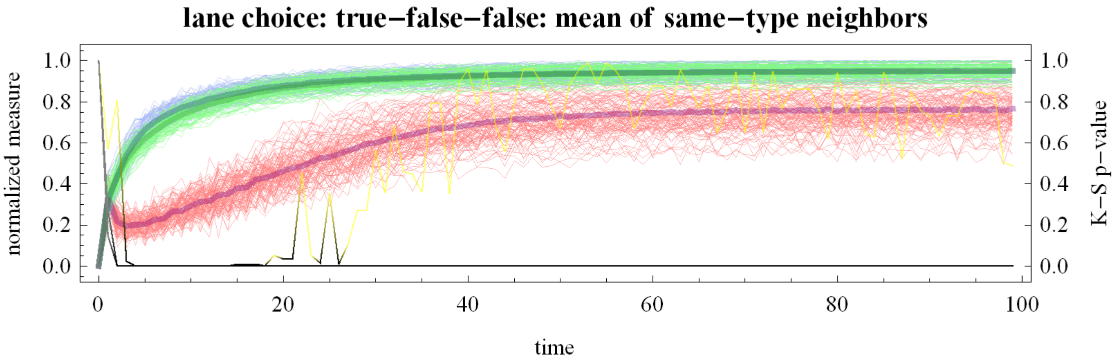
3.9.7 Lane Choice

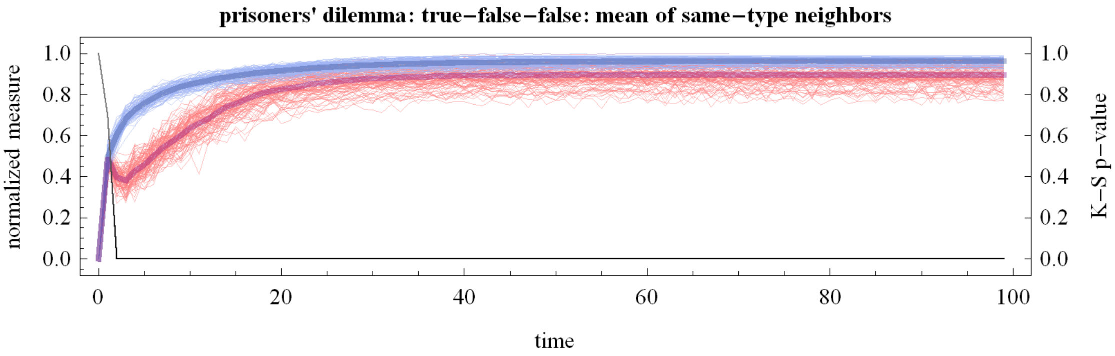
Up to this point we have seen the behavior of preferential detachment in every possible strategic context involving two types of agents. Before moving on to the summary and comparison of these results I analyze the two 3 × 3 games introduced in this project. Recall from 1.8.9 that the Lane Choice situation is a combination of the Coordination Game and Prisoners' Dilemma. For the payoffs chosen in this implementation each type of agent only recognize neighbors as "same" or "other" insofar as the payoffs are the same for both "other" types. In this respect the agents do not have information revealing the full details of the situation. The A- and C-types both seek similarly typed agents, but the B-types seek A- and C-typed (i.e., other-typed) agents.

Given this description, we may think it possible to extrapolate system behavior in

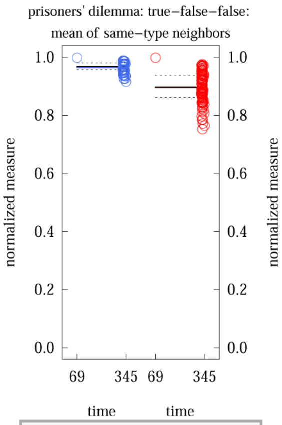
Lane Choice from the component games. To do so we have to adapt expectations to the new proportions of agents in the following way: in the base model the population is fixed at sixty-seven agents of each type, and hence there are two "cooperators" for every "defector" in the Lane Choice game compared to the one-to-one proportion in Prisoners' Dilemma. As a result of this difference we expect to see more B-types in connection with A- and C-types in both transitory cycles and stable heterogeneous communities – in fact twice as many. It is even possible for A- and C-types to remain connected to each other (in addition to B-types) as long as they never connect to a same-type agent, and this heterogeneous outcome likelihood is *increased* because same-types agents make up a smaller percentage of the total population here.

The plots of mean same-type neighbors for the Lane Choice and Prisoners' Dilemma dynamics foster the comparison between the features just discussed. The green and blue lines follow indistinguishable trajectories that also match the shape of the cooperators in the Prisoners' Dilemma. Defectors in both cases experience an initial drop and then a steady climb as the cooperators become saturated with same-type neighbors, but the the steady state level in the Lane Choice game is much lower (0.763 compared to 0.896) and more variable across the runs. This is partly explained by the fact that it is an average over 67 agents rather than 100 agents, but it also reveals the difference that there are more opportunities for defectors when the proportion of cooperators is double their own.

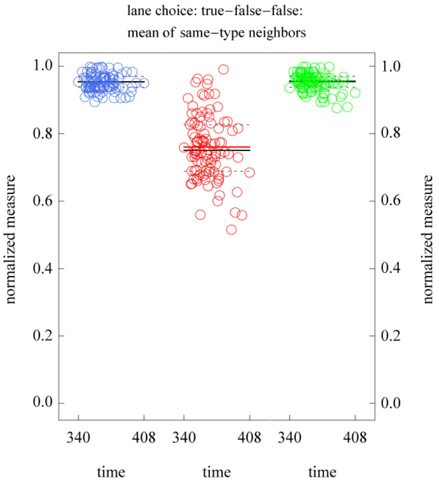




Further along these lines of analysis we can see from the outcome plots that the Lane Choice game produces greater variance in the time to completion as well. Also that the assortativity of each type of cooperator is nearly as successful in establishing a community of same-type neighbors. Therefore the increased success of defectors in exploiting cooperators is best explained by (and is proportional to) the the increase in the proportion of cooperator-type agents.



Mean:	0.965	0.896
Maximum:	1.000	1.000
Upper Quartile:	0.980	0.938
Median:	0.968	0.896
Lower Quartile:	0.958	0.862
Minimum:	0.918	0.754
K–S p–value:	0.000	

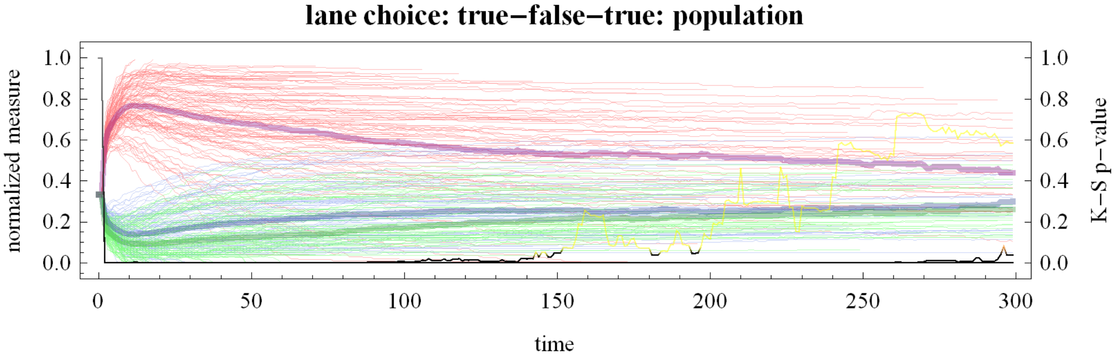


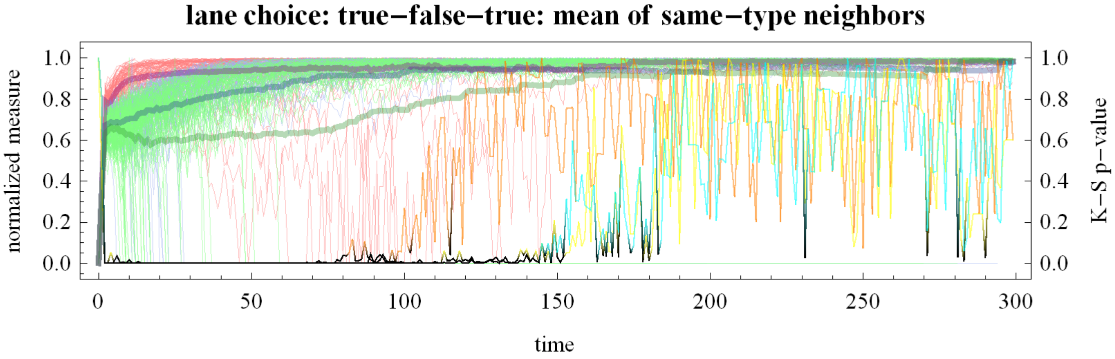
Mean:	0.954	0.763	0.954
Maximum:	1.000	0.993	1.000
Upper Quartile:	0.970	0.828	0.970
Median:	0.954	0.753	0.955
Lower Quartile:	0.937	0.692	0.937
Minimum:	0.896	0.522	0.878
K–S p–value:	0.000	0.000	0.525

The result is that while defectors are indeed more successful, it is only in a predictable linear increase given the changes in the number of agents of each type (since the defectors cannot distinguish between the two types of cooperators). That Prisoners' Dilemma-related result is only half of the Lane Choice's parallel; the Lane Choice also includes Coordination Game structure. With respect to the assortativity of the two types of cooperators, Lane Choice is entirely successful. As was already identified, the trajectories of the A- and C-types are statistically indistinguishable – as are the outcomes levels of assortativity. The payoff table for Lane Choice was designed to reflect both the cooperative and coordinative structure of some social problems, and the result is that using the mechanism of preferential detachment the prosocial outcome of cooperative agents reaching coordinated configurations is achieved at the 95.4% success level for both types of cooperators.

We now turn to the analysis of Lane Choice dynamics when the agents engage in success-based imitation. Looking back at the population trajectories for the Prisoners'

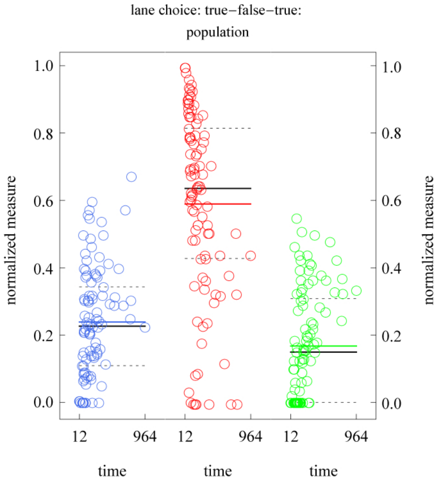
Dilemma we can recall that the defectors experienced an initial boost in numbers that then ratcheted down until the preferential detachment behavior separated the communities (so no more learning could happen). Even though we'd like to see a result of cooperator domination, the actual outcome level is typically less than 50%. In Lane Choice there are twice as many cooperators as defectors, and so (as shown above) defectors have more opportunity to have a cooperative neighbor. When learning is part of the rules this translates into greater defector success.



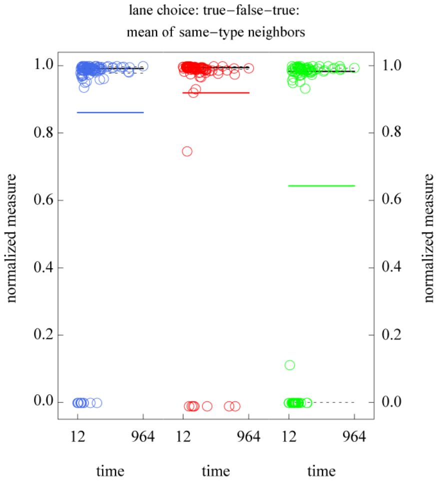


The population plot demonstrates that 1) in some runs the defectors dominate the whole population early on; 2) the population levels for A- and C-types are indistinguishable; and 3) the overall shape is close to the Prisoners' Dilemma results adjusting for the greater success of the defectors due to the greater proportion of exploitable cooperators here. As has been covered already, heterogeneous groups are not stable when learning is operating, so whatever the final population distribution is, it must be the case that each community is composed of agents of the same type. This is revealed in the outcome plot of same-type neighbors which shows each pairwise comparison of same-type neighbor figures is statistically indistinguishable and nearly 1.0.[3]

[3] The delay in the ascent of C-types (green) is best explained by the agent-ID bias. The model is implemented such that when an agent has multiple neighbors providing minimal utility it disconnects



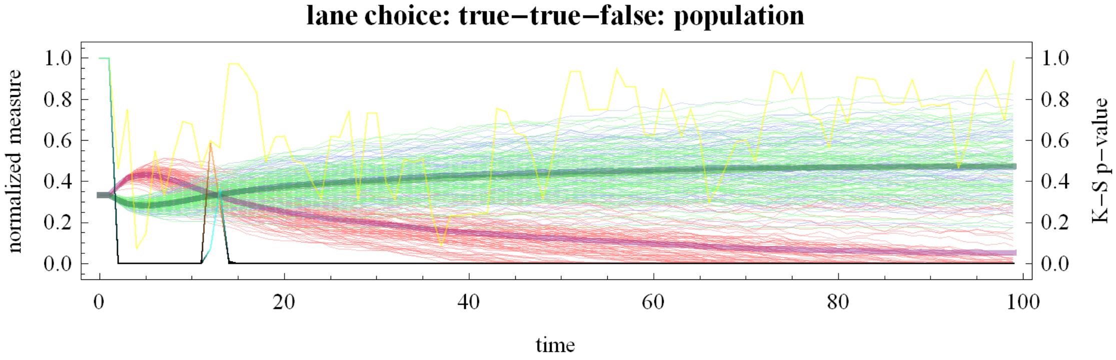
Mean:	0.239	0.593	0.168
Maximum:	0.672	0.995	0.547
Upper Quartile:	0.343	0.816	0.308
Median:	0.226	0.639	0.149
Lower Quartile:	0.109	0.433	0.000
Minimum:	0.000	0.005	0.000
K–S p–value:	0.000	0.000	0.005

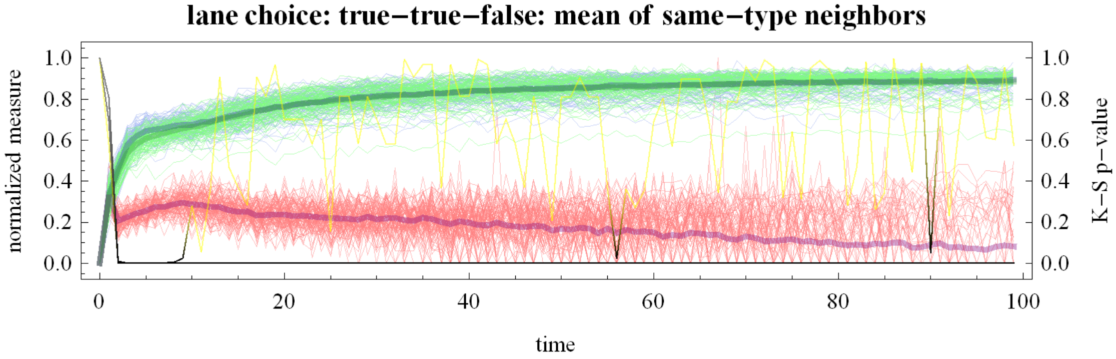


Mean:	0.861	0.920	0.643
Maximum:	1.000	1.000	1.000
Upper Quartile:	0.997	0.997	0.992
Median:	0.992	0.995	0.983
Lower Quartile:	0.977	0.989	0.000
Minimum:	0.000	0.000	0.000
K–S p–value:	0.015	0.000	0.004

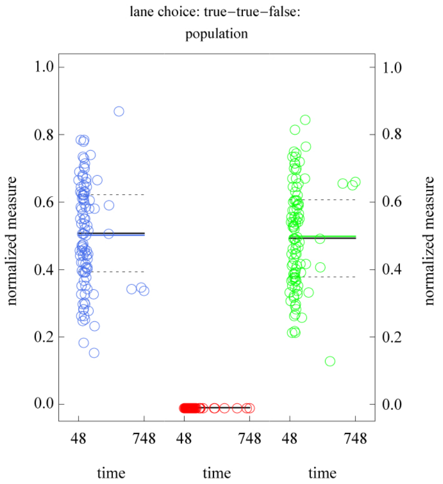
Population dynamics behavior also mirrors the Prisoners' Dilemma behavior with allowanced for the proportions of agents and the existence of two types of cooperators. For example, again the defectors get an initial boost before steadily declining in number toward zero. The A- and C-types then split the population roughly equally between them. This is expected since they each receive the same payoffs in homogeneous communities and (as shown in the plot of same-type neighbors) the stable portion of each type's group (i.e., not the part churning through death/replication) is highly homogeneous (91% assortativity on average).

the one with the lowest ID number. Agents are created in order by type, and hence all A-Type agents initially have lower ID numbers than B-Type agents, which are lower than C-types. This affects detachment whenever neighbors of two different types are providing some agent with equally low utility that is also lower than some third type of agent. In 2×2 games this is impossible because there is no third type to be lower than. In the Lane Choice games, A-Types detach B-Types before C-Types, but C-Types detach A-Types before B-Types. From this observation we should expect more heterogeneous clusters involving B- and C-Type agents than B- and A-Type agents.

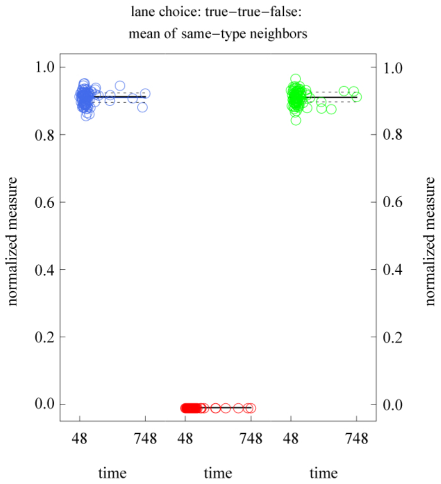




The behavior plot and the outcome plot for population both reflect a high variability in the numbers of each kind of cooperator. Because homogeneous groups of these types each earn the same payoffs, there is no systematic pressure distinguishing the ascent of one over the other, and once one gains population share, it will maintain it with just slight volatility due to noise. The reason (as explained before) is because the 5% of new agents will typically have the fewer connections, hence lower utility, and will therefore be the ones selected for removal in the very next step. Variation in outcome populations, therefore, is explained by the temporary connections with defector agents which reduce the utility of exploited neighbors in the early periods. Since connections are made uniformly at random, the noise in the outcome is evenly distributed such that the mean population level over the 100 runs is 50% for both cooperators despite run-by-run variation from thirteen to eighty-seven percent.



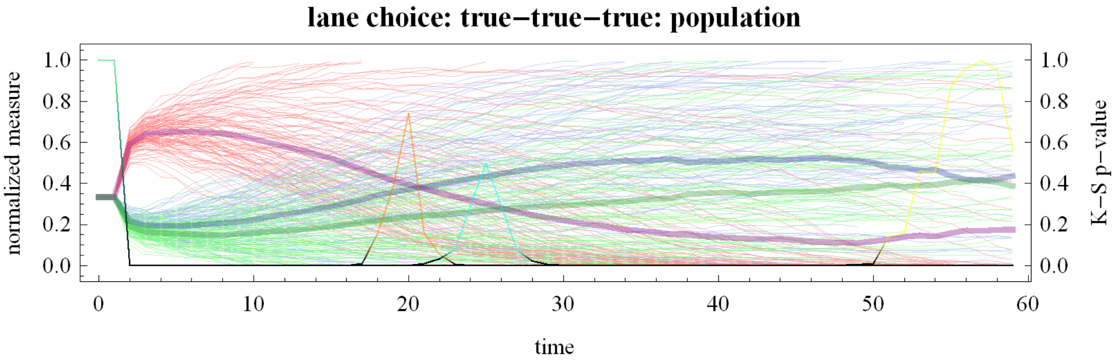
Mean:	0.502	0.000	0.498
Maximum:	0.871	0.000	0.846
Upper Quartile:	0.622	0.000	0.607
Median:	0.507	0.000	0.493
Lower Quartile:	0.393	0.000	0.378
Minimum:	0.154	0.000	0.129
K–S p–value:	0.000	0.000	0.989

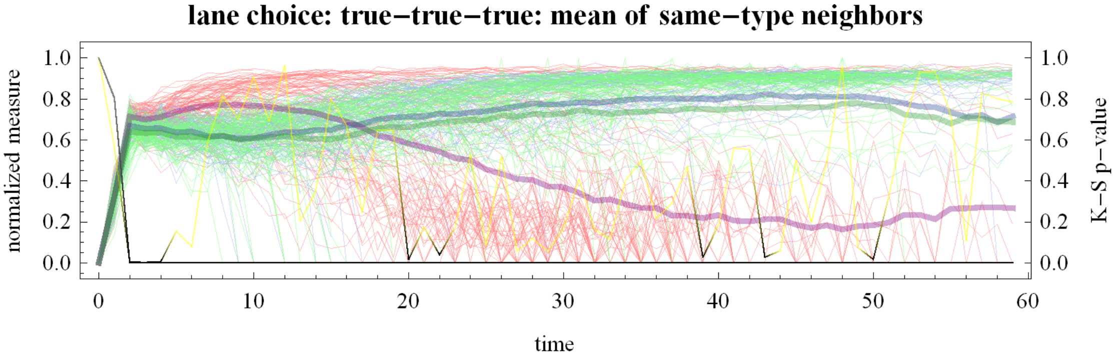


Mean:	0.910	0.000	0.911
Maximum:	0.954	0.000	0.967
Upper Quartile:	0.924	0.000	0.926
Median:	0.913	0.000	0.911
Lower Quartile:	0.896	0.000	0.897
Minimum:	0.857	0.000	0.844
K–S p–value:	0.000	0.000	0.906

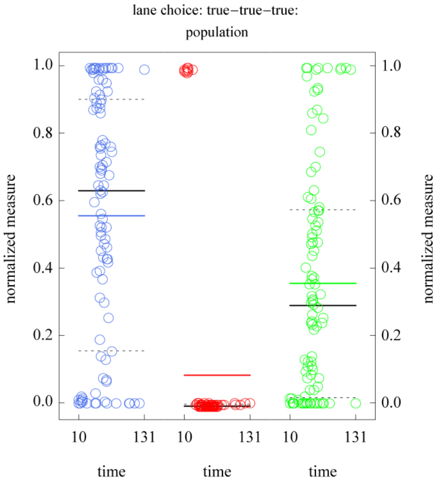
The result of the population dynamics experiment on Lane Choice, therefore, is that it achieves the prosocial outcome identified for the Lane Choice game. The outcome population is a split between the two types of cooperators. Though the proportion of each type is not typically close, each type forms its own segregated community and the process is not biased.

The behavior of Lane Choice with both learning and population dynamics is highly volatile and spans a variety of outcomes and patterns. The behavior shares features from the two previous experiments: 1) learning can lead to early convergence to all defectors; 2) otherwise the initial population boost is gradually diminished until the population is split between the two cooperator types; 3) the outcomes of the two cooperator types are statistically indistinguishable; and 4) the role of the churning of agents through population dynamics adds noise to the distribution of agent types established by learning and detachment, but it also ensures that separated communities of defectors are eventually eliminated. It also seems that the agent ID bias results results in lower assortativity for C-types compared to A-types and this explains the difference in mean population levels across the 100 runs (though still not statistically significant).

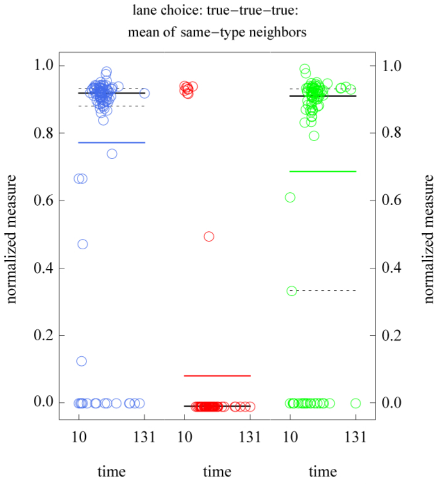

lane choice: true−true−true: population


lane choice: true−true−true: mean of same−type neighbors

The variance in the outcomes is greatly increased in this experiment. The proportion of A-types and C-types fall anywhere along the spectrum from zero to one. Defectors dominate in nine of the runs, and this always during the first 25 time steps. The speed which the other types recover from the defector boost also exhibits high variation: as quickly as cooperator dominance in 31 steps or as long as 131 steps.


lane choice: true−true−true: population

Mean:	0.555	0.091	0.354
Maximum:	0.995	0.995	0.995
Upper Quartile:	0.900	0.005	0.572
Median:	0.629	0.000	0.289
Lower Quartile:	0.154	0.000	0.015
Minimum:	0.000	0.000	0.000
K−S p−value:	0.000	0.000	0.000


lane choice: true−true−true: mean of same−type neighbors

Mean:	0.772	0.089	0.686
Maximum:	0.985	0.942	0.992
Upper Quartile:	0.932	0.000	0.931
Median:	0.919	0.000	0.910
Lower Quartile:	0.880	0.000	0.333
Minimum:	0.000	0.000	0.000
K−S p−value:	0.000	0.000	0.276

The complicated output of Lane Choice with all mechanisms operating is directly related to the complication of the situation and the individual effects of each mechanism. In this experiment it is not enough to consider the behavior as if there were one type of cooperator, because learning allows defectors to exploit cooperators who connect to cooperators of the wrong type – at least temporarily. Population dynamics will not select between the two types of cooperators, but even these agents will switch types in mixed groups through learning. And the result of population dynamics is that mixed groups are constantly forming...and then breaking apart. Yet despite the volatility in the behavior within a run, and across the runs, the aggregate behavior is still largely successful (91%) in creating a cooperator-dominated population of some distribution of *A*- and *C*-types.

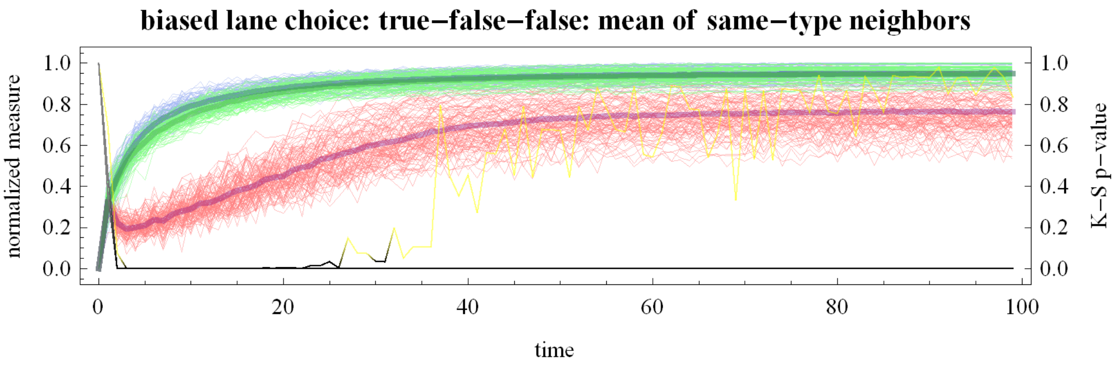
3.9.8 Biased Lane Choice

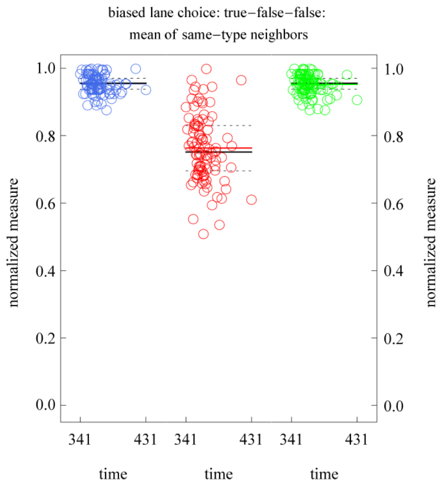
We now add a twist onto the structure of the Lane Choice game; the *C*-type cooperators receive a higher payoff from same-type neighbors than *A*-type cooperators do. In order to maintain the same payoff rankings it is also necessary to increase the exploitation payoff that defectors get from connections to cooperators. The resulting 3 × 3 payoff matrix is displayed below.

