UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Metagames: The Evolution of Game-Changing Traits

by

Adam Jackson

Thesis for the degree of Doctor of Philosophy

11th July 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
Electronics and Computer Science

Doctor of Philosophy

METAGAMES: THE EVOLUTION OF GAME-CHANGING TRAITS

by Adam Jackson

Cooperative social behaviours are ubiquitous in nature and essential to biological theory, yet they present an apparent paradox since cooperators benefit others while potentially incurring a fitness cost. The standard resolution is that cooperation is evolutionarily stable if cooperative behaviours are positively assorted, so their benefits are directed at other cooperators, shifting the problem to explaining the presence of positive assortment. If we view individuals as playing an evolutionary game, then the evolution of assorting traits changes the rules of the game to allow for greater cooperation.

This is one of many ways that individuals can evolve game-changing traits that modify their social niche, since social interactions occur in social environments that are in part the product of evolved traits. We investigate this by introducing a game-theoretic model of metagames where the evolution of individual strategies changes the social game. Because of mathematical equivalences between game-changing mechanisms and pay-off matrix transformations, we can use metagames as a common framework to model the coevolution of social games and social conditions.

Instead of simply identifying the conditions under which cooperation evolves, metagames explain how these conditions arise by identifying the circumstances under which the conditions for cooperation evolve. While positive assortment on social traits is necessary for cooperation to be stable, we show that alone it will only allow game-changing traits promoting cooperation to evolve when cooperation is already favoured. This is insufficient to explain the evolution of assortment in the Prisoner's Dilemma. We find that much as assortment on social traits is crucial to the evolution of cooperation, assortment on game-changing traits is crucial to the evolution of social assortment. Because assortment on social and game-changing traits are connected, this assortment has been hidden in existing accounts. We are able to characterise the relationship between the two types of assortment, and show how assortment on game-changing traits can enable the evolution of social assortment. We can therefore explain the evolution of the conditions assumed necessary for the evolution of cooperation.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Adam Jackson, declare that the thesis entitled *Metagames: The Evolution of Game-Changing Traits* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (Jackson and Watson, 2013) and (Jackson and Watson, In Prep)

Signed:.....

Date:.....

Acknowledgements

I could not have produced this thesis without the support and guidance of many people.

Thanks to Richard Watson, for being a very helpful and patient supervisor, knowing when I needed support and when I needed a more stringent approach. It took a while but you have me convinced on the deep learning analogy.

To Jason Noble, for recruiting me in Odense. Wishing you all the best in your next adventures!

To Seth Bullock and Nicki Lewin, for creating such a great community at the ICSS.

To the many people who have contributed to PhD life at Southampton: Paul Ryan, always a source of expansive insight and fearless refusal to affirm the consequent. Simon Powers, whose work sparked much of what I have gone on to study. To Simon Tudge and all the other members of our group for keeping the environment in the lab so professional.

To Guy; fundamental.

To James, Chris and Maike, excellent housemates, and Rich and Joe and all the other fantastic people at the ICSS – good luck in your future endeavours!

To Jon, Sam, Stu, Mul and Joe, for your sophisticated conversation.

Finally to my family, for being awesome.

Academic Collaborations and Paper Contributions

A number of primary and contributing author papers were produced over the course of this thesis, either directly or indirectly, in collaboration with others at the University of Southampton.

Jackson and Watson (In Prep) reports the entirety of Chapter 5, introducing the motivations and details of the metagame model and presenting the key results. I coded and ran all iterations of the model and wrote the paper, with my supervisor Richard Watson providing invaluable guidance, feedback and direction. It was his original suggestion to model a metagame in which different payoff matrices would compete.

Jackson and Watson (2013) introduces *interaction functions* (§4.6) and applies them to understanding a simulation model in which there is assortment on both social and game-changing traits (§6.2). Richard Watson suggested the idea of a gravitational attraction model to demonstrate assortment in an agent based model. I then developed the idea of interaction functions as a more rigorous generalisation of our previously *ad hoc* geometrical method of calculating assorted matrices.

Jackson and Noble (2012) (extended abstract for ALife XII) reports the results of an evolutionary algorithm built around an agent-based model of the coevolution of a group size preference (inspired by Powers and Watson (2011)) with agent-based models of dominance hierarchies (based on Hemelrijk (1999)). Though the model does not feature in this thesis, it informed the exploration of social niche construction (§1.3, §6.1) and the potential role of dominance hierarchies in evolutionary transitions (§§2.5.4.2). I programmed and ran the model under the guidance of Jason Noble.

In the contributing author papers, Doncaster et al. (2013a,b), lead author Patrick Doncaster combines (via the replicator and Lotka-Volterra dynamics) evolutionary game theoretic models and ecological competition into a model that can account for the evolution of altruism in the absence of population structure, and an evolutionary pathway for altruistic behaviours to evolve from parasitic interactions. I contributed by pushing the idea of changing games on the ST-plane within the group (Chapter 4).

Watson et al. (2015), lead written by my supervisor Richard Watson, presents a broad and ambitious account of the emergence of biological organisation taking advantage of deep analogies between evolutionary and learning systems. This paper synthesises many of the results obtained in our group, including the metagame model (Chapter 5) as a formalism of social niche construction which provides an explanation for evolutionary transitions.

Chapter 1

Introduction

The classical view of nature was conflict and competition, red in tooth and claw – an understanding in accord with the Darwinian maxim of 'survival of the fittest'. The unforgiving mathematics of natural selection left little room for cooperation or altruism: how could it when cooperators benefit others while potentially incurring a cost to their own fitness? Yet now we recognise that not only is cooperation ubiquitous in nature, it is also essential to understanding fundamental questions in biological theory, such as the origins of organismal life. Individuated organisms composed of many parts are such a pre-philosophical part of our experience of the natural world that they can seem unremarkable, but we are increasingly revealing the stunning complexity of the human organism. This is not just the 100 trillion cells that contain our DNA, but the ten times that number of bacteria that make up the human microbiome (Berg, 1996); bacteria that are not just parasites but participants in mutualistic relationships vital to the organism's functioning. The human organism is the result of a staggering cooperative endeavour on such a scale that is hard to fully comprehend, yet despite its vast size and complexity it is considered a individual entity. How could such a system possibly evolve?

We see this same complexity repeated, simultaneously, on multiple levels. The *eukaryotic* cells that form the body of any multicellular organism are themselves the evolutionary product of symbioses with bacteria that now form vital organelles such as mitochondria. The vast number of unicellular lifeforms attest that the cell is a complex living entity in its own right. At a higher level, the colonies of eusocial insects such as *Atta* genera of leaf cutter ants also display characteristics of a biological individual in that colonies beget other colonies. Eusocial colonies consist of millions of sterile worker ants divided into physically differentiated castes, they harvest leaves to grow a fungus that is in a mutualistic relationship with the colony and only the queen caste is physically able to reproduce (Hölldobler and Wilson, 1990). Though it is the queens that reproduce, it is the colony they are reproducing. This parallels the division between germ-line and somatic cells in animals, where only the germ-lines cells have the potential to directly pass on genetic information to the next generation of animal, and the somatic cells contribute

to that reproduction much as the sterile worker caste ants support the reproduction of the queens.

This hierarchical organisation of the biological world, where organisms are composed of organs, that are composed of cells, that are composed of organelles, was for a long time taken as a given – or at least as a feature of the biological world exogenous to the theory of evolution by natural selection. But Buss (1987) and the influential Maynard Smith and Szathmáry (1997) drew attention to the fact that this hierarchical structure is an evolved feature of the biological world, one that has arisen by a series of major transitions in evolution. A major transition has taken place when 'entities that were capable of independent replication before the transition can replicate only as part of a larger whole after it' (Maynard Smith and Szathmáry, 1997).

This perspective implies that the evolution of new levels of biological organisation is a fundamentally social process (Bourke, 2011). Almost by definition it has to be, since major transitions transform a collection of distinct independently-replicating entities into a new higher-level individual. At the heart of this process is the development of altruistic and strongly cooperative traits by the members of a population undergoing a transformation; indeed, a successful transition typically requires one of the most extreme examples of a cooperative trait in nature — reproductive altruism, where somatic cells or sterile workers sacrifice their own direct reproductive opportunities to increase the reproductive success of a germ-line cell or queen. The naive evolutionary account, that a trait spreads through a population when individuals possessing that trait produce more offspring than those that don't, breaks down when the trait is one that causes sterility in its bearers, yet this reproductive division of labour is common to most eusocial insect colonies and multicellular organisms.

So cooperation (and its extreme form altruism) are important to understanding interesting and important questions in biological theory. But cooperation is always vulnerable to selfish cheats, and altruism, which leads to a net loss of fitness to the altruist, is by its very definition evolutionarily unstable. Though it can appear and spread via methods like population bottlenecks, a population of altruists should always be susceptible to invasion by mutants. Because of this apparent paradox, the evolution of cooperation has been the subject of considerable theoretical study (such as Hamilton, 1964a,b; Axelrod and Hamilton, 1981; Eshel and Cavalli-Sforza, 1982; Michod and Sanderson, 1985; Nowak and May, 1992; Lehmann and Keller, 2006; Fletcher and Doebeli, 2009; Powers, 2010; Stewart and Plotkin, 2014). The standard resolution is that cooperation is evolutionarily stable if cooperative behaviours are positively assorted, so their benefits are directed at other cooperators (Godfrey-Smith, 2009). Cooperative behaviours that benefit other individuals at the expense of the actor can only be stable if the benefits of cooperation are directed at other cooperators (Hamilton, 1964b; Michod and Sanderson, 1985; Lehmann and Keller, 2006). Positive assortment can arise from many proximate mechanisms, such as genetic relatedness, signalling ('greenbeard' effects) and repeated

interactions that allow for reciprocation (Axelrod, 1987). Other resolutions to the apparent paradox modify the costs and benefits of the social interaction more directly, through mechanisms like policing or side-payments (Jackson and Wilkie, 2005).

This theoretical work has identified the conditions necessary for cooperation to evolve, but has often left these conditions as part of the biological background against which social organisms evolve, much as used to be the case with the biological hierarchy. While a population's social conditions will be constrained by external factors, such as resource abundance, recent work has demonstrated ways that social populations can change their social context. Individual traits, such as those that affect population structure, provide a way for evolution to alter the incentive structure of social interactions. For example, a genetic trait that affects the dispersal radius of an entity's offspring will change the likelihood that members of a population will encounter relatives (Pepper and Smuts, 2002), while one for group size will affect the composition of social groups by changing the potential for variance within and between groups (Powers and Watson, 2011).

This recognition is at the heart of an emerging approach to social evolution that goes beyond looking for evolutionary explanations for cooperation in isolation to consider the coevolution of cooperation and the social conditions that can support it, a process of *social niche construction* (Powers et al., 2011; Ryan et al., 2016). Niche construction is the idea that individuals can change their environment to change the selective pressures on their own evolution (Odling-Smee et al., 2003; Laland and Sterelny, 2006). Social niche construction translates these ideas to social evolution: social organisms change the social conditions and so modify the selective pressures on their social evolution. Social niche construction models have found that the coevolution of cooperative strategies and traits that modify the social conditions can feed back into each other to enable ever-higher levels of cooperation (Powers, 2010).

If we think of social interactions as a game, and the participants as players with different types of strategy, then these traits are changing the incentives of the game that is being played. From this perspective, part of a major transition is the process of changing the incentives of social interactions so that a strategy such as total reproductive altruism (obligate sterility to aid another's reproduction) becomes evolutionarily viable. Traditional accounts of social evolution that have been satisfied with explaining cooperation as an adaptive response to specific conditions have been akin to seeking the conditions under which various games will result in cooperative equilibria. But it is equally important to understand why populations play particular games and how the rules of those games can be changed.

This raises a new, higher-level question which we investigate in this thesis: when individual adaptations can change the social game is there a general tendency towards new games that favour increased cooperation or increased selfishness? We introduce a game-theoretic model of *metagames*, a formal framework for studying the coevolution

between social games and the social conditions under which they are played in a principled manner. These social conditions are influenced by the evolution of *game-changing traits*: traits which change the incentive structures of social interactions, such as by influencing population structures or introducing policing. This provides a minimal model for the coevolution of social strategies and game-changing traits that affect the social context, and the feedback between the two. We develop the metagames formalism in the abstract and then apply it to a basic class of games that includes the canonical social dilemmas seen in the evolution of cooperation, such as the Prisoner's Dilemma. Using mathematical equivalences between game-changing mechanisms and transformations to the payoff matrix of a game, we can represent the evolution of social games and social conditions in a common framework.

This allows us to use metagames as a formal framework for characterising the effect of natural selection on game-changing traits that modify social conditions that provides new insights into evolutionary processes. Instead of simply identifying the conditions under which cooperation evolves, metagames explain how these conditions arise by identifying the circumstances under which the conditions for cooperation evolve.

We find that when cooperation is already favoured, such as by positive assortment on social traits, this is sufficient to enable game-changing traits that further favour cooperation to evolve. However, this is insufficient to explain the evolution of assortment in social conditions where cooperation requires (strong) altruism. We find that much as assortment on social traits is crucial to the evolution of cooperation, assortment on game-changing traits is crucial to the evolution of assortment. Because assortment on social and game-changing traits is not orthogonal — for instance genetic assortment affects both — the role of assortment on game-changing traits is hidden in existing accounts of social niche construction. We are able to characterise the relationship between these types of assortment under many different conditions. This allows us to demonstrate how assortment on game-changing traits allows populations to change the incentives of their social games when cooperation in those games would initially require strong altruism. We demonstrate theoretical results of metagame models under a wide-range of scenarios that gives new insight into the processes of social niche construction, and the evolution of the conditions assumed necessary for the evolution of cooperation.

1.1 Game-Changing Traits and the Major Transitions

In the natural world, we can see evidence for the subtle yet powerful consequences of the major transitions in the difference between the Portuguese Man o' War and the jellyfish. At first sight it is difficult to distinguish between the two, so it is not surprising that historically the Man o' War was believed to be a type of jellyfish. But they are actually very distinct types of life. Like humans, jellyfish are a single complex multicellular

organism. The Portuguese Man o' War is not. Instead it is a colonial organism, composed of many smaller organisms called zooids. It is the distinction between these two types of life that this thesis is motivated to explore: the processes that transform a group of interacting entities into something that is an individual in its own right; that explain why the biological world is full of complex individuated organisms like the jellyfish and not just aggregates of smaller organisms. These are the evolutionary processes that are at work in a something like the Man o' War — a snapshot of a transition in progress — in which the zooids are sufficiently integrated that they are no longer capable of independent existence, but are still structurally distinct unlike the cells of a jellyfish.

This research programme, substantially sparked by the publication of *The Major Transitions in Evolution* (Maynard Smith and Szathmáry, 1997), has led to an important conceptual shift that has brought the explanation of the observed hierarchical structure of biological life within the purview of evolutionary theory. This is part of what has been a general trend towards endogenising within the theory of evolution by natural selection aspects of observed biological reality that were previously treated as exogenous parameters to the theory (Okasha, 2006). Darwin's initial theory had little to say about the mechanics of inheritance, fair meiosis, mutation rates or sex ratios, but gradually these processes have been incorporated into evolutionary theory, where before they were taken as a background against which evolutionary processes would take place.

It has been argued that these major transitions are in fact social phenomena, extreme examples of social integration (Michod, 2000; Queller and Strassmann, 2009; Bourke, 2011). This perhaps counter-intuitive insight has placed social behaviours, in particular cooperation, at the heart of explaining the evolution of new levels of individuality and at the forefront of contemporary theoretical evolutionary biology. How can transitions in the organisation of life be connected to the evolution of social behaviours? The answer is actually almost definitional; for groups to transform into a new type of entity a whole range of adaptations must take place that are essentially social in character. Individual entities must develop adaptations to form groups, then develop further adaptations to stabilise and maintain those groups so that they will persist for long enough to be affected by evolutionary pressures, and then must develop adaptations that will transform the group to a new level of individual. All of these adaptations are social adaptations that pertain to individuals forming and living in social groups, whether those social groups are composed of microbes or primates.

The greatest obstacle to a social group transformation is the inherent tension between selfish individuals and the good of the group. A vital constituent of a social group transformation then is the evolution of *conflict modifiers* (Michod and Roze, 2001) that reduce within-group conflict, allowing the closer integration necessary for a social group transformation. Indeed Michod (2000) argues that this is a prerequisite for a transition — that a social group cannot have undergone a major transition until it has evolved

adaptations that reduce group conflict at the lower level. Common features of higher-level individuals, such as developmental bottlenecks, reproductive specialisation and high levels of within group relatedness, all serve to align the fitness interests of the members of a social group and enable a social group transformation to take place. When we view the entities undergoing a major transition as playing a social game, the evolution of conflict modifiers represents the evolution of traits that change the social game.

So we see the major transitions in evolution are important to this thesis for three reasons:

- They are an account of a fundamental and inherently interesting problem in theoretical biology that has the evolution of social behaviours at their core. In this way they motivate the development of metagames as a model that can bring new insight to the evolution of cooperation.
- They are an example of the endogenisation within the theory of evolution by natural selection of processes that were previously placed beyond that theory in the same way that metagames make the payoff matrix of a social game a variable rather than a parameter of the model.
- They exemplify the importance of transformations in the social conditions of a population so that one structure of social organisation (a collection of entities capable of independent replication) could become another (a new higher-level entity which the previous entities can only replicate as part of). Features of major transitions, such as the development of unicellular bottlenecks, are examples of game-changing traits.

1.2 Modelling Social Evolution with Evolutionary Game Theory

If the major transitions in evolution provide the motivation for this thesis, evolutionary game theory provides the tools for its analysis of the evolution of cooperation. Evolutionary game theory is the standard mathematical formalism for modelling the evolution of social behaviours (Maynard Smith, 1982), adept at modelling situations in which the fitness of an individual committed to a particular evolutionary strategy depends not just on the inherent qualities of that strategy but on the frequency with which it and other strategies are found in the population. This is the case for the evolution of social behaviours, where the fitness consequences of social actions depend on the social behaviours of others.

The classic example due to Maynard Smith (1982) is the 'Hawk-Dove' (or Snowdrift) game. We imagine a population subdivided into aggressive 'hawks' and pacifistic 'doves' all competing over territory. In this example each individual (all of the same species) is

either of the hawk-type or the dove-type, not a mixture of both. A hawk will not back down in encounters and will always attack another individual, whereas a dove will retreat when threatened. When a hawk encounters a dove it will win the territory. However, when a hawk encounters another hawk they will both engage in a costly conflict. On the other hand, when two doves are in the same territorial patch they will coexist and share the benefits. So in general it is better to be a hawk when the population is composed mostly of doves, but as the number of hawks increases the cost of repeated fighting also increases. The expected population frequency of hawks and doves depends on the exact costs and benefits, but will in general feature both hawks and doves, as the aggressive behaviour of the hawk type has a homeostatic effect on its frequency in the population, leading to a polymorphic population equilibrium in which both types are present. This example illustrates how evolutionary game theory shifts the (expected) fitness of a social individual from something determined against a fixed landscape to a dynamic product of competing strategies (Weibull, 1997).

The general evolutionary game theoretic model consists of a number of genetically determined strategies with associated fitness payoffs for interactions with other strategies. Taken together these fitness payoffs define the payoff matrix of the game. The changing frequencies of the different strategies in the population can then be modelled using the *replicator dynamics* (Taylor and Jonker, 1978). For many games, including all those in this thesis, these frequencies will reach an equilibrium state. Of particular interest are those equilibrium states that are stable — *evolutionarily stable states*(Maynard Smith, 1982) to which the population will return when the frequencies are subject to small perturbations.

Traditionally, evolutionary game theoretic work has taken the payoff matrix of the game as fixed by the nature of the social situation, and then studied the resulting evolution of behaviours. However, this methodology is being challenged by a growing realisation that often it is not obvious what the social game is, and that predefining the payoff matrix can be arbitrary while at the same time having a significant effect on the results (Fort, 2008). Roughgarden (2009) argues for an inversion of the fixed game approach, instead advocating two-level models of behavioural evolution in which the behavioural timescale of goal-driven interactions is separated from the microevolutionary timescale of population genetics. In particular, the payoff matrix of the behavioural game may be subject to evolutionary pressure. Akçay and Roughgarden (2011) found that mutants making side payments can lead a behavioural game with a payoff matrix in the form of a Prisoner's Dilemma to evolve towards a payoff matrix with features of the Hawk-Dove game. Similarly, Worden and Levin (2007) looked at how a population might evolve out of a Prisoner's Dilemma, a social game that does not support cooperation, by periodically introducing additional mutated strategies. Fort (2008) has produced a model to determine social games without specifying the payoff matrix by evolving an initially heterogenous set of games on a spatial lattice, finding that the when the initial

population was drawn from the standard social dilemmas the resultant game would have the form of a Stag Hunt coordination game supporting a non-zero population of cooperators.

At the same time another strand of the literature has drawn attention to the fact that social behaviours do not occur in a vacuum; they occur in population structures that are themselves evolved features. Models of social niche construction (Powers and Watson, 2011) have looked at the circumstances under which population structures supporting increased group selection effects could evolve. It is known that between-group effects can outweigh within-group competition when there is high variance in group composition caused by sampling small groups from a large population (Wilson and Colwell, 1981). When social strategies coevolve with a group size preference, linkage disequilibrium develops, linking a preference for small groups with the cooperative social trait and creating a positive feedback loop. Thus individual-level adaptation drives the population into a state more conducive to cooperation.

In models of the evolution of cultural markers (Snowdon et al., 2011), limited mixing between sub-populations engaged in different behaviours produces a selective pressure in favour of evolving markers correlated with those behaviours; this then has an indirect effect on the level of positive assortment. In models of ecosystems with evolving symbiotic partnerships (Watson et al., 2011a) the evolution of symbiotic partnerships expands the basin of attraction of the globally optimal outcome until it is the only possible outcome, individual-level selection enhancing the chance of a group-optimal result. In the same way that the major transitions programme has begun to endogenize the existence of the biological hierarchy into evolutionary theory, this work has demonstrated the importance of and provided the foundation for bringing population structure into the remit of evolutionary explanation.

Finally, a number of different formal mathematical equivalences have been developed between changing the social context in which a game is played and changing the game itself. Maynard Smith (1978) raised the fact that the fitness of strategies in the Hawk-Dove game would be be different if the game was played by relatives instead of unrelated individuals. Subsequently, formulae for equivalent games have been produced for situations in which there are effective changes to the payoff matrix due to relatedness (Grafen, 1979), reciprocity, kin selection and group selection (Taylor and Nowak, 2007). Other approaches have looked at the way the game changes when populations subdivided into particular group structures affect the replicator dynamics (Van Veelen, 2011). Others have given generalisations of games played in a well-mixed population by considering games played on graphs or networks, such as evolutionary graph theory in which the graph structure evolves (Lieberman et al., 2005), and by explicitly investigating the effective game created by a network structure (Cao et al., 2011). Different network topologies, such as fully mixed or scale free, can change the social equilibria of games played on that network (Santos et al., 2006a), while dynamically changing the

network topology can also be equivalent to transforming the payoff matrix of the game (Pacheco et al., 2006b,a).

We draw three particular conclusions from these strands of work running through the literature:

- That the nature of a social game, represented by its payoff matrix, can be subject to evolutionary control.
- That the population structure in which a social group interacts can be subject to evolutionary control.
- That changes in population structure can be equivalent to changing the payoff matrix of the social game that population is engaged in.

1.3 Metagames: Modelling Social Niche Construction

The synthesis of these three conclusions is the motivation for the introduction of the theory of *metagames*. We use evolutionary game theory to model the evolution of social individuals possessing game-changing traits that influence the conditions of their own evolution. As the population evolves these game-changing traits, such as conflict modifiers, the incentive structure of its social interactions changes, though the physical nature of the interactions might remain the same. The social group members are still faced with the same possible strategies, but because of the evolved conflict modifiers the benefits of the different strategies change. In essence, the social game has changed.

To represent the effect of these game-changing traits, we use *interaction functions*, a generalised way to transform a game to reflect the effect of traits that cause types to interact non-randomly with each other (at different frequencies to those at which they are found in the population). Interaction functions provide a principled method to derive payoff matrix transformations that let us find the effective game the population would be playing without the action of the game-changing trait, complementing other techniques for such transformations (Grafen, 1979; Ohtsuki and Nowak, 2006; Taylor and Nowak, 2007; Boyd et al., 2010; Van Veelen, 2011).

With metagames, we can model the coevolution of social strategies and payoff matrices. The mathematical equivalence between game-changing traits and transformations to payoff matrices means we can use metagames to model the coevolution of social strategies and game-changing traits. We can answer questions about the evolution of social games and game-changing traits in a common framework.

In particular, we claim that metagames provide a model of social niche construction that provides new insight into the conditions necessary for 'runaway' increases in co-operation to take hold. Powers (2010) argues that 'any component of selection on

structure-modifying traits that is due to social behaviour must be in the direction of increased cooperation, rather than increased selfishness.' Metagames provide another way of formally analysing this result. In metagames models of social niche construction, it is the game-changing traits that are modifying the social niches, such as by changing population structure to increase positive assortment of social interactions.

We find that the extent of game-space over which cooperation-promoting game-changing traits can evolve is determined by the level of assortment on the game-changing trait itself. When there is no assortment on the game-changing trait, pro-cooperation game-changing traits (such as those which increase assortment on the social trait) will only evolve in regions of game space where cooperation is already favoured. For game-changing traits that promote assortment to evolve in a Prisoner's Dilemma, there must be a 'second-order' assortment on the game-changing trait itself. In this way, metagames allow us to rigorously characterise the conditions under which social niche construction can take place. We are able to characterise the relationship between the two types of assortment, and show how assortment on game-changing traits can enable the evolution of social assortment. Therefore we can provide explanations for the evolution of the conditions assumed necessary for the evolution of cooperation.