Biased Lane Choice

		player2		
		A	B	C
player1	A	3.3	1,5	1,1
	B	5,1	2,2	5,1
	C	1,1	1,5	4,4

These payoff changes will have no effect on the behavior of the base model (compared to the unbiased Lane Choice experiment just analyzed) because the revealed preference rankings over outcomes are identical for all agents. So once again the behavior of the cooperators taken together is similar to that of the Prisoners' Dilemma adjusting for the altered proportions of cooperators and defectors.

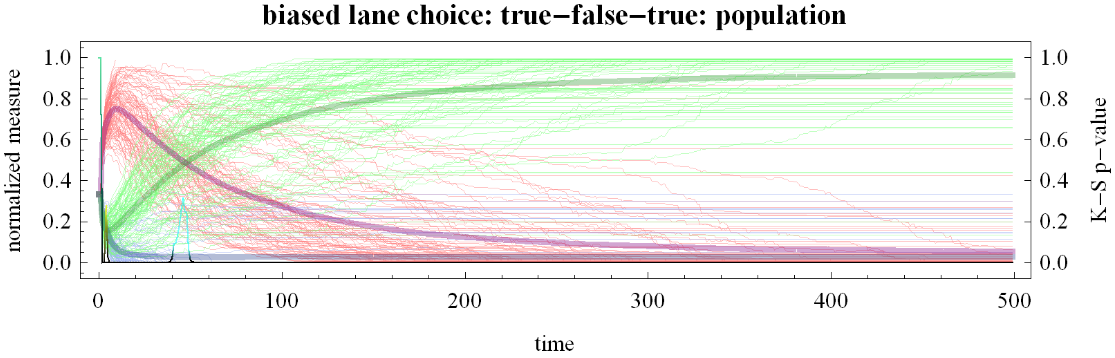


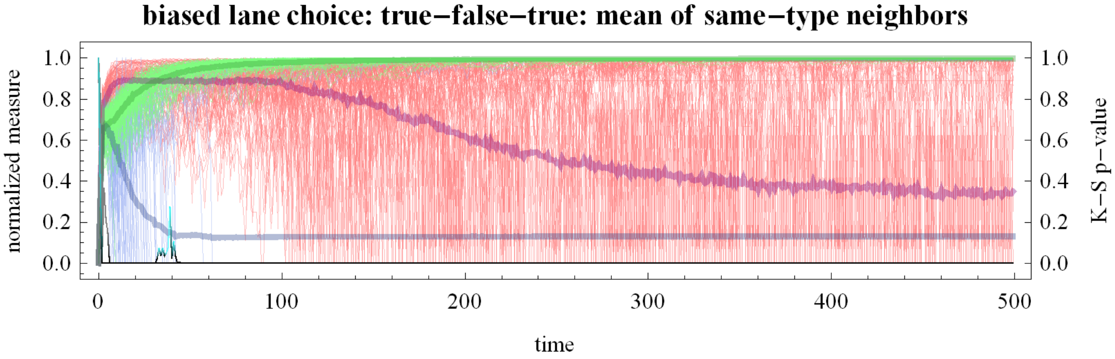


Mean:	0.955	0.766	0.952
Maximum:	1.000	1.000	1.000
Upper Quartile:	0.970	0.832	0.970
Median:	0.955	0.754	0.955
Lower Quartile:	0.937	0.699	0.937
Minimum:	0.878	0.515	0.878
K–S p–value:	0.000	0.000	0.948

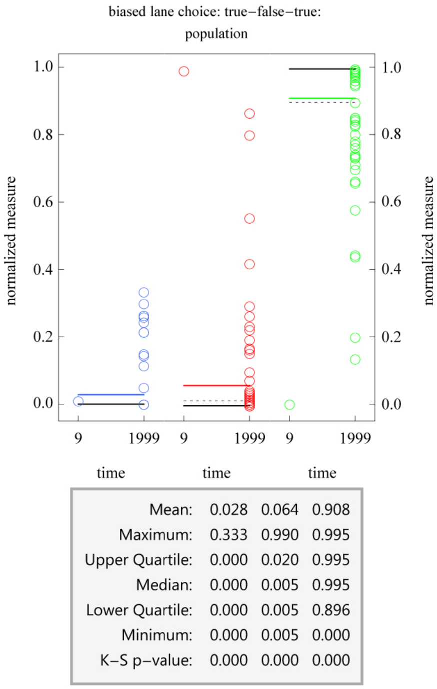
The variation in outcomes is also indistinguishable from the outcomes of the Lane Choice game. Stable heterogeneous groups may persist, and their number is greater here than in the Prisoners' Dilemma, but no different than in the Lane Choice game. For further analysis of the results refer to those for Lane Choice above.

When local success-based imitation is added to the agents' repertoire, the expected effect is indeed observed. Defectors still enjoy an initial jump in population as a result of the much higher payoff they receive in heterogeneous configurations. In only 1 of the 100 runs was the sharp initial rise in defector numbers enough to bring the population to defector convergence. However, because learning is a local behavior, and cooperators detach defectors, a nearly saturated group of defectors forms and persists in many of the runs. That it is "nearly" saturated can be discerned from the plot of same-type neighbors which shows that defectors perpetually transition back and forth from being segregated to high levels of integration. However, the average level of assortativity gradually decreases from an initially high level (because agents are becoming defectors) to much lower sustained levels of 35%. The high volatility is the result of C-type agents with several C-type neighbors imitating B-types, and hence being counted as a B-type with several other-type neighbors. Because unsaturated agents on the fringes of homogeneous communities interact with other-type agents, they go through irregular cycles of type-changing, connection, and detachment. The irregularity is revealed by the fact that in every run except for the one with early defector convergence, the simulation ends via the time-out condition even though aggregate behavior settles by $t = 500$.

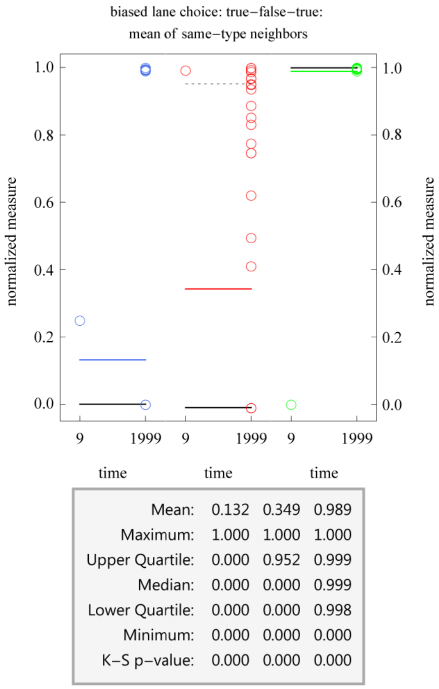




The result is that in more than half the runs the C types gain complete dominance; and achieve a mean value of 91% of the population over all runs (including those with stable communities of A-types and/or B-types, as well as the B-dominated one). Furthermore, the C-types are nearly perfectly segregated in every run, the exception being temporary transitions to defectors discussed above. In more than 75% of the runs the A-types are eliminated completely, more often than even the defectors. So even though there is an easily identifiable “characteristic” behavior for each type, the tails of the distribution of outcomes is quite thick. The high variation in output, though focused on few runs, underscores the role that a rare series of events may have to shape the future development of a complex system.



Mean:	0.028	0.064	0.908
Maximum:	0.333	0.990	0.995
Upper Quartile:	0.000	0.020	0.995
Median:	0.000	0.005	0.995
Lower Quartile:	0.000	0.005	0.896
Minimum:	0.000	0.005	0.000
K–S p–value:	0.000	0.000	0.000

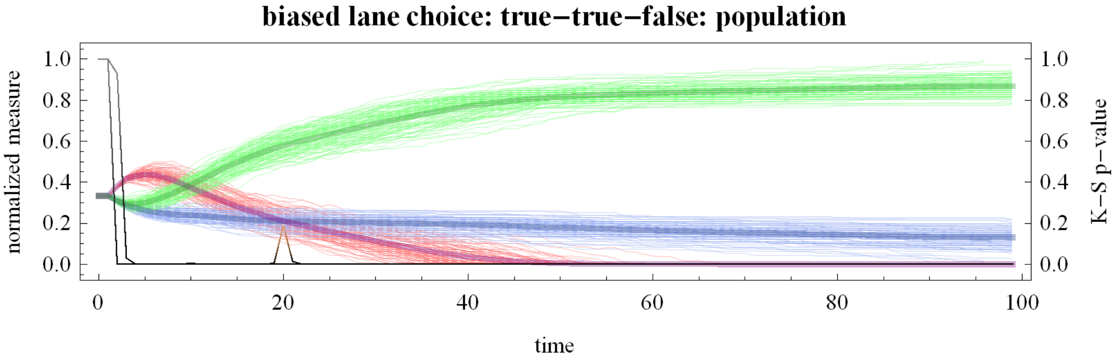


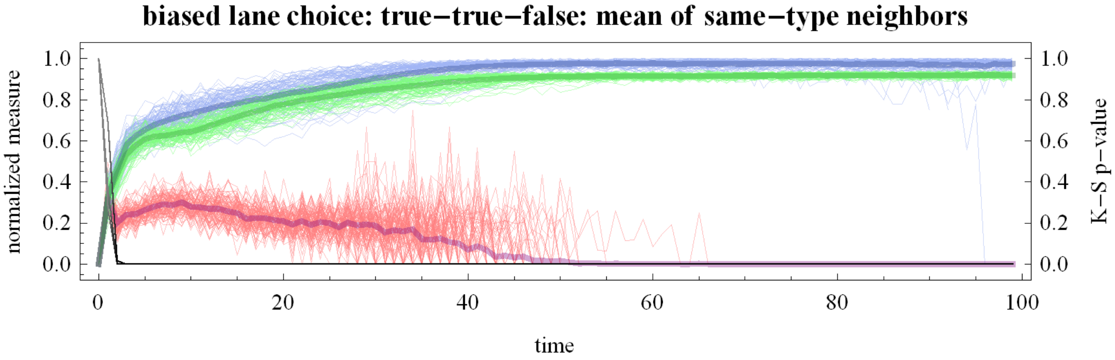
Mean:	0.132	0.349	0.989
Maximum:	1.000	1.000	1.000
Upper Quartile:	0.000	0.952	0.999
Median:	0.000	0.000	0.999
Lower Quartile:	0.000	0.000	0.998
Minimum:	0.000	0.000	0.000
K–S p–value:	0.000	0.000	0.000

The result here is strongly prosocial: the cooperative agents with the higher coordinated payoff succeed in taking over the population in most runs, and still succeed in forming a segregated community (i.e., avoid exploitation and inefficient cooperators). Though defectors do even better in mixed groups than the unbiased Lane Choice game, they cannot get a foothold into population growth once a homogeneous community of cooperators forms – just like in the Prisoners' Dilemma and other relevant situations. The trajectories of population levels reveal that, even though there is a large basin for C-type domination, the outcome is still very sensitive to random connections made during the first few time steps.

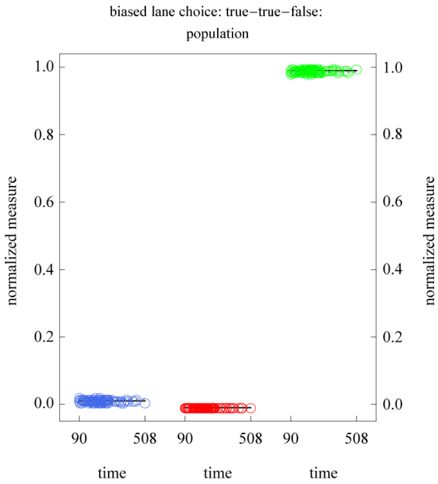
Population dynamics even more clearly favors the advantaged C-types in the Biased Lane Choice. After a short 5-step period of defector superiority, communities of each type begin to form. C-types in a community of pother C-types are doing the best, and groups of B-types are doing the worst, and hence population dynamics trades off these two types while sparing the A-types completely (because they are in the middle of the payoff scale). Once the B-types have been eliminated the A-types start to decline, though more gradually than the B-type demise.

The ability of A-types to hang on longer after the death of the last B-type can be explained by 1) the more assortatively mixed A-type neighborhoods, and 2) the new C-type agents having few connections. This arrangement lets a larger number of A-types receive their max payoff, and simultaneously the C-types (which are at this time approximately 80% of the population) have enough out-of-self-similar connections and missing connections that they fail to achieve their maximum utility. The combined result is that fewer than the full 5% are of A types are being replaced by C-types each period.

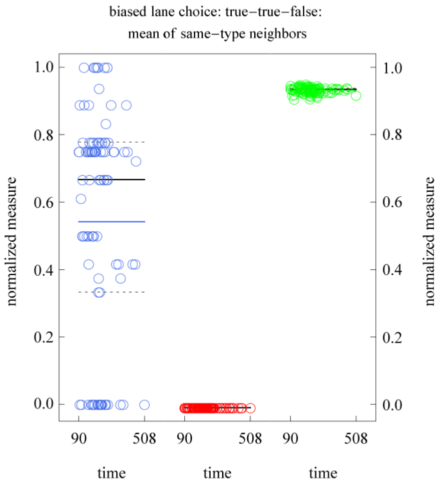




Another recognizable feature of the behavior of population dynamics is that the run-by-run variability is lower; i.e., the behavior is more consistent across the 100 runs which implies that the behavior is much less sensitive to random variation in the connections and particular configurations. The outcome plots below mask the final result of complete C-type dominance because some A-types hold out long enough for some other halting condition to be satisfied before their inevitable removal. For example, once 98% of the agents are C-types, the lowest 5% of the agents are new C-type agents with few connections...and the incumbent A-types and C-types are nearly fully saturated so that the new C-type agents have little hope to reach a configuration that beats out a fully self-similarly connected B-type.



Mean:	0.010	0.000	0.990
Maximum:	0.020	0.000	0.995
Upper Quartile:	0.015	0.000	0.990
Median:	0.010	0.000	0.990
Lower Quartile:	0.010	0.000	0.985
Minimum:	0.005	0.000	0.980
K−S p−value:	0.000	0.000	0.000

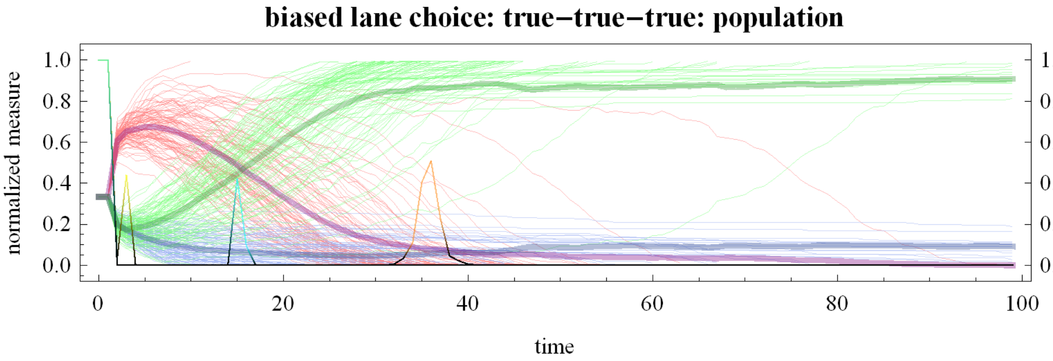


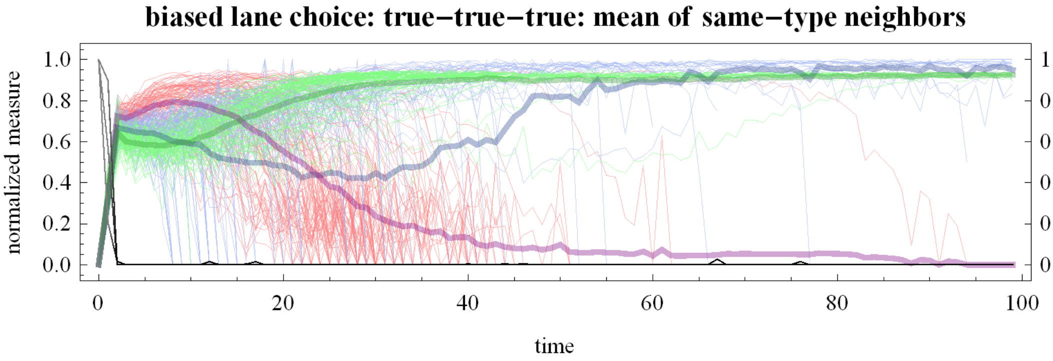
Mean:	0.542	0.000	0.933
Maximum:	1.000	0.000	0.950
Upper Quartile:	0.778	0.000	0.938
Median:	0.667	0.000	0.935
Lower Quartile:	0.333	0.000	0.928
Minimum:	0.000	0.000	0.905
K−S p−value:	0.000	0.000	0.000

The result is again strongly prosocial. Not only do both cooperator types succeed in coordinating and out-performing the defectors, the cooperator type with the greater utility for coordinated, cooperative configurations eventually dominates the population. Hence this is prosocial in two ways: 1) cooperators increase in population at the expense of defectors despite defectors earning a greater payoff in the pairwise connections; and 2) the type of cooperator that yields a greater social benefit wins out over the lower-utility cooperators. This, I will argue in the next section, is an example of group selection (as described in 4.1.3.2) generated by the endogenous processes of preferential detachment in the situation modeled by the Biased Lane Choice game. So the result is both that the prosocial outcome arises in such a context and that this happens even when the context requires an implicit group selection process to achieve it.

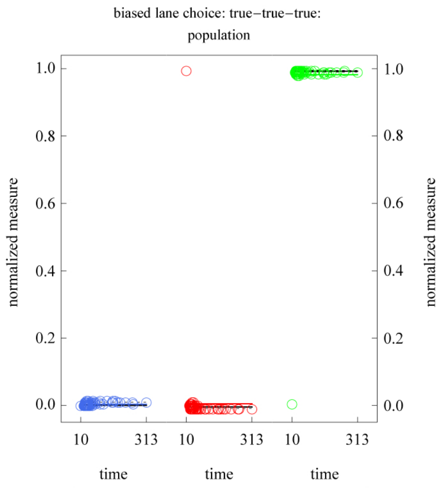
Finally we come to the scenario of the Biased Lane choice with both learning and population dynamics. The mechanisms combined produce a result incorporating features of both the individual mechanisms. The variance among the runs is less tight than population dynamics alone, indicating that the earth-period path dependence generated through learning continues to play an important role in shaping system behavior. The behavior is more consistent than the learning-only experiment in that the trajectories of population figures are similar for each run but spread out over varying lengths of time.

The opening leap in defector numbers is more dramatic (as is the drop in both types of cooperators), but the recovery and domination of the C-types is also quicker and more complete. Despite dropping to fewer than 25% of the population by the 5th time step, the C-types sometimes reach 100% by the 30th. This echoes the effect that we've already seen: the combined mechanisms function to amplify the behavior produced by either one. A side-effect of the more dramatic initial drop in cooperator populations is that the A-types are left with too few agents to build a stable homogeneous group.

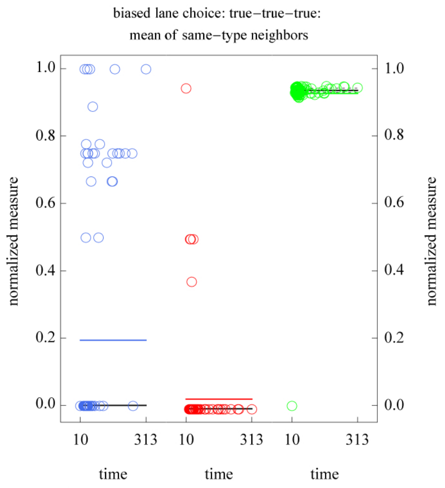




In this situation the amplification means that before communities form, the defectors are performing even better (and they receive a payoff of 5 here for exploitation, compared to 4 in the Lane Choice, so there are more configurations that exploiting defectors can beat). After communities form, but before fully assortative, the C-types benefit from both imitation and population dynamics. Whereas the A-types were spared during this intermediate phase with population dynamics, learning now pushes their numbers lower, often to zero, in these transitional time steps. This is directly related to lower assortativity of A-types; when a perimeter agent connects to a C-type, it typically imitates the C-type, which means all its non-perimeter A-type neighbors become the new perimeter agents. The increase in A-type average assortativity from time 35 to 60 is not the result of them forming tighter communities, but rather that the runs in which they failed to form segregated communities the C-types have already dominated the population and those runs have ended.


biased lane choice: true−true−true: population

Mean:	0.004	0.014	0.982
Maximum:	0.015	0.995	0.995
Upper Quartile:	0.010	0.005	0.995
Median:	0.000	0.005	0.993
Lower Quartile:	0.000	0.000	0.990
Minimum:	0.000	0.000	0.005
K–S p–value:	0.008	0.000	0.000


biased lane choice: true−true−true: mean of same−type neighbors

Mean:	0.194	0.028	0.926
Maximum:	1.000	0.944	0.950
Upper Quartile:	0.000	0.000	0.942
Median:	0.000	0.000	0.935
Lower Quartile:	0.000	0.000	0.930
Minimum:	0.000	0.000	0.000
K–S p–value:	0.000	0.000	0.000

Because the agents in the C-type coordinated cooperative group receive greater utility than the agents that converge on the A-type group, the C-types agents can gain dominance through imitation and population dynamics. Because preferential detachment is so effective at achieving assortative mixing in the appropriate contexts, the agents quickly form type-specific groups and exclude the defector agents. This process is so efficient that C-types gain complete dominance in all but one run; early defector convergence is still possible if the random connection happen just right for them. Most runs end fairly quickly (fewer than 80 time steps), and the runs that take longer than this do so because of the same behavior described in the population dynamics case above: the communities of A-types and C-types are segregated and the agents being removed are the newly-formed C-types.

The result, therefore, is again hugely successful in achieving the prosocial outcome. The socially better type of cooperator gains complete dominance in 99% of the runs. Exploitative agents are the first to die out on average – despite their initial success in every run. Recall that with either just learning or just population dynamics the C-types were already successful; amplifying these effects makes them more successful. Note also that, the Biased Lane Choice game is more successful than either the Prisoners' Dilemma or the unbiased Lane Choice game, especially considering the learning-only experiments. Even though the payoffs to cooperation and exploitation have increased, the net effect benefits the communities of coordinating cooperators more than the perimeter defectors.

3.9.9 Summary of Game Sweep Results

This section highlights the most important results described in detail above. The outcome figures for population levels and percent-similar neighbors are provided for each experiment in table 3.8 below. There are cases (described in the game-specific results) in which the outcome numbers mask distinguishing properties of the system behavior; such as 1) being just one time-slice of a process that is perpetually fluctuating, 2) an experiment that is halted prematurely, or 3) periodic and/or limiting behavior. And, naturally, it is often not the outcome but the process that is really of interest.

Cooperative Games: The cooperative games are highly successful in achieving prosociality except in the case of the Prisoners' Dilemma with imitative learning. This situation still succeeds in segregating the agents, but the segregation prevents cooperators from being imitated after the time at which they form the groups that make them more successful. In the three cases in which cardinal payoffs matter, the Hawk and Dove game produces either more highly prosocial outcomes, or the same level of prosociality more quickly and reliably (despite indistinguishable results for the base case). The difference in results underscores the importance of calibrating the cardinal utilities to produce meaningful results in a particular application.

Coordinative Games: Preferential detachment is maximally successful in producing the prosocial outcome in both the Coordination Game and Battle of the Sexes in all four cases. The difference in payoffs does produce distinct system behavior; specifically, the population trajectories are always indistinguishable in the pure Coordination Game, whereas the A-type systematically enjoy a population boost when any type-changing behavior is included in the Battle of the Sexes. This difference does not affect the level of prosociality, however, because that is defined as assortativity in these coordinative strategic contexts.

Specialized Games: Lichen, the miscoordination or specialization game, requires a system arrangement that is maximally disassortative rather than assortative. The payoffs to agents reflect that they perform better when attached to other-typed agents, and preferential detachment therefore has them disconnecting from agents of the same type. The base model produces the ideal "checkerboard" network arrangement; including population dynamics does not interfere with this drive and merely registers as noise. However, whenever imitative learning is included, the outcome is moderately to highly assortative because local imitation can only reduce the variation in neighbors' types. The outcome in such experiments is not random, however, and exhibits a robust and dynamically stable balance of system properties characteristic of complex systems – not the behavior we were looking for here, but certainly worthy of further study.

Contributive Games: The base preferential detachment model on the Stag Hunt game produces nearly fully segregated groups because having just one type initiating detachment suffices to produce high levels of assortativity (i.e., prosociality) – a result that is not prima facie obvious, but deducible from the theory upon retrospection. Learning produces the prosocial outcome of stag hunter dominance because B-types readily cohere to A-types, who are receiving a higher payoff as long as a few of their other neighbors are A-types. And because preferential detachment alone produces segregated populations, adding population dynamics ensures that the agents with a higher-paying intra-type payoff take over the population.

Commensal Games: In this category the A-types (commensals) prefer connections with B-types (hosts), and the B-types are indifferent. Because B-types connect indiscriminately and stay connected, they absorb half of the available B-type connections among themselves. This outcome denies A-types access to their preferred neighbor and is no better for B-types than any other network arrangement in which their maximum degree is saturated. The result for the base model is that the level of A-type assortativity is at 22% instead of the optimal level of zero, and they maintain fewer than K connections. Adding learning can never increase disassortativity, so the level of prosociality decreases because the B-type agents copy their more successful A-type neighbors. At the system level there is feedback between the number of each type because the commensal A-types only perform better when there are a sufficient number of B-type "hosts" available. This system-level feedback stabilizes the numbers of each type and the assortativity of each type within a narrow band – which is nearly the same in all 100 runs – despite a constantly changing network structure and churning of agents' type. But the process (however fascinating) is inefficient, even to the point that the agents would be better off if they were all B-types.

The outcome produced by population dynamics is qualitatively nearly identical to the base model, but there are slightly more A-types here. The reason is that A-types perform better, and increase in number, but their improved performance depends on the number of B-types available, so the population quickly stabilizes to proportions such that both types earn the same utility on average. Given the cardinal payoffs used here, the A-types perform as well as the saturated B-types despite only utilizing half their maximum degree. Finding a dynamically stable equilibrium proportion of agents is extremely interesting, and the game is novel to the literature, but the level of disassortativity for A-types (prosociality) is 26.7% compared to an ideal value near 5%. The results from learning and population dynamics combined are dominated by the effects of learning; hence local imitation overpowers detachment and the arrangements produced are moderately-to-highly assortative.

Undifferentiated Games: In the base model of Matching Pennies, each run finishes with mean percent similar values near 0.500; as expected when an evenly distributed population connects randomly and keeps all its connections. When learning is added, agents that (by chance) gain more connections than neighbors in the formative periods have greater utility and are therefore imitated. Because all agents are indifferent this continues until all agents are saturated. Learning thus produces high run-by-run variation in outcome populations, but the aggregate value across runs is again equal for both types. Population dynamics cannot produce cascades across the network the way learning can, so the result is the same as the base model plus a 5% noise added at each time-step. When learning and population dynamics are combined, the result is the frozen highly variable run-by-run trajectories produced by learning, but with 5% noise added to the population level forged in the early periods. The most interesting result is that when imitative learning is involved the outcome of a run can be biased as much as 23:77% based on the path dependence of early "lucky" agents; and this despite all agents receiving the same utility from all connections. The variance in the results belies the undifferentiated nature of the setting.

Collaborative Games: The Lane Choice and Biased Lane Choice games represent a more sophisticated situation that combines the needs of coordination and cooperation. The results cannot be directly extrapolated from the constituent games because the proportion of agents in each role is different. Without type-changing behavior the results mirror the Prisoners' Dilemma with more cooperators to exploit, although the cooperators are highly successful in avoiding the heterogeneous groups that allow exploitation. Because the prosocial outcome is reached when assortativity is high, and learning promotes assortativity, adding learning might be expected to produce high levels of prosociality. However, the greater number of initial cooperators in Lane Choice gives defectors an immense early-period boost. Their population ratchets down, but leaves a high number of defectors on average when type-specific groups have formed. Population dynamics favors cooperation over defection, and in the Lane Choice game both types of cooperator taken together dominate. Despite a run-by-run population variation between 20 to 80% for each type, they share an equal proportion of the population on average. With learning and population dynamics both included, the Lane Choice game produces outcomes that range the whole spectrum: e.g., it is even the case that each of the three types dominates in some runs. The defectors can only win if they do so early through lucky connections and learning – 9 of 100 runs. Excepting these runs, the population ends with some proportion of A- and C-types similar to the Coordination Game results on average (but highly variable due to the path dependent defector elimination stage).

Without learning or population dynamics the Biased Lane Choice is identical to the Lane Choice game. With learning, the C-types (who receive a higher coordinated payoff than A-types) nearly dominate the population in most runs, very different from the defector success seen in the unbiased Lane Choice. The aggregate results with population dynamics are very similar to learning, except here the run-by-run variation is small, whereas with learning the individual runs end with very different proportions of agents. The combined rules produces results that combine features of both: some early defector success, then a sharp rise in C-types, followed by a slow death of the B-types as they are selected against via population dynamics despite forming a homogeneous segregated group.

Imitative Learning: In addition to the game-specific results just described, there are general results with respect to patterns in system behavior across games. Learning creates comparatively volatile behavior in most games as well as making the run trajectory/outcome sensitive to the early periods. Both effects result from learning being a local behavioral rule. Starting from initially unconnected agents and random connections, there is a great deal of stochastic variance in the arrangements of agents in these initial periods – some of which greatly benefit one type over another – and this advantage is then amplified through imitation of the lucky agents. The detachment rule operates over the same network as imitation, so when an agent changes its type, it also affects the types of agents it is willing to accept as a neighbor and the types that accept it. Detachment and attachment happen at roughly the same rate in unsaturated portions of the network, and random variation in these processes creates a baseline volatility in system behavior, but adding imitation on top of that allows type/connection changes to propagate through otherwise stable portions of the network structure and disrupt system behavior.

Population Dynamics: In contradistinction to learning, adding population dynamics typically has a smoothing effect on run-by-run trajectories; making it less volatile and less path sensitive. These effects are the result of it being a global operation that does not disrupt large portions of the network. For example, a lucky defector in the Prisoners' Dilemma with four cooperator neighbors will create 1 new defector offspring before being detached by all its neighbors. With learning engaged, those four cooperators would likely all become defectors. 5% of the population is removed and added each time step, so 95% of the network is undergoing standard preferential detachment processes. But the smoothing effect goes beyond just this: even in base preferential detachment there are fringe agents that are constantly connecting and disconnecting, and hence perturbing the network arrangement. These agents, because they have fewer connections, have lower utility and are exactly those selected for elimination by population dynamics. Newly created agents also typically have lower utility and are those selected for elimination for the same reason, and so the saturated part of the network remains completely stable through the population dynamics process. Despite this overall smoothing effect, and whereas learning may (and usually does) create stable arrangements in segregated populations, population dynamics is perpetually removing and adding agents which makes stasis (and sometimes pure prosociality) impossible.

Preferential Detachment

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim	C %-pop	C %-sim
Prisoners' Dilemma	0.500	0.965	0.500	0.896		
Hawk and Dove	0.500	0.965	0.500	0.899		
Battle of the Sexes	0.500	1.000	0.500	0.999		
Coordination Game	0.500	1.000	0.500	1.000		
Lichen	0.500	0.000	0.500	0.000		
Stag Hunt	0.500	0.961	0.500	0.961		
Commensal	0.500	0.220	0.500	0.475		
Matching Pennies	0.500	0.496	0.500	0.496		
Lane Choice	0.333	0.954	0.333	0.763	0.333	0.954
Biased Lane Choice	0.333	0.955	0.333	0.766	0.333	0.952

Learning

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim	C %-pop	C %-sim
Prisoners' Dilemma	0.411	0.990	0.589	0.996		
Hawk and Dove	0.991	0.999	0.009	0.180		
Battle of the Sexes	0.565	0.999	0.435	0.999		
Coordination Game	0.495	0.999	0.505	0.999		
Lichen	0.489	0.694	0.511	0.709		
Stag Hunt	0.991	0.998	0.009	0.188		
Commensal	0.759	0.842	0.241	0.446		
Matching Pennies	0.514	0.667	0.486	0.647		
Lane Choice	0.239	0.861	0.593	0.920	0.168	0.643
Biased Lane Choice	0.028	0.132	0.064	0.349	0.908	0.989

Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim	C %-pop	C %-sim
Prisoners' Dilemma	0.993	0.936	0.007	0.004		
Hawk and Dove	0.993	0.936	0.007	0.029		
Battle of the Sexes	0.583	0.912	0.417	0.905		
Coordination Game	0.503	0.912	0.497	0.912		
Lichen	0.499	0.028	0.501	0.031		
Stag Hunt	0.990	0.936	0.990	0.338		
Commensal	0.514	0.267	0.486	0.488		
Matching Pennies	0.500	0.473	0.500	0.466		
Lane Choice	0.502	0.910	0.000	0.000	0.498	0.911
Biased Lane Choice	0.010	0.542	0.000	0.000	0.990	0.933

Learning and Population Dynamics

Parameter Value	A % pop	A %-sim	B %-pop	B %-sim	C %-pop	C %-sim
Prisoners' Dilemma	0.983	0.927	0.017	0.039		
Hawk and Dove	0.993	0.937	0.007	0.022		
Battle of the Sexes	0.631	0.912	0.369	0.903		
Coordination Game	0.509	0.905	0.491	0.908		
Lichen	0.499	0.663	0.501	0.670		
Stag Hunt	0.990	0.938	0.010	0.263		
Commensal	0.747	0.771	0.253	0.465		
Matching Pennies	0.504	0.621	0.496	0.611		
Lane Choice	0.555	0.772	0.091	0.089	0.354	0.686
Biased Lane Choice	0.004	0.194	0.014	0.028	0.982	0.926

Table 3.8: Results Across Strategic Contexts

3.10 ABM Conclusions

The implications of the agent-based model relevant to substantive questions in morality are left for the next chapter. This section, instead, covers conclusions regarding the results of the ABM model and their relationship to the theory of preferential detachment, the Markov model presented in chapter II, and results of other models from the literature. These include an evaluation of the sensitivity of system properties, the coverage of the strategic contexts, the systemic effects of learning and population dynamics, a comparison of mechanisms, and other notes on the behavior of preferential detachment models that indicate potentially fruitful avenues of future research.

3.10.1 Sensitivity to Parameters

Testing the sensitivity of the results to changes in the input or model parameters serves several aims. When we make claims of the form "preferential detachment suffices to produce a prosocial arrangement for over 90% of cooperator agents", we want to be sure that the conclusion does not depend on the assumptions being confined to narrow bands of parameter values. On the other hand, we can gain the greatest insight into how the system-wide dynamics are produced by the agent interactions by tracking how changes in the agent behavior (controlled by the parameters) affect the results. Even a simple model such as preferential detachment includes enough assumptions that testing several values for each one becomes prohibitively time-consuming and burdensome. Instead we select as control parameters those 1) capturing the least-defensible assumptions, 2) representing features with known variation, and/or 3) about which we have the greatest interest. For the current research those three considerations overlap on population size, maximum degree, and strategic context.

Because the averages over runs aggregate time-by-time values for 100 runs, they are largely insensitive to a small number of outlier runs. When the variation is just stochastic noise, the variance of the individual runs is reliably smoothed out to produce a near-constant valued aggregate. If the aggregate is volatile and the run-by-run variance is high, then the aggregate is likely hiding irregular or patterned (e.g., periodic, lagged) system behavior rather than revealing "typical" behavior. 'Sensitivity' can refer to changes in the aggregate trajectory, changes in the volatility of the runs, changes in the variance of the runs, changes in the timing of events, and other aspects of model behavior. Each of these sensitivities may be important, and the form of sensitivity appropriate to the question at hand is addressed throughout.

Taking the Prisoners' Dilemma payoffs with $N_A = N_B = 100$ and $K = 5$ as the anchor point, experiment branches 1 and 2 explore the effects of changes in initial population levels and maximum degree, respectively – but all the combinations of these two parameters' values are not run and evaluated. It was shown that for the Markov model of the Prisoners' Dilemma an increase in the maximum degree (in 2.6.1) or an increase in (an evenly divided) population (in 2.6.2) enhances the prosociality of the outcome...and that there are no combination effects. We see a similar result in the outcomes of the agent-based models with the base preferential detachment mechanism.

The ability of preferential detachment to segregate the agents by type increases with both population size (see 3.7.1) and with increasing maximum degree (see 3.8.1). For population sweeps we find that although a larger *number* of agents get trapped in heterogeneous clusters with increasing populations, a smaller *percentage* do. Increasing the maximum degree produces increasingly homogeneous populations of both types of agents and with less run-by-run variation in the outcomes. Because these parameter-sweep ABM experiments utilize only the Prisoners' Dilemma payoffs – with a promise that some of the result can be generalized to other strategic contexts – it is worthwhile to drill deeply into how these results are generated.

Increasing the population has a damping affect on rare events (described below), and thus the greater the population level, the greater agreement there is between the ABM and Markov model results. As an example, the stationary distribution for the Markov model of the cooperative games (shown in table 2.16) with $N_A = N_B = 100$ and K=4 is 93.5% of cooperators in a homogeneous group, and the remaining 6.5% stuck in heterogeneous clusters (and the results vary little with equal changes in the population levels). The ABM with the same parameters produces an mean value of 92.7% of cooperators in a homogeneous group – a close match. With $K = 5$, the ABM produces a mean cooperator assortativity value of 96.5%. Sweeping the population levels from $N_i = 10$ to 1000 with $K = 5$, the mean value is nearly constant, but the range of the data (distance between max and min value) decreases from 22.5% to 3.1%. The variation among individual runs shrinks, and their results converge to a value quite close to the value produced by the Markov Model.

In a similar vein we can compare the stationary distributions from the Markov model (with $K = 4$) to the ABM base model results (though with $K = 5$) for each strategic context (with $N_i = 100$) as presented in table 3.9. The populations are constant in both cases, and prosociality is defined in terms of the level of assortativity. From both the generalizations of the Markov model presented in 2.6 and branches

1 and 2 of the ABM, we know that for the Prisoners' Dilemma the results become increasingly prosocial (for both models) and increasingly consistent (for the ABM) as either N_i or K increases.

	Markov Model		ABM Model	
Game Category	A %-sim	B %-sim	A %-sim	B %-sim
Cooperative	93.47	36.99	96.50	89.60
Coordinative	100.00	100.00	100.00	100.00
Specialized	0.00	0.00	0.00	0.00
Contributive	93.47	100.00	96.10	96.10
Commensal	0.00	49.76	22.00	47.50
Undifferentiated	49.76	49.76	49.60	49.60

Table 3.9: Assortativity levels for the Markov and agent-based models.

For the Markov model, I presented the effects of parameter changes for each game category. The close match between the results of the ABM and Markov models for the base parameters indicates that the Markov model's results may be reliable indicators of behavior for the agent-based model as well. This is a form of cross-validation that Richard Levins calls "robustness analysis" for its ability to demonstrate that the essential features captured in both models of a given phenomena are not sensitive to the simplifying assumptions native to each methodology.

> [I]f these models, despite their different assumptions, lead to similar results, we have what we can call a robust theorem that is relatively free of the details of the model. Hence, our truth is the intersection of independent lies. (in [Weisberg 2006])

The only two outstanding discrepancies between the models' results are the assortativities of defectors and commensals. Both of these results were anticipated because the two agent types are pulling in different directions within these games, yet the Markov model (as constructed) cannot capture system-level arrangements of agents. Hence defectors' and commensals' probabilities to connect to cooperators and hosts do not adapt to the number of available agents with which to form connections. The magnitude of the defector discrepancy is not worrisome in light of the explanation just provided and the known scaling properties of the models. It is, however, important to recognize that in the Markov model the commensals achieve a perfect prosocial arrangement, while in the ABM the degree of prosociality is 22% lower. Increasing the population levels equally will not have an appreciable effect on the prosociality in this game, but increasing the maximum degree will *increase* the prosociality of the Commensal game. This would result because with increasing degree, commensals would become more likely to become attached to at least one host. Such a connection is permanent, and incites the commensals to disconnect any future commensal connections made, thus increasing their disassortativity (prosociality). It would never reach 0%, however, because the hosts would quickly become saturated, and the commensals would continue to randomly connect to (and then disconnect from) their own kind.

Main conclusion: From these results we can conclude that both the Markov and agent-based models are in agreement with respect to the results of the preferential detachment mechanism. This agreement indicates that the results are generated by the encoded mechanism in a way that does not depend on the underlying specifics of either modeling technique. What we can conclude from these combined results is that preferential detachment suffices to reach high (but sometimes not optimal) levels of prosociality across all possible strategic contexts, at all evenly distributed population levels, and for any maximum degree. This conclusion pertains to the base preferential detachment mechanism, and is the primary contribution of the modeling component of this project.

3.10.2 Adding Learning

All of the results involving learning and/or population dynamics, and any conclusions based on these results, must be considered preliminary and tentative because only one set of cardinal values is examined for each game. As described below, there are some obvious cases in which changing the cardinal values has a dramatic impact on the the results, as well as the appropriate interpretation of the model. With this in mind, it is still valuable to analyze the experiments including these additional mechanisms and make tentative conclusions.

Starting our analysis (as usual) with the Prisoners' Dilemma, we know that defectors receive a greater payoff than cooperators in mixed arrangements. Because the agents are initialized unconnected and then connect randomly, the first several periods of every experiment are marked with high levels of mixed arrangements. Hence we observe a "defector boost" in these formative periods before preferential detachment segregates the agents by type. The same behavior occurs in other strategic contexts with learning: Hawk and Dove, Stag Hunt, Lane Choice and Biased Lane Choice – all of which share game-theoretic features of the Prisoners' Dilemma. Though the concept of defection does not directly apply in these games, this "defector boost" can lead to antisocial domination in some runs – even when the outcome behavior is prosocial for most runs.

The size and duration of the initial benefit to antisocial agents is sensitive to both parameters explored here. As the population increases, the cooperators are more likely to form the tightly-connected group (quasi-clique) they need to achieve higher payoffs. This is because imitation can only happen at the perimeter of groups, where agents of different types interact.[4] In the first ten periods, random connections dominate the network formation, and so most agents are perimeter agents.

Because all agents have only a few connections during this initial formative period, a cooperator that happens to get a defector neighbor before a cooperator, or more defectors than cooperators, will convert into a defector. This process happens in every experiment with learning (under all parameter values), but when the population is low, the probability that one of the remaining cooperators will next connect to another remaining cooperator (rather than a defector) shifts a great deal more. For example, the probability of a cooperator connecting randomly to another cooperator when just one cooperator has converted goes from $\frac{9}{19} = 0.47$ to $\frac{8}{19} = 0.42$ when $N_i = 10$. When $N_i = 1000$ the effect is a shift from $\frac{999}{1999} = 0.4997$ to $\frac{998}{1999} = 0.4992$...a near-negligible effect.

Not only is defector connection more likely with lower populations, but any cooperators connected to the newly converted agent will now find themselves exploited. This can initiate a cascade of defector imitation that, in certain arrangements, can lead to defector domination in these early periods; i.e., before detachment can effectively drive segregation. So, when N is larger, not only are the initial connection probabilities more conducive to prosociality, but they remain so through a larger number of early-period defector imitations.

For those runs that don't end via early defector domination (which is most runs), preferential detachment then segregates the agents by type, and (as a matter of contingent fact for the simulations presented here) creates a singular giant component of cooperator agents quickly (typically by $t = 10$). A larger population therefore produces a larger number, *but smaller ratio*, of perimeter agents. Once the agents are segregated, there are few opportunities for imitative learning in either direction.

Increasing the maximum degree, however, prolongs the mixed-arrangement phase just described. The defector boost effect is therefore amplified enough to allow more

[4] A network perimeter is just the subset of a set of nodes **S** such that some edges are to nodes outside **S**. Hence this is not a spatial notion and depends only on the network arrangement. With groups being identified as collections of same-typed nodes, a perimeter node is (by definition) one with a connection to an other-typed node.

runs to end with defector domination. However, when that outcome doesn't obtain, the system maintains mixed arrangements after cooperators have formed their quasi-clique[5] – long enough for all defectors to imitate cooperators and produce cooperator dominance. It therefore cannot be concluded that denser social networks improve prosociality in general; the volatility increases and both extreme outcomes become more common.

The increase in maximum degree shares features with increasing the radius of information/interaction in spatial game theory models (i.e., they become more global and less local): the prosocial outcome similarly becomes less common [Grim et al. 2006, Nowak 2006, Alexander 2010]. In both model types, every cooperator with a defector earning higher than any cooperator with its neighborhood will turn into a defector. Yet ,the effect is quite different between the fixed lattices used in those models and the dynamic social networks used here; specifically, the cooperators act to exclude defectors from their neighborhood. Larger maximum degree values make this more difficult and that is the parallel. But the real effect here is to delay segregation, which benefits different types at different timescales, rather than an across-the-board benefit for defection. This process deserves further investigation to better understand the changing influences across time and through greater neighborhood saturation.

Another consideration is that learning occurs after the detachment procedure (see 3.4.2 for the order operations). So any perimeter cooperator with a single defector neighbor will detach that defector before it can imitate the more successful cooperator. For defectors to learn from a collection of cooperators there needs to be two defectors connected to the same cooperator – a configuration that becomes vanishingly improbable as groups become increasingly intra-connected. This result differs from fixed interaction models in that they also require a contiguous group of cooperators for cooperation to become imitation-worthy, but the structure guarantees defector pupils after the time at which this occurs. The order of these steps is motivated by the consideration that an agent wouldn't naturally imitate an agent that it would otherwise detach...so detachment is first. But a more sophisticated set of contingencies may better capture reasonable behavior, and what comes out as more reasonable may also benefit prosociality more.

3.10.3 Adding Population Dynamics

For the purposes of the evolution of morality, the results from population dynamics are more pertinent than those from imitative learning. And although these results are

[5] The number of cooperators required depends on the relative cardinal payoffs.

also tentative on account of the single cardinal utility values used, system behavior is less sensitive to changes in utility values under population dynamics. The reason is that population dynamics only operates on the lowest and highest 5% of agents. For any cardinal assignment corresponding to a particular ordinal ranking, the same profile of agents will comprise these subgroups if all agent connections are saturated.

Imitative learning ensures local (in-group) homogeneity in assortative contexts, but allows population-wide heterogeneity for unconnected groups. Population dynamics, on the other hand, is a global operation that only admits deviations to global homogeneity in contexts with balanced payoffs across types (more below). The local vs global operation reveals itself in clearly in the output proportions and their trajectories (as seen in the results section above).

The preferential detachment mechanism limits the variety of configurations that are stable for any strategic context. Adding the further constraint that all agents in a specific global arrangement are earning equal total payoffs tightly constrains the set of arrangements that are stable under population dynamics. Arrangements from within that constrained set will still yield varying payoffs whenever there is heterogeneity in the agents' number of connections, i.e., along the way from an unconnected arrangement to the stable one. And in some cases (e.g., the Lichen game) the hypothetically stable arrangement is never reached because population dynamics constantly perturbs the system. What we observe is that when the context promotes assortativity and agents earn differently, it produces a population of entirely one type. Otherwise the population converges to a dynamically stable distribution of types such that both types earn equal total payoffs on average. So changing the cardinal payoffs does not effect the qualitative results under population dynamics...with a certain caveat.

The caveat relates to a phenomena similar to the defector boost explained earlier. The way population dynamics are implemented here, each selected agent is reproduced exactly once. So no matter what the cardinal payoff differences are, the maximal sub-population growth rate is 5%. Let's again take the Prisoners' Dilemma as our example. During the initial steps of any run some cooperators are being exploited by defectors; the exploiting defectors are replicated and the exploited cooperators are removed by population dynamics. If the exploitation payoff is greater enough than the mutual cooperation payoff, then a defector with even just a single cooperator neighbor could be outperforming a cooperator saturated with other cooperators.[6]

[6] Using $\psi(\theta ab)$ to represent the cardinal payoff that an agent of type θ earns from having a connections to A-type agents and b connections to B-type agents, the condition can be expressed as $K\psi(AK0) < \psi(B10)$.

Though preferential detachment would eventually completely segregate the types, after which the cooperators would rise in numbers (because if the C-C payoff isn't greater than the D-D payoff then it isn't a Prisoners' Dilemma), the defectors may dominate the population before segregation can occur.

The parameter sweep experiments performed on the Prisoners' Dilemma show us that population dynamics produces cooperator-dominated outcomes across all population and maximum degree levels.[7] Again note the caveat that unless defectors manage to dominate before segregation occurs, these results apply to all cardinal utility assignments for the Prisoners' Dilemma. Similar conditions hold for other strategic contexts, and future work could identify the payoff thresholds outlining the basins of attraction of different outcomes.

The overwhelming success of population dynamics in generating the prosocial outcome across all strategic contexts (given the current cardinal utilities) is mitigated by that 5% of new agents having no connections. No permanent structure is possible among this subset because they are typically the lowest earning agents, and so what we're actually seeking is either a stable distribution with churning agents, or a stable configuration among the other roughly 95% of agents. For domination games this doesn't matter because after the first dozen iterations, the agents coming in are typically all of the dominating type. In the other cases, however, having the same set of agents being added and then removed is at odds with actual population dynamics since in this case many of the agents live forever and only newborns will die. The mechanism was not supposed to be biologically accurate, but rather a mechanistic proxy for population-wide selection pressures of behavior types. The form of population dynamics used here typical among the related literature (as described in chapter I), but when forming conclusions about the model's result it is important to keep this in mind.

3.10.4 Adding Both Learning and Population Dynamics

Each mechanism affects the distribution of agents in a different way, and their combination can produce extreme behavior by amplifying those effects, or interfering with them, or one may be swamped by the effects of the other. As described above, learning acts to make groups homogeneous, whereas population dynamics produces either a globally dominated population or a mixed population of agents with equal

[7] Even more extreme, the sweep of maximum degree shows us that when including a population dynamics mechanism, a maximum degree of twenty is overkill for achieving cooperator dominance – the full number is not even utilized before the simulation ends.

average total payoffs.

The timescales of learning and population dynamics are not explored, and the values currently implemented assume a population turnover of 5% for each opportunity of neighbor imitation. If learning is interpreted as cultural imprinting, assimilation, or other such once-in-a-lifetime operations, then these scales work as they are. But if learning is supposed to be a short-term behavior on the scale of day-by-day personal interactions, then the turnover rate ought to be drastically reduced. Neither mechanism is meant to replicate closely any specific feature of actual human, protohuman, or other animal populations; the level of abstraction is intended to make the model general.

But even considering this point, more exploration into the relative learning and reproduction rates is clearly in order. As already stated, these results must be considered tentative and preliminary, though no more so than the bulk of models in the literature which similarly must make simplifying assumptions, state them clearly, disclose that changes in them would affect the model, and then promise to look more carefully at it in the future. Here are some observations and implications given the current work.

In Cooperative games the effect of the mechanisms combined is very clearly a combined effect: cooperator domination in most (but not all) runs and greater variance among the runs and volatility within each run. The overall effect is the same though: preferential detachment segregates the types and cooperators win in assortative arrangements. The number of runs ending with defector dominance is in between the number for each mechanism alone. When that doesn't happen, the result is a large proportion of cooperators, but straggling defectors do very well in a cooperator dominated environment, so they are more entirely eliminated before the simulation ends. The primary insight gained here is how robust the prosocial result really is for Cooperative games like the Prisoners' Dilemma and Hawk and Dove.

The Coordination-category games differ in that the Battle of Sexes pays A-types more, and in the Coordination Game they are equal: the result is equal proportions in the Coordination Game and more A-types in the Battle of the Sexes. Upon first analysis one might expect the A-types to dominate because types segregate and A-types perform better. This is not so. Once the A-types are consistently in the top 5%, their numbers grow until newly entering A-types are not earning more than connected B-types. At this time A-types and B-types are receiving equal utility on average (over the agents over time within each run), but the specifics change frequently, thus creating volatile population distributions. The balanced payoffs, and the population distribution that balances them, are maintained not by an equilibrium condition, but are rather endogenously generated dynamically stable system-level features. Furthermore, if one changes the cardinal utility values in the Battle of the Sexes or the Coordination Game, the result is a different distribution of agents, but still a distribution in which utility is balanced.[8]

The other games reveal similar patterns and comparisons. In the Lichen, Stag Hunt, Commensal, and Matching Pennies games, the combined dynamics are similar to the learning-only dynamics, which are highly volatile across time, yet with averages over time that are dynamically stable. The Lane Choice game and the Biased Lane Choice games show distinct system behaviors with learning-only, population dynamics-only and combined mechanisms...specifically, the population trajectories are drastically different within each game across these mechanisms.