1.4 Publications

The work in this thesis has contributed to the following publications and manuscripts:

- Jackson, A. and Noble, J. (2012) Coevolution of aggression and group size preferences for social agents in Dominance Hierarchies. *ALife XIII* (Extended Abstract)
- Doncaster, C., Jackson, A., and Watson, R. (2013a). Manipulated into giving: when parasitism drives apparent or incidental altruism. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758).
- Doncaster, C. P., Jackson, A., and Watson, R. A. (2013b). Competitive environments sustain costly altruism with negligible assortment of interactions. *Scientific reports*, 3.
- Jackson, A. and Watson, R. A. (2013). The effects of assortment on population structuring traits on the evolution of cooperation. In *Advances in Artificial Life, ECAL*, volume 12, pages 356–363.
- Watson, R. A., Mills, R., Buckley, C., Kouvaris, K., Jackson, A., Powers, S. T., Cox, C., Tudge, S., Davies, A., Kounios, L., et al. (2015). Evolutionary connectionism: algorithmic principles underlying the evolution of biological organisation in evo-devo, evo-eco and evolutionary transitions. *Evolutionary Biology*, pages 1–29.

- Jackson, A. and Watson, R. A. (2016). Metagames: A formal framework for the evolution of game-changing behaviours. *Under submission.*

In the primary authored papers, Jackson and Watson (2013) introduces *interaction functions* (Chapter 4) and applies them to understanding a simulation model in which there is assortment on both social and game-changing traits (Chapter 6). Jackson and Watson (In Prep) reports the work of Chapter 5, introducing the motivations and details of the metagame model and presenting the key results.

In the contributing author papers, Doncaster et al. (2013a,b) use the development of ST-space and preliminary ideas behind the metagames model more fully developed in this thesis as part of a new account, combining game theory and ecological modelling, of a pathway for altruistic behaviours to evolve from parasitic interactions by effectively changing the game. Watson et al. (2015) take the metagame model as a formalisation of social niche construction, a theory that gives an explanation for evolutionary transitions, as part of a broad synthesis of the emergence of biological organisation.

1.5 Thesis Structure

This thesis is divided into three sections. The first three chapters make up the first section. These chapters introduce the existing technical and non-technical background for the claims of the thesis, explain its motivations and discuss the relevant debates in the literature. In this chapter we have provided an overview of the arguments, introducing the significance of social evolution, its connection to fundamental topics in theoretical biology such as major evolutionary transitions and how this motivates the idea of a metagame.

Chapter 2 is a broad review of the literature on social evolution. It introduces the main terminology and concepts, such as inclusive fitness and Hamilton's rule, and details the major transitions as social phenomena. It discusses the process of a major transition, through which a social group becomes a higher level individual. We detail proposed mechanisms through which a social group transformation is achieved, and how these can be seen as game-changing traits.

Our modelling formalism is evolutionary game theory, which is introduced in Chapter 3, including the importance of equilibria and dynamics in understanding the behaviour of the social game. We emphasise the class of two-player two-strategy symmetric games, which are thoroughly classified, and the projection of these games onto the two dimensional ST-plane and the conceptual benefits this provides.

The second section builds the technical framework used to support the thesis. It is mostly comprised of original modelling work. Chapter 4 looks at the mathematical

correspondence between game-changing traits and transformations of the payoff matrix of that game. We review examples of these transformations in the literature. We also explain how this is motivated by the importance of positive assortment to the evolution of cooperation. We demonstrate that increasing levels of assortment change the effective payoff matrix of the game. Then we generalise this by introducing *interaction functions*, which modify the replicator dynamics for a well-mixed population to represent population structures in which types encounter each other at non-random frequencies; the earlier results for positive assortment emerge as a special case.

Chapter 5 then introduces metagames, a novel evolutionary game theoretic model in which the payoff matrix of the game also evolves. First we situate the model in the literature and examine related approaches, then we define metagames in the abstract and demonstrate the behaviour of metagames over the ST-plane. Because of the connection between game-changing traits and transformations to payoff matrices we can use metagames to model the coevolution of social strategies and game-changing traits that modify social niches. We claim this means that metagames provide a formal model of social niche construction. We analyse the evolutionary dynamics for a range of game-changing traits. We show conditions for Snowdrift and Stag-Hunt games to be transformed into either Prisoner's Dilemmas or Harmony Games, depending on the type of game-changing trait involved, the constraints on the metagame and the initial frequencies of the social strategies. Importantly, we show that changes to the game driven by selfish individuals maximising their own payoff can still result in games that support increased cooperation.

We also apply metagames to studying the evolution of assortment, creating a model where the level of assortment in the game is not an imposed parameter but a variable that can evolve. We find that the evolution of increased assortment is favoured in regions of game-space where cooperative outcomes are already favoured, so cooperation can beget the conditions for even more cooperation. However, this disagrees with results from the theory of social niche construction that the evolution of cooperation promoting traits will be favoured in that assortment does not spread in a Prisoner's Dilemma.

In Chapter 6 we examine and attempt to resolve this disagreement. We hypothesise that assortment on the game-changing trait itself plays an important role in the evolution of game-changing traits that support increased cooperation. In the third section of this thesis we pursue novel modelling work to investigate this hypothesis.

First we determine how to model assortment on game-changing traits. We start by developing a simulation model of an agent-based population that is subject to different kinds of assortment, which demonstrates that the assortment on social and game-changing traits can have complex interactions that reduce the success of cooperation while aiding the spread of the social conditions that could support more cooperation in the future. We then assess the suitability of two candidate techniques to model assortment on the

game-changing trait in a metagame, settling on the interaction functions we introduced in Chapter 4.

With this technique we start to model how assortment on the game-changing traits affects the evolution of those game-changing traits. In Chapter 7 we extensively model the effects of game-changing trait assortment when that game-changing trait is one that creates assortment on the social trait, so there is 'first-order' assortment on social traits governed by the game-changing traits, and 'second-order' assortment on the game-changing traits themselves. This chapter serves as a stepping stone where we examine a lot of different scenarios but focus on understanding the behaviour of the model. We analyse many different variables at different points on the ST-plane and develop summary measures to present this data in an intelligible way.

This lays the groundwork for Chapter 8 where we apply our understanding to study key models of the evolution of game-changing traits. We model the paths the effective social dilemma a population plays can take through the space of games, plotting the trajectories of these stochastic paths through game space. We see the effect of increasing assortment on the game-changing trait by calculating the mean paths from a single point in ST-space and see that sufficient assortment on the game-changing trait means populations engaged in a Prisoner's Dilemma that would evolve game-changing traits to further favour defectors will instead take paths through to the Snowdrift and even the Harmony Game. We assess the significance of assortment on the game-changing trait for the evolution of cooperation by showing how the evolutionary paths of populations through game space changes the social equilibria of the games to allow for increased cooperation across an increased region of ST-space.

Finally in Chapter 9 we draw together the results of all the modelling work and discuss the insights it has provided. We assess how metagames can model social niche construction and conclude that assortment on game-changing traits can drive the evolution of the conditions necessary for cooperation.

Chapter 2

Social Evolution

In this chapter we review the growing importance of the study of social evolution in evolutionary biology. We explain the main terminology and theoretical frameworks used in social evolution theory and in the rest of this thesis. We settle on concrete definitions for terms such as 'cooperation' and 'altruism' which we have used informally thus far.

We then examine in more detail the central role social interactions assume in the major transitions in evolution. We explain the concept of inclusive fitness, which Bourke (2011) places at the centre of the major transitions, and how accounting for inclusive fitness can change the incentives for a social interaction.

We provide a detailed review of the nature and structure of the major transitions, and draw a distinction between those transitions that see a social group transform into a higher-level individual and those that do not. We introduce Bourke (2011)'s decomposition of these social major transitions into three stages: social group formation, social group maintenance and social group transformation. We discuss the philosophical conception of the biological individual, with particular emphasis on Folse and Roughgarden's evolutionary criteria for individuality (Folse and Roughgarden, 2010) and Godfrey-Smith's idea of of a 'Darwinian population' (Godfrey-Smith, 2009).

We then consider some of the specific mechanisms involved in social group transformation, such as bottlenecks, the evolution of reproductive specialisation (the germ-soma distinction) and the size-complexity hypothesis. These mechanisms can be understood as game-changing traits, and serve as an important motivation for the rest of this thesis, where we model the coevolution of game-changing traits and social behaviours.

2.1 The Expanding Domain of Social Evolution

> Social evolution has grown outwards from the study of the beehive and the baboon troop to embrace the entire sweep of biological organisation. It

> claims as its subject matter not just the evolution of social systems narrowly defined, but the evolution of all forms of stable biological grouping, from genomes and eukaryotic unicells to multicellular organisms, animal societies, and interspecific mutualisms. — Bourke (2011, p. 5)

For a long time, the historical study of social evolution was restricted to studying the perplexing appearance of selfless, altruistic acts in a few select species such as ants. These rare cases were viewed as an important test for evolutionary theory; Darwin himself wrote in *The Origin of Species* that the apparent contradiction between altruistic behaviour and selfish natural selection was the most important challenge to his theory. But such examples of altruism were supposed to be rare, the biological world a Hobbesian state of nature, red in tooth and claw. The frequent conflation of cooperation and altruism reinforced a view of nature in which cooperation was an anomaly because of a universal temptation to defect, where the 'Prisoner's Dilemma' and its multiplayer generalisation 'the tragedy of the commons' were the default models for social interaction (Worden and Levin, 2007). Theoretical work focused on how to explain the possibility of cooperation in the face of such pressure to defect (May, 1981).

Now though, sociality is becoming recognised as one of the most pervasive forces in evolution, vital not only in the obvious contexts such as the evolution of complex societies, but also as a driving force behind deep questions like the hierarchical organisation of the biological world (Buss, 1987; Maynard Smith and Szathmáry, 1997). Whether the population of social interactors is one of genes, cells, or organisms, the evolution of behaviours governing the relations between these entities is fundamental to explaining how they form the next level of the hierarchy. Recent work such as *The Principles of Social Evolution* (Bourke, 2011) present broad syntheses of research on the major transitions with social evolution at the centre. The details of this thesis are contentious, such as Bourke's claim that Hamiltonian inclusive fitness theory is sufficiently powerful to explain the breadth of the major transitions, but that social evolution plays a significant role is not.

2.2 Social Evolution Terminology

At this point it is useful to set out definitions of the biological meaning for social descriptors that have been imported into biology from everyday discussion of human interactions. We define a *social interaction* as an interaction between two or more 'individuals' that has an effect on the fitness (expected number of offspring) of both the actor and the recipients. Social interactions encompass a broad range of behaviours, such as fighting over reproductive opportunities or territory, providing food for offspring, and producing (or failing to produce) the materials needed for the maintenance of a microbial biofilm (Velicer, 2003). As ever there is some dispute over this definition. It presupposes

a transactional model of social interactions and overemphasises actor-recipient pairwise interactions. But the definition is simple enough that it is clear and provides theoretical utility: using this definition Hamilton categorised social interactions into four groups based on the fitness consequences for the actor and the recipient (Hamilton, 1964b). There is still considerable disagreement over the classification that extends beyond differences in terminology to reveal semantic differences over the nature and evolution of altruism and cooperation (Rosas, 2010). However, these definitions will be used throughout this thesis as they have gained some measure of traction in the literature, though the major sources of disagreement will be highlighted where they occur.

2.2.1 Altruism

The actor's fitness decreases, the recipient's increases

Altruistic behaviours that appear to decrease the fitness of the altruist for a gain in the fitness of another have traditionally been viewed as a major challenge for evolutionary theory to explain. It seems *prima facie* apparent that such behaviours should be heavily selected against. Nevertheless there are many instances of altruism in nature. Particularly significant is the existence of somatic cells in multicellular organisms and obligate sterile worker castes in eusocial societies, both of which sacrifice their own reproductive capabilities in favour of reproductive specialists. The dilemma in these cases can be resolved by the looking at the level of selection; altruistic behaviours at one level being an evolved characteristic of a higher level group. A broad and congruent explanation from the genic point of view is Hamilton's theory of *inclusive fitness*. This theory is based on the realisation that a gene can spread not just by enhancing the reproductive success of a carrier, but also if the carrier enhances the reproductive success of other carriers. Thus altruistic genes can spread if the cost to the actor is less than the total benefits conferred both on the actor and the recipients weighted by a regression of their relatedness to the actor.

The definitions of altruism are part of the battleground over group selection, sometimes mistakenly characterised as between group selection and kin selection/inclusive fitness — this is not necessarily the case, though the impression is understandable given that the opponents of group selection have typically come from the inclusive fitness school. A standard group selection definition of an altruistic trait is that put forward by Sober and Wilson (1998): any trait that causes altruists to have lower fitness than selfish traits within groups, but where between groups those groups that contain higher numbers of altruists have higher fitness. This definition recasts the nature of altruism in terms of the group-individual distinction. It is also more permissive as it includes social behaviours that would not count as altruism under the definition we use, such as those that result in positive net fitness benefits to both actor and recipient, and reciprocal altruism.

One problem with the group selection definition of altruism is that altruism can evolve between kin in a well mixed population with no clearly defined groups. Sober and Wilson argue that this is in fact another incidence of group selection, but that groups should be understood in a more general sense as something generated by patterns of interactors. This may be too liberal a definition of a group (Maynard Smith, 1998), but as Okasha (2006) points out, this expansive definition of a group to admit kin selection and reciprocal altruism is not just about what to call a group; it recognises an essential similarity between the evolutionary mechanism at work in evolving these behaviours — 'that kinship, group structure and reciprocation are all ways of getting altruists to direct their benefits onto each other.' This is an important point, but if the essential mechanism is the generation of assortment, it makes more sense to base such a definition of altruism on assortment than to excessively widen the definition of a group to encompass all instances of assortment. This is precisely what is done by Rosas (2010) — altruism is defined as any behaviour that requires positive assortment between altruism to evolve. This viewpoint will be discussed more substantially in the later discussion of assortment.

There are also some theorists who argue from an ecological perspective that altruism is not a monolithic behavioural category, and is in fact divided into different subtypes with different evolutionary requirements. Van Dyken and Wade (2012a) gives a breakdown of altruism into four categories: *survival altruism* such as nest defence and alarm calling; *fecundity altruism* as in social insect worker care of queens; *resource-enhancement altruism* like provisioning, agriculture, and livestock rearing; *resource-efficiency altruism* as in pack hunting and communal foraging. It is not clear though if all these behaviours are altruism or cooperation in the narrower sense, and the differences between them may lie more in proximate rather than ultimate mechanisms for the origin of altruism. This is still an important practical classification, but not so relevant to the distinctions considered in this thesis. It is also important to note that there are many potential pathways to altruism and not all are cooperative; ecological models suggest that the origins of some altruistic behaviours may in fact derive from parasitism by the recipient of the altruistic act; the interaction can then take on the characteristics of altruism if the altruist is able to direct this enforced fitness donation towards relatives (Doncaster et al., 2013a).

2.2.2 Cooperation (narrow sense)

Both the actor and recipient increase in fitness.

Cooperation in this narrower sense is also sometimes known as *weak altruism*. Its labelling in this scheme as cooperation, as opposed to weak altruism, deliberately reflects the desire of the inclusive fitness tradition to distinguish between the two types of interaction based on net fitness effects for the individual. Inclusive fitness, group selection and

assortment based definitions of altruism would all agree though that there is a difference between the two. Cooperative behaviours of this form can evolve under a wider range of conditions than pure altruism. The individual fitness of both individuals increases, so this kind of cooperation does not necessarily require inclusive fitness considerations to evolve. However there is still the risk that a defector behaviour could emerge to take advantage of the benefit of the cooperative act without assuming the cost, so cooperation would not always be favoured in a thoroughly mixed population. However it can evolve in a group structure even when group formation is random rather than assortative (though including the actor in the accounting can mean even randomly formed groups have positive assortment (Pepper, 2000). On the other hand others have argued that too often cooperation in this narrower sense has been conflated with altruism, both conceptually and terminologically, and that this has focused too much attention on the apparent 'paradox of cooperation' present in the stricter sense of altruism and away from the many mutually beneficial forms of cooperation (Worden and Levin, 2007).

2.2.3 Selfishness

The actor's fitness increases, the recipient's decreases.

Selfish or exploitative interactions were the paradigmatic understanding of social interactions in nature. These types of behaviour are common, observed, for example, in territorial conflict, dominance interactions and parasitism.

2.2.4 Spite

Both the actor and recipient decrease in fitness.

Spite is rare in nature but an example are bacteria that release individually costly toxins against competing strains (Gardner and West, 2006b). As with altruism, some interactions that might be described as purely spiteful on one level, such as suicidal nest defence in social insects, can also be viewed in the context of a higher-level individual.

2.3 Inclusive Fitness

One of the more important concepts in social evolution theory is inclusive fitness. The motivation for inclusive fitness was the apparent paradox of cooperation that Darwin remarked upon in the *Origin of Species*. For a time the dominant school of thought was what is now termed *naive group selection*. Altruistic and other self-sacrificing traits emerged 'for the good of the species'. But an important conceptual development came in the idea that as well as the fitness of an organism, one can think of the fitness of a gene.

Genes with a positive fitness spread through a population, genes with a negative fitness decrease in frequency. A gene that on average causes a fitness decrease to its carriers must go extinct, so genes that lead to altruistic behaviours in their carriers must have positive fitness even if some phenotypic carriers engage in altruistic behaviours that reduce their own fitness, lest it is to go extinct. Once the possible separation between the fitness of a phenotype and a gene is recognised, it becomes clear that actions can have different effects on the fitness of phenotypes and on genes. In particular, an altruistic gene can have a negative fitness effect on a particular carrier phenotype, but will have positive fitness as a gene if the positive fitness benefit to the recipients falls on other carriers of the altruistic gene in such a way that the total benefits outweigh the cost. For a single altruistic gene in a single population model of a diploid sexual individuals, this entails weighting the benefits falling on each recipient by a regression on the relatedness by descent of the actor and recipient (Hamilton, 1975). This then gives *Hamilton's rule* (Hamilton, 1964b):

$$rb > c \tag{2.1}$$

Where r is the regression coefficient, b the benefit to the recipient and c the cost to the actor. Following on this scheme, we can think of dividing the personal fitness of an individual — the number of offspring — into a component of *direct fitness*, that might include characteristics of the individual like size, and an component of *indirect fitness* contributed to by the social acts of others. From this decomposition one can than define an individual's *inclusive fitness*: it is the individual's direct fitness plus the sum of its indirect fitness contributions to other individuals, weighted by the relatedness regression.

What motivates this method of partitioning fitness? One advantage is that it is believed to be the fitness quantity that is under an individual's control. If one thinks of individuals as attempting to maximise something, then inclusive fitness is what they are maximising (West et al., 2011). This is part of a school of thought that is using the mathematics of optimisation theory to claim that individuals behave *as if* they are optimising their inclusive fitness (Grafen, 2006); the argument goes that even if the language of 'individuals maximising fitness' is an inaccurate agential metaphor for the process of evolution, if the theory is mathematically equivalent then the metaphor can be safely used accompanied by the qualifying 'as if'. Whether this represents a useful project or a substantial commitment to rescuing a bad but seductive metaphor is open to debate. A more concrete objection to the maximisation paradigm comes from Rosas (2010), namely that it requires a very liberal definition of what is under an agent's control to count all inclusive fitness effects in this. The strength of a donated indirect fitness contribution to the inclusive fitness accounting is not determined purely by the act of helping, but also by how that help falls. This distribution of contributions often depends on traits that, for instance, control assortment with kin.

However, Rosas' proposed alternative accounting technique, neighbour modulated fitness (NMF), is not necessarily an improvement. NMF is calculated by taking the direct fitness of the individual in question and adding the indirect fitness benefits received by the individual weighted by the regression coefficient as in indirect fitness. So NMF is actually just the same as traditional personal fitness. It is unclear that any accounting based on NMF avoids what Rosas terms the illusion of the paradox of altruism — that "if you think that the altruistic gene will spread in spite of the fact that donors are losers, you are the victim of an illusion: although some donors will be losers, they are not losers on average, and cannot be, if altruism evolves." There is an apparent confusion here between the fitness of genes and the fitness of phenotypes. Inclusive fitness accounting (or indeed any accounting) is clear that a gene will never spread if it has an average negative fitness effect, but phenotypic expressions can have markedly different fitness. Rosas acknowledges this heterogenous distribution and that it can give the appearance of paradox, but claims it is only the average that matters. But the average of personal fitness over all carriers of a gene is essentially just the fitness of the gene; on this particular point Rosas seems to be just restating Hamilton's original point that phenotypes can be subject to an apparent paradox of altruism, but it is an illusion on the level of genes. Nevertheless the fact that there is an apparent paradox given a phenotypic description is significant, as it is this that motivates the need for mechanisms such as assortment and fitness contributions to or from other individuals to resolve it.

Inclusive fitness is also an important to the idea of changing evolutionary games, because it has been recognised since Maynard Smith (1978) that the fitness of strategies in evolutionary games can differ if games are played between relatives than if they are played by unrelated individuals. Grafen (1979) produced an early model for how the inclusive benefits accrued from the whole population will change the fitness benefits to a player of the Hawk-Dove Game. We will see in Chapter 4 how this coincides with our own models of game transformation.

2.4 The Major Transitions in Evolution as Social Phenomena

The *major transitions in evolution*, as defined by Maynard Smith and Szathmáry, are a series of events in evolutionary history after which 'entities that were capable of independent replication before the transition can replicate only as part of a larger whole after it' (Maynard Smith and Szathmáry, 1997). For example, the cells of a multicellular organism are descended from unicellular protists that could survive independently; today these cells can only survive as part of the larger whole. Similarly, in advanced eusocial insect colonies such as the leaf-cutting *Atta* ants, individual worker ants are infertile, able to reproduce only in the sense of aiding the reproduction of the colony. In the language of inclusive fitness, the direct fitness of the sterile worker is zero, but there

is a non-zero indirect fitness contribution to the fitness of members of the reproductive caste — though this neat decomposition is challenged by the argument in Birch (2012a) that the extreme redundancy of large colonies means that each worker may contribute a vanishingly small amount to the inclusive fitness of the whole colony for the same cost of its direct fitness, so mechanisms like coercion may be necessary to sustain this arrangement.

Table 2.1: The major transitions in evolution, reproduced from Maynard Smith and Szathmáry (1997)

Replicating molecules	⇒	Populations of molecules in compartments
Independent replicators	⇒	Chromosomes
RNA as gene and enzyme	⇒	DNA + protein (genetic code)
Prokaryotes	⇒	Eukaryotes
Asexual clones	⇒	Sexual Populations
Protists	⇒	Animals, plants, fungi (cell differentiation)
Solitary individuals	⇒	Colonies (non-reproductive castes)
Primate societies	⇒	Human societies (language)

The complete list of transitions identified by Maynard Smith and Szathmáry are given in Table 2.1. One important caveat is that some of the major transitions, like the enclosure of replicating genes to form the first cells or the evolution of sex, are believed to be unique events in the history of life on Earth; while for others like the origins of multicellularity and eusocial societies there is evidence that the transition has occurred many times independently. Drawing together all the major transitions into a single conceptual framework is an elegant and heuristically satisfying idea, but we must be cautious not to overstate the inevitability of such transitions; there is no guarantee that were the history of life on Earth to be replayed they would emerge again in the same manner.

When looking at major transitions as examples of social evolution, a useful distinction can be made between *fraternal* and *egalitarian* transitions, terminology introduced by Queller (1997) in his review of *The Major Transitions in Evolution*. A fraternal transition is one in which the base level entities are related, while an egalitarian transition involves unrelated entities. The transitions to multicellularity and eusocial societies are examples of fraternal transitions, while the origin of the eukaryotic cell is an example of an egalitarian transition. Inclusive fitness theory predicts that different social interactions are likely to evolve between related and unrelated organisms (Bourke, 2011) — in particular that altruism will not occur in fraternal transitions. This is supported by the fact that participants in an egalitarian transition do not lose their ability to reproduce, an example being the organelles inside eukaryotic cells (Queller, 1997). On the other hand kin selection effects provide a pathway for the evolution of altruistic non-reproductive castes in the fraternal transitions.

2.5 The Process of a Major Transition

The complete list of major transitions given by Maynard Smith and Szathmáry includes transitions that do not result in the creation of a new entity that would be termed an 'individual', such as the introduction of language in the transition from primate to human societies. There is some consensus on the subset of the major transitions that do result in the emergence of a higher order individual (Bourke, 2011; Godfrey-Smith, 2009) — these are given in Table 2.2:. In this thesis it is these transitions we are most interested in. Bourke (2011) breaks the process of these major transitions into three stages:

- *Social group formation* in which the social group is established
- *Social group maintenance* through which the social group is stabilised
- *Social group transformation* in which a stable social group becomes an integrated whole that can be termed an individual.

The key steps entailed in each process for the different major transitions are shown in Table 2.2.

2.5.1 Social Group Formation

The social group formation stage of a major transition is the stage in which social groups are created from non-social populations, such as through the spread of genes for social behaviours. The mechanisms for social group transformation differ depending on whether it is an egalitarian or fraternal transition.

Egalitarian transitions between unrelated partners depend on the evolution of mutualistic relationships. Mutualisms can be classified into *open* or *closed* according to if new partners are sought in each generation or if a partnership persists across generations through vertical descent (Leigh Jr, 1995). The mutualism between male and female gametes and the reproduction of fungus growing termite colonies are all open mutualisms, as male and female gametes seek out new partners each generation and termites seek out a new source of fungus. By contrast, the mutualism between mitochondria and eukaryotic cells and the reproduction of fungus growing ants where virgin queens carry samples of the parent colony's fungus are examples of closed mutualisms. It is hypothesised that closed mutualisms derive from open parasitic relationships that switch from horizontal to vertical transfer, leading to a greater coincidence of fitness interests (Thompson, 2005).

In fraternal transitions based on groupings of related organisms there are two known pathways: subsocial and semisocial groupings. Subsocial groupings arise when offspring

Table 2.2: The major transitions that lead to a new level of individuality and the main processes of social group formation, maintenance and transformation entailed, reproduced from Bourke (2011)

Major evolutionary transition	Examples of phenomena at each stage		
	Social group formation	Social group maintenance	Social group transformation
Separate replicators (genes) ⇒ cell enclosing genome	Origin of compartmentalized genomes	Control of selfish DNA	Evolution of large, complex genomes
Separate unicells ⇒ symbiotic unicell	Origin of eukaryotic cells	Control of organellar reproduction	Evolution of hybrid genomes through transfer of genes from organellar to nuclear genome
Asexual unicells ⇒ sexual unicell	Origin of zygotes	Control of meiotic drive	Evolution of obligate sexual reproduction
Unicells ⇒ multicellular organisms	Origin of multicellular organisms	Control of selfish cell lineages (cancers)	Evolution of segregated, early-diverging germline
Multicellular organisms ⇒ eusocial society	Origin of societies	Control of conflict with dominance, punishment, or policing	Evolution of dimorphic reproductive and non-reproductive castes
Separate species ⇒ interspecific mutualism	Origin of interspecific mutualisms	Control of cheating (sanctions)	Evolution of physically conjoined social partners

do not disperse, remaining near their parents. On the other hand, in a semisocial grouping individuals from same generation form groups together. Single cell bottleneck stages in the development of multicellular organisms and eusocial insect colonies indicate the predominance of the subsocial pathway, as these force subsocial groupings, though there are examples of the semisocial path being followed in slime mould where cells aggregate.

The formation of social groups is likely to be due to a combination of ecological pressures and inclusive fitness benefits. Ecological conditions that affect dispersal or food supply may provide incentives for (or even enforce) group living. The widespread occurrence of highly related social groups demonstrates that inclusive fitness effects are another important driver of group formation; individuals receive increased indirect fitness benefits from social traits by grouping with other related individuals.

2.5.2 Social Group Maintenance

During the social group maintenance phase of a transition, adaptations develop that maintain group stability, a necessary precursor to the integration of that group during social group transformation. The processes of social group maintenance include those that protect the group against external threats such as parasitism, and those that stabilise it against internal selfish cheats. To counter external threats, the components of the social group must develop recognition of group members versus non-group members, either directly through mechanisms like external chemical cues for nestmate recognition in social insects, or indirectly by engineering environments in which interactions will be with other members of the social group by default. Inevitably these systems will be imperfect because they are in an evolutionary arms race against exploiters (Dawkins and Krebs, 1979).

Meanwhile internal group maintenance is a key process mediating between the benefits to the individual of selfish behaviour and the negative effect on the group. For example cancer cells can be understood as selfishly reproducing somatic cells (Buss, 1987). These are the kinds of behaviours that were once explained by invoking naive group selection arguments. As selfish mutations increase in frequency within the group the fitness of the group as a whole reduces. Though it is now recognised that group selection effects cannot be taken for granted, it is possible for social group to evolve adaptations that increase the efficacy of group level selection — the fact that major transitions occur proves this to be the case.

The theory of *social niche construction* presented in Powers et al. (2011) provides one argument for how individuals can come to live in a population structure that both supports high levels of cooperation within the group and makes it individually advantageous to do so, assuming individuals have genetic preferences for group structure. When there exist two possible population structures, one of which supports higher levels of cooperation than the other then, as long as individuals with a genetic preference for one structure or another assortively form groups of their preferred structure, the cooperation-favouring structure trait will spread. Linkage disequilibrium will evolve, as individuals with cooperative traits will be more successful in the cooperation-favouring structure, linking cooperative traits with cooperation-favouring population structuring traits.

This process of social niche construction gives a general argument for how social groups can evolve to support increasingly greater levels of cooperation. The theory of metagames we develop in this thesis replicates this same process, and indeed we claim that metagames can serve as a formal model for social niche construction. We expand on the connection between metagames and social niche construction in Chapter 6.

The evolution of a range of more specific mechanisms is also important to social group maintenance. When group members are not related, the development of a shared reproductive fate enhances the stability of both the groups themselves and the group living arrangement, seen in the development of the vertical transmission of mutualisms, or fair meiosis in the reproduction of the genome. This discourages selfish mutations because defectors in groups with a shared reproductive fate incur greater fitness penalties. Policing and coercion are also important in maintaining many social groups; analysis of social insect colonies shows that when relatedness is not complete, such as if queens are multiply-mated, policing plays a vital role in reducing the amount of egg laying by workers (Wenseleers et al., 2004).

2.5.3 Social Group Transformation

2.5.3.1 What Constitutes an Evolutionary Individual?

The study of evolutionary transitions in individuality raises the immediate question —- how do we know when a social group has undergone such a transformation? Is there a discrete point where the generations before are all social groups and the generations after are individuals, or is the process part of a continuum? If evolutionary transitions in individuality are continuous processes it would mean that individuality must be a continuous concept, an idea at odds with the everyday understanding of individuality. So it is important to consider more carefully exactly what constitutes an individual.

The traditional view of individuality in biology could be loosely defined as 'one genotype in one body'. Santelices (1999) identified three traditional attributes of individuality: physiological unity, genetic homogeneity and genetic uniqueness. However, this intuitive conception of individuality is very much based on a relatively small number of example taxa such as humans and other large animals (Buss, 1987); it breaks down when extended to less familiar branches of the tree of life. For example, eusocial insect colonies lack physiological unity though they are widely accepted as evolutionary individuals. In many plants the processes of growth and reproduction are tightly linked; what might look like a new individual 'quaking aspen' tree is in fact part of hundreds or thousands of trees all linked by a common root system from which they grew, and all possessing the same genotype (leaving aside any mutations) (Godfrey-Smith, 2009).

Problematic cases such as these have led many authors (eg. (Folse and Roughgarden, 2010; Bourke, 2011; Godfrey-Smith, 2009)) to propose an evolutionary definition of individuality. Folse and Roughgarden (2010) examined different ways researchers were understanding individuality and proposed a set of three nested criteria for an evolutionary individual — alignment of fitness, export of fitness from lower to higher levels by germ-soma specialisation and adaptive functional organisation. Their argument is that

genetic relatedness plays an important role in facilitating cooperation but is not logically necessary or sufficient for individuality.

Alignment of fitness is the broadest criterion. As discussed earlier, the alignment of fitness is a vital step in social group formation, whether it occurs through shared reproductive fate, genetic bottlenecks or high levels of relatedness between group members. However, though a definition of an individual that allowed for parts of the individual to have differential fitness would be too broad, alignment of fitness is not a sufficient criteria for individuality, otherwise it would admit clonal colonies like the Portuguese Man o' War where almost perfect relatedness means fitness interests are closely aligned.

The second nested criteria is that the components of the social group are interdependent on each other for reproduction leading to fitness export from lower level group members to the higher level individual. The paradigm example of this comes from the evolution of reproductive specialisation, where a non-reproductive caste exports all their fitness to the reproductive of the group as a whole. This process shall be covered in more detail later in this chapter. While obviously related to the first criterion, this criterion marks the difference between the potential for cooperation inherent in aligned fitness interest to the realisation of that cooperation. For example a mass of genetically and phenotypically homogenous but undifferentiated cells that would pass the first criteria fails this one.

The third criteria is functional integration at the level of the group sufficient that it can display coordinated action and adaptations at the level of the whole group. Adaptation at the level of, for example, an organism, can be taken as evidence that the whole organism is a locus of fitness (Folse and Roughgarden, 2010). This property is stronger than the existence of group level behaviours where the locus of fitness remains the individual. Consider a hypothetical example of multi-level selection operating on deer herds. The deer herd has a number of properties such as herd speed that could be said to have a fitness benefit for members of the herd. However, it would be difficult to justify the claim that a deer herd is an individual. Godfrey-Smith makes the point that it is important to consider the level at which reproduction is occurring — if a herd level process is being selected for it should contribute to differential reproduction of deer herds (Godfrey-Smith, 2009). In this case, the locus of fitness remains the deer despite the many emergent properties of deer herds.

These criteria for individuality also suggest the different paths that must be taken during egalitarian and fraternal transitions to create a new evolutionary individual. In a fraternal transition between related individuals, alignment of fitness is already present to some extent due to relatedness, and kin selection effects provide a pathway for reproductive interdependency to evolve. However the related individuals are likely to initially be homogenous (due to the high degree of relatedness), so the evolution of division of labour within the group leading to functional integration is a key step. By contrast, in

an egalitarian transition division of labour would be expected to be present from the outset; it is the benefits of functional integration between disparate entities that leads to the mutualisms necessary for an egalitarian transition. Here then the important step is the evolution of mechanisms to align fitness interests.

The volvocine algae are a source of excellent test cases for definitions of individuality because they include species ranging from unicells to complex groups with sufficient reproductive integration to meet the criteria for individuality (Michod, 2007). Figure 2.1 pictures different types of volvocine species at different points along the transition to multicellularity. *Gonium pectoral* and *Eudorina elegant* group in clumps of related cells that pass only the first criterion, *Pledorina californica* and *Volvox carterie* groups feature germ-soma differentiation meeting the second criterion, while colonies of *Volvox aureus* possess sufficient functional integration to be considered evolutionary individuals.

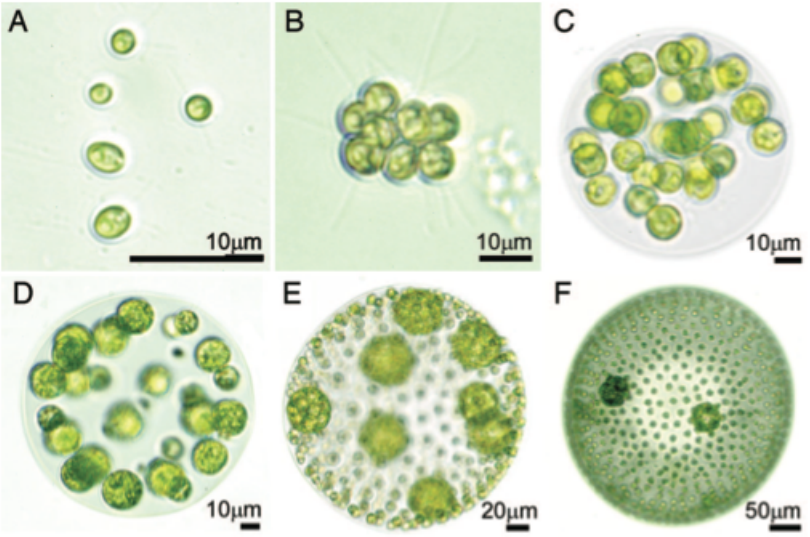


Figure 2.1: Volvocine species at different stages in the transition to multicellularity, reproduced from Michod (2007). A. — *C. reinhardtii*, a unicell. B. — *Gonium pectora*l, 8-32 undifferentiated cells. C. — *Eudorina elegant*, a spherical colony of 16-64 undifferentiated cells. D. — *Pledorina californica*, a spherical colony with 30-50% somatic cells. E. — *Volvox carterie*. F. — *Volvox aureus*

2.5.3.2 Darwinian Individuals

As seen in the example of deer herds and in accord with the major transitions viewpoint, there is often the potential for individuality to exist simultaneously at multiple levels. A useful conceptual tool in these borderline cases of individuality is Godfrey-Smith's idea of a *Darwinian population* — a population subject to evolution by natural selection (Godfrey-Smith, 2009). A *Darwinian individual* is defined as a member of a Darwinian population. One of the attractive features of Godfrey-Smith's formalism is that it is not prescriptive, giving a rigid definition of individuality that would inevitably fall to some

example of biological peculiarity, but breaks down the processes needed for evolution by natural selection to work into along a number of axes. In this 'spatial' definition, different regions correspond to different intensities of Darwinian processes, from marginal cases through to paradigm instances of a Darwinian population. The major axes identified are:

- H — fidelity of heredity.
- V — abundance of variation
- α — competition for reproduction
- C — continuity or smoothness of fitness landscape
- S — dependence of different reproductive outcomes on intrinsic properties

As a Darwinian individual is defined to be a member of a Darwinian population, there are similar marginal and paradigm instances of Darwinian individuals. Darwinian populations can also be nested, so Darwinian individuals on one level can be composed of a Darwinian population at a lower level. This formalism helps to make sense of the example of the deer herd; while the herd is a group, the population of herds does not meet the standards for being a Darwinian population — for instance H is very low and S does not depend on intrinsic properties of herds. Similarly in the Volvocine algae example, *C. reinhardtii* cells have high H, V, α, C and S and so are paradigmatic Darwinian individuals. On the other hand, *Volvox aureus* cells are marginal Darwinian individuals; reproductive outcomes depend heavily on extrinsic features like in-colony location, and there is low variation or competition for reproduction. On the other hand the *Volvox aureus* colonies are paradigm cases of Darwinian individuals.

2.5.4 Mechanisms of Social Group Transformation

The discussion of evolutionary definitions of an individual highlights those characteristics of an individual sufficiently important that they are definitional. These properties must occur as part of a social group transformation. This section highlights these key properties. Many of these mechanisms arise from traits that change the conditions for social interactions, and can therefore be understood as game-changing traits.

2.5.4.1 Reproductive Bottlenecks

Multicellular organisms and eusocial insect colonies all pass through a unicellular bottleneck during their reproduction, growing from a unicellular stage or a founding queen in the case of ants (or a group of founding queens in some multiply-founded colonies where

fierce ecological competition incentivises initial cooperation between unrelated queens (Wenseleers et al., 2004)). The existence of these unicellular bottlenecks between generations has many consequences that are important to the transformation of a social group into an individual. In particular it assists in fulfilling Folse and Roughgarden's first criterion by ensuring a high degree of relatedness throughout the organism that develops, ensuring that genetic heterogeneity arising from mutations will be distributed evenly among the group. This aligns the fitness interests of the group; it is an example of the evolution of what Michod (Michod and Herron, 2006) terms a *conflict modifier* — an adaptation that reduce the conflict between higher and lower levels.

A pathway to the evolution of such reproductive bottlenecks can be seen in some of the sophisticated ways that individual level traits evolve to control the assortment of cooperators in a social group. A strong mechanism for generating indirect assortment is limited dispersal, so relatives are naturally more likely to encounter other relatives. There is an ecological drawback though due to increased competition for resources between locally distributed progeny. This can by countered if progeny are instead distributed in buds (Gardner and West, 2006a). Local competition for resources is reduced while kinship is maintained. It is not difficult to extrapolate this behaviour into extreme examples where the bud becomes a propagule and a developmental process then creates the entire higher level individual.

By making interactions between highly related individuals more likely, the evolution of reproductive bottlenecks changes the conditions of a population's social interactions to reduce conflict and promote cooperation. Therefore we can see that traits for the evolution of reproductive bottlenecks are a type of game-changing trait.

2.5.4.2 The Germ-Soma Distinction

If there had to be one single definite marker between a social group and an individual, the best candidate would be the evolution of reproductive specialisation — the germ-soma distinction. It is through this separation that Maynard-Smith and Száthmary's central characteristic of a major transition is realised — entities that were previously able to reproduce independently must now reproduce together. The main examples of altruism in biology are found in non-reproductive castes, somatic cells in multicellular organisms or sterile workers in social insects. By sacrificing their own direct fitness, these somatic entities raise the fitness of their entire group, fulfilling Folse and Roughgarden's second criterion of reproductive interdependency.

The evolution of specialist germ and somatic cells has been extensively studied in Volvocine algae by Michod (eg. (Michod, 2007)). Based on these studies Michod developed a 'fitness isocline' model of germ-soma evolution arising because of the trade off between cell viability (survival capacity) and fecundity (reproductive capability) (Michod et al.,

2006). When the fitness curve of this trade-off is concave, the overall fitness is maximised by all group members investing in both viability and fecundity. However, when the fitness curve is convex the overall fitness is maximised by specialisation. Michod argues in that small groups the curve is concave because excessive somatic specialisation leads to diminishing returns on investment — increasing somatic specialisation can only raise the group fitness to a certain extent, so a small group cannot afford too much specialisation. On the other hand, as group size gets larger the cost of reproduction increases, so the fitness curve becomes convex and specialisation is favoured. This is in accord with the empirical evidence — there is a correlation between the size of a Volvocine colony and the degree of reproductive specialisation.

Michod's model gives an example of how germ-soma specialisation can be beneficial for the group as whole, but that benefit is not distributed equally — though somatic cells or sterile workers may benefit overall via indirect fitness benefits, it does not change the fact that the reproductive caste benefits the most from this division of labour. While relatedness is typically high within a group undergoing social group transformation, it is not perfect, and accumulated mutations mean that non-reproductive members will inevitably be less related on average to the offspring of the reproductive group members than the reproductive group members are themselves. This suggests that during the evolution of reproductive specialisation there would be a prisoner's dilemma type situation where it is best for the group that some members specialise in non-reproductive tasks but best for individual social group members not to be the specialists.

One possible solution comes from the social structure of the group. Dominance hierarchies are commonly found in the social groups of vertebrates; they play an important functional role in mediating intra-group conflicts, hence improving group stability (Van Schaik, 1983). The dominants typically receive increased mating opportunities and produce a greater number of offspring. In the case of primitive eusocial insects where workers are still capable of reproduction queens may be physically dominant. Similarly in species of bee and in naked mole rats, the most eusocial vertebrate species, the dominant female chemically suppresses the fertility of lower ranked group members. It is easy to hypothesise that this serves as a pathway to genetically obligate non-reproductive specialisation via a positive feedback process. If the existence of reproductive dominants becomes entrenched, it is in the interests of non-reproductives to convert a prisoner's dilemma type game of reproductive exclusion into a division of labour type game in which they gain inclusive fitness benefits by contributing to the indirect fitness of the reproductive dominants.

Dominance hierarchies may be applicable even in cases where a physical dominance hierarchy does not exist. Reeve and Jeanne (2003) introduced the concept of a *virtual dominant* based on their investigations of the conditions under which members of primate social groups might cooperate to aid a particular group member to reproduce, a model with consequences for the evolution of germ cells in multicellular organisms. In

their model, define π_i to be the physical power held by a group member i, r_{ij} to be the relatedness of group member j with any offspring of i and n to be the group size. Then the model finds that the group dominant is the member with the highest sum:

$$\sum(\pi_i r_{ij}) \tag{2.2}$$

Since this group member has the greatest combination of power and offspring that are maximally related to other group member. When physical power within the group is effectively even, so $pi_i = \frac{1}{n}$, this can be rewritten as:

$$\frac{\sum(r_{ij})}{n} \tag{2.3}$$

The group member that maximises this quantity is the *virtual dominant* because if only one group member can reproduce then every group member favouring the virtual dominant would be an evolutionarily stable outcome. In a social insect colony, the virtual dominant will usually be the queen. Bourke points out that in a multicellular organism where cells descend from a unicellular ancestor via repeated cell division, the virtual dominant will tend to be found in the slowest dividing cell lineages, as these will accumulate the least mutations from the unicellular ancestor (Bourke, 2011). The model is not perfect; as Birch points out it assumes that individuals are continuously attempting to increase the reproductive skew in favour of individuals proportionately to their relatedness, which is a non-trivial assumption (Birch, 2012b). However it is an intriguing hypothesis that may explain which cells are favoured for reproductive specialisation in the transition to multicellularity.

It is clear that traits for germ-soma differentiation or the establishment of social hierarchies will all change the social environment that individuals socially interact in. When we model those social interactions as evolutionary games, these traits will change the incentives of the game.

2.5.4.3 The Size-Complexity Hypothesis

The discussion of the evolution of reproductive specialisation suggested that group size plays an important role in the process of social group transformation. Bourke proposes that increased size may in fact have a causal role in promoting group complexity — the *size-complexity hypothesis* (Bourke, 2011). There are a cluster of traits including high reproductive division of labour, large size, increased dimorphism between germline and somatic cells and the decreased reproductive potential of somatic cells that seem to differentiate simple and complex social groups. Bourke's size-complexity hypothesis then is that external drivers start a positive feedback loop where increased group size

leads to social changes that increase the complexity of the social group which in turn leads to further group size increases.

There is a range of evidence that links group size with social complexity. In eusocial insects, there is a link between the number of workers and social complexity (Wilson et al., 1971); ant genera such at *Atta* with largest number of workers also have most morphologically distinct castes (Hölldobler and Wilson, 1990). Similarly in multicellular organisms there is a correlation between cell number and organism complexity (Bonner, 1988). Volvocine algae with small numbers of cells have single cell type, though some cells may have a lower probability of dividing, suggesting the emergence of a virtual dominant. Larger Volvocene algae with on the order of 10^5 group members are divided into morphologically different germ and somatic cells (Herron and Michod, 2008).

There are a number of proposed ecological and evolutionary drivers of large cell size. These include those processes that drive group formation in the first place. It is also typically the case that the next size class up is an empty ecological niche (Bonner, 1988). The size-complexity hypothesis predicts that increased group size inevitably leads to the development of the characteristics that define a complex group.

It is also worth noting that increased size enables a social group to be involved in a number of complex mutualisms. In many of the paradigm cases of complex individuals, such as the advanced *Atta* leaf cutter ants and large vertebrates, the continuing process of social group transformation has also involved a number of interspecific partnerships, with fungus in ants or with the vast number of microbes that live inside large vertebrates. It is estimated that in the human body there are ten times the number of microbes as there are human cells (Berg, 1996). This is not necessarily a requirement for a transition, but may be required for the complexity of the new individual to increase beyond a certain point.

2.6 Summary

We have explained the importance of social evolution and the key concepts involved, as well as describing the debate over issues such as group selection. These form the conceptual foundation for the modelling work we undertake in this thesis.

We have also tried to explain why the major transition are both biologically interesting and important. The major transitions are the primary inspiration behind our modelling work. The major transitions are an extreme demonstration of how changes to the social environment of a population can have significant consequences, such as enabling reproductive altruism and the creation of entirely new levels of biological organisation.

We claim that metagames can serve as a model for the reciprocal evolution of cooperation and cooperation-promoting traits as in the theory of social niche construction. Social

niche construction offers an answer to the how new evolutionary units evolve (Ryan et al., 2016). Through a major transition, a collection of particles becomes a new evolutionary unit. Before a transition, particles are evolutionary units (Darwinian individuals in Godfrey-Smith's terms) – they possess heritable variation in fitness. After a transition, it is the collectives that are the units of evolution, and tension between particle and collective has been minimised through the evolution of what we would term game-changing traits that suppress conflict within the collective. However, if such game-changing traits are evolved at the level of the collective then selection at the level of the collective is being invoked to explain the evolutionary process that formed that collective.

Social niche construction avoids this problem by explaining how individual particles modifying their own social niches can align their fitnesses and bring about selection on the level of the collective. This means we can use the metagames model to understand the conditions under which game-changing traits for conflict suppression between individuals (thereby promoting cooperation) might evolve, supporting a theoretical understanding of the social conditions necessary for a major transition.

Chapter 3

Evolutionary Game Theory

The standard mathematical tool for analysing the evolution of social behaviours is evolutionary game theory (Maynard Smith, 1982). Evolutionary game theoretic models are appropriate when the fitness of a strategy depends not just on the course of action the strategy entails (such as a strategy of aggressively competing for mating locations), but also on the frequency with which it and other strategies are found in the population. This is the case for social behaviours, where the success of an individual's strategy depends in part on the strategy of the other individuals it is interacting with.

Historically, game theory was developed as a tool for the mathematical analysis of strategic behaviour in economics, first by von Neumann and Morgenstern (1944). The most important early development came with Nash's equilibrium solution concept (Nash, 1951). However, the economic application of game theory relied on assumptions about the rationality of the players: first to justify their playing a Nash equilibrium strategy, then, via numerous competing refinements, to account for why one particular Nash equilibrium might be chosen over another (Weibull, 1997).

While these problems frustrated the economic application of game theory, Maynard Smith (1982) demonstrated the power of game theoretic models when applied to problems in evolution. Assumptions about rationality were no longer required (and indeed would have been absurd if the players were, for instance, bacteria). Instead, the players of an evolutionary game might reach an equilibrium state by the trail-and-error process of evolution.

This chapter gives an overview of the key concepts and mathematical results of evolutionary game theory. In particular it introduces two of the most important tools for understanding an evolutionary game: equilibrium states and the replicator dynamics. These answer the two obvious questions about any evolutionary game: what are the possible outcomes of the game (the equilibrium states) and how might the players reach those states (the replicator dynamics).

The equilibrium states we are most interested in are those that are robust under small perturbations in the nature and distribution of strategies, and hence unlikely to be displaced by natural selection — the *evolutionary stable states* (ESS). The *replicator dynamics* provide a straightforward way to model the evolution of the frequency of strategies used in a population over time. We use these two techniques in all the game theoretic models in the rest of this thesis.

We are interested in the evolution of game-changing traits that change the social game a population is engaged in. To understand how a population playing one game might transition to playing another game, we need a complete understanding of the types of game the population might be playing — to characterise the 'landscape' of the space of possible games so we can understand why a population might take particular paths through it. This chapter classifies the symmetric two-player games, the class of games used as the basis for this thesis. This includes the canonical social dilemmas: the Prisoner's Dilemma, Stag Hunt and Snowdrift games.

In particular, we introduce a geometrical way of understanding these games: the ST-plane (Santos et al., 2006a) (and ST-space, the region of the plane of particular interest). This is a representative subset of the full space of possible symmetric two-player games with the important property that every game in the complete space of possible two-player games has a corresponding game on the ST-plane with the same equilibria and dynamics. The ST-plane is the setting for most of the game theoretic content in the rest of this thesis. The dynamics of all possible two-player symmetric games are classified and given a complete spatial categorisation into four fundamentally different games on the ST-plane. Finally we present an even smaller circular subspace of the ST-plane (θ-space) that preserves the range of possible equilibria and will serve as a minimal setting for our initial models of metagames in Chapter 5.

3.1 Two-Player Symmetric Games

For most of this thesis we restrict our attention to a particular type of evolutionary game — symmetric two-player finite games in normal form. This is a rather involved technical classification that requires some unpacking, so first we will run through each of the component parts of this definition.

A *finite game* is one played between a finite number of players, each player pursing one of a finite number of potential *pure strategies*. We might draw this finite number of players from a larger, possibly infinite, population. We let $I = 1, 2, \ldots, n$ be the finite set of players in a particular game. For each player $i \in I$, we let S_i be the player's (finite) set of pure strategies. A game can be defined so that the players have different roles, such as two-player game in which the player in the 'player one' role has a different

set of available strategies to the player in the 'player two' role. If instead the player roles are all interchangeable, so $S_i = S_j$ for all $i, j \in I$, then we say the game is *symmetric*.