The last lesson should not be understated: this indicates that as games become more complicated, mechanisms that are equivalent in simpler contexts diverge in the behavior they produce. The agent rules are identical, but a more complicated environment can separate mechanisms that simpler environments cannot. This has major implications for a field in which simple 2×2 games dominate the literature, and there are a profusion of distinct mechanisms that produce similar results.

Expanding the game library to the full range of 2×2 strategic contexts relevant to prosociality (which I argue above may be all of them) is a first step and I have already discussed in chapter I how mechanisms such as punishment and reciprocal altruism fail in the broader context. But this too may be insufficient to adequately understand the role of different mechanisms in bringing social groups together in sustainable, competitive arrangements. Even more complicated environments are clearly relevant to the evolution of prosociality, and the 3×3 games explored here demonstrate that these social problems can be combined into a single environment. Further exploration along these lines, with preferential detachment and other mechanisms as well, ought be explored. Whatever mechanisms fostered the evolution of prosocial behavior and social arrangements, they clearly had to solve complicated and changing sets of social problems.

[8] If the utility that an A-type with one connection (even a connection to a B-type) receives exceeds the utility that a B-type saturated with other B-types receives ($\psi(Aab) : a + b = 1 > \psi(B0K)$), then A-types would dominate the population. Determining the cardinal utility values that operate as thresholds and boundaries is beyond the scope of the current work, but is identified for future explorations of the cardinal utility space.

3.10.5 Efficiency and Intervention

Though the base case preferential detachment model produces highly prosocial outcomes across the range of strategic contexts, it does not produce the ideal/optimal social arrangement in every case. For example, cooperators in the Prisoners' Dilemma and contributers in the Stag Hunt get caught in heterogeneous groups that, within the rules of preferential detachment, are completely stable. If one were to bend these rules in specific ways, one can engineer an improved social arrangement. The present model aims to be maximally general, and the set of interventions that are possible/reasonable/acceptable depends on the substantive interpretations of the agents. Despite that, some specific interventions are worth bringing up to reveal the sufficiency of small changes to even further enhance prosociality in the appropriate situations.

The aforementioned cases of Prisoners' Dilemma and Stag Hunt offer some simple alterations with big beneficial impact. Almost any sort of mixing that 1) lasts through the segregation phase and 2) then peters out would have the desired effect – it just needs to free up some connections to existing cooperators in the quasi-clique and breakup the hold that B-types have on those A-types. There are many ways to achieve these ends. A simple mutation rule that randomly switches agent types at a rate that anneals over time would suffice for Stag Hunt. Or we could specify K as a *soft limit* on the number of neighbors such that there exists some probability of temporarily exceeding this number. Or there could be random global rewiring, or specific "outreach" agents that connect (either randomly or purposefully) to agents in network components other than their own. Each of these sufficient mechanisms has a plausible interpretation relevant to the evolution of prosociality. Some of these appear in the literature already, and there are certainly addition plausible ones.

In the Commensal game, a sub-optimal outcome is reached because the B-types are using up available connections with other B-types, and these are sought by A-types who are then frustrated (i.e., they disconnect from less desirable neighbors but are unable to connect to more desirable ones). Because the B-types are just as satisfied with A- and B-type neighbors, the outcome produced by preferential detachment is Pareto inferior to another system arrangement reachable through a monotonicly utility-improving rewiring. The conclusion here is that for situations in which the Commensal game is an appropriate representation, the mechanism of preferential detachment fails to fully utilize the social network's potential. In such a situation the indifference of B-types must be tipped toward preferring A-types; this converts the game into an asymmetrical version of Lichen, which then produces the perfect disassortative outcome. This could be achieved through an institutional intervention, environmental manipulation, as side-payments by A-types to free up more connections to B-types, or other such maneuvers depending on the problem domain.

3.10.6 Network Properties

Existing research using fixed interaction structures choose networks expressing properties with mathematical interest, empirical support, or convenience [Abramson and Kuperman 2001, Sobkowicz 2003, Assenza et al. 2008, Arapaki 2009, Bilancini and Boncinelli 2009, Perc 2009, Roca et al. 2009, Yang et al. 2009]. The networks generated by preferential detachment, however, do not typically match the network structures in the literature: rings, lattices (regular k-graphs), complete graphs, scale-free networks, random graphs, or networks with a fat-tailed degree distribution (aka power-law networks).

The maximum degree assumption is based on research regarding humans' ability to maintain close and loose social ties [Dunbar 1998] (see 1.4). To keep things simple I assumed that all agents' maximum degree is the same. For all values $K = 3$ to 20 the agents all fill their maximum allotment of neighbors (if possible) because the preferences are set up that way. I could instead make the maximum degree values heterogeneous (say) according to a power law distribution and produce that degree distribution among the agents, but that assumption is not directly supported by evidence...it's just an ad hoc way to to get my network to match some existing structure. But even then, a comparison between the networks here and the standard list is not possible because the resulting network arrangements depend on the strategic context. As indicated in the future work section below, the network structures that *are* generated deserve greater attention, but the lack of a close match between the generated networks here and the standardly analyzed networks has two major implications.

First, there isn't just a single network structure produced by preferential detachment. Each context yields distinct detachment behavior (e.g., assortative vs disassortative) and clearly these generate distinct network structures. Within a particular strategic context there are regularly appearing patterns, but the run-by-run variation still admits to quite different interaction networks (e.g., in the number of stable heterogeneous groups). Furthermore, the network measures that must be compared would have to include the detail that agents are typed.

The second major implication points to a need for more detailed interaction data. The networks in preferential detachment represent actual interactions, not just person-to-person familiarity. What the results imply is that we should see *dis-*

tinct social interaction patterns in distinct problem environments. This belies the oversimplified notion that people just have their social networks and those networks have particular properties. Whatever the agents are conceived to be, they more likely have shifting and problem specific sets of interaction partners, and the empirical data (and even methods) capturing such fluid social structures are few (though see [Flack and de Waal 2000]). A by-product is that a result demonstrating that cooperation can spread on a network with some property P is of little value to us here when we think that property P is unlikely to be representative of interaction structures along the evolutionary path to prosocial social arrangements.

3.10.7 Mechanism Comparison and Complementarity

Often a proposed mechanism is only demonstrated to work on a single game and is not generalizable; what works for Prisoners' Dilemma may not work for the Battle of the Sexes. Even more limiting is that for some mechanisms in the literature, their application to other games is truly nonsensical. Though detachment looks like punishment in the Prisoners' Dilemma game, when an *Opera* (A-type) agent disconnects from a *Boxing* (B-type) agent in the Battle of the Sexes, to consider this action as punishment is very unnatural. Punishment is a specific interpretation of behavior in cooperation games, but the preferential detachment mechanism applies unaltered to cooperation, coordination, mis-coordination, zero-sum, and all other type of strategic situations. It doesn't satisfy definitions of punishment mechanisms as they are described [Binmore 1998, Boyd et al. 2003, Sripada 2005, Mendes and Aguirre 2009, Nikiforakis 2010]. And furthermore, even if the resulting *behavior* is considered punishment in the Prisoners' Dilemma context, it needn't be, and doing so removes all the thick considerations (such as intent and acting fairly) that are loaded on the term 'punishment'.

But even this detachment-as-punishment-behavior version is not a supported interpretation in this model. Defectors also detach defectors if they are connected to any cooperators. Defectors are detached because they yield less value than cooperators and not because agents are lowering the value of defecting [Axelrod 1984]. The prosocial outcome of cooperator dominance is achieved (when population dynamics are present) because preferential attachment fosters assortative mixing among the cooperative agents. Assortative mixing is one of two ways that Rapoport proved suffices to achieve cooperation in the Prisoners' Dilemma [Rapoport and Chammah 1970]; the other way is to change the payoffs and **that** is how punishment achieves the cooperative outcome in the models in which it is shown to be sufficient. Punishment is certainly not a necessary mechanism for cooperation, and in fact requires much more restrictive assumptions and greater agent capabilities than many assortative mixing methods do.

Kin selection shows how small groups of related individuals would benefit fitness-wise by helping others. Reciprocal altruism is also a viable mechanism for achieving cooperation in small groups. But these two mechanisms don't scale beyond a few dozen individuals unless further assumptions are made about individuals' ability to keep track of and/or communicate past actions (memory, trust, language, etc.). Preferential detachment, on the other hand, starts working well at about 20 individuals and is increasingly likely to produce prosocial outcomes as the population climbs. Thankfully, we are not in a mechanism competition, and these plausible small-scale mechanisms may be complementary mechanisms for first generating a population of cooperators and defectors to interact. So in an application to a particular creature (like humans, proto-humans, or primates) we may benefit from generating initial configurations using simple rules based on these mechanisms rather than the implausible unconnected initial setup used here.

Another consideration therefore becomes whether preferential detachment acts as a good theory for post-domestication human societies. As hunter-gatherers people and protohumans could easily wander to different groups and find new interaction partners. Once agriculture became prominent, a people's geographical location to some degree determined who they could interact with. If one is forced by location to share or not share a common grazing ground with physical neighbors, then preferential detachment is no longer viable (taken as a literal mechanism for selecting neighbors). This leads to a necessity to regulate interaction in a more abstract form. Even in these land-pegged practices there is still some mobility and some choice among interaction partners, and preferential detachment can be seen as an institutional framework for participation and/or inclusion in social organizations. It would also be interesting to see whether we can identify failures of achieving prosocial outcomes and tie those to circumstances for which the assumptions of preferential detachment fail to obtain (e.g., land leasing practices, non-market goods allocations, political representation).

Keep in mind that the motivator for preferential detachment, and the subject of the next chapter, is the evolution of morality. And for this subject the above-mentioned possible limitation to nomadic or quasi-nomadic medium-scale bands is quite acceptable. The bounds of its applicability in modern societies (e.g., to institutional problems) will depend on the particulars of the system being modeled/interpreted as a preferential detachment model. The question of applicability and generality is no more or less worrisome here than any other model, and though it impacts the value and explanatory power of the model, these questions are not obvious detractors for preferential detachment vis-à-vis other theories of the evolution of prosociality.

3.10.7.1 Assortativity without Homophily

Throughout the text I refer to the property of similar agents being connected as "assortativity" in accordance with the standard definition. There is a convention in social network literature to use 'assortative' and 'homophilic' interchangeably, but this usage fails to appreciate that there are mechanisms for generating assortativity that do not imply or require the nodes to connect through a *pursuit* of sameness [Newman 2003]. I am endorsing a convention by which "assortativity" or "disassortativity" are properties of network arrangements, while "homophily" is a property of network generating processes; i.e. homophily is a particular explanation of *why* nodes with similar properties are more likely to be connected. Homophily, then, becomes a special case of *preferential attachment* in which nodes prefer other nodes like themselves.

We have already noted (in 1.9.1) that assortatively mixed arrangements coincide with the prosocial arrangements in Cooperative, Coordinative, and Contributive games.[9] These games represent the bulk of previous work in prosociality. Mechanisms such as kinship and markers in which individuals actively pursue kin and/or other individuals with similar markers are directly homophilic, and generate the prosocial assortative outcome this way.

Static network models (including grids and rings) in which agents change their action (behavior, type) also require assortativity to achieve the prosocial outcome in the range of games analyzed. Regardless of whether the action change happens through following a strategy, imitating neighbors, receiving a communication, or whatever else, it cannot be the case that the action is changed in order to associate with agents similar to itself...it cannot be homophily because nodes do not change their neighbors.

Preferential detachment operates on dynamic social networks, but it is also not homophilic. Attachment is random, and who detaches whom depends on the strategic context. In those contexts in which an agent prefers to have its own type as neighbors, the process is still not homophilic – rather it's heterophobic. Under the constraints of ordinal preferences relationships, a *preference* toward the same type is equivalent to a preference against a different type.[10] However, generating social networks by

[9] This is also true for the Collaborative (Lane Choice) games explored here and the Ultimatum Game excluded from the current version.

[10] This is especially clear when we consider the preference orderings for 2 × 2 games. When the

preferentially attaching A-types produces distinct structures from preferentially detaching B-types. The import of this distinction is to foster a rethinking of social network theorists' explanation of the assortatively mixed arrangements detected in many kinds of networks along several different features.

3.10.8 Evolutionary Approach

To more fully understand the evolution of prosociality with formal models, the models must include the spectrum of features of evolutionary processes: endogenously generated communities, intercommunity competition, intracommunity structural changes, and individual selection pressures. Each feature can be explored independently and offered as a partial explanation to be combined with other models to fill in the full evolutionary explanation [Boyd and Richerson 2005]. This modular approach has its benefits; however, a model that can cover the full spectrum of evolutionary processes is a more powerful explanatory tool, especially if such a model is as simple and general as an overlapping partially explanatory model. Creating such a model is the aim of this project, and the results demonstrate a preliminary success in fulfilling many of these desiderata.

In light of the features of evolutionary models just identified, the cellular automata and fixed network models should be interpreted as intracommunity models; i.e., changes in behavior among individuals within a community who do not have the power to alter their community members or structure. The results from such models therefore address adaptiveness of cooperative behavior, but not a form of adaptability directly relevant to questions of the *evolution* of prosocial behavior because they assume as fixed some features of the problem space that are unreasonable to consider static on evolutionary timescales. This is not to say that preferential detachment (with population dynamics) succeeds in capturing all the necessary and sufficient features to make it a thoroughly evolutionary approach...it's not even clear yet what those features are. But one thing it does succeed in is helping to demonstrate that as we move closer to an evolutionary approach, prosociality may reveal itself to be *more* reliably achieved and maintained.

I have made the point in chapter I that evolutionary fitness is measured in behaviors rather than the mental (or other) machinery that generates those behaviors. That behavior-based claim comes off stronger than I really want (or need) to make

relationship $A \succ B$ holds for A-type agents it means that they detach B-types when A-types are also present. This can be interpreted equivalently as having a preference toward A-types (compared to B-types) or a preference against B-types (compared to A-types).

here. Fitness can only be determined in a larger context including the behaviors of other individuals, the environment, etc. The point is that the individuals' component hinges on actual behaviors rather than actual or potential thoughts and feelings...or the contingent (but not actualized) behaviors they may support.

This is analogous to the phenotype vs genotype line of reasoning. If one is interested in the persistence of particular genes, proteins, or other specifics of the biological construction an animals happens to actually have, then one needs to push the analysis down to those features and relate them to the phenotype to understand how they affect fitness. These microstructures bound the expression of phenotypes and how they can change, but if one is interested in which traits persist, then one must examine the dynamics of a population exhibiting those traits. The microstructural details are just contingent facts at that level, and if one were to introduce an individual that was phenotypically identical (regardless of its genotype) then its fitness would also be identical.

Among the reasons that we care about this distinction are 1) the tension between homo economicus vs simple rule-following agents and 2) the focus on the evolutionary psychology/neurology of prosociality. The latter topic will be a focal point in the next chapter because it ties closely with the topics of emotion, cognition, and morality. The former point, I want to emphasize, is a methodological point rather than a substantive one. As discussed in section 1.6.5, the rule-based behavior of the agents in these models can just as well be captured by agents optimizing a preference function.[11] The motivation to have done that is to underline the fact that the internal mechanism driving the agents is not a salient topic of the current project. Any mechanism that produces the same agent behavior in all the same instances, as specified by the theory of preferential detachment, is equivalent for the purposes of exploring system behavior.

[11] Note, however, that the full context including network effects, randomized connections, etc. cannot be *conveniently* represented in a form that allows it to be solved for steady states. My model does not violate rational choice assumptions in the sense discussed in section 1.6.5, it could be made into a game theoretic model in which the choices made are over links rather than over action type. Assuming action type is fixed (as I do unless learning is operating) it should be possible to identify sets of Nash equilibria based on best response dynamics by solving a set of equations involving link formation choices. The question is how to represent link formation benefits and costs (preferences) so that the analytical model accurately represents the theory. That translation task is tractable, but the resulting formulation may not be computable. Furthermore, we are interested in the trajectories of system behavior as well as the outcomes, so the benefits may be outweighed by other considerations.

3.10.9 Future Work for Preferential Detachment

The analysis performed in the current work focuses on demonstrating preferential detachment's sufficiency in generating prosocial arrangements. Deeper analysis can reveal other features of the model which may be of interest. Some of those possibilities are discussed below.

3.10.9.1 Network Structure

Preferential detachment uses a dynamic social network for agent interaction, which is in contradistinction to the vast majority of networked agent models of prosociality [Nowak and May 1992, Sobkowicz 2003, Grim et al. 2006, Arapaki 2009, Bergin and Bernhardt 2009, Yang et al. 2009]. Rather than having a fixed network structure with known (set) structural properties, the network structure here changes dramatically across time and is only constrained by the maximum degree of the agents and the generative process of preferential detachment. Analyzing the generated network structure across the distinct strategic contexts, along with any correlation of structure and outcome, may yield additional insights into the fixed structures of other models thus widening the implications of results there. This would allow us to test the importance of specific structures for the achievement of prosocial outcomes.

There are models in which the agent interaction structure changes, and this technique is becoming increasingly popular [Ashlocka et al. 1996, Skyrms and Pemantle 2000, Tesfatsion 2001, Sobkowicz 2004, Zimmermann and Eguíluz 2005, Santos et al. 2006, Szolnoki and Perc 2009, Tomassini et al. 2010, Wu et al. 2010]. Teasing out the relationship between structure, structural dynamics, and prosociality may help focus the development of dynamic interaction models. We may also want to address the connection of various initial network structures to prosociality (rather than the unconnected agents used here). Doing so, however, may require inventing new representational and analytical methods capable of capturing and comparing model dynamics.

3.10.9.2 Parameters and Heterogeneity

The parameters swept here, population and maximum degree, were both done only for the Prisoners' Dilemma and combined effects were not simulated. There are reasons (discussed above) that many of the combined effects can be extrapolated from the experiments performed, but not completely. So some further experimentation is required to fully generalize the results. Sweeps for some of the other strategic contexts seems the most interesting next step.[12] Filling in the range of swept parameters (e.g., between $K = 10$ and 15) may help identify phase transitions in some of the system behaviors identified earlier. Of particular interest is a further exploration of the effects of learning and population dynamics across other cardinal utility value assignments.

In addition to the expansion of the search space just presented, introducing heterogeneity in the parameters is an important direction of exploration: uneven initial populations, diverse maximum degree values, and multiple games played simultaneously. Particularly promising are experiments in which the maximum degree and/or game(s) played adapt to the actions of agents. Such an extension would then allow us to determine whether preferential detachment could self-organize not only prosocial agent arrangements, but ones reflecting network structures and strategic contexts observed in actual animal (including human) societies. Doing so requires specifying further rules for choosing games and increasing/decreasing the number of maximum edges, and hence reduces the generality of the theory, but may provide more useful and interesting results in this more specific domain.

3.10.9.3 Applications

Though the next chapter presents implications of preferential detachment for morality and its evolution (which is the primary motivator for the formal work), there are more direct applications of preferential detachment theory to more concrete social problems. The technique can be adapted to address questions in group formation among heterogeneous collections of agents (for which the agents may be cells, people, social collections, organizations, institutions, nations, etc.). For any of these problems, the primary difficulty is in mapping the strategic aspects of the situation into the proper payoff matrix (or set of payoff matrices). Many of these will require more than two types of agents (i.e., a payoff matrix larger than 2×2), and/or different preference rankings over configurations, and/or changing preferences – all of which are easily implementable within the preferential detachment theory presented in chapter I

[12] The analysis of the results demonstrate that altering the parameters for Lichen or Matching Pennies would produce qualitatively identical results. Again, many of the extended results could be extrapolated from expanded sweeps of the Prisoners' Dilemma and/or other games, hence shifting the burden from simulation to analysis. But some parameter-game combinations are sui generis and require independent simulation.

3.10.10 Conclusions Summary

The primary result of these experiments is that preferential detachment suffices to achieve high levels of prosociality in all social problems representable as a 2×2 payoff matrix. This means that it can be added to the list of mechanisms that produce and sustain cooperation in the Prisoners' Dilemma and Hawk and Dove games, contribution in the Stag Hunt game, and coordination in the Battle of the Sexes and Coordination game. Preferential detachment goes beyond this by generating and sustaining beneficial social arrangements in the remaining payoff structures as well.

These extended results are important because cooperative behavior did not evolve in isolation of other forms of social problems. Mechanisms that only work, or even only sensibly apply, for a proper subset of social problems only provide a partial explanation to how prosociality evolved. Furthermore, demonstrating that a single, simple mechanism suffices in this broader scope presents advantages in both modeling and explanatory power over positing a collection of problem-specific mechanisms. Even more powerful, preferential detachment can be extended without alteration to more complicated games (such as the 3×3 collaborative games included and the Ultimatum Game[13]) which bolsters its plausibility, usefulness, and appeal as a generative and explanatory mechanism.

The results of including imitative learning and population dynamics, though preliminary, are also encouraging. Both additional mechanisms are observed in creatures we naturally wish to consider as the subjects of our prosociality research, and so the result that these mechanisms (separately or in conjunction) tend to enhance prosociality across strategic contexts lends further support for preferential detachment as a viable mechanism for solving social problems.

Achieving these results with a dynamic interaction network is an important innovation for research in this field. Neither the static nor random interaction structures characteristic of models deployed to explain the evolution of prosociality appropriately capture the changing social arrangements likely produced by agent reactions to others' behavior. They also fail to simultaneously explore or explain the relationship among community structure, community competition, and agent success. Earlier models make simplifying assumptions to either ignore and/or hardwire these at least some of these considerations, and so this is another dimension in which predecessor

[13] The analysis of the application of preferential detachment to the Ultimatum Game was left out of the current work, but when encoded as a discrete payoff space (e.g., [Binmore et al. 1995, Nowak et al. 2000] the results of both the Markov model and agent-based model produce high/perfect levels of prosociality.

models offer only partial explanations for the questions of evolution central to our understanding of prosocial behavior.

Despite these advantages, the current work should still be seen as tentative foray into these complicated and difficult issues. Though the mechanism of preferential detachment is itself simple, the modeling environment makes the system behavior challenging to analyze and understand. Methodology for agent-based models is still in its infancy, and there is much left to explore in analyzing dynamic network structures, tracking agent behavior, testing sensitivity of parameters, and representing model behavior and outcomes. And the model itself deserves further investigation into various forms of heterogeneity, more complicated strategic contexts, and finer grained explorations of imitative learning and population dynamics.

CHAPTER IV

Evolution of Prosociality and the Moral Experience

> If you prick us, do we not bleed? if you tickle us, do we not laugh? if you poison us, do we not die? and if you wrong us, shall we not revenge? If we are like you in the rest, we will resemble you in that.
> –William Shakespeare, *Merchant of Venice*

In this chapter I build a bridge from the evolution of prosocial behaviors to various psychological and sociological phenomena associated with morality. The thesis is that moral attitudes, which admit to a variety of forms and expressions (urges, perceptions, emotions, intuitions, affect, judgments, etc.), are experiences with a particular characteristic: these attitudes correlate with behaviors which are necessary for achieving and sustaining a population of individuals who similarly behave appropriately for the perpetuation of that population. If we accept that moral attitudes are the result of (or significantly shaped by) an evolutionary process, then a better understanding of the behaviors required for sustaining prosociality can provide insights into important features of our moral attitudes: which behaviors elicit them, how plastic they are, how universal they are, and how reliable they are as guides.

An important implication of the behavior-based, adaptive-social-arrangement-grounded account of moral *attitudes* is that they are only contingently and coincidently adaptive: they happen to correlate with specific behaviors, which are adaptive in the evolutionary sense. Because the evolutionary process is constrained by tweaking and selecting biological structures, it produces a continuity of *mechanisms* and a series of incremental changes in behavior through time. This near continuity in nature, combined with the universality of critical social problems, implies that the drivers of moral attitudes operate beyond the domain of humans and their societies. The evolutionary lens can therefore help explain the manifold of moral experiences of humans as well as other animals.

Though explaining the evolutionary origins of specific moral behaviors, attitudes, beliefs, and social structures has become a commonplace research agenda, these typically take the form of providing plausible post hoc explanations for particular moral (or morally relevant) phenomena. The project here is to make inferences from general features of evolutionary processes over sustainably adaptive behaviors (and the collective structures they generate) to our moral experiences (and the social phenomena that elicit them). This is distinct from showing how, or explaining why, certain forms of morality, or some set of normative propositions, are (or had been) adaptive. The inferences reveal which features we would expect moral experiences to have if they resulted from the evolution of prosocial behavior. Then we explore evidence to test the strength of the connection between existing features of our moral experience and implications of the theory that they are derived from the mechanisms driving adapted prosocial behaviors.

None of the substance in this chapter depends on the specific results of previous chapters. The identification of prosociality with actual behavioral repertoires in the previous chapters is still in play here, and it builds into aspects of evolvability and moral experience that are key to this chapter. This construction is covered again here. Additionally, I use my preferential detachment model for examples in many cases because those results demonstrate the essential features of interest here (which are lacking, or less clear, in other models). Insights from other models, experimental data, observational research, and philosophical insights will naturally be called upon at many stages in the argument flow.

4.1 Evolution of Prosocial Behavior

This section defines prosociality and draws out the details of how prosocial behavior evolves. Like most evolutionary explanations, there is a quasi-cyclical nature to the explanandum and explanans: a behavioral repertoire that de facto enhances the relative reproduction rate of practicing individuals becomes more prevalent precisely because that behavioral repertoire de facto enhances their relative reproduction rate. We can, as biologists and computer scientists typically do, use *fitness* as a placeholder term for a relative change in prevalence: this can occur through biological reproduction, imitation, group growth and splitting, and other processes discussed below. It is important to remember throughout that "fitness" means nothing beyond these differential rates of replication and survival.

A behavior can be considered *adapted* to a context on an individual basis; meaning that individuals enacting that behavior, or the behavior itself, will gain prevalence in the current context. The context of an individual's behavior, however, includes the environment, the social problem being faced with, and the behaviors of all its interaction partners. Therefore the fitness of an individual behavior likewise depends on the behaviors of the individuals with which it interacts – its group. The fitness of a group (i.e., the ability of the individuals in that group to collectively sustain themselves, produce new groups, and out-compete other groups) therefore depends on how adaptive the entire behavioral repertoire is in the current context, not just on there being many high-fitness individuals.

The position taken here (and detailed below) is that *prosociality* is the property of behavioral repertoires such that groups of individuals enacting those behaviors sustain the generated social arrangement (the set of individuals' behaviors and their interaction structure).[1] Sustenance implies both the ability to maintain a population through shifting environments and internal changes as well as the ability to out-compete other groups (in terms of fitness). Through this evolutionary lens, we can investigate general properties of systems capable of sustaining themselves, and in section 4.2 relate this to the spectrum of moral experiences.

4.1.1 The Primacy of Behavior

The evolutionary account begins with explaining why behaviors – rather than attitudes, physical traits, abilities, mechanisms, genes, individuals, groups, etc.– are the primary unit for selection and adaptation in this approach. *Behavior* here is the broad description of what an individual does. It includes the action taken in a game[2] (the type in preferential detachment), the interactions engaged in, the imitation of another individual, and any other thing that actually happens. Deciding not to

[1] Though the individuals need not be the kind of objects typically considered social (e.g. molecules), because the property applies to the behaviors of collections of individuals (including all their interactions) rather than mere aggregates of individuals, the closest idea to the sort of arrangements the behaviors produce is a society. For the subject matter we care about with respect to the moral experience (e.g. mammals), considering these interacting collectives as societies is unproblematic; for the rest we can consider it metaphorical (along the lines of Minsky's "Society of Mind").