A vector $\mathbf{s}$ of pure strategies $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, where each s_i is a pure strategy for the player i, is called a *pure-strategy profile*. A strategy profile is essentially a selection of strategies by each player playing the game. The set of all pure strategy profiles for the game is the cartesian product of all the player's pure strategy sets.

The game is in *normal form* when there exists a function f that maps a tuple of strategy choices by each player (that is, a strategy profile $\mathbf{s}$) to a tuple of real numbers that correspond to each player's payoff given these strategy choices. In evolutionary game theory the payoff from a game is typically interpreted as mean individual fitness. The other typical formulation of a game is *extensive form*, where the game is specified by a tree of possible choices. We do not consider extensive form games here.

The most basic class of games of interest to the evolution of cooperation are simultaneous games between players with two strategies, 'cooperate' and 'defect'. A *simultaneous game* is a game where all players make their moves at the same time, rather than in sequence (or at least each player selects their own strategy without observing any other player's strategy choice). Although these are the simplest types of game they include all the canonical social dilemmas, such as the Prisoner's Dilemma and the Snowdrift Game, the traditional focus of game theoretic work on social evolution.

Two pure strategy choices for two players means there are four possible pure strategy profiles. When the payoffs are linear functions of the frequencies of the strategy-types the complete payoff functions for the game can be specified by the four entries of a single 2×2 matrix (or in general by a $n \times m$ matrix for n players with m pure strategies). An arbitrary symmetric two-player two-strategy game G for a focal player F and an opponent O is given by the matrix:

$$\begin{array}{c} \\ C_F \\ D_F \end{array} \begin{array}{c} \begin{array}{cc} C_O & D_O \end{array} \\ \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array} \tag{3.1}$$

Because the game is symmetric the payoff matrix is the same for both players. Each row corresponds to a pure strategy for the focal player, each column to the pure strategy used by the opponent. If the game were not symmetric then it would require a pair of payoff matrices (G_1, G_2) to fully specify the game. The labelling of the coefficients adopts the standard convention (Axelrod and Hamilton, 1981) — R is the reward for mutual cooperation by both players, P the punishment for mutual defection, T the 'temptation to defect' received by the focal player for defecting when the other player cooperates, and S the 'sucker's payoff' for cooperating when the opponent defects. The two strategies are labelled C for 'cooperate' and D for 'defect', though for some games it

is not strictly the case that the 'cooperative' strategy is actual cooperation. We use these labels consistently though we identify where they may take on an alternative meaning.

We say a strategy is *dominated* if there is an alternative strategy that will give a higher payoff whatever strategy is chosen by the other player. For example, if $R > T$ and $S > P$ then defection would be a dominated strategy. If an alternative strategy will only give a greater or equal payoff, rather than a strictly greater payoff, the original strategy is *weakly dominated*.

These definitions give a natural spatial structure to the set of all two-player two-strategy symmetric games, where we identify each game with the point $(R, S, T, P) \in \mathbb{R}^4$. It is customary for the study of social dilemmas to restrict the range of these four parameters. In our analysis of social dilemmas we take the position that mutual cooperation is always more advantageous than mutual defection, so we assume that $R > P$. The set of all 2×2 matrices with $R > P$ is a linear subspace of the set of all 2×2 real matrices (if you also include the degenerate game $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$). We call this subspace *RSTP-space*; it is the largest set of symmetric 2×2 matrices suitable for the study of social dilemmas in game theory. Varying the relative magnitudes of R, S, T and P leads to a diverse set of social interactions.

There are a number of ways to interpret evolutionary game theoretic models; one obvious method might be to use a direct analogue with games in an economic context, each player corresponding to an individual animal. Here though we make use of the population interpretation of evolutionary game theory, the roots of which go back to Nash's unpublished PhD thesis (Weibull, 1997). In the population interpretation we think of a large population divided into n strategy-types $\tau_1, \ldots, \tau_n$. Each type is genetically committed to using a particular (pure) strategy. We then imagine that individuals are drawn randomly from a large (essentially infinite) population to play the game. This interpretation connects evolutionary game theory with population dynamics through the replicator dynamics (or alternative dynamical formulations). Instead of talking about a player using a particular mixed strategy, such as 'cooperate half the time, defect half the time', we talk about a polymorphic population composition of 'half the cooperator type, half the defector type'.

3.1.1 Nash Equilibria and Evolutionarily Stable States

The *Nash equilibrium* is considered the most important concept in game theory (Binmore, 1992, page 12). In the economic context in which it was developed, a game is in a Nash equilibrium when no player can benefit by unilaterally changing its own strategy given the current strategies of the other players. Given the strategies of all the other players each player's own chosen strategy is the best possible response. Nash proved that a mixed strategy Nash equilibrium exists for every finite game (Nash, 1951), which

includes all the games in this thesis. It is important to note that there can be multiple Nash equilibria for a single game, and that a Nash equilibrium is not necessarily the best possible outcome for all players, just one from which they have no incentive to deviate given the strategies of the other players.

The Nash equilibrium concept transfers to the context of evolutionary game theory through the expectation that natural selection can lead the frequencies of the strategy-types to some (possibly optimal) equilibrium frequencies through a trial-and-improvement process.

However, many Nash equilibria are unstable and so any deviation in the strategy-type frequencies will throw the game out of equilibrium again. Over evolutionary timescales such perturbations are to be expected. Therefore we are particularly interested in *evolutionarily stable states* (ESS), the equivalent in the population interpretation of evolutionary game theory of the *evolutionarily stable strategy* concept introduced by Maynard Smith and Price (1973). An evolutionarily stable strategy is a (possibly mixed) strategy which, if adopted by most of the members of the population, cannot be displaced by natural selection — no small mutations in the rate at which the component pure strategies are played would lead to a mixed strategy with higher fitness.

This does not mean that an evolutionarily stable strategy is necessarily an optimal strategy. Other strategies conferring higher fitness could exist, but they could not be the result of a mutation to any evolutionarily stable strategy. While the ESS condition does not guarantee global optimality, it is a stronger condition than the Nash equilibrium of no incentive to deviate as it implies local optimality. The relation between the two conditions is in fact even stronger: though the concept of an evolutionarily stable strategy was developed independently from that of a Nash equilibrium, the set of all evolutionarily stable strategies of a given game is a subset of the set of all Nash equilibria (Weibull, 1997).

An evolutionarily stable state is the same mathematical concept restated in the population interpretation of evolutionary game theory: a population state that is stable under small perturbations in the strategy-type frequencies, so when these frequencies are slightly changed the population will return to the evolutionarily stable state.

3.1.2 The Replicator Dynamics

Nash equilibria and evolutionary stable states are vital tools for understanding the behaviour of evolutionary systems based on games but say nothing about the behaviour of the system out of equilibrium. There are a number of possible ways to introduce dynamical behaviour into evolutionary game theoretic models, such as playing games in a spatial structure where more successful strategies physically displace less successful

ones (Nowak and May, 1992), or playing games on graphs (Lieberman et al., 2005) or networks (Pacheco et al., 2006a) where the game results affect the graph structure.

The basic model of the dynamics of an evolutionary game are the *replicator dynamics* introduced by Taylor and Jonker (1978) and derived from differential equation models in population dynamics. The replicator dynamics are based on the intuitive principle that, at any given point in time, successful strategies — strategies that at that point result in greater than average fitness — are more likely to be employed in the future as they will be inherited by a greater than average number of offspring. This is a logical step from the interpretation of payoffs as fitness. A system evolving under the replicator dynamics models the change in the frequencies of different traits within a population as proportionate to the difference between the fitness conferred by the trait and the average fitness of the population.

We let $\mathbf{x} = (x_1, \ldots, x_n)^T$ be the *population state vector* representing the proportions of all the types in the population, so x_i is the frequency of the strategy-type τ_i (for example, 'C' or 'D'). This state vector $\mathbf{x}$ is function of time t. Because the population is completely partitioned into the different strategy-types, the sum of all the frequencies $\sum_{i=1}^{n} x_i = 1$, so $\mathbf{x}$ is defined on the simplex S_n (the set of all n-tuples that sum to 1). Let $f(\tau_i)$ be the fitness of type τ_i (some sources use $\omega(\tau_i)$ for fitness instead). This is defined to be the expected payoff for a member of τ_i given the current state of the population. We then define $\bar{f}(\mathbf{x}) = \sum x_i f(\tau_i)$ to be the mean fitness of a population with type frequencies represented by the vector $\mathbf{x}$. Then the replicator dynamics are determined by the *replicator equation*:

$$\dot{x_i} = x_i(f_i(\mathbf{x}) - \bar{f}(\mathbf{x})) \tag{3.2}$$

Note that the replicator dynamics make three major simplifications from a real evolutionary system: there is no mutation so strategy frequency changes only occur due to differential reproduction; no strategy can go extinct, or recover from exogenously imposed extinction; the population size is infinite and fixed (or at the least makes no allowance for ecological pressures).

One of the advantages of the replicator dynamics lies in the simple connection between the replicator equation and the Nash equilibria and evolutionarily stable states of the game. Two important theorems make this connection clear, as well as clarifying the relationship between Nash equilibria and evolutionarily stable states:

- If $x \in S_n$ is a Nash equilibrium state of the underlying game then it is a fixed point of the replicator equation (Hofbauer and Sigmund, 1998, page 69).

- If $x \in S_n$ is an evolutionary stable state of the game then it is an (asymptotically) stable fixed point of the replicator equation (Taylor and Jonker, 1978; Hofbauer and Sigmund, 1998, page 70).

When a strategy is dominated, that strategy will (asymptotically) tend to extinction under the replicator dynamics (though it cannot actually become extinct). However, weakly dominated strategies may survive if those strategies against which they always give a lower payoff than the dominating strategy go (asymptotically close to) extinct, so the weakly dominated strategy gives the same payoffs against the remaining strategy types as the dominating strategy.

For two-player two-strategy games we can write the replicator equation explicitly in terms of R, S, T and P and the frequency of cooperators by x_C:

$$\dot{x_C} = x_C(1 - x_C)((R - T + P - S)x_C - (P - S)) \tag{3.3}$$

Note the occurrence of the terms $(R - T)$ and $(P - S)$, which have an important interpretational significance which will be discussed later. For games defined by a 2×2 matrix, the replicator equation takes the form of a one-dimensional first-order ordinary differential equation. From the general theory of ordinary differential equations it is known that there are only three possible behaviours for trajectories in a vector field on the real line — the trajectory approaches a rest point or diverges to $\pm\infty$ (Strogatz, 1994, page 28). As the value of every x_i subject to the replicator equation is restricted to the interval $[0, 1]$, with rest points at 0 and 1, this rules out the possibility of the replicator equation diverging to $\pm\infty$. This means that in every such game the frequency of cooperators must always reach an equilibrium for any initial conditions. In asymmetric two-player games or games with more strategies a broader range of behaviours are possible, such as limit cycles or strange attractors.

3.2 ST-Space

A two-player symmetric game is fully determined by the four entries in the payoff matrix, so the space of possible games is a four-dimensional space parameterised by R, S, T and P; varying their relative magnitudes leads to a diverse range of social scenarios. This is the most general space of 2×2 symmetric games. In this general setting there are a large number of similar games. For example the game $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is qualitatively identical to the game $\left(\begin{smallmatrix} 2 & 2 \\ 2 & 0 \end{smallmatrix}\right)$. We can simplify this space of games. By normalising R to 1 and P to 0 we project this space onto a two-dimensional plane parameterised by S and T that preserves the important features of $RSTP$-space (Santos et al., 2006a). We call this the ST-plane. This projection is done via the transformation:

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} \rightarrow \begin{pmatrix} \frac{R-P}{R-P} & \frac{S-P}{R-P} \\ \frac{T-P}{R-P} & \frac{P-P}{R-P} \end{pmatrix} = \begin{pmatrix} 1 & S' \\ T' & 0 \end{pmatrix} \tag{3.4}$$

This transformation has the important property that every game with $R > P$ is equivalent to a game on the ST-plane: there is a corresponding game with the same equilibria and same stability conditions for those equilibria, though the speed at which the equilibria are reached may change (roughly equivalent to changing the intensity of selection). This is because the replicator equation for a transformed game is the same as the replicator equation for the original game multiplied by a constant, so the roots of the equation do not change, nor does the stability of those roots. Because of the connections between the equilibria of the game and the roots of the replicator equation, the Nash equilibria and evolutionarily stable states determined by the replicator equation's fixed points are also unchanged. The projection is valid whenever $R \neq P$, which is always true for the games where $R > P$ that we are using in this thesis.

This makes the ST-plane an extremely useful conceptual tool: a representative subset of the two-strategy social dilemmas in which the relationship between different games can be shown spatially. This lets us represent changes to the payoff matrix of a game geometrically.

For simplicity, from here we will define $R = 1$, $P = 0$ and omit the prime notation S' and T' for games on the ST-plane. The replicator equation for the evolution of the frequency of cooperators for any game on the ST-plane can be written as:

$$\dot{x_C} = x_C(1 - x_C)((1 - S - T)x_C + S) \tag{3.5}$$

The conventional definition of a social dilemma assumes that a cooperator benefits more from mutual cooperation than from cooperating while the other player defects ($R = 1 > S$), and that mutual cooperation is also preferable to an equal probability of unilateral cooperation or defection ($2 > T+S$), so we typically restrict our interest in the ST-plane to the region $-1 \leq S \leq 1$, $0 \leq T \leq 2$ (Macy and Flache, 2002). We call this bounded region *ST-space*. It is the standard setting for the game-theoretic models in the rest of this thesis.

We select this privileged region of the space based on these preferences for how the payoffs to a game should be understood, but we should note that the behaviour of the games does not change at the boundaries of this square region. However, there may be changes in the appropriate interpretations of games in different regions of game space. Games in the Snowdrift Region where $2R < T + S$ are often excluded from the study of social dilemmas since they can be best thought of as 'division of labour games' in which the population-optimal outcome is not pure cooperation but a balance

of cooperation of defection (Tudge et al., 2013). As such this is a region of game-space in which 'cooperation' and 'defection' may be inappropriate names for the two strategies. Though it is often not thought of as a 'social dilemma' as cooperation is not the population-optimal outcome, we include this region in ST-space so that we can consider how social dilemmas might transform into 'division of labour' games.

Note that we assign a payoff of $P = 0$ to mutual defection, implying that mutual defection is neutral with respect to fitness. There are circumstances when in fact it might be better not to play at all (in a non-obligate game), or where mutually defecting has a fitness cost. If ecological constraints result in negative payoffs for one or both of R and P (so interactions carry fitness costs for all involved) it may open up alternate pathways for the evolution of altruism in the absence of population structures (Doncaster et al., 2013a).

In social dilemmas there is conflict between the rational outcome and the Pareto-efficient outcome, the individually rational choices for each player leading to a deficient outcome for both. This happens because of *greed*, when there is a temptation to defect because unilateral defection is more advantageous than mutual cooperation ($T > R = 1$), or *fear*, the risk of payoff loss when facing a defector for maintaining a cooperative strategy instead of also defecting ($S < P = 0$). Social dilemmas arise when the game is one that features fear, greed, or both (Santos et al., 2006a). These two factors correspond to the two axes of the ST-plane. The lines $S - P = 0$ and $T - R = 0$ split the ST-plane into four quadrants (Figure 3.1), corresponding to four fundamental two-player games that define the most common types of conflict, in each of which the problem of cooperation takes a different form (Archetti and Scheuring, 2012).

3.3 The Four Fundamental Games

It is an immediate consequence of the three roots of the replicator equation that the Nash equilibria of any symmetric two-player game corresponds to one of three states: a population of no cooperators (at $x_C = 0$), all cooperators ($x_C = 1$) or a polymorphic population ($x_C = \frac{P-S}{R-S-T+P}$). The mixed equilibrium only exists when its value lies between 0 and 1. The stability of the fixed points determines which of these equilibria are evolutionarily stable states. The existence and stability of these fixed points divides the ST-plane into four qualitatively different regions (Figure 3.1) — the quadrants split by the lines $S = P = 0$ and $T = R = 1$. These four regions correspond to four fundamental two-player games that define the most common types of conflict. In each one the problem of cooperation takes a different form (Archetti and Scheuring, 2012):

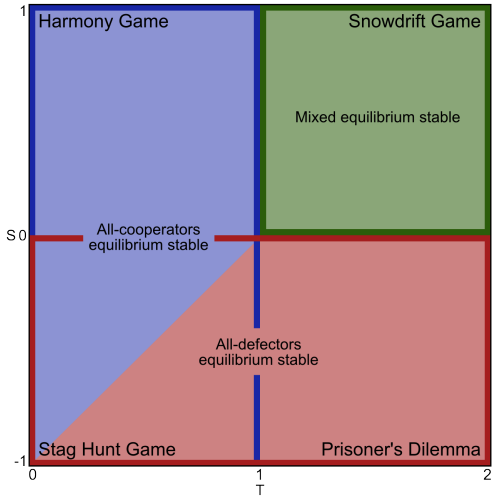


Figure 3.1: ST-space showing the four different game types corresponding to the regions in which the different equilibria are stable. Cooperation is stable in the blue-bordered area, defection in the red-bordered area; these overlap in the Stag Hunt region where both strategies are stable and the equilibrium depends on the initial conditions. The mixed strategy equilibrium $x_C = \frac{S(T-1)}{1-S-T}$ is stable in the green shaded area. Adapted from Fig. 2 of Santos et al. (2006a).

3.3.1 The Prisoner's Dilemma

- Defined by $T > R > P > S$
- No-cooperators is an ESS
- All-cooperators is an unstable equilibrium
- The polymorphic equilibrium does not exist

In the Prisoner's Dilemma unilateral defection is more advantageous than mutual cooperation and unilateral cooperation worse that mutual defection, so there is both fear and greed in the game and cooperation is a dominated strategy. The evolutionarily stable state is a population of no cooperators, while all cooperators is an unstable equilibrium.

The Prisoner's Dilemma is the most widely seen social dilemma. It models the tension between a 'rational' selfish strategy and a mutually optimal outcome, a key scenario in the evolution of cooperation. However it has been argued that too often the Prisoner's Dilemma is used as the default model of social interaction, overstating the prevalence of the apparent paradox of altruism (Skyrms, 2004; Worden and Levin, 2007).

In ST-space, the act of unilateral cooperation in a Prisoner's Dilemma is an altruistic one, according to the definitions in Section 2.2.1, since a defector recipient will experience a gain in fitness ($T > R = 1$) while the cooperator experiences a loss in fitness ($S < P = 0$). This is why the Prisoner's Dilemma is so prevalent: it models conditions when cooperation requires altruism. Off the ST-plane this may not be the case: if $0 > R > P$

(so both are negative) all parties to such a social interaction would lose relative to inaction. Such conditions may occur due to ecological constraints (Doncaster et al., 2013a). In this thesis, however, we identify altruism with cooperation in a Prisoner's Dilemma (and, conditionally, the Stag-Hunt Game).

3.3.2 The Harmony Game

- Defined by $R > T > S > P$
- All-cooperators is an ESS
- No-cooperators is an unstable equilibrium
- The polymorphic equilibrium does not exist

The Harmony Game is the least seen of these four games as there is no social dilemma — cooperation is the most successful strategy for the individual and for the collective welfare. A population of all cooperators is the evolutionarily stable state (while a population of no cooperators is an unstable equilibrium).

3.3.3 The Snowdrift Game

- Defined by $T > R > S > P$
- A polymorphic population with $\frac{S-P}{S+T-R-P}$ cooperators is the ESS
- All-cooperators and no-cooperators are both unstable equilibria

As described earlier, the Snowdrift Game (also called the *Hawk-Dove Game* or *chicken*) is based on the example of a population divided into aggressive 'hawks' and pacifistic 'doves'. It is an anti-coordination game where it is advantageous to use a different strategy to the majority of the population; unilateral defection is better than mutual cooperation while unilateral cooperation is better than mutual defection. The combination of these two pressures has a homeostatic effect on the frequency of cooperators, hence the Snowdrift Game is significant as the only game that sustains a stable polymorphic population — the evolutionarily stable state has $\frac{P-S}{R-S-T+P}$ cooperators. There are also unstable equilibria at populations of all cooperators or no cooperators.

When $2R < S+T$, the best outcome for the 'group' is not for both players to cooperate but for one player to cooperate and the other defect. As stated earlier, in such circumstances a Snowdrift Game is a model of the need for the division of labour (Tudge et al., 2013), and 'cooperate' and 'defect' cease to be strictly appropriate labels for those strategies.

3.3.4 The Stag Hunt Game

- Defined by $R > T > P > S$
- All-cooperators and no-cooperators are both ESS
- A polymorphic population with $\frac{S-P}{S+T-R-P}$ cooperators is an unstable equilibrium

The Stag Hunt is based on a parable by Rousseau (Fort, 2008). Two individuals go hunting and each can either choose to hunt a stag or a hare. Each hunter can successfully kill a hare by himself, but they would have to both choose to work together to hunt the stag. However, half a stag is worth more than a single hare. Thus the Stag Hunt is an example of a coordination game. It models interactions in which it is best to be using the same strategy as the majority of the population. There are two stable outcomes — both hunt a stag (cooperation) or the both hunt their own hare (defection).

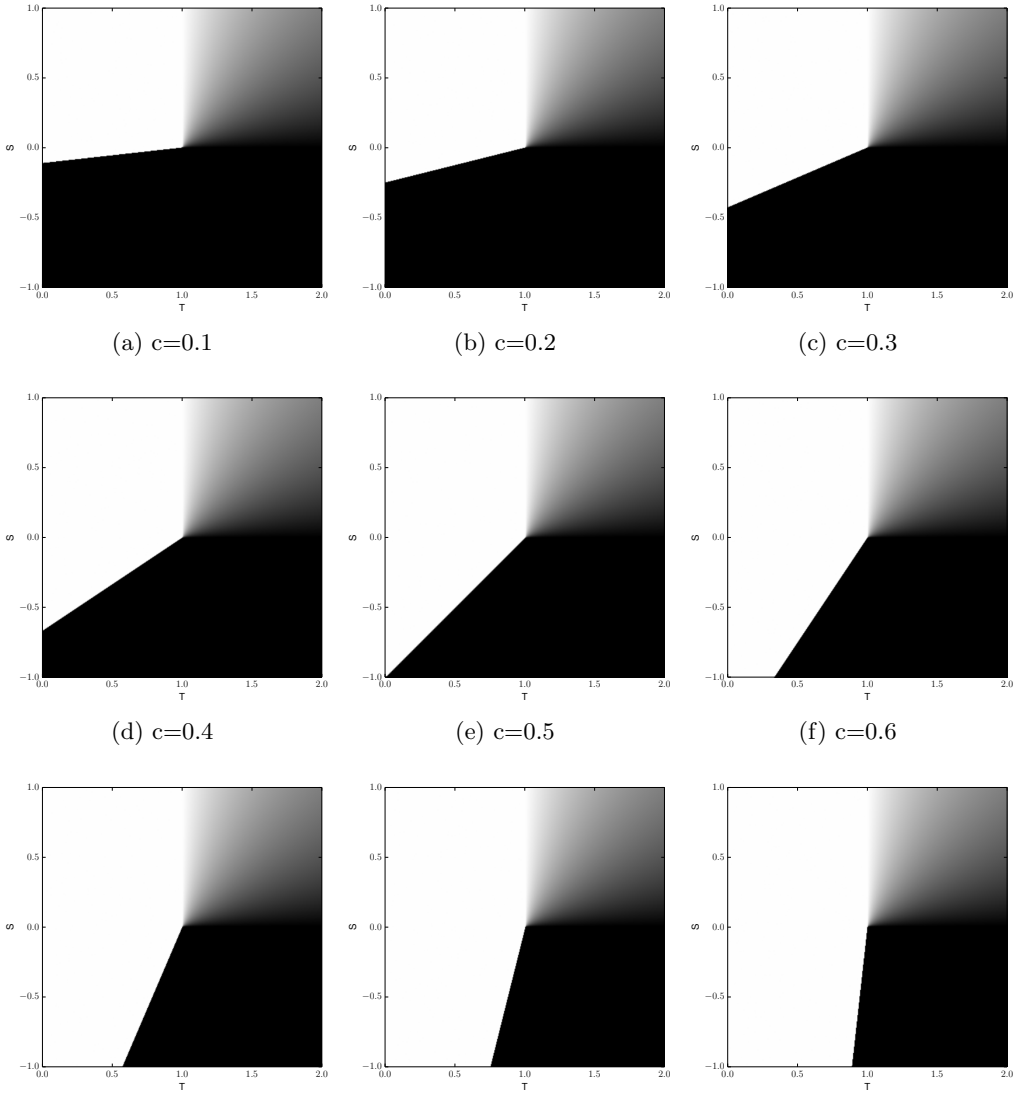
Depending on the initial frequency of cooperators, the population will be driven towards an equilibrium of all-cooperators or all-defectors — $x_C = 0$ and $x_C = 1$ are both stable attractors. There is a Nash equilibrium at $x_C = \frac{P-S}{R-S-T+P}$ too, but it is not stable. The significance of this Nash equilibrium is that is divides the basins of attraction of the two evolutionarily stable states. If the initial frequency of cooperators is greater than the unstable polymorphic Nash equilibrium then the frequency will increase to all cooperators, and if it is less then the population will end up at all defectors. This is the only one of the four fundamental games in which the initial frequency of cooperators is significant in determining the ESS.

3.4 Visualising ST-Space

One of the benefits of projecting the space of games onto the ST-plane is that as it is two dimensional, it is easy to visualise, which helps us develop a spatial understanding of the relationship between different types of game. Figure 3.2 illustrates the frequency of cooperators at the stable attractor for different games in ST-space from a range of initial frequencies of cooperators obtained by evaluating the replicator equation to equilibrium at each point in the space. Black indicates no-cooperators at equilibrium, white all-cooperators. We do not plot the single strategy initial conditions of $c = 0$ or $c = 1$ as with only one strategy present there will be no change under the replicator dynamics.

The figure clearly shows the differences between the four fundamental games. The upper left and lower right quadrants are the Harmony Game and Prisoner's Dilemma respectively with their single stable equilibrium value. The upper right quadrant is the Snowdrift Game, displaying a continuous change in the frequency of cooperators present in the evolutionarily stable state that does not vary with the initial conditions. This

contrasts with the Stag Hunt game in the lower left quadrant where there is a sharp dividing line between the games that end up at the two different attractors. This is the only region of ST-space sensitive to the initial frequency of cooperators (aside from the extreme single strategy conditions).



(a) c=0.1 (b) c=0.2 (c) c=0.3

(d) c=0.4 (e) c=0.5 (f) c=0.6

(g) c=0.7 (h) c=0.8 (i) c=0.9

Figure 3.2: The equilibrium frequency of cooperators under the replicator dynamics across the ST-plane from different initial frequencies of cooperators ($c = 0.1$ to $c = 0.9$). Black indicates 0% cooperators at equilibrium, white 100% cooperators. The figures illustrate the split into different games — the Harmony Game (upper left) and Prisoner's Dilemma (lower right) quadrants with their single stable equilibrium value, the Snowdrift Game (upper right), displaying a continuous change in the equilibrium, and the Stag Hunt (lower left) in which there is a sharp dividing line between the initial social conditions that end up in populations of all-cooperators or all-defectors.

In the modelling work in this thesis we will frequently plot the behaviour of metagames over ST-space. Figure 3.2 demonstrates the prevailing social conditions in ST-space

in the absence of any game-changing traits and serves as a point of comparison for subsequent models.

3.5 Games with Constant Selection Strength (θ-Space)

It is possible to further compress this space of games while still retaining information about the equilibria. The ST-plane (less the point at $S = 0, T = 1$) can be projected onto a circle centred at this point. Like the projection onto the ST-plane, this transformation is equivalent to multiplying the replicator equation by a constant (the radius of the circle), so the equilibria are preserved. This is evident purely from inspection of Figure 3.2, which shows that the equilibrium frequency of cooperators is a constant on the ST-plane along half-lines emanating from $S = 0 = P, T = 1 = R$. Increasing the radius of the circle increases the rate at which the replicator equation approaches equilibrium; this could be interpreted as equivalent to increasing the intensity of selection.

The significance of this point is that, given $P = 0$, $R = 1$, the circle is centred at the point $S = P$, $T = R$, so it is the locus of all points such that $(S - P)^2 + (T - R)^2$ is a constant — in other words where there is a constant trade-off between the amount of 'fear' and 'greed' in the game (with the interpretation of $S-P$ as the fear of encountering defection, $T - R$ the greed favouring defection). Thus any game with coordinates (T, S) on the ST-plane can be written in radial coordinates as $(r, \theta) + (1, 0)$ where θ determines the equilibrium frequency of cooperators and r determines the rate at which the game reaches equilibrium (as in Tanimoto and Sagara, 2007).

So when an even simpler space is useful for the study of symmetric two-player two-strategy games, the ST-plane can be projected onto a one dimensional circular space in which every game is represented by a single number $\theta \in [0, 2\pi)$ since r is irrelevant in determining the equilibria. Such a circle is shown in Figure 3.3. This is the most compressed continuous representation of the space of two-player symmetric games that preserves the full range of equilibria, which each circle maintaining a constant strength of selection as r will be constant. The existence of this projection reveals that there is a circular structure underlying the space of possible symmetric two-player two-strategy games. We use games on this circle as an appropriate minimal setting to introduce metagames in Section 5.4.1. The payoff matrix for a game G_θ in θ-space is given by:

$$G_\theta = \begin{pmatrix} 1 & \sin\theta \\ 1 + \cos\theta & 0 \end{pmatrix} \tag{3.6}$$

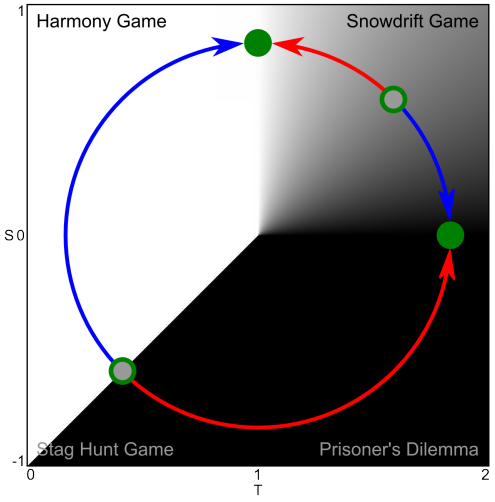


Figure 3.3: ST-space with the circle defined by θ shown.

3.6 Discussion

We have now reviewed the necessary game-theoretic material to proceed with the primary modelling work in this thesis. In particular, we have presented ST-space (Santos et al., 2006a), the setting for the models of game-change we will be presenting. When we model the changes to a population's social game brought about by the evolution of game-changing traits we can represent this as movement within ST-space.

This is an appropriate setting to investigate the evolution of game-changing traits since it is representative subset of all the symmetric two-player two-strategy games where mutual cooperation is more advantageous than mutual defection: every such game corresponds to a game on the ST-plane. We then further restrict the plane down to ST-space that respects the standard restrictions on the social dilemmas (Macy and Flache, 2002) (though we also include the division of labour region of the Snowdrift Game where $2R < S + T$). Modelling game-change means evaluating the differences between different games; drawing these games from a consistent space of possible games keeps these comparisons principled. ST-space also has the heuristic benefit of being easy to visualise which allows us to develop geometric intuitions about the behaviour of landscape of games where our modelling work will take place.

When we wish to restrict the space of games we are interested in even further, such as when we want a simple setting to introduce the mechanics of metagames, we can choose subsets of ST-space like θ-space, the circle where the strength of selection is constant.

Chapter 4

Changing the Game: Representing Game-Changing Traits

In this chapter we introduce a core concept in this thesis: game-changing traits, which change the incentives for social interactions. We define the concept of a game-changing trait and discuss how we can formally represent the effect of game-changing traits as transformations to the payoff matrix of the evolutionary game the population is playing. This lays the groundwork for Chapter 5 where we introduce metagames to study the evolution of these game-changing traits: the formal equivalence between the actions of various game-changing traits and changes to a payoff matrix will allow us to use the metagames model of the evolution of payoff matrices to talk about the evolution of game-changing traits themselves.

We focus our attention on a key class of game-changing trait: traits that alter population structures to positively assort interactions between cooperators. This is because assortment is widely considered to be the ultimate cause for the successful spread of cooperative behaviours (Hamilton, 1975; Eshel and Cavalli-Sforza, 1982; Michod and Sanderson, 1985; Nowak, 2006; Okasha, 2006; Godfrey-Smith, 2009; Rosas, 2010), though the proximate mechanisms that generate it can vary. We review why assortment is so important to the evolution of cooperation, and some of the proximate mechanisms that can create it. Relatedness due to common descent is a particularly important method of creating positive assortment, and indeed is promoted by some theorists as the only relevant mechanism (Fletcher and Zwick, 2006), but we discuss why this is not necessarily the case, and then discuss other mechanisms for creating positive assortment.

We then present a method to represent the effect of game-changing traits that affect population structure. It has long been recognised that social strategies have different fitness consequences when are games played between relatives (Maynard Smith, 1978;

Grafen, 1979). Instead of looking at relatedness we use the example of directly increasing assortment in two-player two-strategy games to demonstrate the principles behind changing the payoff matrix of a game, using our method to derive the same transformation formula as Grafen (1979). We review other existing examples of transforming the payoff matrix of a game to account for the influence of mechanisms such kin selection, group selection, direct and indirect reciprocity, network structure (Ohtsuki and Nowak, 2006; Taylor and Nowak, 2007; Boyd et al., 2010; Van Veelen, 2011). We then introduce a generalisation of our method, *interaction functions*, which provide a consistent and justifiable procedure to determine transformations to a game when there are non-random interaction patterns due to game-changing traits that affect population structure. We show that the game transformations we developed for positive assortment were in fact a simple application of interaction functions. This provides us with a mathematical way to represent the effects of a wide class of game-changing traits.

4.1 Game-Changing Traits

A theme we have repeatedly come back to in this thesis is that the social game a population is playing is not just an exogenous fact of its circumstances, but also an endogenous product of evolved traits. In this thesis we develop the evolutionary interpretation of models in which the game being played is not a parameter but a variable tied to the evolution of *game-changing traits* (GCTs). These are traits that affect the social conditions of the population, altering the payoffs for different strategies of a social game without changing the physical nature of the interaction. To make this idea clear consider a simple example where the game is food donation — an individual may choose to give or not give food to another individual. Then a game-changing trait would be a trait which affects the costs and benefits of giving the food, but the game would remain one of food donation. For example, the game-changing trait could alter population structure so that individuals are more likely to be related, or it could be a trait that leads to the punishment of non-givers.

In the language of social niche construction, a game-changing trait is a like a *social niche modifier* (Ryan et al., 2016). Game-changing traits alter a population's social niche by changing the effective game being played by its bearers. When we represent the action of game-changing traits as transformations to the payoff matrix of a game, we distinguish between the hypothetical original game that the population would be playing were there no game-changing traits acting on the social context (the original payoff matrix), and the *effective game* that is actually being played as a result of the game-changing traits (the transformed matrix). Thus game-changing traits modify the social niche a population is in, affecting the strength and the direction of selection on social behaviours.

In this sense a game-changing trait is similar to a modifier gene, because it changes the social conditions that would be in place if it had not been in effect, much as modifier genes change parameters determining the effects of genes at other loci. The parallel is instructive because modifier genes are selectively neutral, yet the modifier frequencies evolve to maximise the mean fitness of the population (in random mating systems) (Karlin and McGregor, 1974). In this thesis we will investigate how the evolution of game-changing traits that change social parameters determining the effects of social traits at other loci in the model changes the mean fitness of the population by increasing or decreasing the success of cooperative behaviours.

Population structuring traits that indirectly modify the payoffs of the social game by changing the frequency at which different strategies encounter each other are an obvious category of game-changing trait. Population structuring traits come in many forms including: limiting dispersal in reproduction hence changing the social context of interactions by increasing the likelihood of interacting with relatives (Pepper and Smuts, 2002); greenbeard markers that signal the bearer is committed to a cooperative social strategy (though such a marker may be subject to imitation) (Gardner and West, 2010); group size preferences that can bias group formation to create high levels of within- and between-group variance(Powers and Watson, 2011).

But game-changing traits need not change population structure to change strategy incentives. Cell-cycle synchronisation in fungi, where nuclei divide in union, is a mechanism that acts to control variation (Buss, 1987). This trait greatly reduces the potential for parasitic defectors because all nuclei divide at the same time, so no variant can increase at a rate greater than the ancestral nuclei. In particular, this control turns potential defector mutants that benefit themselves to the detriment of the group into deleterious mutants that are detrimental to themselves because they are detrimental to the group.

Policing, punishment, and coercion around social interactions — like the policing of egg-laying ant workers — are also game-changing traits. These change the incentive structure for social behaviours by imposing additional costs or benefits on the base interaction. Models of punishment reduce the temptation to defect in a Prisoner's Dilemma so the ESS is to cooperate (Boyd et al., 2010). More broadly games with side payments (Jackson and Wilkie, 2005) can also be seen within the context of game-changing traits. In games with side payments, games are played in two stages. First, players announce binding transfer functions, 'payments' of utility they promise to make to each other player conditional on the choices of strategies when the game is played. Then they play the game with a payoff matrix that has been effectively rewritten. The perspective taken in games with side payments of a game embedded in a larger game is in the same vein as the metagames concept in Chapter 5.

4.2 The Importance of Assortment

> An altruistic act will be favoured if the actor's behaviour leads to benefits falling on individuals who tend to bear the heritable basis for that same action. — Godfrey-Smith (2009)

> ... Kinship, group structure and reciprocation are all ways of getting altruists to direct their benefits onto each other. — Okasha (2006)

> A cooperative behaviour is altruistic whenever it requires positive assortment between altruists to evolve. —Rosas (2010)

Traits that cause the positive assortment of cooperative behaviours are a particularly important way of changing the game for the evolution of cooperation. As we discussed in Chapter 2, there are many proximate explanations for the evolution of altruism and cooperation. Kin selection, reciprocity, signalling, and group structure have all been advanced as explanations for the origin of altruism under certain conditions, or even as general explanations.

The common feature of these mechanisms is that they create positive assortment on cooperative behaviours, though they do so in different ways. Theoretical work has shown that all can lead to the evolution of cooperation. Increased relatedness assorts genetically-specified cooperative traits together since a cooperator interacting with another individual related by common descent is more likely to be interacting with an individual that shares the cooperative trait (Hamilton, 1964a,b; Maynard Smith, 1978; Grafen, 1979; Hines and Smith, 1979). Reciprocity can assort cooperative acts over time if one cooperative act is rewarded by another one (Axelrod and Hamilton, 1981). Small groups formed from large populations will have a high variance in group composition, with some groups assorting cooperators due to the sampling process (Wilson and Colwell, 1981).

Since all these proximate mechanisms enable the evolution of cooperation and all generate positive assortment between cooperators, it implies that it is not that any particular mechanism is essential to the evolution of cooperation but that the mechanism is one that generates assortment. The assortment of cooperative behaviours is the ultimate explanation for the evolution of cooperation, while the mechanisms that generate it, such as group structure, relatedness, reciprocity or non-random spatial distribution, are proximate explanations (Godfrey-Smith, 2009). Godfrey-Smith (2009) unpacks the role of assortment in the evolution of cooperation (specifically altruism) into two components: altruistic individuals have higher fitness within generations if behaviours are correlated, and altruism can survive and spread across generations if benefits fall on those with a tendency to pass it on.

Given this fundamental importance of assortment to the evolution of altruism, Rosas (2010) goes even further and argues that assortment should actually be the defining characteristic of altruism — that the definition should be that "a cooperative behaviour is altruistic whenever it requires positive assortment between altruists to evolve". This approach is in contrast to that of the inclusive fitness theorists behind the definition of altruism used in this thesis (Section 2.2.1) which is based on fitness consequences to the individual, with an altruistic act one that provides a benefit to the recipient at net cost to the actor (West et al., 2007). They reject including 'weak altruism' (our narrow-sense cooperation) and reciprocal altruism as true instances of altruism since the actor still gains in fitness from their social act, though not necessarily as much as the recipient of the act. Rosas' assortment-based definition is more permissive and closer to the wider sense in which we use cooperation in this thesis. There is serious semantic disagreement as to what constitutes altruism, and which is more primitive, altruism or assortment. However, there is also a sense in which both are right — there is an important difference between those social behaviours that have a net cost to the actor and those that do not, but there is also a very important way in which assortment supports the evolution of both types of cooperative behaviour. Ultimately, since cooperation can be achieved by other means such as policing, while Rosas' definition is provocative we use the more 'standard' definition of altruism in this thesis.

However, Rosas' account makes the important point that the mechanisms that enable altruism (in our sense, cooperation) must reliably emerge as traits of the individual altruist if altruism is to evolve. Typically work in the evolution of cooperation focuses on how assortment is achieved, as assortment itself is not seen as a mechanism but as the consequences of one (Nowak, 2006). However the importance of assortment means it is valuable to investigate it in the abstract, as well as considering the different mechanisms that are capable of creating or maintaining assortment to support altruism (Fletcher and Doebeli, 2009). This supports the methodology we adopt in this thesis of investigating the coevolution of social behaviours (such as altruism) and the game-changing traits that affect them (such as assortment).

4.3 Mechanisms for Assortment

4.3.1 Assortment versus Relatedness

Perhaps the most important proximate mechanism for generating assortment is relatedness due to common descent. If a cooperative trait is genetically determined, then a cooperator's relatives are more likely to share that cooperative trait than a random member of the population (Hamilton, 1964a,b; Grafen, 1979). But while it is not disputed that relatedness generates assortment, there is disagreement as to whether assortment

is important beyond relatedness. It is argued that if the only practical manner of generating assortment is relatedness by common descent then it is relatedness that is the key to the evolution of altruism and assortment is a property generated by relatedness. If, on the other hand, assortment is a more general phenomenon than relatedness, then relatedness is a (significant) example of a mechanism for generating assortment. The former view is advocated by some theorists who give particular primacy to explanations on the level of the gene, as relatedness is an inherently genetic concept while assortment does not have to be. The latter case is argued by those who claim that assortment can promote altruism without relatedness (Fletcher and Zwick, 2006).

There is certainly a very strong parallel between explanations of relatedness such as that in Bourke (2011) and accounts of assortment. Relatedness is defined as a measure of the chance of sharing a gene at a given locus above the average in population, while positive assortment is a measure of the chance of encountering an individual with a given property above the average in a population. In kin interactions, positive assortment and gene frequency overlap — the assortment on a gene at a given locus is the same as the relatedness of individuals bearing it. Though it is an abstract point, those overly in thrall to the gene risk conflating the type-token distinction found in philosophy — namely that there is a difference between a type (such as a letter) and the physical marks that instantiate it in the world.

On their own, genes as they physically exist in the world are just markers that have no inherent semantic meaning. The appropriate interpretation of a gene, such as a gene for altruism, is heavily context-dependent. Interpreting what a gene 'means' requires the developmental process that may express the gene, the regulatory mechanisms that can turn the gene on and off and the social environment of the bearer. Social niche construction affects the evolution of gene frequencies, but it can also change their semantic meaning by aligning genes for social strategy traits with genes for social structuring traits. Traits that cause individuals to share their reproductive fate will alter the semantic meaning of selfish genetic traits by changing the consequences of their expression. A selfish trait that benefits one individual at the expense of another would become a deleterious trait if other mechanisms cause both individuals to share their reproductive outcomes.

Using a linguistic metaphor, inheritance transmits both the 'syntax' of sequences of genes and the 'semantics' of how the genetic information is phenotypically expressed. Seen this way, relatedness is a syntactic concept. It is about the chance of sharing particular markers, like the chance of sharing sequences of letters in a sentence. Assortment is a semantic concept. We can still talk in a limited sense about the assortment of individuals according to the value of certain alleles as in relatedness, but individuals can also assort on phenotypically expressed behaviours like altruism. In our linguistic metaphor, whereas sentences would be related by shared markers, they could be assorted by shared meaning.

This interplay between 'syntactic' genotypes and 'semantic' phenotypes has been analysed in mathematical models. Queller (1985) reformulates Hamilton's rule in terms of the covariance between the *phenotype* of the actor and the *genotype* of the recipients. In the model, altruistic phenotypes survive if the benefits they provide fall disproportionately on those with the ability to pass the behaviour on, but does not require that the altruist and recipients are related by common descent. This is important to the ultimate explanation for the evolution of cooperative behaviours since if relatedness is not essential to the explanation but can be replaced by correlations between phenotypes and genotypes then it is correlation not relatedness that is the key to the mathematical models like kin selection (Godfrey-Smith, 2009).

However, recognising that assortment is an ultimate explanation and relatedness a proximate one is not in opposition to inclusive fitness accounts of the evolution of altruism. Hamilton (1975) puts assortment at the heart of inclusive fitness, with simple measures of assortment in asexual single-population models replaced by the regression coefficients when moving to sexual models with overlapping generations. It is no coincidence that the extreme altruism involved in a major transition requires the evolution of game-changing traits such as unicellular bottlenecks that create essentially clonal relatedness among the transitioning entities (Section 2.5). Relatedness due to common descent is a particularly powerful mechanism for assortment because it is less susceptible to suppressor mutations. This becomes clear when you consider the difference with other mechanisms for creating assortment like signalling.

4.3.2 Assortment due to Signalling

The difference between assortment in the abstract and assortment due to relatedness is shown in the 'greenbeard' problem that can undermine a simpler signalling mechanism for cooperation. The problem was first posed by Hamilton (1964b) and named by Dawklns (1976). Imagine all carriers of a cooperative gene possess a marker — for instance a green beard — that enables them to recognise other cooperators and hence interact preferentially with them, assorting the interaction of cooperative behaviours. Then there will be an evolutionary incentive both for defector genotypes to evolve their own greenbeard marker, essentially engaging in parasitism (Okasha, 2002), and also for intragenomic conflict with the cooperator genotype evolving suppressor traits at other *loci* so that they retain the green beard but do not expend the fitness cost of cooperating.

For this reason it is believed that assortment due to common descent is more stable as relatedness leads to the same average assortment across all the genes of an organism whereas greenbeard traits only lead to correlation on a few genes (Ridley and Grafen, 1981; Bourke, 2011). The problem arises because the marker is not necessarily tied to the cooperative behaviour. Some have argued then that the greenbeard problem is not inevitable as selection for greenbeard and greenbeard-imitation traits may be

aligned (Gardner and West, 2010; Biernaskie et al., 2011). In the extreme case where the marker is the cooperative behaviour itself, the situation is essentially equivalent to one of reciprocal altruism. There is evidence for the existence of greenbeard traits in ants (Keller and Ross, 1998) and social amoebae (Queller et al., 2003), so greenbeard mechanisms can exist even if they may be unstable.

A further consequence of the fact that assortment is not uniform across the genome where assortment is generated by signalling but is when it is due to relatedness, is that when altruistic behaviours coevolve with the game-changing traits that support assortment, relatedness should entail assortment on both the social and game-changing traits, which a mechanism such as signalling would not do.

4.4 Transforming Games

Game-changing traits then, such as assortment, can alter the social conditions of a population's interactions. Maynard Smith (1978) brought this idea into evolutionary game theory, discussing how the fitness of strategies in evolutionary games can differ according to the social environment of the players, specifically the fact that fitness will be different if games are played between relatives than if they are played by unrelated individuals. Maynard Smith proposed to recognise the kin selection benefits of games between relatives in a Hawk-Dove game (which we call the Snowdrift Game) by adding to the payoff of a strategy the payoff of the opponent's strategy multiplied by the coefficient of relatedness r.

Grafen (1979) argued that this method was invalid because a player accrues inclusive benefits from the whole population, not just those members it directly plays, and that in a related population an individual is more likely to be playing against its own strategy than in a well-mixed one (because relatedness creates assortment). Grafen mathematically derived a different solution for the ESS frequency of cooperators in the Snowdrift Game. Rewriting the payoff labels to match those in the rest of this thesis, the ESS frequency of cooperators in a Snowdrift Game in a population of relatedness r is:

$$x_C = \frac{S - P + r(R - S)}{(1 - r)(R - S - T + P)} \tag{4.1}$$

Further theoretical models analysed how cooperation could evolve in conditions where it was otherwise impossible due to explicitly represented population structure, such as spatial clustering of cooperators (Nowak and May, 1992). This has developed into an extensive body of literature in games played on networks. Different network topologies, such as fully connected or scale free degree distributions, can change the social equilibria of games played on that network (Santos et al., 2006a), while dynamically changing the

network topology can also be equivalent to transforming the payoff matrix of the game (Pacheco et al., 2006b,a).

It is this connection between the effect of changing social conditions and transformations to the payoff matrix of the evolutionary game that we make use of in this thesis. Instead of explicitly representing population structure, there have been derivations of the effective payoff matrix for games played with reciprocity on networks (Ohtsuki and Nowak, 2006), direct and indirect reciprocity and kin and group selection (Taylor and Nowak, 2007), punishment (Boyd et al., 2010) and non-random matching (Van Veelen, 2011). Taylor and Nowak (2007) present a number of these transformations.

4.4.1 Direct and Indirect Reciprocity

Direct reciprocity can occur in games where interactions are repeated (Trivers, 1971). In a basic model of direct reciprocity, defectors defect each round while cooperators play tit-for-tat (cooperating at first, then reciprocating the opponents move) (Axelrod and Hamilton, 1981), with w the probability of another round in the game. For this model, Taylor and Nowak (2007) derive the payoff matrix transformation:

$$\begin{pmatrix} \frac{R}{1-w} & S + \frac{wP}{1-w} \\ T + \frac{wP}{1-w} & \frac{P}{1-w} \end{pmatrix} \tag{4.2}$$

All-cooperators is an ESS if $w > \frac{T-R}{T-P}$, so cooperation can evolve if there are enough rounds in the game.

Indirect reciprocity extends the idea of direct reciprocity so that an individual's strategy choice is based not just on the behaviour of the opponent in previous games between the two, but also the opponent's behaviour in all previous games – in essence, the development of a reputation system. If q is the probability of knowing the opponent's reputation, the payoff matrix transformation is:

$$\begin{pmatrix} R & (1-q)S + qP \\ (1-q)T + qP & P \end{pmatrix} \tag{4.3}$$

All-cooperators is an ESS if $q > \frac{T-R}{T-P}$ much as in the case for direct reciprocity.

We can use the ST-plane as a tool here to geometrically interpret Taylor and Nowak (2007)'s transformations to the game. If we restrict ourselves to an arbitrary game $\left(\begin{smallmatrix} 1 & S \\ T & 0 \end{smallmatrix}\right)$ on the ST-plane, then for indirect reciprocity the transformed game matrix is still in the ST-plane:

$$\begin{pmatrix} 1 & (1-q)S \\ (1-q)T & 0 \end{pmatrix} \tag{4.4}$$

This is not the case for direct reciprocity due to the $\frac{R}{1-w}$ term. However, we can project the transformed matrix back onto the ST-plane:

$$\begin{pmatrix} \frac{1}{1-w} & S + \frac{0}{1-w} \\ T + \frac{0}{1-w} & \frac{0}{1-w} \end{pmatrix} = \begin{pmatrix} \frac{1}{1-w} & S \\ T & 0 \end{pmatrix} \to \begin{pmatrix} 1 & (1-w)S \\ (1-w)T & 0 \end{pmatrix} \tag{4.5}$$

We see that both direct and indirect reciprocity induce the same geometric transformation of the plane (substituting w and q). This is a scaling of the ST-plane towards the point corresponding to the game $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ on the boundary of the Harmony Game and the Stag Hunt Game. We illustrate this transformation in Figure 4.1 for a selection of starting points across ST-space. The arrows show the direction of the transformation along lines that are the continuous image of the transformation as w (or q) ranges from 0 to 1.