[2] There is some jargon incompatibility in this paragraph. In game theory the row or column played in a game is called the "action". Philosophy has a tradition of reserving "action" for intentional performances of behavior [Davidson 1980]. Because intentionality does not play a role in the current project, it will be convenient to use the conventional game theoretic usage.

interact with a current partner is not itself a behavior, nor is the resulting lack of interaction; but the interactions actually engaged in are behaviors, so detachment counts as a *change in behavior* rather than a behavior.

Individuals' behaviors may be contingent upon interaction structure, environmental features, the behaviors of others, performance, memory, signals, etc. It may be the case that an individual performs action A in situation X, and action B in situation Y, and that both situations arise. So an individual's behavior need not be a single, constant set of interaction partners, actions, etc. – an individual's behavior is all the things it actually does. It matters that A is performed in situation X and not Y, so in capturing behavior we also need to include its context. Of course, it could also be the case that situation Y never occurs, and hence what an individual would do in that situation is not part of its behavior. This marks the difference between capturing behavior (which is retrospective) and capturing behavioral tendencies (which is contingent).

Some mechanism provides the contextual contingency (e.g. an exhaustive map from contexts to particular behaviors, or a general rule like preferential detachment, or game strategies[3]), but that mechanism is not directly selected for or against. It is the performance of A and B behaviors in the appropriate contexts that is selected for. All mechanisms that have generated the same behaviors in the same contexts are necessarily equally fit because the replication of an individual (or the behavior itself through imitation) is also a behavior. The further point is that a mechanism that *would* do better in a situation that never actually arises, but is equivalent to all other mechanisms in all realized situations, does not increase in prevalence vis-à-vis the other mechanisms.

4.1.1.1 Individual Interest

I have already described that by *fitness* I mean the standard evolutionary placeholder for the relative success in replication. An individual's *interest* is determined by what is in line with that individual's behavioral tendencies. In this way we can refer

[3] In game theory, a strategy is a mapping from each possible input an agent may face into a behavior. One common way this is done in the prosociality literature for 2 × 2 games is to specify what behavior to enact on the opening move, which behavior to enact next time if action A is observed, and which behavior to enact next time if action B is observed [Axelrod 1984, Grim 1995]. As should be clear, this is a very different form of contingency than preferential detachment because 1) it ranges over actions rather than neighbors, 2) it is sensitive to each interaction partner rather than the set of partners, and 3) it is not sensitive to payoffs/context. Of course, this is not the only way to specify strategies, but it is common enough to mention that specifying contingent behavior in this way is fully compatible with everything here.

to the interests of molecules, insects, rats, humans, and even groups of such individuals using the same terminology regardless of the mechanism. This usage of 'interests' follows a precedent of biologists referring to animal behavior and to "reproductive interests" of species [Alexander 1987]. Alexander further points out that "nonhuman organisms live out their lives serving their interests without knowing, in the human sense, what those interests are." Humans also often do not have full epistemic access to their own interests, thus people behave contrary to their self-considered preferences or without being cognizant of the drivers of their own behavior on a regular basis.

Behavioral tendencies capture motivations, preferences, urges, reactions, and whatever other social, psychological, biological, chemical, or physical, process is considered the driving mechanism. Here we are merely identifying that whatever the mechanisms, individuals behave according to their interests. These interests can be captured by a system of contingent rules (as is done for preferential detachment and many other agent-based models) or through revealed preference rankings (see 1.6.5), or a cardinal utility function. So one can consider individuals as consciously optimizing their interests à la rational choice theory, but that is not essential.

We may talk of individuals "wanting" to perform some behavior or "preferring" some outcome or local configuration, and these can be considered loose and/or metaphorical uses of these terms: wanting and preferring X just means that X is according to the individual's interests. Behaving to serve one's interest therefore does not require that individuals are consciously aware of their interests, experience a sense of satisfaction when their interests are served, or that they are cognizant of hypothetical configurations that better serve their interests. So, for example, in the agent-based model using just preferential detachment in the Prisoners' Dilemma, a cooperator connected to both cooperators and defectors will serve its interests by detaching a defector – the behavior rule captures this. A cooperator with all defector or all cooperator neighbors serves its interests by (or "prefers") making a new connection (unless its maximum degree is fully utilized). If such an agent were unable to make a new connection (e.g., there are none available), then it would be *frustrated* with respect to its interests. It would also be frustrated as long as it had both cooperator and defector neighbors (remember that in the model the agents can only detach one at a time).

Even a long-term plan that requires sacrificing short-term gains for long-term benefits will be revealed as behavior over time, and so the individual's interest was served by the entire stream of behaviors engaged in. Or, for example, a rule to cooperate four times and then defect also plays out through behavior. If individuals are considered capable of creating intricate contingency plans, setting expectations, and deceiving others, then past behavior may not be a good indicator of future behavior, but the evolutionary process works on actual (and therefore past) behaviors, not plans or unrealized contingencies. Evolutionary explanations of past events and existing features are much stronger than evolutionary predictions, a aspect of this account that will be explored more thoroughly in section 4.1.4.4 below.

4.1.1.2 Individual Selection

As noted in 1.7, interest and fitness are distinct features in general: the former reflects individual behavioral tendencies and the latter is the contextually determined increase in prevalence. For example, in preferential detachment an agent will detach a defector when it has at least one cooperator neighbor. However, keeping defector neighbors increases fitness (reproduction and learning are based on utility, and defector neighbors are worth one util), but they are detached (which indicates that having those neighbors is against a cooperator's interest). Some models equivocate interest (preference, utility, etc.) with fitness, but doing so precludes discovering the importance of aligning (or failing to align) group fitness and individual interest (discussed in 4.1.4.3).

Individuals behave according to their interest; however, it is the fitness of an individual's behavior that determines whether that behavior persists. Behaviors can be eliminated because the enacting individuals are themselves eliminated or because the individuals change behavior tendencies through imitation or other means (recall that behavior includes what an agent did in each context[4]). Individuals whose interests align with those behaviors that, in the prevailing context, foster a higher rate of replication are the ones selected for increased prevalence.

Even if some interest-aligned behavior increases in prevalence in the prevailing context, it might do worse through the wider lens of the evolutionary process. An example of this phenomenon is defectors' exploitation of cooperators in the Prisoners' Dilemma; once a cluster of cooperators forms, they squeeze out defectors and dominate the population (see 3.9.1). As environments change, groups change composition, groups compete, and new behaviors come into existence, the fitness of any

[4] An individual may seem to change behavior without actually changing its behavioral tendencies. In the previous example giving the rule to cooperate four times and then defect, to somebody recording actions it may seem to be a change in behavior. But if we consider history as part of context (which we should), then the context for triggering the defection includes that the individual cooperated four times. Capturing behavior generation with rules, equations, patterns, etc. brings in indeterminacy (too many ways to capture the same behavior and no way to choose among them), but actual behavior is just the complete description of what has actually occurred.

particular behavior can fluctuate. Fitness must therefore be contextualized in time as well, because the reproduction rate changes across time. In consideration of this, actual all-things-considered fitness can only be determined historically – behaviors that de facto survived and became more prevalent than competing behaviors.[5] For this evolutionary account, we are therefore concerned with the full series of contextual changes, and the behaviors adapted to them, that resulted in current arrangements.

4.1.1.3 Behavior-Generating Mechanisms

For any behavior, there is some mechanism that produces it. One can imagine an individual that has a singular behavior that is always active, not contingent, and completely insensitive; many formal models contain agents like this. But even individuals as simple as molecules behave differently in different contexts (e.g., in different temperatures or in the presence of different mixes of other molecules). In 3.10.8 I draw the parallel between genotypes/phenotypes and mechanisms/behaviors, respectively. Behaviors are actually selected for via their relative ability to persist and replicate, but it is typically the behavior-generating mechanism that gets passed on (either in whole are a part thereof). Biological reproduction, at least, typically works this way because the relevant mechanisms are hereditary. When behaviors such as "Do X in context Y" spread via imitation, that contingency can be incorporated into individuals' behavior-generating mechanisms in distinct ways and still produce appropriate behavior.

Let's take another example from preferential detachment to home in on a particular aspect of the mechanisms important to individual selection (other points will be made below). We start with a Coordination category game under the assumptions of chapter I, so over time the agents change their interaction structure to have only like-typed neighbors. Now imagine that the strategic context switches to Lichen. All the agents are assortatively mixed, which is the worst-case anti-social outcome for Lichen. In the model there is no mechanism capable of shifting behavior because agents only use information from their current local interaction partners. However, if some individuals had a method by which those agents could detect the game change and then adjust their behavior appropriately, then those individuals would be more fit, and this higher level of contingency would become more prominent.[6] It is hav-

[5] In this light, all behaviors that persist in some extant behavioral tendency have higher fitness than all extinct behaviors.

[6] What would become more prominent in this case is actually the ability to specialize appropriately when the problem is like Lichen, and coordinate appropriately when the problem is like Coordination Game. The switching behavior is only more fit if the problem keeps changing like that. But my

ing a mechanism that generates behavior that is contingent in the appropriate ways that allows an individual to behave adaptively in shifting problem contexts, but it is still the actual behavior in realized situations that determines behavioral (and hence individual) fitness.

4.1.2 Prosociality in Collectives

Groups are composed of individuals, the social arrangement of a group is the set of individuals' behaviors and interactions, and a group's behavior (including movement, solving social problems, replication, and imitation) is an aggregate of individuals' behavior; but the fitness of the group is not a direct aggregate of the fitness of its members even though group fitness does supervene on the individual fitness levels. The meaning of fitness is the same for individuals and groups: the ability to increase in prevalence. Because individuals' behaviors affect other individuals, there is a complex interplay among the individual behaviors that generates group fitness. The behaviors of a group of individuals considered together is called a *behavioral repertoire*, and the degree to which a behavioral repertoire results in greater group fitness is its *prosociality*.

Fitness can be divided into two components: endogenous and exogenous. Endogenous fitness is the ability of a unit to stay alive and reproduce – endogenous aspects that affect unit prominence. Exogenous fitness captures aspects that affect unit prominence through competition, environmental change, accidents, etc. The next section addresses exogenous fitness of groups, while this section tackles the endogenous aspects. Boyd and Richerson identify the two primary means that individual behaviors effect endogenous group fitness as:

> The group beneficial trait can increase the productivity of the group so that it produces more emigrants, called "differential proliferation" or it can reduce the extinction rate of the group, called "differential extinction".[Boyd and Richerson 2005]

However, it is not the case that a particular individual's behaviors can be prosocial in isolation, and there is no specific behavior that is always necessarily prosocial. It is only the set of collectively adapted behaviors that bear prosociality properties. The individual behaviors (actions and interactions) involved in this repertoire are only derivatively prosocial. Each individual's behavior plays an insufficient but necessary part in a behavioral repertoire that is unnecessary but sufficient for achieving

point here is the general one that meta-contingencies for changing the behavioral contingencies are adapted to contexts with changing features.

a certain level of prosociality.[7] Multiple repertoires may achieve the same level of prosociality; and for each repertoire, each individual behavior is an essential element of the repertoire.

Making each behavior strictly necessary may be too constricting. This implies that any change in the set of individuals (aside from immediate identical replacement) is also a change in the behavioral repertoire. If this were the case then no repertoire could be robust or adapted. This restriction can be tempered by categorizing behavior. The behaviors making up the repertoire can be proportions of the total population, and the values can be flexible or fuzzy to admit minor variations that still maintain the overall structure.[8] The ability of groups to handle behavioral niches and mistakes will depend on the size of the group, the fitness effects of the niche or mistake, and the starkness of the context. This is not a conceptual issue since the repertoire, as just defined, applies unambiguously to the preferential detachment and other computational models. There is a familiar difficulty in getting the specification of a behavioral repertoire to bend in the right way to cohere with our intuitions of sameness, but this difficultly will not trouble us in what follows.

4.1.2.1 Examples from Preferential Detachment

Let's again look at some examples from preferential detachment to nail down what these definitions imply. First consider the action indicated by agents being A-type. In the Prisoners' Dilemma this action is associated with cooperation and a repertoire filled with A-types is prosocial. If the situation were Lichen then the prosociality of being an A-type depends on the numbers of A-types and B-types in the population because the most prosocial behavioral repertoire involves equal numbers of both types. When the context is structured like a Coordination Game then the number of each type is irrelevant as long as each type interacts only with individuals of the same type. As already explained in chapter I, what it means to

[7] This clearly relates to Mackie's INUS conditions for causality: insufficient but non-redundant parts of a condition which is itself unnecessary but sufficient for the occurrence of the effect [Mackie 1988]. The connection could be used to open a discussion about whether the individuals' behaviors can be understood as a cause of the prosociality. There is a relationship, and it is being explained in this section, but I will not attempt any causation arguments in this text.

[8] More formally, we could categorize together all repertoires that have both the same fitness *and* that can be reached via a series of fitness-neutral unit changes.This is similar in idea to being on the same plateau of a neutral mutation network[Fontana 2003]. Remember that the repertoire may include a specification for different token behaviors at distinct population levels, so the repertoire itself can stay the same through population changes. But as a group grows, its fitness will likely change even though the behavioral repertoire may not, and since the repertoire is the focus, fitness-based categorization is inferior.

be prosocial is unified across all situations, but what satisfies prosociality depends on the context. To maintain prosociality across changing contexts requires the ability to adjust behavior to the situation at hand, but it's the resulting behavioral repertoire that will be prosocial (or not, to some degree).

So there isn't a particular behavior (agent type and interaction network) that is prosocial, but rather behavior which is appropriate given the context and the behaviors of one's neighbors. This really underscores the point that being prosocial depends necessarily on the interactions of the individuals (each enacting some behavior). Studies that only examine contexts in which prosociality is achieved through the domination of one type of behavior (such as in the Prisoners' Dilemma, Stag Hunt, or Ultimatum Game) will likely miss this nuance of prosociality (and as a result may make unsupported generalizations). As already noted, the actual social problems faced by creatures are rarely as simple as the 2×2 games used to highlight a specific strategic feature. Complicated social problems will admit to a range of niche behaviors that are collectively adaptive. But the answer to whether some niche contributes to prosociality is that it depends on the whole behavioral repertoire.

The realization that multiple, distinct existing behavioral repertoires may achieve equal levels of prosociality implies that the modal status of particular behaviors may vary as well. In some contexts being B-type is forbidden (Prisoners' Dilemma), in other contexts it may be obligatory (Lichen), and in others it is permissible (Coordination Game). Since different regions likely present different social problems, different behavioral repertoires will be prosocial in different regions, but each may be equally prosocial given the repertoire's adaptiveness to the local situation. In some cases there may not be any repertoire capable of sustaining a population in the area, it can go extinct because it fails to "solve" the local social problem (in some cases no solution will exist – no life is sustainable).

Given a social problem, some behavioral repertoires are going to be more or less fit than others. Or some behavioral repertoires that are adaptive may preclude exploratory behavior, thus creating a sustainable insular group that fails to spread and proliferate – it reaches an endogenous fitness limit. These endogenous fitness effects are the first order considerations for prosociality. The repertoire differences yield variable fitness level distributions for members of different collectives faced with the same social problem (e.g., autonomous groups in the same region). Distinct individual behavior fitness level distributions result in varying proliferation rates. When the region is bounded, competition results. So the various repertoires generate varying group-level fitness levels as well. These fitness variations among collectives is another vehicle for the group selection process discussed in the next section.

4.1.3 Intergroup Competition

I use the general category of *intergroup competition* to label any process through which autonomous groups become more or less prevalent with respect to other groups. Intergroup competition does not imply *direct* competition (such as conflict); e.g. the relative reproduction rates of group members under constraints discussed in the previous section generates a competition for prevalence – possibly without any group interaction. And there are processes by which a group's characteristics may become more or less prevalent without requiring the extinction of any group members; e.g., Boyd and Richerson's version of "cultural group selection" (group members imitating the behaviors of other groups) is such a process [Boyd and Richerson 2005]. I restrict my use of the term *group selection* for a process that is analogous to individual selection: "Extinction, coupled with recolonization by a single other group, means that groups become crude 'individuals' that reproduce their own group characteristics [Boyd and Richerson 2005]." This section completes the evolutionary account of prosociality by explaining the role of these various forms of intergroup competition in producing prosocial collectives.

4.1.3.1 Relative Reproductive Rates

The simplest form of competition is comparative replication speed. Endogenous group fitness determines how successful an autonomous group is in increasing its number of members. There can be multiple autonomous groups in the same resource- or geographically-constrained region that nonetheless do not interact directly. They do, however, interact indirectly through their impact on the mutual constraints. Let's say that a region can only sustain N_{max} individuals, and contains three groups: N_1, N_2, and N_3. The relative reproductive rates among these three groups determine the prevalence of each group's characteristic behavioral repertoire.

The relationship between population growth and prevalence over the medium- and long-term may not be as simple as "faster growing group's features become more prevalent." There may be threshold effects, such as a nonlinear fitness by population relationship. For example, the fitness cost of maintaining a population may increase until the population reaches a certain number, and then fall off. The fastest growing group (say N_1) may have the greatest proportion when N_{max} is first reached, but it could still be too low to breach the threshold. At that point it has actually become less fit through its early fast growth, and now N_2, and N_3 grow as N_1 shrinks. The fitness dynamics may churn through periods of dominance by one group followed by collapse and the rise of another, or some one group may eventually dominate over the threshold. Or the threshold may be above N_{max} so that no population can stabilize.

That is just one example of how complicated the relative reproductive rate scenarios can be. All these features can be captured in formal models, and though even these simple models can produce complex and chaotic dynamics, they are gross simplifications of evolutionary dynamics in nature. The point here is that over evolutionary timescales the behavioral repertoire that actually fostered dominance (i.e., that succeeded in perpetuating groups utilizing that repertoire) is the prosocial one. This may not be the group behavior that produced the greatest short-term growth (and we'll keep this in mind when we address the tentative nature of adaptive features in 4.1.4.4). We have also just seen that in some cases there may not be a prosocial repertoire because none of them are sustainable. Evolution pursues many paths that end up being dead ends, though not problematic for the account of prosociality presented here, it does raise considerations for using adapted moral attitudes as guiding or motivating that will come up later.

4.1.3.2 Group Selection

Group selection is an evolutionary process in which groups as a whole play the role that individuals do in "standard" versions of natural selection. The idea of group selection has had a rough history, but it plays an important role in many discussions of the evolution of prosociality. The core idea is that a group-level property that varies across groups plays a role in the evolutionary fitness of those groups – they expand, spread, duplicate, and out-compete other groups. Attempts to identify social properties having this effect, and that can't be traced directly to the individual fitness of the group members, have not been widely successful or accepted [Wynne-Edwards 1986], but versions of group selection that tie directly to individual fitness have seen a resurgence [Smith 1964, Wilson 1987, Bergstrom 2002].

Depictions of group selection typically describe a scenario of interspersed periods of group isolation and competition. Such situations may only occasionally occur in natural systems, but the formal models showing how, when, and why such a scenario works to promote prosocial behaviors are well understood. And there is nothing contentious about the proposition that the survival of individuals depends on behavior most naturally described at the group level: natural selection of multicellular organisms already has this form when cells are considered the individuals and animal bodies are the groups.[9] The threat of slippage occurs when trying to individuate interacting groups. It is easy to consider isolated groups as autonomous, but when two collections of (say) people meet, and some individuals within the groups interact, group boundaries (and hence autonomy) can become fuzzy.

For the theory of preferential detachment I was careful to define groups as implicitly organized sets of contiguously interacting agents possessing some specified property (e.g. being the same type, having their maximum degree filled, or matching some pattern – see 1.1.1.6). This definition continues to suffice here because it 1) identifies groups directly through individual composition, 2) fosters differentiation through individual properties and/or interaction patterns, and 3) encapsulates social-level entities with social-level properties that can be selected for or against. Thus abstract organizations such as companies, regions, and nations which encompass collections of individuals do not count as groups; nor do those abstract collectives of which people consider themselves members. Among the virtues of this definition are that 1) since interaction partnership is symmetric, group membership is transitive, and so groups are partitions of agents; and 2) this maintains the behavior-based approach essential for evolutionary processes (see 1.1.1.4).[10]

With this account of groups in hand, we can see that problems of scope can now depend on interaction timing. If two tribal bands, wolf packs, orca pods etc. meet only once a year and swap out a few individuals, then that brings into consideration whether these bands, packs, or pods are distinct groups (with the swapping seen as a group-level behavior) or one large group. If each band has the same behavioral repertoire and exists in the same environment, then the distinction makes no difference. But otherwise those behavioral repertoires are in competition and the groups encapsulating them would need to be considered distinct in order to track relative prevalence. These distinctions are only necessary, however, for "local" evolutionary explanations: explanations of the prevalence of some particular feature in comparison to other features over comparatively short time-spans, small geographical regions, and

[9] There are complications in considering metazoa as a society of individual cells; for example, the germ line cells necessary to reproduce whole organisms are cells that cannot reproduce themselves. And a multicellular offspring is not composed of cells from the parents in the way that a new city is founded by migrants from another city. These complications notwithstanding, the prevalence of the pattern of somatic cells changes due to selection pressures most easily discussed at the level of the body.

[10] This avoids problems such as people considering themselves as members of multiple groups with conflicting agendas. What may be understood as multiple distinct interaction networks with overlapping members can instead be captured as a single encapsulated group with interactions representing different specific behaviors. And what a person considers is anyway irrelevant – only actual behavior matters.

small population scope. For our purposes of better understanding the evolutionary origins of our moral experiences, a large scale and scope is more fruitful, therefore this general characterization of group autonomy and interaction suffices – especially remembering that group behavior can always be described in terms of individual behaviors.

Getting back to selection processes: we have seen how endogenous group fitness determines how quickly a group's size increases (or decreases); it also determines the rate at which groups grow to the point of splitting. When a group grows beyond a particular threshold (which varies widely with context and creature) it may schism to produce two or more groups. There are numerous observations of this process occurring with human and primate groups forced by resource limitations or competitions for group leadership [de Waal 1997]. The resulting groups may differ in the set of mechanisms and behavioral repertoires because the group divisions can occur along lines that vary in their features (e.g. familial lines, individual abilities, geographic preference). This feature fosters some degree of fitness exploration, analogous to the novel genetic makeup of diploid offspring. However, also like diploid offspring, the range of variation is constrained fostering the exploitation of what made the original group successful enough to split. This conservatism is achieved by the fact that the individuals in the new groups have mechanisms that also generated the behavioral repertoire of the original group.

Pulling this together we have an general description of groups that can survive longer, grow faster, and split into new groups with the same or similar features. These features allow groups to be seen analogously to individuals: even without direct conflict the differential rates of growth and schism determines the prevalence of groups' characteristic feature (behavioral repertoires in the role of individuals' mechanisms). Given resource constraints, the relative rates of growth and schism select for behavioral repertoires such that groups of individuals enacting those behaviors grow faster and split more often. This is a standard evolutionary account for the prevalence of a feature, in this case it is an evolutionary account of the prosociality of a group.

Examples from Preferential Detachment The results of both the Markov and agent-based model are compatible with the individual-grounded group selection descriptions found in the related literature [Smith 1964, Wynne-Edwards 1986, Bergstrom 2002, Boyd and Richerson 2002]. Furthermore, the Biased Lane Choice strategic context represents an especially clear form of individual-benefit driven group selection (see section). The groups become isolated through the assortative force of preferential detachment, and when population dynamics are added the utility benefit of intramixed C-types ensures their eventual domination of the whole population. The interpretation is that there are two coordinated cooperative outcomes, but one set of behaviors is slightly more beneficial (in terms of relative replication capability) to agents who engage in it. The model shows that not only do both cooperative types outcompete the defectors, the more fit group of cooperators out-competes the social arrangement with the same interests but lower fitness.

This formal result is compatible with, and perhaps an example of, different conceptions of group selection. It can be seen as a version of equilibrium selection among stable outcomes [Dawkins 1976]. It closely resembles the features of systems presented by Gintis [Gintis et al. 2003], but preferential detachment requires no feedback from social factors to achieve the same characteristic behavior. It also mirrors aspects of social learning and cultural feedback, but again preferential detachment does so without requiring agents to process the properties of its environment. Let us be clear that some of this explanatory success is possible because agent types are outside the agents' choice set, but it is also in virtue of this that the evolutionary explanation provided is more straightforward.

The conclusion tying preferential detachment to these other effective particular mechanisms is that the success of those more specific, multi-level, and social feedback systems can be equally explained by a simple, purely agent-level mechanism. Groups are nothing more than the realized collections of agents expressing a chosen property. At the group level we can see the success of one group allows it to dominate the population, but the group-level description can be translated into a agent-level description in a transparent way. This transparency fosters the later discussion of competing cultures, norms, institutions, and moral attitudes because a reduction of that discussion to the individual level is, in light of this discussion, unambiguous.

Cultural Group Selection: The idea that culture, rather than genetic differences, are driving selective reproduction practices has also been interpreted as a form of group selection [Boyd and Richerson 2002, Fehr and Henrich 2003, Gintis et al. 2003, Boyd and Richerson 2005, Bell et al. 2009, Boyd and Richerson 2009]. Cooperation, altruism, promise-keeping, and other proposed specific prosocial behaviors persist because the individuals in groups that can maintain these practices outperform the individuals composing groups that cannot maintain these prosocial activities. Note that in all these cases the success of the group can be described in terms of individual success, but it is the success of individuals in the context of the other individuals of the group. Because the behavioral repertoires also capture behaviors such as those considered "cultural" by these authors, we can consider within-group learning a special form of inheritance and address it appropriately when it comes up below.[11]

4.1.3.3 Behavior Diffusion

Finally, we address a group-level behavior analogous to individual learning: a collection of individuals adopting the behavioral repertoire of another group of individuals. This process, like all others considered, operates at the individual level, and a behavior's fitness is still "determined by the composition of the local group [Boyd and Richerson 2005]." If we limit our individuals to animals, then across evolutionary time it is often the case that groups meet and interact. In some cases individuals disperse from their home group specifically to join, form, or investigate other groups. The perimeter interactions that ensue can infuse a group with novel behaviors.

In some cases the new behavior is a bad match and the individual leaves, does poorly, or adopts a behavior to fit in to its new context. In some cases these novel behaviors maybe inimitable, though the practicing individuals achieve greater reproductive success in the new context, and over many generations it can become an important component of the behavioral repertoire. In other cases the novel behavior may be replicable by existing individuals, and the fitness of the borrowed behavior in the new context can allow it to spread rapidly. In this way it is the behavioral repertoire itself, rather than the group of individuals enacting it, which has the greater fitness.[12] As was already covered in 4.1.1, it is the behaviors that are primary, so whether the distribution of behaviors changes through population dynamics or learning makes no difference.