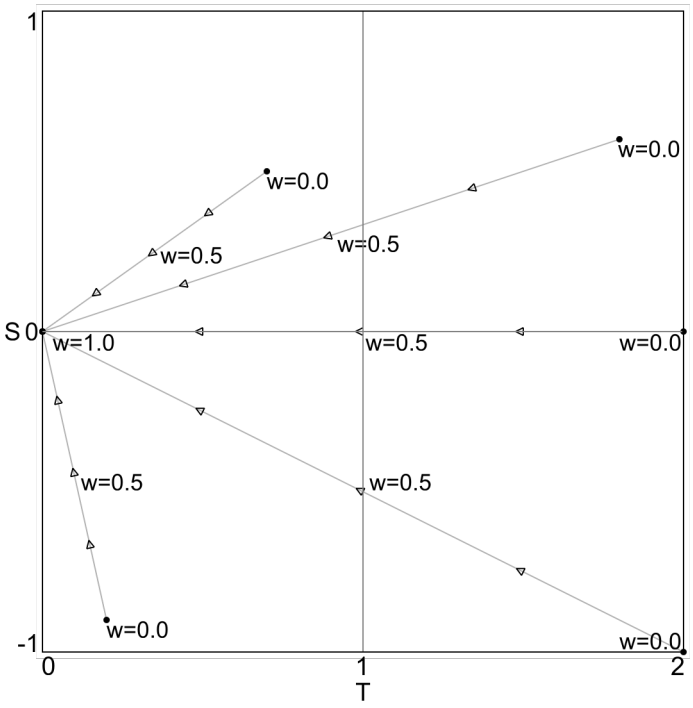


Figure 4.1: The transformation of a selection of points in ST-space due to the effect of reciprocity (labelled with w, though q would be equivalent). The arrows show the direction of the transformation along lines that are the continuous image of the transformation as w (or q) ranges from 0 to 1.

This clearly shows why higher w or q will lead to an increased chance of cooperation. Harmony and Stag Hunt Games will remain in the same game quadrants: Harmony Games will become ones in which defector receive a lower payoff, while Stag Hunt Games

move further into the basin of attraction of the all-cooperators ESS, reaching an all-cooperators equilibrium over a wider range of initial population compositions. If w or q are sufficiently large then Snowdrift Games will become Harmony Games and Prisoner's Dilemmas will become Stag Hunt Games.

4.4.2 Kin Selection

In Taylor and Nowak (2007)'s model of the effects of Kin Selection, when there is relatedness r the transformed payoff matrix is:

$$\begin{pmatrix} R & \frac{S+rT}{1+r} \\ \frac{T+rS}{1+r} & P \end{pmatrix} \tag{4.6}$$

All cooperators is an ESS if $r > \min\left\{\frac{T-R}{R-S}, \frac{P-S}{T-P}\right\}$. Note however that this formulation of Kin Selection is derived from that in Maynard Smith (1978), and so is subject to the same critique by Grafen (1979): it does not account for assortment and does not recognise the fact that kin selection benefits accrue even between non-playing members of the population.

4.5 Transforming a Game by Assortment

Direct and indirect reciprocity and kin selection are all proposed methods for generating assortment. As well as considering proximate mechanisms, it is instructive to model assortment in the abstract. Here we outline a mathematical equivalence between games played by a population with an increased level of assortment and the effective game that would have to be being played in a freely mixed population to have the same equilibria. The payoff matrix transformation for assortment we describe has a simple geometric interpretation on the ST-plane, as it acts as a linear transformation of the plane. We will then formalise and generalise the method we demonstrate here for assortment to provide a general way of calculating payoff matrix transformations to represent the effect of game-changing traits that lead strategy-types to interact with each other at non-random frequencies.

All the games here are two-player two-strategy games representable by matrices. As set out in the previous chapter, we denote an arbitrary such game as $G = \left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$, with strategy set C, D corresponding to cooperators and defectors. The population vector is $x = \left(\begin{smallmatrix} x_C \\ x_D \end{smallmatrix}\right)$, so x_C is the frequency of cooperators in the population, x_D the frequency of defectors, and $x_C + x_D = 1$.

Recall that the fitness of a type is calculated as the expected payoff for a member of that type:

$$f(x_C) = Rx_C + Sx_D \tag{4.7}$$
$$f(x_D) = Tx_C + Px_D \tag{4.8}$$

The average fitness of a member of the population is $\bar{f}(x) = \Sigma^n_{i\in\{C,D\}} x_i f(x_i) = x_C f(x_C) + x_D f(x_D)$. For the type C we can rewrite the replicator equation in terms of fitness functions as:

$$\begin{aligned} \dot{x_C} &= x_C(f(x_C) - x_C f(x_C) - x_D f(x_D)) \\ &= x_C((1 - x_C)f(x_C) - x_D f(x_D)) \\ &= x_C(1 - x_C)(f(x_C) - f(x_D)) \end{aligned} \tag{4.9}$$

Substituting in the fitness functions gives:

$$\dot{x_C} = x_C(1 - x_C)((R - S - T + P)x_C - (P - S)) \tag{4.10}$$
$$\dot{x_D} = x_D(1 - x_D)((R - S - T + P)x_D - (R - T)) \tag{4.11}$$

Let $\alpha \in [0, 1]$ be a measure of positive assortment. The definition of assortment needs to be clarified. It is unambiguous what is meant by a fully assorted game ($\alpha = 1$) — cooperators always play against cooperators, defectors against defectors. Similarly, a completely anti-assorted game ($\alpha = 0$) is one in which cooperators always encounter defectors. No assortment ($\alpha = 0$) means encounters are random, and so happen with the same frequencies as the population composition. Following Hamilton (1975), we understand an intermediate level of α to mean that a strategy type will encounter an individual of the same type with probability α or else play a random member of the population with probability $1-\alpha$ (so the actual probability of type i playing an individual of the same type is $\alpha + (1 - \alpha)x_i$). This change in the frequency at which members of the population with different social behaviour traits interact leads to a change in the expected payoff from playing the game for each strategy and hence to the fitness for that strategy.

For an assortment coefficient of $\alpha \in [0, 1]$ we can write the changed fitness functions f^α:

$$f^\alpha(x_C) = \alpha R + (1 - \alpha)f(x_C) \tag{4.12}$$
$$f^\alpha(x_D) = \alpha P + (1 - \alpha)f(x_D) \tag{4.13}$$

Substituting the new fitness functions into the replicator equation gives:

$$\begin{aligned}
\dot{x_C} &= x_C(1-x_C)\Big(f^\alpha(x_C) - f^\alpha(x_D)\Big) \\
&= x_C(1-x_C)\Big((\alpha R + (1-\alpha)f(x_C)) - (\alpha P + (1-\alpha)f(x_D))\Big) \\
&= x_C(1-x_C)\Big(\alpha(R-P) + (1-\alpha)(f(x_C) - f(x_D))\Big) \\
&= x_C(1-x_C)\Big(\alpha(R-P) + (1-\alpha)((R-S-T+P)x_C - (P-S))\Big) \\
&= x_C(1-x_C)\Big((1-\alpha)(R-S-T+P)x_C - ((1-\alpha)(P-S) - \alpha(R-P))\Big)
\end{aligned} \tag{4.14}$$

To simplify the algebra, we let $A = R-T$, $B = P-S$. These correspond to the measures of greed and fear in the game discussed in the context of the ST-plane. We can then see the transformation to the replicator equation caused by a positive assortment level α:

$$\begin{gathered}
\dot{x_C} = x_C(1-x_C)((A+B)x_C - B) \mapsto \dot{x_C} = x_C(1-x_C)((1-\alpha)(A+B)x_C - ((1-\alpha)B - \alpha(R-P))) \\
A \mapsto (1-\alpha)A + \alpha(R-P) \\
B \mapsto (1-\alpha)B - \alpha(R-P)
\end{gathered} \tag{4.15}$$

This is equal to the replicator equation for the game

$$\begin{pmatrix} R & S + \alpha(R-S) \\ T + \alpha(P-T) & P \end{pmatrix} = (1-\alpha)\begin{pmatrix} R & S \\ T & P \end{pmatrix} + \alpha\begin{pmatrix} R & R \\ P & P \end{pmatrix} \tag{4.16}$$

So the dynamics of the game $\left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$ played with assortment level α are the same as those of the game $(1-\alpha)\left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right) + \alpha\left(\begin{smallmatrix} R & R \\ P & P \end{smallmatrix}\right)$ in a freely mixed population.

For any game $G = \left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$ there exists a corresponding fully assorted game $G^{assort} = \left(\begin{smallmatrix} R & R \\ P & P \end{smallmatrix}\right)$. As assortment increases, the effective game played moves linearly through the space of possible games from the point (R, S, T, P) to the point (R, R, P, P). If we define G^α to be the effective game given an assortment level of α (so $G = G^0$, $G^{assort} = G^1$) then it is the linear combination:

$$G^\alpha = (1-\alpha)G + \alpha G^{assort} = (1-\alpha)\begin{pmatrix} R & S \\ T & P \end{pmatrix} + \alpha\begin{pmatrix} R & R \\ P & P \end{pmatrix} \tag{4.17}$$

Note that if G obeys the rule that $R > P$ then so does G^{assort}, and if G is on the ST-plane then so is G^α for all $\alpha \in [0, 1]$. The effect of this transformation on a number

of points across ST-space in shown in Figure 4.2 (compare to Figure 4.1 showing the transformation due to reciprocity). This transformation scales the space towards the most extreme Harmony Game in ST-space.

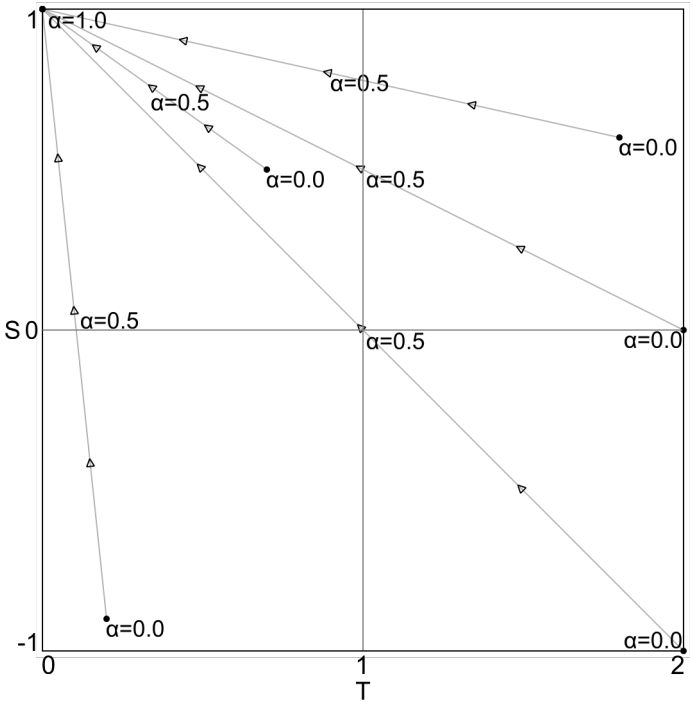


Figure 4.2: The transformation of a selection of points in ST-space due to the effect of assortment. The arrows show the direction of the transformation along lines that are the continuous image of the transformation as α ranges from 0 to 1.

The polymorphic ESS for a transformed payoff matrix in the Snowdrift Game is:

$$x_C = \frac{S - P + \alpha(R - S)}{(1 - \alpha)(R - S - T + P)} \tag{4.18}$$

This is precisely the same result as that for the ESS derived by Grafen (1979) in the 'discrete' case where an individual could only play 'Hawk' (defect) or 'Dove' (cooperate) (Equation 4.1, as we do here, but with relatedness r replaced by our assortment index α.

4.5.1 The Effect of Assortment on the Evolution of Cooperation

How does this transformation affect the evolution of cooperation? When $\alpha = 0$ then this is just the well-mixed case and there is no transformation to the game. When $\alpha = 1$ then the effective game is $G^{assort} = \left(\begin{smallmatrix} R & R \\ P & P \end{smallmatrix}\right)$; as we required that $R > P$ this means that cooperation dominates defection so all cooperators will always be the ESS — recall from Chapter 3 that dominated strategies always go asymptotically extinct under the

replicator dynamics. This means that no matter what social dilemma G represents, G^{assort} is a Harmony Game. For intermediate values of α, then given the transformed payoff matrix we know that cooperation dominates defection if both:

$$R > (1-\alpha)T + \alpha P \tag{4.19}$$
$$(1-\alpha)S + \alpha R > P \tag{4.20}$$

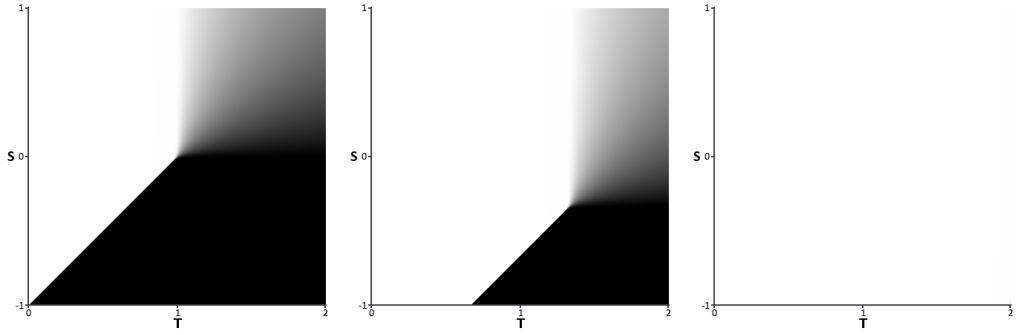
On the ST-plane this means that all cooperators will be the ESS if:

$$1 > (1-\alpha)T \Rightarrow T < \frac{1}{1-\alpha} \tag{4.21}$$
$$(1-\alpha)S + \alpha > 0 \Rightarrow S > -\frac{\alpha}{1-\alpha} \tag{4.22}$$

As α increases from 0 to 1, $\frac{1}{1-\alpha} \to \infty$ and $-\frac{\alpha}{1-\alpha} \to -\infty$, so the basin of attraction of the all-cooperators ESS increases to encompass the entirety of the ST-plane. In particular, cooperation is the ESS over the whole of ST-space (the region of the ST-plane where $-1 \leq S \leq 1$ and $0 \leq T \leq 2$) if:

$$2 < \frac{1}{1-\alpha} \Rightarrow 2 - 2\alpha < 1 \Rightarrow \alpha > 0.5 \tag{4.23}$$
$$-1 > -\frac{\alpha}{1-\alpha} \Rightarrow 1 - \alpha > \alpha \Rightarrow \alpha > 0.5 \tag{4.24}$$

So an assortment level of $\alpha = 0.5$ is sufficient to make all-cooperators the ESS outcome over all the social dilemmas. Geometrically we see that the payoff matrix transformation due to assortment induces a transformation of the ST-plane, shrinking it towards the point $S = 1, T = 0$, and in the limit of full assortment collapsing the entire space into this point. This is shown in Figure 4.3, displaying the equilibrium frequency of cooperators on the same white-black scale as Figure 3.2, where white is a population of all-cooperators as equilibrium. This geometric understanding makes it clear that assortment is the most favourable continuous transformation of ST-space for cooperators, since it is a scaling of the space towards the single point in ST-space most favourable to cooperators and least favourable to defectors.



(a) $\alpha = 0$ (b) $\alpha = 0.25$ (c) $\alpha = 0.5$

Figure 4.3: The equilibrium frequency of cooperators (white) across ST-space where the games are played with assortment level α (initial frequency of cooperators $c = 0.5$, determining the boundary of the all-cooperators equilibrium in the lower right Stag Hunt quadrant). Above $\alpha = 0.5$, the all-cooperators equilibrium covers all of ST-space.

4.6 Interaction Functions

This transformation of the game for the case of assortment illustrates the advantages of this way of understanding changes to a game wrought by changes in population structure. The results agree with other derivations of this transformation for assortment, such as Van Veelen (2011) that uses group partitioning to derive the same result, and with the formula for the equilibrium frequency of cooperators in the Snowdrift Game presented in Grafen (1979) when assortment due to relatedness is replaced by assortment in the abstract.

However, as presented the technique may seem ad hoc. Here we expand on the mathematical principles behind this transformation in the form of *interaction functions*, a method for transforming the game to reflect non-random interaction frequencies with assortment as a special case. Evolutionary game theory models using the replicator equation determine the evolutionary success of a strategy type based on the difference between the fitness of the type and the mean fitness of the population. The fitness of an individual committed to a given strategy is modelled as its expected payoff from the social dilemma it is engaged in — the payoff for its interaction with each other strategy type multiplied by the probability of interacting with that type. When the population is freely mixed, these interaction probabilities match the frequencies at which the types occur in the population, but game-changing traits can change these interaction probabilities, for instance by creating positive assortment on behavioural traits.

Interaction functions are a simple mechanism for modelling such a situation. They have the useful feature that if the mappings between the population frequencies and the encounter frequencies satisfy certain properties then any game with a payoff matrix A played with the changed encounter frequencies is equivalent to a different game with

payoff matrix A' played in a well-mixed population. When the encounter functions do not possess this property the resulting dynamics can still be modelled using the replicator equation, but because of non-linear effects the resultant payoff matrix of the game can no longer be neatly specified.

To accurately define the transformations determined by interaction functions, we need to be more precise about the mathematical setting of this work. Formally, when we talk about the space of possible symmetric two-player n-strategy games (in normal form) we are talking about a matrix $A \in M_{n\times n}(\mathbb{R})$. There is a vector space isomorphism $\Phi : M_{n\times n}(\mathbb{R}) \to \mathbb{R}^{n^2}$, that maps the matrix entry $a_{i,j}$ to the $(ni + j)$-th element of the vector. As the spaces are isomorphic it is generally unnecessary to differentiate between the matrix form and the n-tuple of payoffs that define it, and so throughout this work we have switched freely between talking about the game as given by the matrix and as a point in a spatial representation of the game. But to be precise, when we transform the game payoff matrix of the game here we are actually operating on the column vector of payoffs, and using the vector space isomorphism to switch between the two.

In the replicator equation, the population state vector $x \in S_n$ is a vector of the frequencies of different strategy types in a population. But what actually matters when determining fitness under the replicator equation model is not the frequency of the different types in the population but the frequency of *interactions* with the different types in the population. The sum of the frequencies at which the different types are encountered must by definition be one, so the interaction frequency vector is also defined on the simplex $S_n = \{(x_1, \ldots, x_n : \sum_i^n x_i = 1\}$, the set of all n-tuples where the entries sum to one.

We can define a family of n *interaction functions* $e_i : S_n \to S_n$ that map the population state vector to the actual frequencies at which the i-th strategy type interacts with other strategy types. In the case of a freely mixed population these are the same, so all encounter functions are just the identity transformation (they leave the population state unchanged).

The fitness of the i-th genotype is modelled as the expected payoff for that genotype given the actual frequency at which it encounters other genotypes. We let the fitness of the i-th genotype be $f_i(x)$. Essentially the set of fitness functions $f = \{f_1, \ldots, f_n\}$ define a game played against a mixed population. If there is a matrix A such that $f_i(x) = (Ax)_i$ then we say that A is the payoff matrix for the game f. In the generalised case where the population is not well-mixed the fitness of the type is the composition of the fitness and interaction functions $f_i \circ e_i(x)$, and the mean fitness is $\sum_j^n x_j f_j \circ e_j(x)$. This gives the new replicator equation:

$$x_i = \dot{x_i}\Big(f_i \circ e_i(x) - \sum_j^n x_j f_j \circ e_j(x)\Big) \tag{4.25}$$

The set of fitness functions $g_i : g_i = f_i \circ e_i$ then defines a new game g. g played in a well-mixed population is equivalent to the original game f played with the changed population structure. We can further show that if each e_i is an affine function and f can be represented by the payoff matrix A then g can be represented by a payoff matrix A' with $x_i = \dot{x_i}\Big(f_i \circ e_i(x) - \sum_j^n x_j f_j \circ e_j(x)\Big) = \dot{x_i}\Big((A'x)_i - x^T A' x\Big)$.

It is possible to construct a matrix $E \in M_{n^2 \times n^2}(\mathbb{R})$ such that $A' = (\Phi^{-1} \circ E \circ \Phi)(A)$. This allows comparisons between games played under the effects of different game-changing traits to be recast in terms of comparisons between equivalent effective games played in freely-mixed populations.

4.6.1 Affine Interaction Functions

Mathematically an *affine function* is one that maps each entry v_j of a vector $v = (v_1, \ldots, v_n)^T$ to $k_0 + k_1 v_1 + \cdots + k_n v_n$ — a linear combination of the elements of the vectors plus a constant. Affine functions can in general be represented in homogenous coordinates by a $(n+1) \times n$ matrix. However, because interaction functions map the simplex S_n to itself, the constant term $k_0 = k_0(v_1 + \cdots + v_n)$ so $v_j \mapsto (k_1 + k_0)v_1 + \cdots + (k_n + k_1)$. This means any affine function from the simplex S_n to itself has a representation as a linear function, hence any affine encounter function e_i can by represented by a matrix E_i.

Example: Positive Assortment

To demonstrate the use of interaction functions, we return to the example of positive assortment in two-player two-strategy games we looked at before. The assortment coefficient α is the minimum probability that a genotype will encounter itself, the rest of the time the genotype encounters others at their frequency in the population (including possibly itself). So if $i = j$, $e_i(x_j) = \alpha + (1 - \alpha)x_j$; otherwise $e_i(x_j) = (1 - \alpha)x_j$. For two-strategy games this means both e_1 and e_2 are affine functions as they are linear combinations of the entries of a vector and possibly a constant α. We can therefore use the fact that $x_1 + x_2 = 1$ to represent the encounter functions as 2×2 matrices.

$$e_1 : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \alpha + (1-\alpha)x_1 \\ (1-\alpha)x_2 \end{pmatrix} = \begin{pmatrix} x_1 + \alpha x_2 \\ (1-\alpha)x_2 \end{pmatrix}, E_1 = \begin{pmatrix} 1 & \alpha \\ 0 & (1-\alpha) \end{pmatrix} \tag{4.26}$$

$$e_2 : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} (1-\alpha)x_1 \\ \alpha + (1-\alpha)x_2 \end{pmatrix} = \begin{pmatrix} (1-\alpha)x_1 \\ x_2 + \alpha x_1 \end{pmatrix}, E_2 = \begin{pmatrix} (1-\alpha) & 0 \\ \alpha & 1 \end{pmatrix} \tag{4.27}$$

For two-strategy games the population state can be represented by a single number x_1 (since $x_1 + x_2 = 1$) so E_1 and E_2 must represent the same transformation with a different basis. Indeed E_1 and E_2 are similar matrices with $E_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) E_2 \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

4.6.2 Transforming the Payoff Matrix

When the encounter functions are affine and so can be written as transformation matrices, we immediately have:

$$g_i(x) = (f_i(e_i(x)) = (AE_i x)_i \tag{4.28}$$

Let A' be the matrix of the transformed game. We show it exists by constructing it. If A' does exist it means:

$$\forall x \in S_n, (AE_i x)_i = (A'x)_i \tag{4.29}$$

In particular, since matrix multiplication is associative this means that the i-th row of A' is equal to the i-th row of AE_i. If we write $a_{i,j}$ and $a'_{i,j}$ for the i-th row and j-th column entries of the payoff matrices A and A' respectively, then $a'_{i,j}$ is equal to the i, j-th entry of AE_i. If we write E_i as:

$$E_i = \begin{pmatrix} c_{i,11} & \cdots & c_{i,1n} \\ \vdots & \ddots & \\ c_{i,n1} & & c_{i,nn} \end{pmatrix} \tag{4.30}$$

The i, j-th element of AE_i is $\sum_{k=0}^{n} a_{i,k} c_{i,kj} = a'_{i,j}$, which gives the complete matrix for A'. Then there is some transformation E that sends $a_{i,j}$ to $\sum_{k=0}^{n} a_{i,k} c_{i,kj}$. To construct the matrix of the transformation E we consider the n^2-tuples $\Phi(A)$ and $\Phi(A')$ that are the images of the payoff matrices under the vector space isomorphism defined previously. This then gives a formula for the matrix E - it is the $n^2 \times n^2$ matrix formed by placing the transposes of the n encounter matrices $E_1, \ldots, E_n$ on the diagonal with zero elsewhere.

$$E = \begin{pmatrix} E_1^T & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & E_n^T \end{pmatrix} \tag{4.31}$$

And we have $\Phi(A') = E\Phi(A) \implies A' = (\Phi^{-1} \circ E \circ \Phi)(A)$ (using the vector space isomorphism to switch between spaces).

Example: Positive Assortment

Returning to the example of positive assortment in a two-player two-strategy social dilemma, we have the fitness functions:

$$f_1(x) = \left(\begin{pmatrix} R & S \\ T & P \end{pmatrix}\begin{pmatrix} x_1 + \alpha x_2 \\ (1-\alpha)x_2 \end{pmatrix}\right)_1 = \begin{pmatrix} Rx_1 + \alpha R x_2 + (1-\alpha)Sx_2 \\ Tx_1 + \alpha T x_2 + (1-\alpha)Px_2 \end{pmatrix}_1 = \alpha R + (1-\alpha)(Rx_1 + Sx_2) \tag{4.32}$$

$$f_2(x) = \left(\begin{pmatrix} R & S \\ T & P \end{pmatrix}\begin{pmatrix} (1-\alpha)x_1 \\ x_2 + \alpha x_1 \end{pmatrix}\right)_2 = \begin{pmatrix} (1-\alpha)Rx_1 + \alpha S x_1 + Sx_2 \\ (1-\alpha)Tx_1 + \alpha P x_1 + Px_2 \end{pmatrix}_2 = \alpha P + (1-\alpha)(Tx_1 + Px_2) \tag{4.33}$$

From equation 4.31 we have the value of the transformation matrix:

$$E = \begin{pmatrix} E_1^T & 0 \\ 0 & E_2^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \alpha & (1-\alpha) & 0 & 0 \\ 0 & 0 & (1-\alpha) & \alpha \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.34}$$

Applying this transformation:

$$\begin{aligned} (\Phi^{-1} \circ E \circ \Phi)(A) &= \Phi^{-1}\Bigg(\begin{pmatrix} 1 & 0 & 0 & 0 \\ \alpha & (1-\alpha) & 0 & 0 \\ 0 & 0 & (1-\alpha) & \alpha \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ S \\ T \\ P \end{pmatrix} \Bigg) \\ &= \Phi^{-1}\Bigg(\begin{pmatrix} R \\ \alpha R + (1-\alpha)S \\ (1-\alpha)T + \alpha P \\ P \end{pmatrix} \Bigg) \\ &= \begin{pmatrix} R & \alpha R + (1-\alpha)S \\ \alpha P + (1-\alpha)T & P \end{pmatrix} \end{aligned} \tag{4.35}$$

Which is the same as the transformed matrix calculated directly earlier.