Because group-level fitness depends on the interplay of all its individuals' behavior, a change in one individual's behavior can alter the fitness of the whole group – for better or for worse. It is natural to expect that any group's behavioral repertoire includes how to handle these intergroup interactions and how to receive visiting individuals. In contexts where disruption of the status quo can be highly damaging to

[11] The distinction between "biological" and "cultural" selection is questionable to start [Alexander 1987], but furthermore it is often deployed to draw a divide between human and other animals' ability to achieve prosocial arrangements. In section 4.2 I present arguments and evidence for a continuity across species that runs counter to this divide.

[12] Meme theory naturally comes to mind at this point [Blackmore 2000]. Though there are similarities, behaviors as I use them do not qualify as memes and so the parallels would have to be stretched to fit. Furthermore, though the idea is interesting, building the bridge to the meme theory literature would not provide any clear advantages over the current evolutionary account.

the group's cohesion and survival we would expect rebuffing. In highly robust and fertile contexts we would expect more mixing to be in the group's interest. It can be in an individual's interest to disperse a novel behavior even in cases where doing so hinders the receivers' fitness (a suicide cult being an extreme version of this). However, relying on population dynamics to change the repertoire of behaviors limits the adaptiveness of the population.

Several formal models produce the relevant dynamics, each highlighting the power of imitation in behavior diffusion. "Consequently, behaviors can spread from groups at high payoff equilibria to neighboring groups at lower payoff equilibria because people imitate their more successful neighbors. Such spread can be rapid because is depends on the rate at which individuals imitate new strategies, rather than the rate at which groups become extinct[Boyd and Richerson 2005]." This describes the dynamics of evolutionary and spatial game theory models, and the same results were seen in preferential detachment, indicating that relationship is robust to differences in modeling approaches. This is not at all surprising, but it has important and unsettling implications for the appropriateness of behaviors and attitudes generated by parts of mechanisms that change at different timescales (a straightforward version is the Mismatch Hypothesis – the idea that Paleolithic instincts trigger responses that are maladapted to modern social structures). So, while the imitative form of behavior diffusion may be more rapid and more fluid (i.e. rather than incremental by generations) than biological reproduction-based changes in prevalence, learning behavior also faces the same exploitation/exploration tradeoff.

4.1.3.4 Group Fusion

So far we have discussed group growth, splitting, and competition; however, the possibility of group combination also deserves a few words. Group fusion implies that two autonomous groups with distinct behavioral repertoires initiate prolonged interactions such that the constituting individuals must now be considered a single group. Recall that groups are sets of individuals along a contiguous interaction chain having some specific property. So we can consider subgroups by separating along appropriate properties (e.g. history, physical variations, diet), and behavioral contingencies can maintain their subgroup sensitivity in the way that the original intergroup interactions played out.

However, as the interactions among subgroups become increasingly critical for the perpetuation of the larger set of individuals, it becomes more natural to describe the situation as interactions among subgroups rather than between separate groups.

Behaviors adapt to the presence of a different set of other behaviors, and what was two distinct characteristic behavioral repertoires eventually blends into one characteristic behavioral repertoire of the larger groups – possibly maintaining some features of each constituent group, or one may dominate, or the combination may result in something completely new. Because all behavior can be tied to individuals, the difference is only a matter of ease of reference.

The process is analogous to colliding galaxies: the stars follow the same laws of physics regardless of whether we consider the clusters two interacting galaxies or one new galaxy. Autonomous galaxies have characteristic shapes: spirals or globs. Recently fused galaxies have amorphous shapes, but over time the forces that forge those regular shapes pull the larger cluster into a new shape. The new shape will be some mix of the magnitudes and velocities of the two combined groups, perhaps resembling neither.

4.1.3.5 Summary of Group Interaction

There are still other forms and means of intergroup competition. I've mentioned in 1.7.1 that criteria besides perceived success may form the basis of imitation behavior: popularity of the behavior, similarity of target to oneself, target's role in the social network, innovativeness of the behavior, etc. All of these are learning behaviors in the sense above because the agents adopt new behaviors and hence can change their behavior-generating mechanism. Furthermore, though what it means for a group to grow or go extinct is tightly bound by the meaning of those words, what counts toward the fitness of individuals and groups can vary a great deal across time and environment. Thus it is important to remember that fitness can only be ascertained retrospectively, can be described entirely in terms of actual individual behaviors, and explains the prevalence of behaviors (and behavioral repertoires) observed on an all-inclusive basis. Smaller scales, scopes, and time spans can be partially evaluated when properly contextualized, and the group interaction material just covered must be understood in that light.

4.1.4 Implications of Evolved Prosociality

I have heretofore provided a detailed, though fully general, account of the evolution of prosociality: prosociality evolves because those behaviors that are conducive to sustaining a group of practicing individuals maintain that behavioral repertoire. The behaviors enacted by individuals as part of the repertoire will match the behavioral tendencies of the individuals (aka interests) *in the current context*.

However, a behavior exists in a context of other behaviors, and series of best-interest responses to interaction partners' behaviors can lead the repertoire to an arrangement that serve their interests worse than other possible arrangements. So even though a group's behavioral repertoire can sustain itself and out-compete other groups, this does not directly imply that those behaviors align with the individuals' overall interests in the sense that, given the collective choice, the individuals wouldn't collectively behave differently. This consideration is the familiar problem of suboptimal equilibria common in genetic algorithms and game theory. The clearest examples of how this can happen is from the literature on coalitions in spatial voting models and Condorcet paradoxes [Saari 2003] in which a series of incremental coalition adjustments produces an outcome far outside what is desired by any participant.

Below are some further considerations and implications of the evolution of prosociality that will play an important role in our discussion of the evolution of the moral experience in the next section. These specifically deal with prosociality considered over spans of time in different ways.

4.1.4.1 Progressive Prosociality

As the mix of behaviors, size of the population, available resources, external pressures, etc. vary across time, the context in which individuals adapt likewise changes and behaviors must adapt. Thus we would expect that prosocial behavioral repertoire adapts over time through a series of contextual shifts– even in cases in which the problem environment does not change over time. The progression of behavioral repertoires serves the individuals' interests at each step, and thus the individuals are adapted to either 1) each step with changing repertoires or 2) the shifting contexts via a repertoire with appropriately contextualized behaviors for each step. The succession of adapted behavioral repertoires gives the appearance of progress toward a better (i.e., more sustainable, more competitive, faster growing, etc.) prosocial arrangement, though there may not be any equilibrium for the process, and many paths will turn out to be evolutionary dead ends. Furthermore since fitness can only be ascertained retrospectively, this appearance of progress is illusory (see 4.1.4.4 for more on this).

An analogy to trees repopulating a denuded forest may provide some addition intuition about the sort of progression to which I am referring. The soil and weather at some patch of land makes a certain mix of vegetation best adapted (prosocial), but just after a forest is cleared, those trees' seedlings are not successful competitors against weeds and smaller plants that grow more quickly. Once those small plants are there, larger shrubs that go deeper for nutrients are better adapted, and drive out the small plants. The shrubs clear the ground and provide cover for seedlings that eventually replace the underbrush through competition for sunlight, water, and nutrients. Through a series of progressive stages the trees, which were best adapted to the environment, retake the land. Their ability to do so depends on a context of shrubs, though; their's is not an all-things-considered best fitness.

The story also underlines the crucial role that a temporarily adapted behavior (e.g., the weeds) plays in the intertemporal social structure. If not for the weeds, there would never be the trees; so the ability of weeds to persist in the initial stages must also be considered part of the prosocial repertoire since those types are essential to the repertoire of behaviors that *generates* the sustainable arrangement of trees – even though they are not included in the sustainable final stage. The prosocial behavior in a context where there is no vegetation is to be a weed. In some environments there is no better type to be, no other behavior would serves the interests of an individual deviating from weed behavior in a context of weeds. If the weeds manage to arrange themselves in such a way as to prevent the ascent of these other types in an otherwise fertile environment, then the weed type will seem part of an antisocial arrangement when we know a forest could thrive if a forest were there. But insofar as the weed arrangement is sustainable and continues to serve the interests of the individuals within it, it is prosocial within that context.

4.1.4.2 Sufficient Variety

Recall that to be sustainable, and hence to be prosocial, a behavioral repertoire must also out-compete other groups' behavioral repertoires. Intergroup competition thus further promotes the all-things-considered fitness of groups by ensuring that a persisting behavioral repertoire is also more fit than alternative repertoires in competition range.

Thus if one group converges, by chance, on a behavioral repertoire that does not provide as high fitness as some other possible repertoire (e.g. the weeds that keep out other behaviors), then the inferior-fitness group may lose in competition with a group employing a higher-fitness repertoire (e.g. shrubs from a neighboring patch). The ability of the global community to find more fit prosocial repertoires depends on there being a *sufficient variety* of repertoires for more fit ones to be among them.

This is not always the case though; in fact it is infrequently the case that repertoires would be de facto optimized for their environment through an evolutionary process. We cannot conclude or assume that observed arrangements are the best they could be in terms of fitness or in terms of serving the all-things-considered interests of the comprising individuals. We can only conclude that the series of contextual changes resulted in social arrangements such that the interests of individuals extant in each context were served by behaving in a manner that also fostered the adaptiveness of those individuals in that context.

4.1.4.3 Aligning Fitness and Interest

As has been stated already, only repertoires that are capable of sustaining themselves (i.e., are more fit than extant others) are prosocial (by the definition of prosocial). And only behaviors that serve the interests of individuals are enacted (by the definition of interests). It can happen that some behavior serves an individual's interest while decreasing the fitness of the enacting individual, but it cannot happen that some behavior that increases fitness will trump an individual's behavioral tendencies. Thus a behavior cannot gain prevalence unless it is in the interest of extant individuals, and it cannot maintain prevalence unless it is prosocial. This combination of factors implies that *over evolutionary time, the individuals are adapted such that their behaviors align with prosociality.* This claim derives from the premises of evolution, behavioral tendencies, and prosociality; and though intuitive, several caveats must be made.

Individual-level deviations from this pattern exist; exploiters, cheaters, murderers, etc. pursue their own interests through behaviors that are antisocial. If too many individuals have antisocial interests then the resulting behavioral repertoire is more likely to be out-competed and/or endogenously less fit. For example, it is possible for a group of defectors to become locally dominant and not invadable by a small number of cooperators (this is like a patch of weeds keeping out shrubs). However, with population dynamics included over two isolated groups of cooperators and defectors, the cooperators will take over the population. Mutant defectors within a population of cooperators gain short-term benefits from that behavior, but a group adapted to the prosocial cooperator behavior will either include in their repertoire a means to eliminate those defectors, or that group will be out-competed by some group that does.

Hence the evolutionary account predicts that even if we observe groups with antisocial behaviors in the interest of individuals, and even though those behaviors are adapted to the context in terms of having higher short-term fitness for those individuals, those behaviors are not part of the all-things-considered best prosocial social arrangement. The theory also predicts that the most fit actual arrangement at any time in any environment will not be the all-things-considered best prosocial social arrangement. It doesn't have to be the absolute best, it only has to be better than extant competitors.

4.1.4.4 Adaptive Features Are Tentative

An interesting implication of this account is that since fitness is not a forward-looking feature of systems, we can only explain the prosociality of a system in terms of how it came to be and persist. It has limited value for projecting properties for future arrangements and further contextual shifts. Determining whether some extant social arrangement is adaptive or not requires predicting its sustainability and resistance to being out-competed by groups with other repertoires, but as Yogi Berra famously said, "It's tough to make predictions, especially about the future."

As will be shown in section 4.2, the general historical process suffices to explain why humans (and other animals) have the moral attitudes they do, and why they would *appear* to be universal immutable truths (hint: because them seeming that way is adaptive). But we are still left in a position in which prosociality assessments of the current system can be only tentative and uncertain. The norms established through the evolutionary process are those that brought about and maintain the current arrangement, and have out-competed other norm systems against which the current system has actually competed, but they are only valuable as guides as long as the future is very much like the past.

We'll see in 4.2.2 the role that stability plays in establishing general solutions to universal social problems, but for now consider that the future is never actually like the past with respect to the specifics of the social arrangement. The conditions for which a group's moral attitudes are adaptive are not quite the ones they are likely facing at present...societies of living creatures are not equilibrium systems. By the time the norms have adapted to extant conditions, the conditions will again be different because the adaptation process itself changes the context for behaviors. This idea segues nicely into the evolutionary explanation of moral attitudes as those appropriate for individual mechanisms to produce the prosocial behavioral repertoires.

4.2 Evolution of the Moral Experience

I have already explained why behaviors, rather than psychological or other behavior-generating mechanisms, are the currency of evolutionary explanations for prosociality in 1.1.1.4. The corollary for this section is that if the way we experience morality likewise evolved, then it too depends on the presence and/or absence of specific behaviors rather than depending on any specific underlying mechanisms. With respect to morally relevant activities, the drivers of behavior are merely contingent; however, the restrictions of biology allow us to expand the domain of morality (or, at least, of moral experiences) well beyond humans and human societies.

Once we accept that the moral intuitions humans carry are indeed the result of an evolutionary process, and that they are indirectly adapted to produce the appropriate behaviors that are directly selected for, we also must accept several consequences implied by this proposition. Some do not accept that moral features themselves evolved, but only that certain features (e.g. cognitive abilities, mindreading, large-scale social networks) which are necessary prerequisites for morality to have evolved, and that morality (and our experience of it) has been built from these evolved building blocks. I will here argue on the opposite side: not only are humans' manifestations of morality a result of an evolutionary process, but similar manifestations obtain in any system that undergoes an evolutionary process.

The thesis is that moral intuitions are a particular flavor of attitude that correlates with the mechanism that enacts those behaviors which are contingently necessary to achieve and sustain a population of individuals who behave in that way. This section unpacks each step of that thesis, some parts of which were already covered in the description of the evolution or prosociality above. This section also shows that given a plausible evolutionary explanation for the prosocial function of moral intuitions, the universality of the problems moral attitudes are attuned to solve, and many similarities in prosocial behaviors across species, that we can extend the assignment of moral attitudes to systems of individuals other than human individuals.

4.2.1 Whence the Moral Attitudes

The dominant tradition in moral theory is the rationalist model by which individuals consciously think through the facts of a situation, and through that process determine the moral value of each option [Mikhail 2011]. This moral judgment approach does not eliminate the role of moral emotions – indeed the long-standing "naturalist fallacy" and "is-ought gap" imply that a non-reasoned moral sentiment of some kind must lie at the base of all evaluative lines of reasoning [Hume 1998 (1777]. In contradistinction, an intuitionist (in the psychological rather than philosophical sense) approach claims that people have more automatic moral reactions to situations generated by innate instincts or internalized norms, and moral reasoning provides (at best) post hoc rationalizations of these revealed sentiments. Though some recent psychological/neurological research has shown that emotional responses correlate better with behavior than the stated conclusions of moral reasoning [Haidt 2001], the strongest candidate psychological explanation is a mixture of the two mechanisms [Greene et al. 2009].

As already made clear in 4.1.1, the prosociality of a behavior is ambivalent about what mechanisms are responsible for producing that behavior. Both reasoning and intuition are admissible as drivers of the behaviors that produce prosocial arrangements (as well as physical reactions, trained habits, and any other behavior-generating mechanism). The demands of prosociality, combined with the constraints of behavioral mechanisms in evolved biological creatures, imply that a narrow range of experiences would be expected to correlate with behaviors producing prosocial and antisocial behavior. The character, force, and ubiquity of these attitudes reveals, and results from, their import for group sustenance, cohesion, growth, and replications – aka fitness and prosociality. What follows is a set of explanations for how certain moral attitudes are expected to evolve as prosocial adaptations to specific environments and social problems.

The idea that behaviors become entrenched as moral norms when they are part of a system of self-reinforcing behaviors is not unprecedented.

> “Sugden (1986) maintains that a convention acquires moral force when almost everyone in the community follows it, and it is in the interests of each individual that people with whom he or she deals follow the rule providing that the individual does too. What evolves according to Sugden is a ‘morality of cooperation’ [North 1990].”

Keeping in mind that norms can be implicit, this quote translates to the following in the language of this project: behaviors that are an essential part of a prosocial behavioral repertoire become associated with moral attitudes. These moral attitudes form a part of behavior-generating mechanisms that gives priority to group-promoting behaviors in appropriate contexts. This project makes an even stronger proposal: a process akin to norm development, carried across billions of years of evolution, is responsible for the existence of moral experiences.

4.2.1.1 Moral Attitudes

I use “attitude” to refer to the broadest, most inclusive class of phenomena captured under the moral experience. An attitude can be an emotion simply felt or coloring cognition or perception; a feeling accompanying a judgment or belief; an affective state corresponding to mood or intuition; a visceral motivation to enact a specific behavior; an urge to follow a norm; and/or other similar psychological responses. Moral attitudes may be undirected (e.g., raw feels), directed toward specific targets (e.g. indignation toward other individuals or others' behaviors, guilt over one's own thoughts or actions) or generally/abstractly projected (e.g. toward norms, institutions, organizations, or other ideas).

Though inclusive of all these variations, clearly not just any attitude can count as a moral attitude. Moral attitudes possess a particular "flavor" that will be explained in the material that follows. But first it is worth mentioning now (to keep the reader in the right frame of thinking) that I will not define this moral flavor in a functionalist way; i.e., an attitude isn't moral *because* it is coincident with and/or generates prosocial behavior. Also, a moral attitude is *not itself* the content of any proposition, though it can be attached to such a proposition. It is a non-linguistic, stimulus-responsive, behavior-inductive, emotionally loaded mental state.

The moral attitude I am invoking is the phenomenologically familiar one: a cold chill upon hearing a tale of physical harm against innocents, a warm glow when the unfortunate are helped, an inner pain when caught in a lie, indignation when receiving an unfair share, rage upon catching a thief, and so on. I will not attempt to catalog or categorize the spectrum of moral attitudes because one implication of the explanation posed here is that it is a perpetually shifting continuum. I claim that there is a unique feel or sense to these experiences, which seems supported in the literature from moral philosophy and psychology, but it is also clearly experienced simultaneously (and perhaps mixed) with other emotions, urges, moods, motivations, feelings, etc.

4.2.1.2 Theories of Emotion

Emotions come into play in two ways here: 1) the specifically moral attitudes, and 2) the relationship between the broad spectrum of emotional/attitudinal responses and the moral attitudes. The distinction is that emotions such as anger, fear, joy, and love play a role in generating and regulating behavior, and the resulting behavior has an effect on prosociality and moral responses. The moral attitudes are those that specifically relate to moral goodness and badness, right and wrong, approbation and disapprobation, etc. It covers experiences of "X is morally wrong." or "Y did the morally right thing." or even "Yay!" or "Boo!" with moral sense attached. Here I tease apart what these moral emotions are like, how they relate to other emotions, and how they fit into the evolution of moral experience more generally.

Hearing and cogitating facts can prompt feelings of (for example) injustice and jealousy alike, though it may never trigger any thought equivalent to “That’s unjust.” or “I’m jealous.” One need not be aware of one’s own moral attitudes, even as they affect behavior. This is analogous to hunger affecting behavior with low energy and agitation even when the person is 1) not aware of a feeling of hunger 2) not thinking “I’m hungry.”, and 3) not able to reflectively attribute behavior as resulting from hunger. It is an effective state that colors ones perceptions and actions as well as (at times) ; and moral attitudes can operate similarly. Another parallel is that moral attitudes, like hunger, affect behavior as a matter of degree; one can be more or less hungry, and also hunger’s effect can be mitigated by other factors (e.g. stress, need to focus, and pain). So, some stimuli produce mildly aversive moral responses and others overpoweringly intense ones. The moral response can sometimes be overcome by concerns of personal gain or fear and at other times overcome the drive to preserve one’s own life, respectively.

Though there are theories of emotion that focus on each aspect of the emotional experience [Prinz 2004], the current work is not committed to any of these because it is agnostic on the details of the behavior-generating mechanism. The closest ties can be made to the behaviorally related theories of Nico Frijda [Frijda 1987] and Paul Ekman [Ekman 2003] presenting emotions as tendencies to act in certain ways in specific circumstances. These theories posit that emotions embody disposition to respond to stimuli, and as such can combine cognitive, sensual, physical, and other aspects of experience. For an evolutionary explanation of the behavior, it must be the case that the behavior itself is selected for. A completely different apparatus that also produces the same behavioral repertoire (e.g., a twin Earth or body switching scenario in which creatures’ internal biology is drastically different) would be equally fit within a given environment. Though the associated feelings are conceptually and evolutionarily contingent, they are de facto biologically essential to appropriate behavior generation in animals as they have actually evolved.

4.2.1.3 Neurobiology of the Moral Experience

I have claimed that moral attitudes are feelings, reactions, colorings, etc. with a particular “flavor” and there is neurological evidence to support the claim that there is cognitive activity unique to the moral experience. fMRI research over the past decade by Jorge Moll and colleagues, as well as other teams, have uncovered that particular regions of the brain (e.g. the medial orbitofrontal cortex, the temporal pole, and the superior temporal sulcus of the left hemisphere) become activated when a social situation involves moral judgments but not when it is emotionally evocative in general [Moll et al. 2002]. They had shown previously that different regions are activated for tasks of moral versus factual discrimination [de Oliveira-Souza and Moll 2000, Moll et al. 2001]. Though the limits of fMRI analysis makes the details of their results preliminary, they suffice to demonstrate that moral experiences correlate with a unique signature in brain activity. And this is precisely the evidence I require to support the existence to a specific flavor to those experiences that we categorize as moral.

Of course, finding a brain area unique to moral attitudes is neither necessary nor sufficient for the overarching evolutionary account. Morality could be experienced (for example) through a unique combination of areas with ulterior purposes, and it certainly doesn't matter which particular brain areas are involved (because it's the resulting behavior that matters). The evidence is helpful in supporting my argument that the critical importance of solving social problems over long enough time spans selected for a distinctive experience with regard to the appropriate behaviors in those social contexts. This line of research offers another helpful neurological fact; the brain regions associated with moral experiences in humans, activate similarly in non-human primates and rodents as well. This point will come up again soon in 4.2.3.

A separate branch of neurological research by Joshua Greene, Fiery Cushman and others have uncovered distinctive features regarding people's physical responses, anxiety reactions, and decision outcomes in variations of the famous trolley problems.[13] Among their conclusions is the presence of more than one neural/mental mechanism to solve moral problems, and the mechanisms do not always proffer the same solutions [Greene et al. 2009]. Their research separates the responses into gut reactions and reasoned decisions.

This reasoning needn't be the type of moral reasoning associated with Kantian systems of moral judgment. It is clear that even rats reason through problems before acting in some cases, and just react instinctively in other cases [Bekoff and Pierce 2009]. This does not imply that they are weighing learned norms against their gut feeling either, only that they have to figure out what to do in their current context. This may require an imaginative exercise, accessing memories, priming muscle memory, or other such cognitive activities. None of that actually matters for our present purposes; what *does* matter is a distinction between a moral experience immediately prompting an action, and situations requiring more complicated balancing and comparing of factors. The presence of distinct neural regions for these functions, and

[13] Trolley problems are any situation in which a person must decide to do nothing and let several people die, or do something that kills (or results in the death of) a single person but saves the others.

their activation in different problem environments, and the broad range of creatures that share these features, combined with the existence of levels or degrees of strength in moral experience all establish a strong foundation for the continuity of the moral experience discussed below.

4.2.1.4 The Moral/Conventional Distinction

Another point needing addressing along the path of explaining the evolutionary origins of moral experience is the distinction between moral transgressions and transgressions against conventions, etiquette, laws, rules, and the like. There are many experiments in the developmental psychology and the psychopathy literatures exploring the differences in the timing, brain regions, and cognitive capacities necessary and associated with each ability. The fact that "... the attempt to draw an analytic distinction between morality and convention is fraught with controversy [Nichols 2004]" actually plays well into the account given here. Behaviors essential to sustainable arrangements attach to moral experiences over evolutionary time, those that are important yet not essential gain weaker moral force, and those that are inessential (mere conventions) gain an attitude that is similar in operation but distinct. The moral/conventional associations shift over time and vary by context – they can change in degree, be triggered by different stimuli, and generate new response behaviors – in ways consistent with variations in the essential elements of the behavioral repertoire.

Conventional norms, like moral norms, are sets of behaviors that are varyingly appropriate in specific contexts, and the appropriateness shows the same contextual contingency as moral norms as well. Furthermore, the proper behavior reinforces what are typically solutions to the same kinds of social problems that morality adapts to solve: cooperation, coordination, specialization, contribution, etc. The important difference between moral and conventional norms and responses is the length of time that a violation of the norm precipitates a decrease in the fitness of the group members or the group as a whole. The longer the de facto fitness detriment of a behavior, the more likely it is to become associated with a moral attitude; and the larger the detriment the stronger the moral force of that attitude.

As contexts change, and the fitness effect of a behavior becomes more or less intense, it will become more or less strongly associated with a moral attitude. In some cases this can only happen over generational time because the behavior (or the physiological connection between behavior and the mechanisms that generates it) is not plastic, while in other cases an individual can realize (and this need not be a conscious realization) that a behavior (for example) causes harm and only then associate it with a moral attitude. The ability to change the behavioral/moral association requires greater, higher level behavioral contingency and thus a more sophisticated behavior-generating mechanism. The flexibility produced by being able to utilize hierarchal structural thinking also admits to smooth or incremental variation across species and time.

Conventional and moral transgressions typically trigger different parts of the brain, and some psychopathic criminals (who have violated moral norms) appeal to conventions (rather than moral attitudes) when explaining why moral transgressions are wrong. They are behaviorally similar, and even similar in mechanism; however, even young children and individuals with autism are able to make the distinction and see moral transgressions as universal and independent of authority, while conventional transgressions depend on the disapproval of other people [Nichols 2004]. These distinctions, and the early age in which they can be made, are exactly as one would predict from the evolutionary account being presented. And yet the theory entails a more complicated and highly contextually contingent relation of behaviors and moral attitudes.

Observing one individual injuring another individual without provocation is a standard example of a scenario that prompts a negative moral response; and it does so across human cultures, in children as young as three years old, and (as far as the evidence suggests) in species ranging from rats to apes (at least) [Nichols 2004, Bekoff and Pierce 2009]. This reaction seems a strong candidate for an antisocial behavior that has persisted long enough to be deeply encoded across most vertebrates to elicit the negative moral reaction. Of course there are cases in which physical violence of a similar form fails to prompt the negative moral response: punishment and intergroup conflict being the clearest examples. These injurious behaviors famously promote group fitness (sometimes at a cost to the enacting individuals) and are thus expected to prompt feelings of approbation rather than disapprobation...which they typically do in humans, primates, and dogs (at least) [de Waal 1997, Bekoff and Pierce 2009]. So even something as straightforward, constant, and universal as norms against physical harm admit to contextual contingencies across species and far back in evolutionary time.

One corollary of these results is that changes in context, changes in the conditional mapping of behaviors to attitudes, and/or changes in the attitudes (and their mechanisms) themselves may all effect the expression of moral attitudes. But evolutionary processes on individuals and their social arrangements rewards behavior-generating mechanisms that promote prosociality. Behavioral repertoires that failed to purge behaviors that reduce individual and/or group fitness in almost all cases are those that would have been out-competed and eliminated by those that had. Behavioral aversions and the appropriate features in the behavior-generating mechanisms are therefore expected to persist in all societies that have sustained themselves.