4.6.3 Non-Affine Interaction Functions

The transformation induced by interaction functions is very useful for games where the interaction functions are linear. However, there are many more possible interaction functions that are not linear. This frequently happens in modelling due to the need for the interaction functions to result in valid population states. A simple example of a non-linear encounter function is one in which a member of the population is twice as likely to interact with an individual of the same strategy type as it would be given the population frequencies.

If $A = \left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$ then naively we might put $e_1\left(\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)\right) = \left(\begin{smallmatrix} 2x_1 \\ x_2 \end{smallmatrix}\right)$. Clearly though if $x_1 = 1$ this would define an invalid population state, as the frequencies of the types in the population must sum to 1. In general any map of S_n can be normalised to one of S_n to itself by dividing the resulting vector by the sum of its entries. In this example the correct interaction function is:

$$e_1 : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \frac{1}{2x_1 + x_2} \begin{pmatrix} 2x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{2x_1}{x_1+1} \\ \frac{x_2}{2-x_2} \end{pmatrix} \tag{4.36}$$

However, because this interaction function is not linear, it cannot be expressed as a transformation matrix. Hence the game defined by the family of fitness functions does not have a corresponding game matrix. However, the fitness functions can still be used in the replicator equation, as all that is actually needed to model evolutionary outcomes with the replicator dynamics are functions to calculate the fitness of each strategy type given a population state. We will use this later in this thesis when modelling the effects of non-linear changes in type interaction frequencies.

4.7 Discussion

In this chapter we have defined our concept of game-changing traits. We have reviewed the special importance of game-changing traits that generate assortment, and why relatedness is an important type of assortment but not the only type. We have reviewed how evolutionary game theoretic models can represent game-changing traits as transformations to the payoff matrix of the game, and presented how to do this in the case of assortment. This method agreed with the results of similar derivations (Grafen, 1979; Van Veelen, 2011). We generalised this method into the idea of interaction functions which provide a principled method to derive transformations to a payoff matrix for game-changing traits that affect interaction frequencies.

This has let us see geometrically why assortment is such an important transformation to the game for the evolution of cooperation. When we consider games in ST-space,

assortment scales the ST-plane to the full assorted game $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, the game where cooperation is most advantageous and defection least advantageous. Such a scaling is the most advantageous possible continuous transformation of the plane for cooperators. Other transformations of the plane, such as those for reciprocity, still provide advantages to cooperators but not to the same extent. This supports the biological evidence that assortment is so hugely beneficial to the evolution of cooperation.

With interaction functions and the examples of other transformations to the payoff matrix of games for other game-changing traits, we are ready to introduce metagames. In metagames, we model the coevolution of social strategies and payoff matrices. The mathematical equivalence between game-changing traits and transformations to payoff matrices means we can use metagames to model the coevolution of social strategies and game-changing traits.

Chapter 5

Metagames

In Chapter 4 we demonstrated how to represent game-changing traits that affect the social conditions of a population's interactions as transformations to the payoff matrix of an evolutionary game. In this chapter we consider what happens when these game-changing traits can evolve. Is there a general tendency towards new games that favour increased cooperation or increased selfishness? To investigate this we introduce *metagames* in which the evolution of individual strategies can change the game being played. This provides a minimal model for the coevolution of individuals' social strategies with traits that affect the social context, and the feedback between the two.

We provide the mathematical definitions of metagames and metagame interactions. First we do this in a general, abstract sense, but throughout we illustrate the use of metagames by looking at metagames in ST-space where we can illustrate the overall dynamics of the metagame. These are mapped and the selective pressures on the game visualised using analogs vector field diagrams. We also develop a control model to examine what inherent selective pressures operate due to the nature of game space. While determining the behaviour of the metagame over the whole of ST-space is conceptually useful, the results are more interesting and more relevant when the metagame is constrained, the range of games restricted to particular sets corresponding to particular ways that the game can change. We demonstrate this by looking at metagames on the circular θ-space, the set of games where there is a constant trade-off between the amount of 'fear' and 'greed' in the social dilemma.

Because of the connection between game-changing traits and transformations to payoff matrices we can use metagames to model the coevolution of social strategies and game-changing traits that modify social niches. We claim this means that metagames provide a formal model of social niche construction (Powers et al., 2011; Ryan et al., 2016). Our formalism allows us to comprehensively describe the evolutionary dynamics of game-changing behaviours. We show conditions for Snowdrift and Stag-Hunt games to be transformed into either Prisoner's Dilemmas or Harmony Games, depending on the

type of game-changing trait involved, the constraints on the metagame and the initial frequencies of the social strategies. Importantly, we show that changes to the game driven by selfish individuals maximising their own payoff can still result in games that support increased cooperation.

Finally given the importance of assortment to the evolution of cooperation and the mathematical transformations to payoff matrices reflecting increased assortment developed in Chapter 4, we apply metagames to studying the evolution of assortment. Here we are addressing the thesis research question of the evolution of game-changing traits by creating a model where the level of assortment in the game is not an imposed parameter but a variable that can evolve. We find that the evolution of increased assortment is favoured in regions of game-space where cooperative outcomes are already favoured, so cooperation can beget the conditions for even more cooperation. However, our initial model disagrees with results from the theory of social niche construction that the evolution of cooperation promoting traits will be favoured (Powers, 2010; Powers et al., 2011) in that assortment does not increase in a Prisoner's Dilemma. Indeed, the level of assortment will decrease. We identify assortment on the game-changing trait as a potential explanations, an analysis which leads into Chapter 6.

5.1 Introduction

We have seen how social behaviours, in particular cooperation, are increasingly prominent in evolutionary accounts of the biological world (Chapter 2). They are the key to understanding fundamental processes such as the major transitions, the evolution of independently replicating entities into new higher-level individuals (Buss, 1987; Maynard Smith and Szathmáry, 1997), which are argued to be extreme examples of social integration (Michod, 2000; Queller and Strassmann, 2009; Bourke, 2011). Historically the occurrence of cooperative and altruistic behaviours was seen as a paradox since they benefit others at the possible expense of the actor.

However, social interactions do not occur in the abstract but in social contexts that are themselves subject to evolution. The social conditions of a population can be influenced by the evolution of game-changing traits such as those that affect population structure or establish punishment or policing. When populations are engaged in a social dilemma, it is natural to investigate the potential outcomes of the dilemma – the social equilibrium that might be reached. But it is equally important to understand why the population is playing that particular social game and how the rules or incentive structure of the game might be changed.

Game changing may even be necessary for a social behaviour to persist. Considered in isolation altruistic behaviours should be evolutionarily unsuccessful strategies since the actor incurs a net fitness cost, but the evolution of game-changing traits such as limited

dispersal create social contexts in which altruists assort with other altruists, preferentially directing the benefits of altruism onto the bearers of altruistic traits, changing the nature of the game into one where altruism is sustained. This resolution shifts the problem to explaining the presence of positive assortment (Hamilton, 1964b; Michod and Sanderson, 1985; Lehmann and Keller, 2006). This can also come from genetic relatedness, signalling (greenbeard effects) or where interactions are repeated (Axelrod, 1987). Therefore a complete account of social evolution must consider how altruism and assortment-promoting traits coevolve.

Similarly, evolutionary transitions require the development of conflict modifiers (Chapter 2), adaptations that reduce the inherent tension between two levels of biological organisation (Michod and Herron, 2006). High relatedness, bottlenecks and reproductive specialisation are all ways in which conflict is reduced at the level of the social group. By changing the population structure they change the social context and hence the incentives for social behaviours. If the members of a social group undergoing an evolutionary transition are playing an evolutionary game, the evolution of conflict modifiers changes that game, for instance changing a host-parasite interaction to a mutualism.

Traditionally, evolutionary game theoretic analyses fix the payoff matrix of the game and then study the resulting evolution of behaviours. However, the payoffs of a social game are rarely obvious. This poses a real challenge, since any attempt to specify a particular game runs the dual risk of being arbitrary while significantly affecting the results (Fort, 2008). Roughgarden (2009) has argued for an inversion of the fixed game approach, instead advocating two-level models of behavioural evolution in which the relatively rapid behavioural timescale of goal-driven interactions is separated from the relatively slow evolutionary timescale of population genetics that might alter the nature of the social game. In particular, the payoff matrix of the behavioural game may be subject to evolutionary pressure. This moves beyond the idea of transforming the payoff matrix of a game we discussed in Chapter 4 to considering a dynamic payoff matrix subject to evolutionary change.

Here we follow this general principle. We are interested in the space of possible social games and how a game can be modified through evolution by natural selection. Cooperation is not inevitable, asocial entities and unicellular organisms abound. External conditions, such as resource availability, constrain the future amenability of a social scenario to increased cooperation. But individual traits can alter the context in which the game occurs. For example, a genetic trait that affects the dispersal radius of an entity's offspring will change the likelihood that members of a population will encounter relatives (Pepper and Smuts, 2002). Thus the coevolution of social behaviours and social context can lead to a process of *social niche construction* (Powers, 2010). Prior models have found that population structures supporting increased group selection effects could arise from the coevolution of social behaviours and a genetic preference for group size. A

group size preference can affect the composition of social groups by increasing the variance in group composition (Powers and Watson, 2011), allowing between-group effects to outweigh within-group competition (Wilson and Colwell, 1981). So when social behaviours coevolve with a group size preference, linkage disequilibrium develops between a preference for small groups and the cooperative social trait. A social context supporting increased cooperation coevolves with increasing levels of cooperation, creating a positive feedback loop.

In this Chapter we address the general question: when individuals have traits that can change the game they play, what game does individual natural selection cause them to play? We might assume that because individuals are selfish, when individual traits can change the game, the ever-present opportunity for defectors to gain an advantage by obtaining the benefits of cooperation without incurring the costs will inevitably result in any game ending as a Prisoner's Dilemma. On the other hand, because mutual cooperation is more advantageous than defection, we might anticipate that changes to the game that make the social context more conducive to cooperation will be favoured, or that the tension between individual interests will result in a series of Pareto improvements and ultimately a Pareto-optimal outcome.

There have been scattered investigations of the way that a game can change, typically focused on the specific case of the Prisoner's Dilemma. In a model with learning rather than evolutionary mechanics, Worden and Levin (2007) looked at how a population might escape a Prisoner's Dilemma by periodically introducing additional strategies with mutated payoffs, resulting in a Snowdrift Game. The message being that where suitable mechanisms for transforming social conditions exist, scenarios that start in the form of a Prisoner's Dilemma do not have to end as one. Akçay and Roughgarden (2011) showed the same result by introducing mutants that made side payments. Fort (2008) produced a model to determine social games without specifying the payoff matrix by evolving an initially heterogenous set of games on a spatial lattice, finding that when the initial population was drawn from the standard social dilemmas the resultant game would have the form of a Stag Hunt coordination game.

Other work has explicitly looked at the effects of types of population structure on the evolution of social behaviours, such as how subdivision into particular group structures affects the replicator dynamics (Van Veelen, 2011), or by considering the effective game created by a network structure (Cao et al., 2011). Different network topologies, such as fully connected or scale free degree distributions, can change the social equilibria of games played on that network (Santos et al., 2006a), while dynamically changing the network topology can also be equivalent to transforming the payoff matrix of the game (Pacheco et al., 2006b,a).

In all these works, individuals have evolvable traits that change the nature of the game they are playing, its equilibria and dynamics. But at present there is no unifying theoretical framework to make sense of these diverse examples or to predict how a game changes when its parameters are under individual natural selection. This is the motivation for our introduction of a model of *metagames* in which, as well as governing the frequencies of the social strategies, the evolutionary dynamics affect the payoffs of the social game through the spread of game-changing traits (GCTs) that alter the social context of an individual, changing the incentives for a game without changing its essential nature. We showed in Chapter 4 how many changes to population structure induce changes in the effective payoff matrix of a game. That is, a game with payoff matrix G played in a structured population is mathematically equivalent to a game G' played in a well-mixed population. This work has demonstrated how population structure can be a parameter that changes the game. Here we show, when this parameter is in fact a game-changing trait, how the trait evolves.

Because of the mathematical equivalences between transformations to the game and the effect of game-changing traits, questions about the evolution of social games and game-changing traits can be united in a general framework that predicts how individual selection modifies the social game given any game-changing trait that can be represented by a payoff matrix transformation. We characterise conditions under which Snowdrift or Stag Hunt games are transformed into the Prisoner's Dilemma, and conversely, where they are transformed into the Harmony Game. Further, we show transformations from the Prisoner's Dilemma to the Harmony Game.

5.1.1 Evolving Payoff Matrices

> "I'm not very happy with most of the applications of games theory, because it tends to perpetuate the rules of the game as perceived by the players . . . Nobody knows a thing about changing the rules of the game." — Gregory Bateson Brand (1974)

In Chapter 1 we identified three strands from the literature that serve as the ultimate motivation for the introduction of metagames:

- That the nature of a social game, represented by its payoff matrix, can be subject to evolutionary control.
- That the population structure in which a social group interacts can be subject to evolutionary control.
- That many changes in population structure are formally equivalent to changing the payoff matrix of the social game the population is engaged in.

This can be seen as part of a growing dissatisfaction in evolutionary game theoretic modelling with imposing the parametrisation of the social game by fiat. It is not always obvious what game is the most appropriate to use when using evolutionary game theory models, so models suffer from the twin problem that empirical determination of payoffs can be difficult while model predictions can be very sensitive to the payoffs used (Fort, 2008). These difficulties have led to social scenarios often being modelled as instances of the Prisoner's Dilemma by default. Worden and Levin (2007) dispute this practice, arguing that the overemphasis on the Prisoner's Dilemma has exaggerated the apparent paradox of cooperation. In other games, such as the Snowdrift Game, some level of cooperation is always a viable strategy. In particular they argue that "Beyond comparing different cooperation scenarios, the question of how one scenario might transform into another is much more sparsely investigated". Where such mechanisms exists, social conditions that start in the form of the Prisoner's Dilemma may not end as them. The parallel is clear with the processes we have discussed throughout this thesis where the evolution of population structure and other forms of conflict modifier can change the effective population dilemma.

The work by Fort and by Worden and Levin cite two rare instances where this is examined in the literature, each highlighting the need to better understand the way that social games can change. Mesterton-Gibbons (1991) obliquely raises the issue in a model of insect reproduction where insects must choose locations for egg-laying. Laying eggs in already occupied sites reduces the viability of both eggs, though obviously it is better for the insect than laying no egg at all. The key lies in whether or not the insects can recognise their own eggs from those of others. If this discrimination is possible, then one would expect the insects to freely lay eggs in sites already occupied by unrelated eggs. In its absence though, game theory would predict insects sticking to unused sites, assuming solitary eggs have a sufficient advantage over multiply occupied sites. Thus the evolution (or lack thereof) of egg discrimination capabilities would change the social dilemma.

More concrete is an *in vitro* experiment in Turner and Chao (2003) in which an RNA phage can develop a variant that reproduces prolifically in the host, causing fitness loss to all phages present. This creates a Prisoner's Dilemma scenario, and the defector strain replaces the original. However, it is only a Prisoner's Dilemma when a large number of phages infect a host. If multiple infection is not possible then cooperator phages predominate, and furthermore these cooperator phages, evolved in the absence of multiple infection, are capable of coexisting with the defector phages that evolved in the multiple infection case. Thus limits on coinfection can reduce the incidence of the Prisoner's Dilemma in this instance. This practical work also demonstrates how the payoff matrix of an interaction can vary even for relatively simple subjects such as viruses. Given these problems, Fort and Worden and Levin demonstrate different approaches to a resolution.

5.1.2 Fort (2008) - Evolving heterogeneous games

Fort is motivated to minimise the dependence of model predictions on correctly determining the model parameters, given that in practise this may be impossible. He proposes a minimal model for determining definite payoff matrices by creating a pool of potential payoff matrices seeded with a heterogenous initial distribution, and letting the final matrix be determined by natural selection. When the initial set of games is drawn from the social dilemmas, he finds the resultant matrix is a type of Stag Hunt coordination game.

The model is a cellular automaton on a square $L \times L$ lattice ($L = 100$ to 600) with periodic boundary conditions. Each cell hosts a strategy type — a cell will either always play C or always D. These strategy types are initially distributed randomly over the lattice but with an even split between C and D cell types. There are no payoff parameters as inputs, instead there is a spatial distribution of heterogenous games across the cells, with each game randomly determined by generating the four payoff values R, S, T and P. These are taken from the interval $[0, 1]$, with further constraints imposed to give four different starting conditions for the model: either all initial games are a type of Prisoner's Dilemma, or a social dilemma (so including the Stag Hunt and Snowdrift Games), or a game is a more general dilemma, or is entirely random.

In each update, every cell chooses just one of its neighbours to play against, with the opponent cell drawn at random from the full Moore neighbourhood of all eight surrounding cells. The interaction between two cells with different game types is played as a bimatrix game, where each player (cell) gets the payoff from the payoff matrix it specifies given it and its opponent's choice of strategy. Using a bimatrix game is an important model detail. It means that individuals have more control over the payoff they receive as it is determined by their own payoff matrix in combination with their strategy choice and that of their opponent. Each cell collects a score and then the cellular automaton updates using a "best takes over" update rule, cells adopting the payoff matrix and social strategy of their most successful neighbour. There is no mutation.

It is then possible to calculate average payoffs $\langle R \rangle$, $\langle S \rangle$, $\langle T \rangle$ and $\langle P \rangle$ over all cells and many runs of the model, given the initial starting conditions.

- Prisoner's Dilemma start: the averages converge to a game in the PD but near the Stag Hunt region: $\langle T \rangle = 1 >= \langle R \rangle > \langle P \rangle = 0.9 > \langle S \rangle = 0.45$.
- Social dilemma start: a Stag Hunt with $\langle R \rangle = 0.96 > \langle T \rangle = 0.85 > \langle P \rangle = 0.59 > \langle S \rangle = 0.36$
- Dilemma start: Stag Hunt where $\langle T \rangle$ is approximately equal to $\langle P \rangle$
- Random start: Stag Hunt with $\langle R \rangle$ approximately equal to $\langle P \rangle$, $\langle S \rangle$ and $\langle T \rangle$ very low.

It is unclear whether this approach is really as parsimonious as Fort claims. Fort suggests it offers robustness as well as realism and simplicity, but for it to actually be used in any situation requires justifying assumptions about cellular automaton setup, lattice structure and initial game distribution, amongst others. Nevertheless the motivations are well supported, and there is a clear conclusion: the resultant payoff matrices are all Stag Hunt coordination games, or as close to it as possible, since as there is no mutation when the model is seeded with Prisoner's Dilemma games the resultant game must always be a Prisoner's Dilemma.

Fort's other conclusion, that it is remarkable that the initial heterogeneity of payoff matrices is almost eliminated, is less surprising as there is no mutation in the model and hence no possibility of generating new strategies to replace those eliminated from the lattice. Given this and the stochasticity introduced by playing against a neighbour at random, it is always likely that the model will lead to a reducing set of active strategies. It also does not investigate the importance of the feedback between the social strategies and the payoff matrices as payoff matrix and social strategy are set as a fixed pair. It is this feedback that is fundamental to processes such as social niche construction (Powers et al., 2011).

5.1.3 Worden and Levin (2007) - Evolutionary escape from the Prisoner's Dilemma

For Worden and Levin, on the other hand, the problem is the over-reliance on thinking of all social interactions as forms of the Prisoner's Dilemma. Their model thus investigates whether the addition of mutant strategies can lead the social game out of an initial Prisoner's Dilemma scenario into other regions of game space, though it is not phrased in this way. This is similar to the approach taken in the metagames model, but metagames broaden the analysis to encompass the entirety of game space. As with metagames, Worden and Levin's paper approaches the problem not by demonstrating the conversion of one particular biological scenario into another in a particular species, but in a general, abstract setting.

The Worden-Levin model has players repeatedly playing against each other, adjusting their behaviour by learning what choices lead to higher payoffs as mutant variants of existing strategies are successively added to the model. Only the behaviours change; the players stay the same - they do not evolve, die or reproduce. However, since the addition of mutant strategies creates an 'ecology' of strategies, this learning model can be interpreted in an evolutionary context.

The model starts with all players engaged in a pre-set Prisoner's Dilemma with two possible strategies: cooperate (C) and defect (D). Each player has a probability vector

determining how likely they are to make each strategy choice, starting with a 90% probability of playing C and a 10% probably of playing D.

At each time step players are randomly paired and play the game. Their strategy vectors are then updated so that a strategy is more likely to be used in the future if it led to a positive payoff and less likely to be used if it resulted a negative payoff.

After a million time steps a new mutant strategy is added to the matrix. One of the existing strategies is chosen at random to be the 'parent' strategy, with the chance of each strategy being chosen proportional to the current frequency at which it is used in the population (determined by the strategy vectors). The mutant strategy is created by taking each possible payoff specified by the parent strategy and adding some small Δp drawn from a continuous distribution centred on zero. For instance, if the D strategy were to be mutated, the resulting payoff matrix would be:

$$\begin{pmatrix} R & S & S + \Delta p_{13} \\ T & P & P + \Delta p_{23} \\ T + \Delta p_{31} & P + \Delta p_{32} & P + \Delta p_{33} \end{pmatrix} \tag{5.1}$$

There is no pruning of strategies so all introduced strategies are possible plays through to the end of the model. This means the space of strategy types can become very large, though Worden and Levin found that in general only a few strategies are actually in use at any time. The learning dynamics typically converge to a steady state, but occasionally persistently oscillate.

Worden and Levin observe two common patterns in the results:

- Most of the time there is an early diversification of strategies followed by increasing off-diagonal payoffs, leading to a Snowdrift Game. The mean payoff increases steadily as off-diagonal payoffs increase.
- Less often there is rapid evolutionary increase in the payoff for a single strategy appearing in isolation.

The general results are that the players shift the dynamics of the game into one in which they either achieve positive average payoffs or uniform positive payoffs. Worden and Levin argue this shows that social harmony can be achieved by introducing new options that 'remove the disjuncture between selfishness and generosity'. Essentially this reduces the meaning of 'temptation to defect' as players will often all receive positive benefits even when they are all behaving selfishly, aligning individual and group interests.

However Worden and Levin's assertion that mechanisms like indirect reciprocity and kin selection are therefore unnecessary for the successful cooperative outcomes since

mutually optimal outcome can be reached without them is far from obvious. These mechanisms may be the proximate causes behind the small mutations in strategies that enable the social dynamics to change.

5.2 Introducing Metagames

These are particularly relevant models, but other work has also motivated the introduction of metagames. The term itself is, perhaps, an overloaded one. 'Metagame' is in wide popular usage with a variety of definitions such as 'the highest level of strategy in a complex game that involves not just playing the opponents in the game itself, but also taking into account the opponents expectations of your game play' (Carter et al., 2012).

As such, game theoretic models of 'metagames' have appeared before in the literature, in particular Howard's concept of a metagame introduced in *Paradoxes of Rationality* (Howard, 1971). This metagame theory aimed to subsume 'classical' game theory in a new theory that combined mathematical rigour with practical utility.

A Howard metagame is derived from a particular game by allowing one of the players to choose their strategy with knowledge of the strategy choices of the other players. Formally, if G is a (normal form) game and k a player in the game, the metagame kG is the (normal-form) game that would exist if k chose their strategy knowing the strategies of the other players. This can be recursively extended to form the set of all metagames $k_1 \dots k_r G$ for game players k_i (Howard, 1974).

The point of such successively expanding metagames is not that players can actually predict each other's behaviour, but that they attempt to. In following through on their predictions they play (first-order) metastrategies conditional on the predicted or observed strategy choices of their opponents, or go further and play second-order metastrategies conditional on the choice of first-order metastrategy of their opponent.

As a result, Howard metagames may have equilibria that result in outcomes that are not equilibria in the original game – for instance, a Prisoner's Dilemma can have equilibria in a Howard metagame at all-cooperate and all-defect (Howard, 1976).

Howard's metagames are classified as political science models (Olinick, 1981) and still see applications in business and sociology (Johnson and Leydesdorff, 2015), but they can also have applications to the evolution of cooperation. Howard metagames can purportedly be used to circumvent the need for binding agreements in cooperative games. Human players in cooperative games bargain and make these binding agreements before playing in order to secure mutually beneficial outcomes, but it is claimed a Howard metagame could suffice when games are between non-human participants (Rapoport, 1985). Whether this assumption switch is justifiable is another matter, since the number

of $(n+1)$th-order metastrategies is the square of the number of nth-order metastrategies, so the strategy space becomes very large for nonhuman entities to navigate.

Unlike the Howard metagame, which expands the game through additional metastrategies that enable alternate solutions, we want our metagames to model a dynamic process through which social behaviours and game-changing traits coevolve. The process of evolution is one of gradual change, and like Worden and Levin (2007) we wish to have the possibility of this kind of dynamic account present in the model.

Our approach to metagames is distinct, though not without antecedents and parallels. It is a novel model rooted in an emerging approach to understanding the evolution of cooperation that recognises the importance of the social environment to social actors. There is existing work in the literature that might be, if not identifiable as a metagame, at least analogous in principle or congruent in aims.

In particular, the theory of social niche construction is a powerful mechanism for promoting the evolution of population structures more supportive to cooperation, or in the metagames way of thinking, the evolution of social games that result in higher levels of cooperation (Powers et al., 2011). As metagames model the coevolution between social traits and game-changing traits that modify social niches, we claim that metagames provide a formal model of social niche construction.