The theory here marks conventional norms such as etiquette as being similar to, but less critical than, moral norms in maintaining social arrangements. The resulting urges to act can be just as strong and can sometimes overwhelm the moral feelings evoked.[14] If there were greater selective pressure on (say) holding the door open for others, then it too would become attached with moral force over time. Hence this is another example of how the alignment of individuals' interests and de facto fitness requirements is essential to forging behaviors as moral.

Either mismatch of 1) attributing moral attitudes to behaviors that do not de facto reduce fitness or 2) failing to attribute moral attitudes to those that do will result in selection pressures to correct those discrepancies. Insofar as conventional norms are also behavior-guiding, the selection pressure may be weak. Which behaviors fall into each category change over time, and so we can expect a constant fluctuation over time between behaviors having moral or conventional status...and hence variation among regions and other contextual features. However, the combination of weak selection pressure and the specific import of moral attitudes implies that only those sets of behaviors that have consistently merited moral attitude association across evolutionary time will be associated with a strong moral attitude. The strongest attitudes, therefore, can be expected to solve problems that have faced groups of individuals the longest.

4.2.2 Universality of Social Problems

The behaviors that produce and resolve survival-critical social problems are those that are most tightly bound to moral attitudes, and these problems are faced by creatures great and small, ancient and modern, simple and complicated. As explained in the evolution of prosocial behavior, the general features required to maintain and increase the prevalence of systems/societies of individuals are shared regardless of what manner of object counts as an individual. Narrowing the focus to animals on Earth allows us to be more specific about what those features are and to what degree they are shared across animal groups. What we find is that the commonality of social

[14] As an example, consider the case of Alexander Hamilton's justification of his duel with Aaron Burr. Though against the practice of dueling, he felt obliged to uphold the social norm.

problems across animals has produced similar solutions as well, and we can detect these as shared patterns in behavioral repertoires. Because 1) the environments across the globe and across time have been largely similar, 2) the demands of keeping living things living are largely uniform, and 3) the behavioral repertoires that are prosocial and antisocial are similar, we would anticipate that the key features of behavior-generating mechanisms directed toward prosociality (i.e., moral attitudes) are also similar.

Evolution by natural selection requires both 1) the presence of variations to select among and 2) a high degree of environmental stability relative to the rate of adaptation. The variation requirement was already discussed in 4.1.4.2 and is covered in more detail in 4.2.2.1 below. The stability requirement derives from the fact that biological adaptation proceeds at the pace of generational turnover. Thus we would predict that longer-lived creatures 1) live in stable contexts, 2) have behaviors (and other features) that are robust across contexts, or 3) have behaviors that are appropriately contingent across multiple contexts. Though this may at first indicate a wide range of behaviors, and they do differ in their superficial details, they are functionally similar means to the same procreatory ends.

The "game of life" that animals must play to survive and replicate is quite constant across much of the animal kingdom (and certainly from rats to humans): e.g., they need to cooperate, compete, contribute etc. to eat, get shelter, find mates, raise children, etc. The need to solve the same forms of social problems likely explains why all mammals (at least) have similar moral responses (both behavioral and neural). So the social problem space is stable enough to produce recurring general patterns of solutions employing the same underlying mechanisms. Though the specifics of the environment, and which specific behaviors are adaptive, constantly change, the need to cooperate to maintain resources, contribute to achieve large goals, coordinate on group-wide conventions, obtain nourishment, resist environmental stress, select mates, and produce offspring are omnipresent. Individuals, and groups of individuals, that don't play this prosociality game and don't solve these social problems are outcompeted by those that do.

Prosocial behaviors such as cooperation, contribution, punishment, food sharing, and whistleblowing have been widely observed in diverse animal species [15] and, given

[15] Though the life of animals has been considered "red in tooth in claw" by philosophers and scientists, this seems to have been for lack of actually looking. It is clear that cooperation in animals is the norm rather than the exception. With the obvious exception of predator-prey interaction, there are very few animal interactions (whether intra- or inter-species) that are aimed at causing harm. Mating and territorial competitions do sometimes result in serious injuries, but not typically

the explanations for the attachment of moral attitudes to critically prosocial and antisocial behaviors, the explanation for human moral attitudes run parallel for other species. That is, if the interplay between emotions and the importance of sustaining certain prosocial behavioral repertoires does explain the occurrence of moral feeling, emotions, urges, intuitions, etc. in humans, then it also does so for other species.

This is good news for understanding human morality because in order to understand the evolution of morality in humans we must inevitably breach many issues in animal morality. This is because 1) we have nearly zero data about the lives of our human and pre-human ancestors and 2) there exists mounting evidence in support of the evolutionary continuity among species (discussed in section 4.2.3). In light of the first reason I will use the social behaviors of various animals as a proxy model for the social behaviors of early humans and protohumans, as well as in support of arguments made about the evolution of morality all the way down our evolutionary history. The second reason indicates that this proxy model is quite appropriate.

4.2.2.1 Variation among Groups

Some of the points made above may lead one to think that we should expect convergence in social arrangements, behavioral repertoires, and moral attitudes across time and across species. However, full convergence is not a likely outcome for many reasons: multiple stable outcomes, local environments, intermediate timescales, or changing environments.

The actual social problems "solved" by the adaptation of behaviors have many possible stable outcomes, some of which may even be equally beneficial, so we would expect variation in which outcome segregated groups may reach. Such a result is hinted at in Coordination Game experiments, but is really nailed by the unbiased collaborative game Lane Choice. Whether learning and/or population dynamics are in operation, the population splits into two segregated groups of equal fitness; they are equal in size on average, but in any given run widely varying (20-80%) proportions of each type are observed. Not only that, but defectors dominate in some runs and that is also stable. If we scaled up to problems with dozens, hundreds, or thousands or types, then the number of possible stable outcomes explodes, and each isolated population may reach a different one.

Environments in different geographic locations present individuals with different

and only as a means to an end. Even primarily solitary animals send and receive social cues and have a "way of doing things" regulating mating, territorial disputes, sibling play, etc. [Dugatkin 1997].

social problems. In terms of the modeling tools, it means that different games are played in different regions. Given that's the case, individuals that find an adaptive repertoire of behaviors in location A may find that they perform quite poorly in location B...even without any external competition. These considerations lead us to topics in sociobiology, but these are not addressed for now. The point here is that swamps and tundra make different behaviors adaptive, and so we would expect animals (including humans) to find different social arrangements sustainable. That does not require, however, that they have distinct *contingent* behavior rules – that is, distinct mechanisms. It only implies that different behaviors are prosocial, the environments prompt different behaviors, and hence distinct specific behaviors are associated with moral attitudes in distinct environments.

There are regional differences in behavior that cannot be attributed to differences in local environments or multiple points of convergence, and some of these may be explained by the timescales required for convergence. It may be the case that one of two competing social arrangements is strictly better performing, but if the two groups are sufficiently isolated it may take extremely long stretches of time for one behavior to successfully invade or replace the other. And groups can be de facto isolated regardless of physical proximity if the individuals avoid those other individuals with different behaviors.

Though it is clear that imitated behavior can spread more quickly than genetically modulated behaviors, it is also the case that within a group, imitation reduces variation and can lock in behaviors. Cross-generational learning reinforces a narrowing range of expressed behaviors such that behaving differently may become costly. As a result, "Some elements of culture likely still have time scales of change measured in millennia [Boyd and Richerson 2005]." A clear example of what is likely a timescale issue is tool use by orangutans. Tool use differs across communities according to patterns that indicate (slow) cultural diffusion [van Schaik et al. 2003]. For example, the use of a specific tool to open the hard-shelled Neesia fruit has clear fitness (and happiness) enhancing features for the apes, yet movement of the technology has had a difficult time crossing wide rivers within the orangutan habitat [van Schaik and Knott 2001]. However, given enough time, the spread of the behavior throughout the orangutan range seems inevitable.

Inevitable, that is, unless something changes in the environment that makes the behavior no longer adaptive.[16] The environment (which includes all other species

[16] And also ignoring issues regarding the loss of orangutan habitat, and depletion of their numbers, toward extinction in the wild.

with which the focus species interacts) is constantly in flux. There are the obvious changes brought on by external events (meteors, volcanoes, invading species, and disease outbreaks), but more interesting and more relevant are the coevolutionary changes. Predators and prey are adapting through a constant arms race, flowers and incests become increasingly specialized, and in general behavioral changes alter feedback from fauna and flora, as well as water, heat, and nutrient flows. Life is adapting toward a moving target, and it may reasonably be the case that there are no stable behavioral repertoires to converge on.

In summary, the literature on the evolution of prosocial behavior seems to imply that behavior will converge to a small set of adaptive, sustainable behaviors, but this is a simplification along many dimensions. The set of sustainable behaviors may be extremely large, and multiple collections may coexist. Even if there is convergence of mechanisms, variation in environments can produce variation in deployed behaviors. Even if the environments are similar, the timescale for convergence can be long enough that several intermediate repertoires are observed. And even if adaptation can happen quickly enough to converge on a best social arrangement, having reached that point may make a different arrangement more advantageous.

Despite all this, the similarities outweigh the variations and the stability anchors the range of variation. For all the reasons just described we should not expect convergence, but we also know to expect dramatic similarities across regions, times, and species. The paths of biological evolution only leave narrow trails to follow; to much deviation and the organism isn't viable. The human condition is not appreciably different than the life history of an ape, wolf, or rat. Facing the same social problems, and built with common functionality from similar modules and materials, we should also expect similarities in our experiences while solving these common problems. Furthermore, it is reasonable to expect not just a few similarities when interests, contexts, and requirements for prosociality overlap, but rather a widespread similarity in social orders, behavioral responses to threats to society, mechanisms driving those behaviors, and experiences accompanying those mechanisms.

4.2.3 Continuity in Nature

The expected widespread similarity in moral experiences just mentioned is merely one facet of shared features across animal species on Earth, and it is importantly a consequent of other dominant commonalities. The near-continuum of features in nature is driven by the need for all living things to solve the same social problems – at least in part (and probably a large part, though I won't argue for this directly).

Though many naturalists and philosophers of the past have focused on the differences among species, a closer and more parsimonious look at these differences reveals them to solve the same problems, but in contextually adaptive ways. Conceiving of moral experiences as universally shared contingent cognitive adaptations to foster prosociality has significant implications for both moral theory and application, which are addressed in the final section.

Applying terms describing human cognitive attitudes, emotions, and sometimes even behaviors to animals has historically met with skepticism and even indignation. To propose that other animals could have inner or social lives as complex and nuanced as our own was deemed sloppy and merely analogistic. Much of our scientific tradition has stepped through stages of progress by cutting nature at its joints in different places over time. I contend that nature has no joints. Dividing properties and phenomena into categories, as well as recognizing distinct objects and moments in time, all play useful – indeed essential – roles in our understanding of the world. But these are only useful fictions and we must keep reminding ourselves of this, or we risk adhering to an obsolete and limiting scientific and philosophical mindset. This is one sense in which nature is continuous, and it is important for our investigation here, but there are other, less grandiose, ways in which continuity plays into the explanation here.

Specifically, these concerns on differentiating phenomena apply to our understanding of moral experiences and its evolution as well. Oscar Wilde once said, "Morality, like art, means drawing a line someplace." This quote seems to capture an accepted notion, and researchers have been satisfied in drawing that line between humans and all other animals. But the differences in physiology, perception, cognition, memory, emotional range, social structure, behavior, and most features in the animal kingdom are matters of degree – and in many cases the differences are quite small. It is possible that small quantitative differences can precipitate large qualitative differences. Take, for example, certain single point genetic mutations which can make the difference between a viable organism and a stillborn mutant. These cases, though critical for establishing the "humans are unique" agenda, are scientifically disingenuous because they are the exceptions rather than the rule. To have a strong naturalistic foundation be must recognize that Darwinian evolution implies a continuity of organisms across time, and likewise a vagueness of boundaries between extant species.

If one decides to bin organisms into categories (e.g. vertebrates, mammals, dogs, or even particular animal tokens) we can be sure that those categories change over time, admit to intermediate cases, and only serve some of our purposes in categorizing animals. There are no joints in nature except the ones we draw. Appreciating the continuity will play an important role below in seeing differences as quantitative rather than qualitative. Furthermore, to take seriously the evolutionary project we must consider both variation and similarity across living organisms and the evolutionary history of organisms across time. The usefulness of separating our species (and social organization) from other animals in the categorization of extant species will not serve us well under the evolutionary account.

4.2.3.1 Building Blocks of Morality

Philosophers and anthropologists have in recent accounts considered the obviously norm-guided social behavior of animals as “building blocks” of morality [Flack and de Waal 2000, Thierry 2000, Sober and Wilson 2000]. This phrase can be taken to mean (at least) two things: 1) that morality does not apply to animals, but their behavior includes some features that are part of a genuine morality; or 2) that the behaviors observed in animals reflect only parts of modern human morality. Both take human morality as the yardstick against which to measure moral systems, but the second makes a weaker claim. The first claim depends on a definition of morality that would preclude animal behaviors from ever satisfying it. It is a claim that animals fail to embody some of the necessary conditions for morality to apply. The latter claim is different from, but compatible with, the “qualitative rather than quantitative differences” position proposed here; and in both cases the interpretation of the claim depends on what is meant by morality.

In this project we are limiting our focus to morality as a cognitive and emotional capacity adapted to generate behaviors appropriate to a species’ social environment: i.e., the individuals’ experiences when addressing issues essential to prosociality. The more complex the interaction structure, and the greater the contextual fluctuations, the more complex the behavior-generating system needs to be to cope with it. Insofar as human social structures are presumably the most complex, we would also expect that humans’ moral experiences are the deepest and most nuanced.

But social complexity among animals is a matter of difference in quantity rather than quality. We can accept that simpler animals with simpler social structures also have simpler cognitive and emotional capacities, and hence less deep moral senses with fewer contingencies and less subtle nuances. But the attitudes formed when addressing issues critical to social survival will still be of the same *kind* of experience. After all, the same proteins are responsible for deficiencies in these capabilities across mammals (at least), the same parts of the brain light up, Not necessarily the same experience, and it seems we don’t have access to what it feels like to be (say) a wolf confronted with a thieving wolf. The experience moral And it is a morality that is adapted to that animal's social structure and environment just as human morality is adapted to ours.

Though Jesse Prinz also supports a division of conventional and moral norms by the accompaniment of moral sentiments to the latter, he also holds that morality is not a native biological feature of humans (or other animals):

> Morality, like all human capacities, depends on having particular biological predispositions, but none of these, I submit, deserves to be called a moral faculty. Morality is a byproduct–accidental or invented–of faculties that evolved for other purposes[Prinz 2008].

Prinz's own account of what it takes to embody a moral attitude is in line with the story I've told:

> . . . consider a person who feels rage at someone for performing an action and would feel guilty if she herself had performed that action. On my intuitions, such a person does thereby believe that the action is morally bad [Prinz 2008].

I disagree that the feelings of rage and guilt imply the attribution of a moral belief, and instead claim that the combination and context elicit a uniquely moral flavor of feeling. Making such a judgment, and being able to reason about what other behaviors are appropriate given such a judgment, may indeed be a uniquely human, non-innate capacity for which the moral attitude is an insufficient building block. And if reasoning using moral beliefs is taken as a necessary element of any morality faculty, then I can accept Prinz's claim that this faculty is neither innate nor the direct result of an evolutionary process.

However, in the evolutionary account pursued here it is the evoking of the moral attitude, and the behavior it generates, that constitutes a *moral sense*; and the dispositions toward moral attitudes that a group of individuals making up a prosocial collective have that constitutes a *moral system*. So, I agree with Prinz on the following point: "It is an open question whether the things we emotionally condemn are really wrong, but it is not an open question whether emotionally [sic] condemnation constitutes a moral attitude [Prinz 2008]." A normative moral theory may not have evolved; in fact that may not be the kind of thing for which an evolutionary explanation is even possible. But here we are offering an explanation for why certain kinds of behaviors would become associated with moral attitudes in various contexts, and that does have a plausible evolutionary explanation.

4.2.3.2 Violence Aversion and Moral Rules

Jesse Prinz also brings up an interesting point that touches upon a distinction not made up to now. Some species are naturally not aggressive toward conspecifics (squirrels, deer, . . . , and perhaps humans) and so their lack of violence toward each other cannot be considered as following a moral rule because they have no disposition toward acting in that way. Prinz then proposes that norms against harm, a cornerstone of native morality theories [Nichols 2004], may not be innately moral at all. Rather, the practical implications of harming conspecifics sufficiently explains the aversion: "Harm prohibitions are not universal in form; they can be explained without innateness, through societal needs for stability . . . [Prinz 2008]."

There is no deep disagreement here. Any specific harming behavior's moral association can vary across contexts as appropriate for its affect on prosociality – harming itself need not be automatically and innately morally loaded. However, as already discussed, a lack of conspecific violence is typical in nature and I contend that the lack of violence is part of most group's prosocial behavioral repertoire. A corollary of that claim is that a failure to act in accordance with peaceful coexistence would trigger negative moral attitudes among the members of that group, and the perpetrators would suffer in terms of fitness and become less prevalent.

At the level of behavior and prosociality there is no difference between acting through practical implications for prosociality and through moral rules, but for Prinz the distinction is again humans' ability to reason about the payoffs of "cheating" the system. On Prinz's account, for humans to be following a moral rule they need to be doing so despite interests to act otherwise, and hence need additional non-innate driving factors to prevent violence. I will not address whether that is an essential feature of morality for a moral theory, but it certainly is not essential for explaining the evolutionary origin of a moral attitude. Behavior coheres to a norm against conspecific intragroup violence in most species, and behavior that violates this behavioral pattern has consequences for the prevalence of that behavior. This does not, as Prinz and I agree, make violence aversion moral or immoral, but it does explain why humans across cultures, and at a very young age, experience a morally flavored aversion against harm [Nichols 2004]. An attitude that, as far as we can tell from behavior, expressions, and communication, carries across non-human animals as well [Bekoff and Pierce 2009].

4.2.3.3 Shared Features across Species

We are seeing more and more that people's behavior has more in common with animal behavior than received dogma would have us believe. Simultaneously, experiments are revealing that animals' cognitive abilities are more like humans' (and each others') than superficial differences reveal. For example, many animals mourn their dead: humans, chimpanzees, gorillas, elephants, whales, dolphins, etc. This reveals an important aspect of their inner emotional life. Each of these creatures is part of a lineage that traces back to a creature that did **not** mourn its dead. Considering any one limb of the evolutionary tree (e.g., primates or cetaceans) it is reasonable to suspect that all currently existing species within that family share a common ancestor possessing this capacity, though this is not necessary. Its seems *un*reasonable to suspect that there existed a common ancestor of primates, cetaceans, and pachyderms that was emotionally and cognitively capable of mourning the dead because tracing each of the lineages back that far leads to a quite primitive animal (something akin to a rat or raccoon). The implication is that both 1) descent from a common ancestor, or 2) independent adaptation to similar contexts can explain the evolution of certain socially related emotions.

Laughter is another example. All primates laugh, and this can be traced to a common ancestor of all primates that also laughed. Elephants and cetaceans have a similar emotional response in similar circumstances, but it is behaviorally quite distinct from primate laughter. These differences stem from having evolved independently in animals that are morphologically quite different: capable of different sounds, gestures, and facial expressions. Yet the laughter reflex seems to have evolved to fill in the same niche in these social animals' behavioral repertoire. And though laughter itself may not exist in any common ancestor of these long-ago divergent lineages, some qualitatively similar, but quantitatively less developed capability exists in simpler, highly social animals: dogs, cats, monkeys, meerkats, etc. Tying back to an earlier point, it is not the case that these simpler animals have the building blocks of laughter; they have a emotional reaction that is less specialized than laughter that covers several emotions that have become distinct in primates, elephants, and cetaceans.

Social behavior among non-human animals, and the cues and contingencies they respond to, are also complex in the same ways that we say human behavior is complex. One stark example of complex social behavioral processing in animals is research that shows that domesticated dogs eavesdrop on human food-sharing communication to determine who is more generous, and then use this information to assess cooperative tendencies [Marshall-Pescini et al. 2011]. The behaviors are complex in the same *way*, just quantitatively less complex in terms of the number of social interactions, depth of contingency, variety of social problems, etc.

Even grammar, the requirement that communication signals follow specific syntactic rules, which has been long considered a uniquely human feature, has recently been shown to apply to birds as well. Furthermore, the language skills are similarly acquired postnatally and spontaneously, and a similar brain region is activated upon hearing ungrammatical sequences [Abe and Watanabe 2011]. It needn't be the case that these birds (or other animals that may share this feature) consciously reason about the grammar rules. The important part is that they are able to detect patterns, process hierarchal structures, and react to them in the appropriate way. These behavioral contingencies are of the same kind exhibited by humans, despite that these birds do not share a common ancestor capable of grammar with humans...even other primates lack these grammar skills. There are other connections between language and moral attitudes, but the point here is that a general capability to handle complexity 1) admits to degrees, 2) evolves as context makes appropriate, and 3) incrementally morphs capabilities of evolutionary predecessors.

4.2.3.4 Shared Underlying Mechanisms

In the introduction to this chapter I stated the claim that moral attitudes are those which correlate with mechanisms that generate behaviors which are necessary for achieving and sustaining a population of individuals who behave according to similar mechanisms. With respect to generating an appropriate repertoire, whether the same or drastically different mechanisms are operating makes no difference – the behavior is primary. The sameness of mechanisms is therefore not a logical requirement, or really a requirement at all, but rather a way to define group membership for prosociality that is satisfied by the constrained biological processes shaping behavior-generating mechanisms. Nature is a frugal tinkerer, and as a result the mechanisms that generate behavior are *nearly* identical across humans (regardless of whether it is analyzed at the level of proteins, genes, cellular structures, tissue structure, body plans, etc.). Furthermore, nature's conservatism results in a great many similarities throughout the animal kingdom and surprisingly little variation across the vertebrate phylum.

The high degree of similarity implies that common ancestry is a better explanation for shared mechanisms than convergence. Evolution works partly by establishing stable, modular building blocks and then combining and tweaking these building blocks to vary distinct phenotypes and behaviors [Holland 1996]. More often than not, these stable modules are carried through branching speciation events precisely because the individuals carrying them are more successful. They are tweaked in the level of expression or number of copies – matters of degree – and sometimes combined in new ways to generate phenotypic and behavioral differences. However, even when novel combinations of building blocks produce novel behavior, the underlying mechanisms are conserved (just repurposed). What this means is that even when some behavior or trait is not present in the common ancestor of two species that do possess it, the independent evolution of the trait or behavior has similar building blocks to work with, and is adapting to similar contextual pressures. Therefore, even when independent adaptation to a similar context generates shared features, similar underlying mechanisms may be generating those features.[17]

The level of mechanism consideration does matter for assessing samesness. Take as an example the disposition toward being left and right handed (laterality), a feature which is shared across (at least) many vertebrates. At one level, left-handedness and right-handedness are distinct mechanisms for generating specific behaviors, and there is a distribution of such mechanisms. Furthermore, over evolutionary time the prevalence of each disposition has been roughly maintained through adaptations to different contexts. So again it is the distribution of behaviors which is adaptive and self-sustaining. For our evolutionary purposes, then, we are concerned with the deeper mechanism that generates laterality in individuals for it is this mechanism that is robustly adaptive, continuous across species, inheritable, and also generates the particular handedness behaviors.

With respect to behavior-generating mechanisms, nothing about prosociality depends on whether we consider a dappled collection of surface-level mechanisms, or a smaller set of more basic ones, or even some single meta-mechanism. And as a consequence of this mechanism-independence it also doesn't matter if the mechanisms are described in terms of psychology, neurology, biochemistry, or particle physics. The functional description of what behaviors the mechanisms generate acts as a lossy compression of whatever is taken to be the relevant empirical behavioral description. However, for those researchers who explore these mechanisms in moral psychology, neurology, genomics, etc. it is important to keep in mind the various levels at which the mechanisms can be described. Focusing on specifics may unduly make differences salient, but by taking the evolutionary approach seriously we see that it is critical to

[17] The exception to this scenario is the truly independent evolution using distinct building blocks. The eyes of cephalopods and vertebrates are structurally very distinct and use different light-sensing proteins. However, the eyes of all vertebrates (across fish, birds, and mammals) share many similar features with only minor variation in structure, proteins, pigments, etc. because the basically same function is adaptive across the contexts faced by creatures across evolutionary time.

focus on the inheritable unit that generates behavior. Preliminary investigations that do so reveal that similarities across species are common and close – moral attitudes and their generated behavior also follow the pattern of similarity.

4.2.3.5 Interspecies Prosociality

So far we have seen that though the individuals of a species express variation in their traits (and this includes the specific mechanisms that produce behaviors), the variation is constrained by the fitness of those variations. Over evolutionary time the differences are smoothed out, leaving only those behavioral variations which play a role in the prosocial arrangements. That repertoire of behaviors, and the mechanisms supporting them, turn out to be fairly constant across groups – even of different animals. This is to be expected since they are adaptive ; i.e., must persist through competition with other similar groups. But this last point admits to one important caveat: that the individuals covered are of similar species.

I have made little mention of prosociality involving societies consisting of multiple species. Certainly nothing in the formulation prevents the individuals from being of different species, orders, or even kingdoms. It seems to be the case that human notions of morality are limited to (or at least focused on) other humans. The evolutionary foundations for the arguments here support maintaining that bias: extra-species impacts on reproductive fitness can be considered as part of the environment for intra-species selection. The important caveat arises in consideration of symbioticly adaptive multi-species groups.

All animals serve as host to a panoply of bacteria without which they cannot survive, but a clearer example of the phenomenon under consideration are lichen. The constituent organisms of lichen (fungi and either algae or cyanobacteria) can and do live separately, but not in the same forms or same conditions. The specialized arrangements among the fungi and algae cells allow both organisms to survive in locations and conditions inhospitable to either alone. The multi-species arrangement is selected for, with each member playing its specialized role, and that *arrangement* therefore qualifies as prosocial. It is exactly this scenario that inspired the name "Lichen" for the miscoordination game in section 1.8.5. So, it is not a requirement of the evolution of prosociality or moral attitudes that all included individuals operate via the same mechanisms (fungi and algae certainly do not), but it is a matter of contingent biological fact that within a species all the individuals are nearly biologically identical.

4.2.3.6 Intelligence and Moral Attitudes

Darwin understood that different environmental/social contexts make different repertoires of behaviors adaptive, and that the behaviors and arrangements that would seem moral to practitioners would track whatever social structure had been persistently adaptive. And furthermore, even though the adaptive behaviors of one group may seem bizarre and immoral to any other particular human group's sensibilities, they would still seem moral for the individuals enacting them.

> It may be well first to premise that I do not wish to maintain that any strictly social animal, if its intellectual faculties were to become as active and as highly developed as in man, would acquire exactly the same moral sense as ours. In the same manner as various animals have some sense of beauty, though they admire widely-different objects, so they might have a sense of right and wrong, though led by it to follow widely different lines of conduct. If, for instance, to take an extreme case, men were reared under precisely the same conditions as hive-bees, there can hardly be a doubt that our unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering. [Darwin 2000(1874]

However, I take issue with Darwin's assumption that the human moral experience derives from human intellectual power. This line of thinking connects with the philosophical tradition of prioritizing moral reasoning and judgment over moral sentiment and the ability to solve social problems (which need not require self-awareness, grammar-based language, or other aspects of higher intelligence). Though we can agree that humans apply their moral concepts to the manifold of experiences they encounter through reason and argument, there are few moral experiences unique to these intellectually derived prompts.

It is clear that imagining scenarios triggers moral responses, and that this process plays a role in acting in accordance with the moral (and non-moral) norms of a group. Higher intelligence allows individuals to play through more scenarios, and more complicated scenarios, to reveal the attitudes various behaviors, if actually enacted, would prompt. It seems that the tradition in moral philosophy and moral psychology is to assume that only humans are capable of such imaginative acts, but this is not the case. Many other animals (including rats, dogs, whales, elephants, and apes) clearly mentally process a situation and consider multiple courses of action before acting [de Waal 1997, Bekoff and Pierce 2009, van Schaik et al. 2003, Marshall-Pescini et al. 2011]. The degree of complication is certainly lower for rats than humans, but the contexts of rat behaviors are also simpler. This mental capacity, and the behaviors they generate, are adapted to the contexts in which these species live.

For example, a full theory of mind (the ability to put oneself in the epistemic position of another individual) is not widespread in the animal kingdom. Clearly this ability facilitates making more accurate predictions about which behavior will elicit the desired moral attitudes, but just as clearly this particular ability in inessential for having *some* moral reaction to scenarios of rats pushing a lever, sharing food, etc. to help another rat. Animals have a range of abilities to correctly assess the relative value of resources, but even quite simple animals react in an obviously morally negative way to unfair distributions of resources. Intelligence and prosocial behaviors are likely coevolutionary, but advancement in either is tightly constrained by the underlying biological mechanisms and the pressure to enhance intelligence depends on competition with other extant groups in the same niche.

4.2.3.7 Communicating Moral Attitudes

A particular intellectual capability that has played prominently in the morality literature is humans' use of language. This again may be largely accounted for by the assumptions made about the forms of moral reasoning required to evaluate and consciously refine moral norms. For our evolutionary account we are not concerned with the machinations of human intellect with no effects on behavior. We are similarly unconcerned with the details of the contingent underlying mechanisms that do generate behavior – even ones as impressive human language. One particular use of language that is of interest to the evolutionary account is the behavior of communicating moral attitudes across individuals.

The ability to communicate information is certainly not unique to humans: bees dance, ants leave pheromones, birds chirp, dogs bark, elephants trumpet, dolphins whistle, and humans speak. These forms of communication vary significantly in their ability to convey content and emotion. And the level of complication in a communication systems also parallels (and is likely coevolutionary with) social complexity. Social hierarchies are maintained through shows of power and fierce vocalizations, thieving animals are ostracized and/or punished by other individuals, and unfair treatment is widely met with disruptive and destructive behaviors. My claim is that if an animal is capable of communicating a moral attitude, then it is capable of experiencing that moral attitude.

We cannot simply ask animals what they are experiencing and communicating, but human thoughts and feeling are largely revealed through facial expressions. Even our ability to empathize with others depends on our ability to mimic others' facial expressions (whether overtly or via mirror neurons). This use of the facial expressions is not uniquely human, but shared by all higher primates. Dissection research by Ann Burrows has shown that primate faces have all the same muscles in the same places as human faces [Burrows 2008]. From rhesus monkeys and macaques to chimpanzees and humans, the same muscle movements even create the same expressions. Insofar as people can read another's emotional state from the face, so too can animals communicate in this way. Primate faces are demonstrably just as expressive as human faces, evidencing that they also posses the same range of emotions – up to the ones that can be facially expressed. Surely this list includes moral attitudes: shame, sadness, anger, indignation, pride, etc.

Other animals may not have all the same (or the same number of) facial expressions as primates, but mammal faces are extremely expressive and birds and reptiles too communicate mood and thoughts with behavior including facial expression. And other species can read the signals from other species with very high fidelity. The biological mechanisms for expressing and communicating emotional states via facial expressions, gestures, and postures is shared by all these species despite the large evolutionary distance among them. This is a primitive capability.

The other means of communication beside facial expressions clearly also play a role in sharing and shaping moral attitudes. Wagging tails, raising paws, swimming in rolls, touching trucks, grooming, mating, etc. behaviors can all play critical roles in maintaining prosociality. In some cases a behavior is both directly related to the sustainability to the group and communicative, and in other cases it is only to express approbation or disapprobation regarding another individual's, or one's own, behavior. The former behaviors can, through the process described above, become attached with a specific moral attitude over evolutionary time. Warning signals, public displays of generosity, community punishment, etc. act both to maintain the group and reinforce the norm. Humans are not expected to be able to perceive and interpret all the communicative signals of other animals, but we are able to discern attitudes from, and express our attitudes to, animals that are sufficiently complicated and/or similar. Humans can communicate emotional states with apes, canines, elephants, and dolphins because these states are shared, and these shared emotional states include (at least some, and maybe all) moral attitudes.

4.3 Summary and Concluding Remarks

This last chapter makes a broad-stroke investigation into what features we would expect of moral experience given that it has evolved. The dominant theme to the evolution of morality research thread in philosophy focuses on the evolution of specific cognitive, emotive, behavioral, and perceptual features of humans that are linked to morality. Such an approach takes some notion of human morality as fixed, and then puts forward various explanations of how humans evolved such that they can support and/or embody those moral notions. These research projects are therefore less about the evolution of morality itself, and more about the evolution of human beings' capabilities to sense, feel, discuss, analyze, and evaluate different moral notions as we have them. Research in evolutionary psychology, behavioral economics, cognitive science, and neurology are often pulled into these discussions that combine the topics of evolution and morality, but the subject is not an explanation of why moral attitudes themselves evolved.

The project here is the reverse line of thinking: determining which features of individual and group behavior promote reproductive success and investigate the role of moral experiences as a mechanism for driving such behaviors. To gain the necessary scope for prosociality one must consider the whole behavioral repertoire in an environment while facing a range of social problems. This system-level thinking about the origins of moral experiences leads us to consider which behavioral features are essential for prosociality, and what mechanisms may set those behaviors apart from other, less critical, behaviors. Because it is an evolutionary account it must cover a range of species, many of which went extinct billions of years ago. Though particular data points were used through out to provide evidence for singular claims, the form of the work is to derive general expectation about the nature of moral experiences from basic claims about the prevalence of features in competing groups.

Because natural selection operates on actual behaviors, the mechanisms that produce those behaviors are carried forward only insofar as they produce the adaptive behaviors – as a matter of contingent historical fact. Thus the psychological and neurophysiological research cannot approach the question of the evolution of moral experiences directly because those aspects are not under *direct* selective pressure. However, this research reveals that animals share similar mechanisms to produce similar behavior, and thus implies that there exists a great deal of commonality in our moral experiences. Though the creatures' behaviors and brains are less complicated, and thus likely produces less nuanced moral experiences, they are identical in purpose and associated with analogous behaviors. Seeing moral experiences as a natural feature of any evolutionarily successful group has potentially deep and important ramifications for moral theory.

4.3.1 Implications

The arguments and conclusions made here, though philosophical in import, yield specific, and empirically verifiable/falsifiable substantive claims. One implication is that human moral experiences should have parallel experiences throughout much of the animal kingdom. To be sure, many such features have already been uncovered in primates, cetaceans, and pachyderms. Still others reveal themselves in the behaviors of canines, felines, bovines, rodents, and many other social animals. The prediction implied by these arguments is that any organism which must solve a social problem of form X will have a behavioral repertoire sufficient to solve the problem with a high success rate, and that behavior will resemble the behavior of other animals that must also solve that problem.

Because other animals must solve the same social problems as humans in similar environments, the behaviors, behavior-generating mechanisms, and phenomenological experiences are also likely similar. Because those animals are cognitively less complicated, and their social lives less complex, it is reasonable that their moral experiences are less nuanced, but this is a matter of degree and not a matter of kind. A wide range of animals are therefore likely to have moral intuitions similar to humans' and react to those intuitions in similar ways. The particulars of which prompts trigger which moral attitudes are expected to differ only as much as is adaptive in a species' environment – a much lower degree of difference than the previously assumed gap in moral capabilities of humans and other animals.

This realization opens up other animal societies as models for human social behavior, and this can help us understand our own evolutionary background. Deeper investigations into the moral experiences and behaviors of other animals can also provide a wealth of material for understanding morality itself. The more we look, the more similarities and continuity we will find, and the few differences that remain will become the foci of new moral questions. This line of thinking may also impact other recurring topics in moral philosophy: the value of moral intuitions as a guide to right action, the role of moral sentiments in decisions and behavior, trends in moral norms, comparing norms among societies, viability of retraining moral intuitions, etc.

BIBLIOGRAPHY

BIBLIOGRAPHY

Kentaro Abe and Dai Watanabe. Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience*, 14:1067–1074, 2011. 270

Guillermo Abramson and Marcelo Kuperman. Social games in a social network. *Phys. Rev. E*, 63:030901, 2001. 129, 221

J. McKenzie Alexander. Evolutionary game theory, May 2003. URL `plato.stanford.edu/entries/game-evolutionary/`. 30, 38, 39

J. McKenzie Alexander. *The Structural Evolution of Morality*. Cambridge University Press, 2010. 215

Richard D. Alexander. *The Biology of Moral Systems*. AldineTransaction, 1987. 235, 246

Eleni Arapaki. Uncertainty of cooperation in random scale-free networks. *Physica A*, 388:2757–2761, 2009. 5, 221, 227

Dan Ashlocka, Mark D. Smuckerc, E. Ann Stanleya, and Leigh Tesfatsion. Preferential partner selection in an evolutionary study of prisoner's dilemma. *BioSystems*, 37:99–125, 1996. 6, 227

Salvatore Assenza, Jesús Gómez-Gardees, and Vito Latora. Enhancement of cooperation in highly clustered scale-free networks. *Phys. Rev. E*, 78:017101, 2008. 129, 221

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984. 6, 13, 29, 30, 45, 47, 222, 234

Robert Axelrod. *The Complexity of Cooperation*. Princeton University Press, 1997. 39

Robert Axelrod and D. Scott Bennett. A landscape theory of aggregation. *British Journal of Political Science*, 23(2):211–233, 1993. 9

Robert Axelrod and William D Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, 1981. 31

Jenna Bednar and Scott Page. Can game(s) theory explain culture?: The emergence of cultural behavior within multiple games. *Rationality and Society*, 19(1):65–97, February 2007. 46

Marc Bekoff and Jessica Pierce. *Wild Justice*. University of Chicago Press, 2009. 7, 31, 45, 257, 259, 268, 273

Adrian V. Bell, Peter J. Richerson, and Richard McElreath. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *PNAS*, 106(42):17671–17674, 2009. 245

James Bergin and Dan Bernhardt. Cooperation through imitation. *Games and Economic Behavior*, 67(2):376–388, November 2009. 5, 26, 227

T.C. Bergstrom. Evolution of social behavior: Individual and group selection. *Journal of Economic Perspectives*, 16:67–88, 2002. 242, 244

Cristina Bicchieri. Norms of cooperation. *Ethics*, 100(4):838–861, July 1990. 26

Ennio Bilancini and Leonardo Boncinelli. The co-evolution of cooperation and defection under local interaction and endogenous network formation. *Journal of Economic Behavior & Organization*, 70:186–195, 2009. 221

Ken Binmore. *Game Theory and the Social Contract: Volume 1: Playing Fair*. MIT Press, 1998. 6, 7, 28, 30, 31, 46, 222

Ken Binmore, J. Gale, and L. Samuelson. Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8:56–90, 1995. 229

Susan Blackmore. *The Meme Machine*. Oxford University Press, 2000. 246

Samuel Bowles and Herbert Gintis. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, 65(1): 17–28, 2004. 30, 31

Robert Boyd. Is the repeated prisoner's dilemma a good model of reciprocal altruism? *Ethology and Sociobiology*, 9(2-4):211–222, 1988. 28, 31, 46

Robert Boyd and Peter Richerson. *The Origin and Evolution of Cultures*. Oxford University Press, 2005. 6, 25, 27, 28, 30, 225, 238, 241, 245, 246, 247, 263

Robert Boyd and Peter J. Richerson. The evolution of ethnic markers. *Cultural Anthropology*, 2:65–79, 1987. 26

Robert Boyd and Peter J. Richerson. Social learning as an adaptation. *Lectures on Mathematics in the Life Sciences*, 20:1–26, 1989. 26

Robert Boyd and Peter J. Richerson. Group beneficial norms can spread rapidly in a structured population. *doi:10.1006/jtbi.2001.2515, available online at http://www.idealibrary.com on*, 215:287–296, 2002. 26, 244, 245

Robert Boyd and Peter J. Richerson. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B*, 364:3281–3288, 2009. 245

Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. The evolution of altruistic punishment. *PNAS*, 100:35313535, 2003. 222

Anne Burrows. The facial expression musculature in primates and its evolutionary significance. *Bioessays*, 30(3):212–25, 2008. 275

Charles Darwin. *The Descent of Man and Selection in Relation to Sex*. Project Gutenberg, 2000(1874). 273

Donald Davidson. *Essays on Actions and Events*. Oxford University Press, 1980. 233

Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976. 245

Ricardo de Oliveira-Souza and Jorge Moll. The moral brain: Functional mri correlates of moral judgment in normal adults. *Neurology*, 54 (Suppl. 3):252, 2000. 257

Frans de Waal. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press, 1997. 244, 259, 273

Michael Doebeli and Christoph Hauert. Models of cooperation based on the prisoner's dilemma and the snowdrift game. *Ecology Letters*, 8(7):748–766, 2005. 30

Lee Alan Dugatkin. *Cooperation among Animals: An Evolutionary Perspective*. Oxford University Press, 1997. 262

R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16:681–694, 1993. 11, 133

Robin Dunbar. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, 1998. 11, 12, 221

Paul Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books, 2003. 256

Jon Elster. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press, 2007. 25

Jon Elster. *Reason and Rationality*. Princeton University Press, 2008. 25

Ernst Fehr and Joseph Henrich. *Genetic and Culture Evolution of Cooperation*, chapter Is Strong Reciprocity a Maladaptation., pages 55–82. MIT Press, 2003. 11, 26, 31, 245

A. Feigel. Evolution of cooperation and communication skills as a consequence of environment fluctuations, February 2003. URL `arXiv:q-bio/0609034v1`. 46

Jessica C. Flack and Frans B.M. de Waal. Any animal whatever: Darwinian building blocks of morality. *Journal of Consciousness Studies*, 7:1–29, 2000. 26, 31, 222, 266

Walter Fontana. Topology of the possible. Working Paper, May 2003. URL `http://tuvalu.santafe.edu/~walter/Papers/top.pdf`. 239

Constanza Fosco and Friederike Mengel. Cooperation through imitation and exclusion in networks. *Journal of Economic Dynamics & Control*, doi:10.1016/j.jedc.:2010.12.002, 2010. 26

Nico H. Frijda. *The Emotions (Studies in Emotion and Social Interaction)*. Cambridge University Press, 1987. 256

Drew Fudenberg and David M. Kreps. Lectures on learning and equilibrium in strategic-form games. CORE Lecture Series, 1990. 36

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998. 26

Gerd Gigerenzer. *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, 2000. 14

Herbert Gintis. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, 2009. 25

Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr. Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24:153–172, 2003. 29, 245

Joshua D. Greene, Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, doi:10.1016/j.cognition.2009.02.001, 2009. 254, 257

Patrick Grim. The greater generosity of the spatialized prisoner's dilemma. *Journal of Theoretical Biology*, 173:353–359, 1995. 234

Patrick Grim, Stephanie Wardach, and Vincent Beltrani. Location, location, location: The importance of spatialization in modeling cooperation and communication. *Interactive Studies: Social Behavior and Communication in Biological and Artifical Systems*, 7:43–78, 2006. 5, 14, 26, 215, 227

Werner Güth and M. Yarri. *Explaining Process and Change Approaches to Evolutionary Economics*, chapter An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game, pages 23–34. University of Michigan Press, 1992. 31

Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. 254

Peter Hedstrom. *Dissecting the Social*. Cambridge University Press, 2005. 7, 26

D. Helbing and W. Yu. The outbreak of cooperation among success-driven individuals under noisy conditions. *Proc Natl Acad Sci USA*, 106:3680–3685, 2009. 99

B.S. Hewlett and L.L. Cavalli Sforza. Cultural transmission among aka pygmies. *American Anthropologist*, 88:922–934, 1986. 26

John Holland. *Hidden Order: How Adaptation Builds Complexity*. Helix Books, 1996. 270

David Hume. *An Enquiry concerning the Principles of Morals*. Oxford University Press, 1998 (1777). 253

David Hume. *A Treatise of Human Nature*. Oxford University Press, 2000. 29, 38

Richard C. Jeffrey. *The Logic of Decision*. University Of Chicago Press, 1990. 25

H. C. Kaner, S. G Mohanty, and J. C. Lyons. Critical values of the kolmogorov-smirnov one-sample tests. *Psychological Bulletin*, 88(2):498–501, 1980. 104

D. Marc Kilgour and Niall M. Fraser. Non-strict ordinal 2 2 games: A comprehensive computer-assisted analysis of the 726 possibilities. *Theory and Decision*, 20(2):99–121, 1986. 15

Steven Kuhn, 2007. URL http://plato.stanford.edu/entries/prisoner-dilemma/. 30

Ju-Sung Lee. The emergence of reciprocity through contrast and dissonance, August 1998. URL www.casos.cs.cmu.edu/publications/papers/asa3g2.pdf. 31

Tarmo Lemola. Convergence of national science and technology policies: the case of finland. *Research Policy*, 31(8-9):1481–1490, 2002. 26

David Kellogg Lewis. *Convention: A Philosophical Study*. Blackwell, 1969. 35

John L. Mackie. *The Cement of the Universe: A study in Causation*. Clarendon Press, 1988. 239

J. March, M. Schulz, and X. Zhou. *The dynamics of rules. In: Change in Written Organizational Codes*. Stanford University Press, 2000. 26

S. Marshall-Pescini, C. Passalacqua, A. Ferrario, P. Valsecchi, and E. Prato-Previde. Social eavesdropping in the domestic dog. *Animal Behaviour*, 81(6): 1177 – 1183, 2011. 269, 273

R. McElreath, Robert Boyd, and Peter J. Richerson. Sharded norms can lead to the evolution of ethnic markers. *Current Anthropology*, 44:122–130, 2003. 30

R. Vilela Mendes and Carlos Aguirre. Cooperation, punishment, emergence of government and the tragedy of authorities. *arXiv:0908.4408v1*, V1:1–13, 2009. 222

John Mikhail. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press, 2011. 253

Jorge Moll, P.J. Eslinger, and Ricardo de Oliveira-Souza. Frontopolar and anterior temporal cortex activation in a moral judgment task: Preliminary functional mri results in normal subjects. *Arq. Neuropsiquiatr*, 59:657–664, 2001. 257

Jorge Moll, Ricardo de Oliveira-Souza, Ivanei E. Bramati, and Jordan Grafman. Functional networks in emotional moral and nonmoral social judgments. *NeuroImage*, 16:696–703, 2002. 257

Oskar Morgenstern and John Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 1980. 29, 32, 34

M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett. 103, 058701 (2009).*, 103:058701, 2009. 130, 131

M.E.J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126, 2003. 224

Shaun Nichols. *Sentimental Rules.* Oxford, 2004. 258, 259, 268

Nikos Nikiforakis. Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, 68(2):689–702, March 2010. 222

Douglass North. *Institutions, Institutional Change and Economic Performance.* Cambridge University Press, 1990. 254

Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314: 1560–1563, 2006. 26, 27, 215

Martin A. Nowak and Robert. M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, 1992. 5, 227

Martin A. Nowak, Karen M. Page, and Karl Sigmund. Fairness versus reason in the ultimatum game. *Science*, 289:1773–1775, 2000. 229

Martin A. Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Letters to Nature*, 428:612–646, 2004. 25

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory.* MIT Press, 1994. 34, 35

Elinor Ostrom. *Governing the Commons.* Cambridge University Press, 1990. 2

Matjaz Perc. Evolution of cooperation on scale-free networks subject to error and attack, 2009. URL `arXiv:0902.4661v1`. 221

Matjaz Perc and Attila Szolnoki. Coevolutionary games - a mini review. *Biosystems*, 99:109–125, Oct 2010. URL `arXiv:0910.0826v1`. 26, 30

Plato. *The Collected Dialogues of Plato: Including the Letters*, chapter Symposium, pages 526–574. Princeton University Press, 2005. 29

Jesse Prinz. *Gut Reactions: A Perceptual Theory of Emotion.* Oxford University Press, 2004. 256

Jesse J. Prinz. Is morality innate? In W. Sinnott-Armstrong, editor, *Moral Psychology.* Oxford University Press, 2008. 267, 268

Anatol Rapoport and Albert M. Chammah. *Prisoner's Dilemma.* Univeristy of Michigan Press, 1970. 25, 29, 222

Anatol Rapoport and M. J. Guyer. A taxonomy of 2 x 2 games. *General Systems*, 11:203–214, 1966. 15, 32

Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, 2003. 129

Rick L. Riolo, Michael D. Cohen, and Robert Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414:441–443, 2001. 7

Carlos P. Roca, Jose A. Cuesta, and Angel Sanchez. Promotion of cooperation on networks? the myopic best response case, 2009. URL `arXiv:0901.0355v1`. 221

Jean-Jacques Rousseau. *Rousseau's Political Writings: Discourse on Inequality, Discourse on Political Economy, On Social Contract*, chapter Discourse on the Origin and Foundation of Inequality Among Men, pages 3–57. W. W. Norton & Company, 1987. 38

Donald G. Saari. Geometry of chaotic and stable discussions. In *Mathfest talk in Boulder Colorado*, 2003. 249

F.C. Santos, J.M. Pacheco, and T. Lenaerts. Cooperation prevails when individuals adjust their social ties. *PLoS Comp. Biol.*, 2:12841291, 2006. 6, 227

Thomas C. Schelling. *Micromotives and Macrobehavior.* W.W. Norton and Company, 1978. 35

Dinghua Shi, Stuart X. Zhu, and Liming Liu. Clustering coefficients of growing networksstar, open. *Physica A: Statistical Mechanics and its Applications*, Volume 381:515–524, 2007. 129

Charles R. Shipan and Craig Volden. The mechanisms of policy diffusion. *American Journal of Political Science*, 52(4):840–857, 2008. 26

Brian Skyrms. Signals, evolution and the explanatory power of transient information. *Philosophy of Science*, 69:407428., 2002. 26

Brian Skyrms. *The Stag Hunt and Evolution of the Social Contract.* Cambridge University Press, 2004. 6, 27, 28, 30, 38, 39, 46

Brian Skyrms and Robin Pemantle. A dynamic model of social network formation. *PNAS*, 97(16):9340–9346, 2000. 6, 227

J. Maynard Smith. Group selection and kin selection. *Nature*, 201:1145–1147, 1964. 242, 244

J. Maynard Smith. Evolution and the theory of games. *American Scientist,*, 64:41–45, 1976. 32

J. Maynard Smith and G. R. Price. The logic of animal conflict. *Nature*, 246: 15–18, 1973. 32

Elliot Sober and David Sloan Wilson. *Evolutionary Origins of Morality*, chapter Summary of Unto Others: The Evolution and Psychology of Unselfish Behavior, pages 185–206. Imprint Academic, 2000. 266

Pawel Sobkowicz. Analysis of norms game in networked societies, 2003. URL `arXiv:cond-mato310444v1`. 5, 221, 227

Pawel Sobkowicz. Simulations of assortative matching: Complexity of cooperation emergence process, 2004. URL `arXiv:con-mat040222`. 6, 227

Chandra Sripada. Punishment and the strategic structure pf moral systems. *Biology and Philosophy*, 20:767–789, 2005. 222

Attila Szolnoki and Matjaz Perc. Emergence of multilevel selection in the prisoners dilemma game on coevolving random networks. *arXiv: 0909.4019 vl*, v1: 87.23–89.75, 2009. 6, 227

Leigh Tesfatsion. Structure, behavior, and market power in an evolutionary labor market with adaptive search. *Journal of Economic Dynamics and Control*, 25:419–457, 2001. 6, 227

B. Thierry. Building elements of morality are not elements of morality. *Journal of Consciousness Studies*, 7(1-2):60–62(3), 2000. 266

Marco Tomassini, Enea Pestelacci, and Leslie Luthi. Mutual trust and cooperation in the evolutionary hawks-doves game. *BioSystems*, 99(1):50–59, 2010. 6, 32, 227

Robert L. Trivers. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57, 1971. 7, 26, 31

Carel P. van Schaik and Cheryl D. Knott. Geographic variation in tool use on neesia fruits in orangutans. *American Journal of Physical Anthropology*, 114 (4):331–342, 2001. 263

Carel P. van Schaik, Marc Ancrenaz, Gwendolyn Borgen, Birute Galdikas, Cheryl D. Knott, Ian Singleton, Akira Suzuki, Sri Suci Utami, and Michelle Merrill. Orangutan cultures and the evolution of material culture. *Science*, Vol. 299(5603):102–105, 2003. 263, 273

D. J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–442, 1998. 129

Michael Weisberg. Robustness analysis. *Philosophy of Science*, 73:730–742, 2006. 212

R. R. Wilcox. Some practical reasons for reconsidering the kolmogorov-smirnov test. *British Journal of Mathematical and Statistical Psychology*, 50(1):9–20, 1997. 104

D. S. Wilson. Altruism in mendelian populations derived from sibling groups: The haystack model revisited. *Evolution*, 41 (5):105–1070, 1987. 242

Michael L. Wilson and Richard W. Wrangham. Intergroup relations in chimpanzees. *Annu. Rev. Anthropol*, 32:363–92, 2003. 31

Bin Wu, Da Zhou, Feng Fu, Qingjun Luo, Long Wang, and Arne Traulsen. Evolution of cooperation on stochastic dynamical networks. *PLoS ONE*, 5(6): e11187. doi:10.1371/journal.pone.0011187, 2010. URL `e11187.doi:10.1371/journal.pone.0011187`. 6, 227

V. C. Wynne-Edwards. *Evolution Through Group Selection*. Blackwell, 1986. 242, 244

Dong-Ping Yang, Hai Lin J. W. Shuai, and Chen-Xu Wu. Individual's strategy characterized by local topology conditions in prisoner's dilemma on scale-free networks. *Physica A*, 388:2750–2756, 2009. 5, 129, 221, 227

H. Peyton Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993. ISSN 00129682. URL http://www.jstor.org/stable/2951778. 36

H. Peyton Young. Innovation diffusion in heterogeneous populations: Contagion, social influence and social learning. *American Economic Review*, 99: 18991924, 2009. 26

M.G. Zimmermann and V.M. Eguíluz. Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. *Phys. Rev. E*, 72, 2005. 6, 227