Roughgarden (2009) and Akçay and Roughgarden (2011) have argued for a two-timescale approach in which the payoff matrices emerge from behavioural interactions in the faster timescale. And there is a growing body of literature on games played on networks that straddles the evolution of payoff matrices and mechanisms for changing the game, where changing the network structure results in effective changes to the game (Santos et al., 2006a; Pacheco et al., 2006b,a).

There is also the important admonition in Rosas (2010) that, given the role of assortment in most mechanisms for the evolution of altruism, a complete account of social evolution must consider how altruism and assortment-promoting traits evolve. Major evolutionary transitions show the importance of these traits through the vital role played by the evolution of conflict modifiers, adaptations such as bottlenecks and reproductive specialisation that reduce the inherent tension between levels of biological organisation (Michod and Herron, 2006). We can also see in the historical puzzle over the origins of altruism that social behaviours cannot be divorced from the context in which they occur: apparently paradoxical self-sacrificing behaviours, though no less self-sacrificing, can have fitness advantages in the right context, be it a group of highly related individuals or one in which behaviour is reciprocal.

Recent models have began to look at the same questions, such as the coevolution of strategies and payoffs in iterated games where evolving payoffs can lead to a dramatic collapse in cooperation (Stewart and Plotkin, 2014). Zollman (2008) creates a metagame

of a sort by combining the Nash bargaining and ultimatum games into a larger game of incomplete information to study the evolution of a single fairness norm. In a very recent paper, Levin (2014) explicitly states that "Understanding the factors that guide the adoption of such rules [as fairness] must embed individual realisations within a broader framework in which classes of similar games are considered, leading to the necessity of developing a theory of metagames."

These are the motivations for our concept of a *metagame* as a model for the coevolution of social strategy and social environment, for situations where a population is engaged in a social game and furthermore the social game as defined by its payoffs is also subject to evolutionary control. We do not claim that metagames fulfil the need for a general theory; but they can serve as a minimal model for the coevolution of social strategies and social structures that enables us to begin to explore this important coevolutionary feedback process in a principled way. This approach recalls Queller's comment in his review of *The Major Transitions in Evolution*:

> If you imagine the history of life as a giant cosmic card game, previous thought on major evolutionary transitions has focused on particular high scores — how well this or that strategy ... performs. This work focuses instead on the rules of the game and how they come to be modified. — Queller (1997)

5.3 The Metagame Model

5.3.1 Mathematical Definition

The metagame model is a haploid two-locus model with asexual reproduction. The first gene is for a social trait (ST) that determines an individual's social behaviour. There are two alleles which we will consistently call 'C' and 'D' for cooperation and defection — though as we discussed in Chapter 3, when the game is a division of labour game ($S + T > 2R$) it is not strictly appropriate to think of the two strategies as cooperation and defection.

The second gene is for a game-changing trait (GCT) as described in Chapter 4 — a trait that changes the incentive structure of a social interaction without necessarily changing the physical nature of that interaction, such as by altering the population structure or introducing policing. Game-changing traits are broadly defined, to admit many possible traits within the same scheme. Because of this the definition of a metagame must also be flexible, able to represent the many possible types of GCT.

At any moment, a population will be playing games in a social environment that is shaped by, among other factors, the actions of multiple game-changing traits (Ryan

et al., 2016). Since the effect of game-changing traits can often be represented as transformations to a payoff matrix, we make a distinction between the actual game being played given the effects of these game-changing traits and the *effective game* that would be being played were the population still well-mixed. We take this as justification for having each GCT allele X correspond to a particular evolutionary game with the payoff matrix G_X. So whatever the underlying nature of the social dilemma might be, the game G_X is the effective game played by those members of the population that possess the GCT allele X.

Instead of determining an actual social game, and a series of transformations to the game, such as those created by increasing assortment, and then determining which of these heritable transformations to the game will spread, the metagame model reverses this process. This approach is the basis for the model's simplicity. In its purest form, the metagame model is an evolutionary competition between different games. The utility of metagames then comes from carefully choosing the space of games the game-changing traits range over and the way the between-games competition is interpreted to represent a particular scenario. By operating at an abstract level the metagame model can be used for any situation in which the evolution of a behavioural trait is accompanied by the evolution of secondary traits that affect the social dynamics.

There are two components to the definition of a particular metagame: the set of games Γ that the metagame can range over, which defines the GCT allele space, and rules that specify what happens when individuals with different GCTs engage in a metagame interaction. The GCT allele-space can be finite or continuous. The metagames presented here are for two-player two-strategy symmetric games, though the concept can be generalised to more complex games with a corresponding increase in the complexity of the metagame. We write the payoff matrix for the game G_X corresponding to the GCT allele X as:

$$G_X = \begin{pmatrix} R_X & S_X \\ T_X & P_X \end{pmatrix} \tag{5.2}$$

5.3.2 Metagame Interactions

A *metagame interaction* is a particular interaction within the metagame framework between the bearers of different GCT alleles — for instance N and M (informally, for 'normal' and 'mutant' GCT traits). In strategic terms, when we consider a particular interaction within a metagame, we imagine a population of individuals some of whom want to play a game G_N, others a different game G_M. When both players want to play G_N, they will play G_N, and likewise when both want to play G_M. But when one player wants to play G_N and the other G_M, they will have to play some combination of the games G_N and G_M. When cooperator and defector genotypes fare differently in

the local social games G_N and G_M, linkage disequilibrium develops between GCT and social strategy alleles determining which GCT allele is more successful.

The interaction between individuals with different GCTs is determined by an *encounter function* $\phi : \Gamma \times \Gamma \to \Gamma$. The encounter function determines the effective game when bearers of GCT alleles interact. When individuals with different GCTs encounter each other they play the game defined by the given encounter function — that is $G_K = \phi(G_N, G_M)$ or $G_L = \phi(G_M, G_N)$ depending on which player is considered 'player one'. We say a metagame is *symmetric* when $\phi(G_X, G_Y) = \phi(G_Y, G_X) \quad \forall G_X, G_Y \in \Gamma$.

In a metagame interaction there are four genotypes: NC, ND, MC and MD. Therefore a two-player, two-strategy, two-game metagame interaction is mathematically equivalent to a two-player four-strategy game. But the converse is not true: if ϕ is restricted to linear encounter functions then not all two-player four-strategy games are metagame interactions. The payoff matrix for the four-strategy game form is:

$$\begin{pmatrix} G_N & \phi(G_N, G_M) \\ \phi(G_M, G_N) & G_M \end{pmatrix} = \begin{array}{c} \\ NC \\ ND \\ MC \\ MD \end{array} \begin{array}{c} \begin{array}{cccc} NC & ND & MC & MD \end{array} \\ \begin{pmatrix} R_N & S_N & R_K & S_K \\ T_N & P_N & T_K & P_K \\ R_L & S_L & R_M & S_M \\ T_L & P_L & T_M & P_M \end{pmatrix} \end{array} \tag{5.3}$$

Switching between these two perspectives allows us to model the competition between genotypes using the replicator equation in exactly the same way as a two-player two-strategy game (though the increase in complexity of the equations means analytic solutions are difficult to obtain apart from in special cases, so most results must be obtained by numerical methods), and investigate the change in the frequencies of cooperators or GCT-allele mutants by taking the appropriate combination of genotype frequencies. This takes advantage of the fact that a two-locus two-allele model can be viewed both with this extra structure and also just as a competition between four different genotypes. By performing a series of metagame interactions we can determine the overall dynamics of the metagame and the path that a population will take through game space.

There are a number of possible rule choices for this encounter function ϕ, with different interpretations in the metagames model. Though the specific encounter function chosen does have an impact on the behaviour of the metagame, the most important feature of the encounter function is not the actual value but the rank ordering of the payoffs it produces — that is, if K is one of the payoffs $\{R, S, T, P\}$ and $K_X < K_Y$, then is $K_\phi \leq K_X$, $K_\phi \geq K_Y$ or $K_X < K_\phi < K_Y$? This is because one strategy is preferable to another when its payoffs exceed the other, so the rank order of payoffs is more important than their absolute quantities. The absolute quantities do still determine the exact proportion of strategies present in mixed-strategy equilibria.

Here we focus on linear encounter functions $\phi_{\lambda,\mu}(G_X, G_Y) = \lambda G_X + \mu G_Y$. At one extreme, $\phi_{0,0}$ would represent a situation in which no payoff is received for interactions between subgroups in the population with different GCT alleles. In a non-obligate game where there are no fitness consequences for not interacting, it may be appropriate to interpret this as a scenario in which members of the population with different GCT alleles effectively never play each other. Because the most salient feature of an encounter function is the rank ordering of payoffs it produces, metagames for linear encounter functions where $0 \leq \lambda, \mu \leq 1$ and $\lambda + \mu = 1$ are substantially similar. Therefore, unless otherwise stated, the results illustrated here use the encounter function $\phi_{0.5,0.5}$, so the effective game for cross-GCT allele interactions is the mean of the two GCT games. All metagames that use this encounter function are symmetric.

If the set $\Lambda \subset \Gamma$ then the metagame (Λ, ϕ) is a *submetagame* of the metagame (Γ, ϕ). By taking particular submetagames we can investigate the dynamics of the metagame under specific constraints on the way the social game can change, for example when all the possible population structuring traits generate different levels of assortment.

5.4 The Dynamics of Metagames

We are interested in what happens to a social population when individuals evolve traits that change the game. One can imagine a range of different hypotheses for what might happen. The immediate benefits of selfish defection could inevitably lead to all games becoming types of Prisoner's Dilemma, justifying its widespread adoption as the default model for social interactions. On the other hand, given the benefits of mutual cooperation, one could imagine that games would have a tendency to become Harmony Games more supportive of cooperation.

Here we show how Snowdrift and Stag-Hunt games can be transformed into either the Prisoner's Dilemma or the Harmony Game depending on the type of game-changing trait involved, the constraints on how it can evolve, and the initial frequencies of the social strategies. Taking a series of different metagames we see that in fact there is no one particular way a game will change, instead all these outcomes are possible.

First we consider the case of a metagame over the circular θ-space of games in which there is a constant trade-off between the amount of fear and greed in the game, with all possible types of social dilemma within the GCT allele space. Then we consider metagames where the total utility of the payoff matrix is kept constant. These metagames are motivated by principled but primarily mathematical considerations to illustrate the behaviour of metagames in special cases. Then we broaden the analysis to include the behaviour of an unconstrained metagame over the entirety of the ST-plane. Finally we will focus our attention on an important type of metagame where the biological motivation is more

straightforward — where the game-changing trait affects the level of assortment in the population.

We illustrate our analyses by constructing 'vector fields' showing the selective pressure in the metagame. An important comment needs to be made about their intended interpretation. The concept is that the population starts off at a particular point on the ST-plane — that is, playing a particular game defined by S and T. We then investigate what happens when small mutations occur that change the social context in which the game is played, leading to some change in the effective payoff matrix. In an alternative formulation of the model one could imagine the population being uniformly distributed across all points on the ST-plane, all interacting simultaneously (something closer to Fort (2008)). However, we are interested here in the behaviour of the metagame from the viewpoint of a population moving through game space by a series of small incremental changes in the game, rather than as a competition between all the possible games. We see this viewpoint as more biologically useful — that it makes more sense to consider the way that the genetical social strategies and social structures of a population coevolve from a range of starting points than to imagine a large number of coexisting population structures coalescing into a few.

5.4.1 Dynamics Under Constraint of Constant Selection Strength

We start by illustrating the use of metagames in a minimal setting — when the GCT allele space is a circle on the ST-plane centred at the point $S = R = 1, T = P = 0$. Recall from Section 3.5 that such a circle is the most compressed continuous representation of the space of two-player symmetric games that retains all the different equilibria found on the ST-plane. This gives a natural metagame over all the two-player social dilemmas parametrised by a single variable. These games are interesting because a constant radius is equivalent to a constant strength of selection, so there is a balance between the amount of fear $(S - P)$ and greed $(T - R)$ in all the games on the circle.

We define a metagame on the circle by letting the GCT-allele space be $\Gamma = \{\left(\begin{smallmatrix} 1 & \sin\theta \\ 1+\cos\theta & 0 \end{smallmatrix}\right) : \theta \in [0, 2\pi)\}$ and choosing the encounter function to be the radial average of the two games, $\phi(G_{\theta_1}, G_{\theta_2}) = G_{(\theta_1 + \theta_2)/2}$.

We determine the behaviour of the metagame over the whole of the game space by taking a set of points spaced equidistantly around the circle and calculating the results of metagame interactions between every adjacent pair of games. Note that increasing θ is equivalent to moving anti-clockwise around the circle. Given two adjacent points $\theta \in [0, 2\pi]$ and $\theta + \epsilon$ (for some small ϵ; here we use $\epsilon = 0.002$ to give 1000 points around the circle), we can calculate which of these two GCT alleles will increase in frequency as a result of the metagame interaction. Repeating this around the circle lets us calculate a 'vector field' of the flow of the GCT allele around the circle.

Figure 5.1 shows the direction of movement of the population through the metagame of constant selection strength. There are two parameters for each model run: c is the initial frequency of the cooperation allele (C) and m is the initial frequency of the mutant GCT allele (M). Figure 5.1 shows the results when both parameters were initially set to 0.5 with no linkage disequilibrium, so the initial frequency of each of the four genotypes is 0.25. We can interpret Figure 5.1 as showing the direction of the path a population will take through this GCT allele space given a particular initial position.

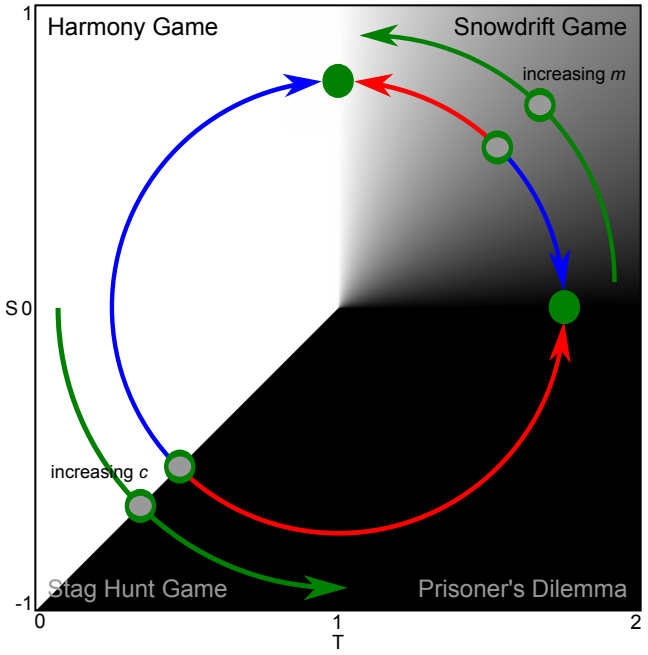


Figure 5.1: ST-space with the direction of selection in the metagame around the circle shown. The initial conditions of all metagame interactions are set to $c = 0.5$, $m = 0.5$. The two filled circles mark the fixed attractors of the metagame, the unshaded circles the boundaries between between the basins of attraction that change position according to the initial conditions c and m. The figure is superimposed over the visualisation of the equilibrium frequency of cooperators in the social game given the same initial conditions.

The movement around the circle shows that there are two attractors in this metagame — the boundaries of the Harmony Game and Snowdrift Game, and the Snowdrift Game and the Prisoner's Dilemma. The circle is superimposed on a plot of the equilibrium frequency of cooperators in the social game from the same initial conditions. It shows that when the metagame interactions take place within a region of ST-space where the equilibrium frequency of cooperators is 0 — that is, in the Prisoner's Dilemma or the no-cooperators attractor of the Stag Hunt — the selective pressure on the GCT allele is in the direction of increasing θ, pushing the population from a Stag Hunt or Prisoner's Dilemma towards the Prisoner's Dilemma-Snowdrift Game boundary. Conversely, where the equilibrium frequency of cooperators in the underlying games is 1, the population is driven towards the boundary between the Harmony and Snowdrift Games. There each of the Stag-Hunt and Snowdrift quadrants there is a point that divides the basins

of attraction of the stable metagame equilibria which shifts position according to the initial frequency of cooperators (in the Stag Hunt quadrant) or the mutant GCT allele (in the Snowdrift quadrant) which we will account for as we break down the behaviour of the metagame on the circle in more detail.

5.4.2 Analysing the Behaviour of the Metagame Under Constraint of Constant Selection Strength

We can understand the behaviour of the metagame on the circle by breaking it down into local cases and applying game theoretic reasoning. Take two GCT alleles M and N from this space, with $\theta_M > \theta_N$. To simplify the analysis we make the assumption that the two games G_M and G_N lie in the same quadrant of the circle. This means both GCT alleles correspond to the same type of social dilemma, and we can analyse the behaviour for each dilemma-region (quadrant of the ST-plane) separately. If we let G_K be the radial average game then this assumption also means that the value of S_K and T_K will be between S_M and S_N and T_M and T_N respectively (so if $S_N < S_M$ then $S_N < S_K < S_M$).

A general metagame interaction on the circle in two-player four-strategy form looks like:

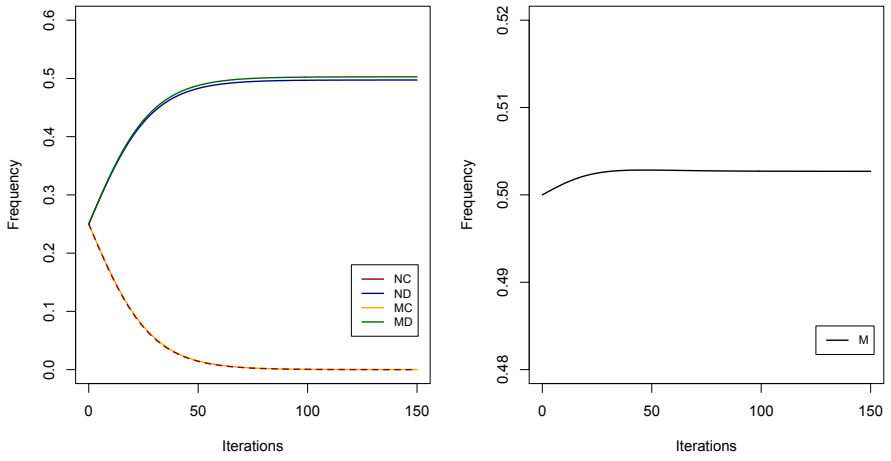
$$\begin{array}{c} \\ NC \\ ND \\ MC \\ MD \end{array} \begin{array}{c} \begin{array}{cccc} NC & ND & MC & MD \end{array} \\ \begin{pmatrix} 1 & S_N & 1 & S_K \\ T_N & 0 & T_K & 0 \\ 1 & S_K & 1 & S_M \\ T_K & 0 & T_M & 0 \end{pmatrix} \end{array} \tag{5.4}$$

5.4.2.1 The Prisoner's Dilemma Quadrant

- $S < 0$, $T > 1$
- Increasing θ means both S and T increase.
- $S_N < S_K < S_M < 0$
- $1 < T_N < T_K < T_M$
- So as θ increases, the temptation to defect increases but so does the sucker's payoff for cooperators that encounter defectors.

Immediately it is clear that MC and NC are dominated by both defector strategies, so the frequency of these types is expected to go to 0 in the replicator dynamics. $T_N < T_K < T_M$ means that ND is weakly dominated by MD. However, when both cooperator types are removed from the population, the reduced game is $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and so the metagame

interaction will have reached equilibrium. However, because ND weakly dominates MD, while the metagame interaction is reaching equilibrium ND will increase in frequency faster than MD. From this reasoning we can state that the equilibrium state of a metagame interaction under these conditions will have MC and NC present at frequency asymptotically approaching 0, and the proportion of MD to ND will have increased relative to the initial conditions. When all types are introduced at equal frequency, $x_{MD} > x_{ND}$, so the M allele spreads and the metagame moves in the direction of increasing θ. This is illustrated in Figure 5.2.



(a) Type frequency (b) M allele frequency

Figure 5.2: The change in frequency of the different types and of the M allele in a metagame interaction in the Prisoner's Dilemma region - here $\theta_N = \frac{7\pi}{4}$, $\theta_M = \frac{7\pi}{4} + \frac{\pi}{100}$

5.4.2.2 The Harmony Game Quadrant

- $S > 0, T < 1$
- Increasing θ means both S and T decrease.
- $0 < S_M < S_K < S_N$
- $T_M < T_K < T_N < 1$

The Harmony Game is the mirror image of the Prisoner's Dilemma. Following the same reasoning as the Prisoner's Dilemma, we can say that in a metagame interaction where both GCT alleles lie in the Harmony Game region, the equilibrium state will have the types MD and ND asymptotically close to 0. NC weakly dominates MC, though the two are neutral with the defector strategies removed, so the proportion of NC relative to MC will increase. This means that the N allele spreads and the metagame moves in the direction of decreasing θ.

5.4.2.3 The Stag Hunt Quadrant

- $S < 0$, $T < 1$
- Increasing θ means S decreases and T increases.
- $S_M < S_K < S_N < 0$
- $T_N < T_K < T_M < 1$

Comparing strategy payoffs, we see that NC weakly dominates MC, and MD weakly dominates NC. The significance of this depends on the basin of attraction of the local games. As G_M, G_N and G_K are all Stag Hunt games we know that there exist three constants c_M, c_N and c_K between 0 and 1 that are the unstable Nash equilibria of these Stag Hunt games, dividing the basins of attraction of the all-cooperators and all-defectors ESSs. These unstable equilibria are at $c_X = \frac{S_X}{S_X+T_X-1}$ for $X \in \{M, N, K\}$. In terms of θ this is at $\frac{\sin\theta}{\sin\theta+\cos\theta+1-1}$. In the Stag Hunt quadrant, $\pi < \theta < \frac{3\pi}{4}$, so $c_M > c_K > c_N$.

If the initial frequency of cooperators c is sufficiently high that $c > c_M$ then all the local games G_M, G_N and G_K will lie in the all-cooperators basin of attraction of ST-space and the reasoning of the Harmony Game will apply and the frequency of the mutant trait will decrease (which we interpret as the metagame to moving in the direction of decreasing θ). If $c < c_N$ then the local games will all lie in the all-defectors basin of attraction of ST-space and the reasoning of the Prisoner's Dilemma will apply, leading the frequency of the mutant trait to increase and the metagame to move in the direction of increasing θ.

This behaviour means there is a break point dividing the basins of attraction within the metagame which is determined by the initial frequency of cooperators c. This point moves from the border of the Harmony Game ($\theta_c = 2\pi$) when c is (asymptotically close to) 0, and as c increases, it moves in the direction of increasing θ round to the border of the Prisoner's Dilemma ($\theta_c = \frac{3\pi}{2}$) at $c = 1$.

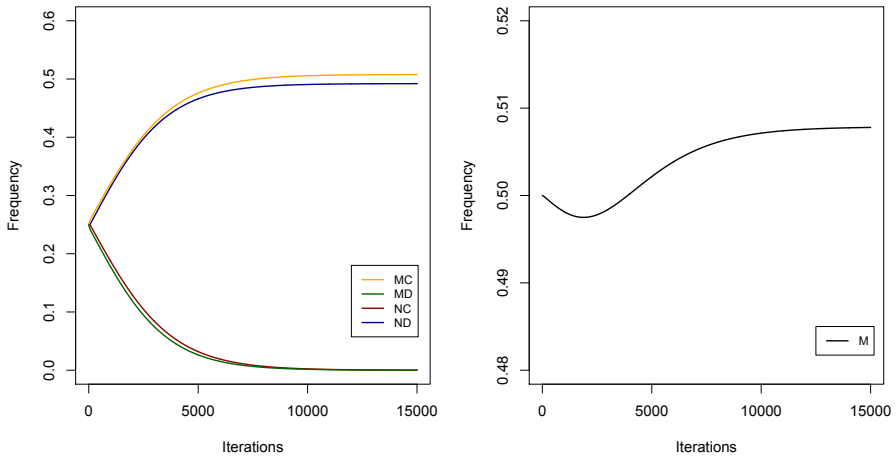
5.4.2.4 The Snowdrift Quadrant

- $S > 0$, $T > 1$
- Increasing θ means S increases and T decreases.
- $0 < S_N < S_K < S_M$
- $1 < T_M < T_K < T_N$

Comparing strategy payoffs, we see that MC weakly dominates NC, and ND weakly dominates MD. Unlike the other regions of the ST-plane both defectors and cooperators are present at strategy equilibrium, so the metagame interaction does not reach an equilibrium when one strategy allele becomes extinct. Because of this, the weakly dominated MD and NC types both become asymptotically extinct. Upon their extinction the reduced game when just ND and MC are present becomes:

$$\begin{pmatrix} 1 & S_K \\ T_K & 0 \end{pmatrix} \tag{5.5}$$

This is exactly the game G_K, which has a polymorphic Nash equilibrium at $\frac{S_K}{S_K+T_K-1}$. The success of the M GCT allele is tied to the success of cooperators, the N GCT allele is tied to the success of defectors. Therefore the equilibrium frequency of both the M and C alleles will be $S_K/S_K + T_K - 1$. Therefore, the metagame moves in the direction of increasing θ if this equilibrium is greater than m, the initial frequency of the M allele, and in the direction of decreasing θ if it is less. For example, when the initial frequency of the M GCT is $m = 0.5$, the dividing line is at $\theta = \pi/4$. This is illustrated in Figure 5.3.



(a) Type frequency (b) M allele frequency

Figure 5.3: The change in frequency of the different types and of the M allele in a metagame interaction in the Snowdrift region - here $\theta_N = \pi/4$, $\theta_M = \pi/4+\pi/100$. Unlike in the other games the frequency change of the M allele is not monotonic, it initially decreases before increasing. Note that in the Snowdrift Game it takes many orders-of-magnitude more iterations for the metagame interaction to reach equilibrium than in the Prisoner's Dilemma

We see then that in the Snowdrift Game, it is the initial frequency of the GCT allele that matters to the increase or decrease of the mutant type. Paradoxically, this is because the Snowdrift quadrant is the one quadrant of the game in which the initial frequency

of the mutant type has no bearing at all on the results of a metagame interaction. The equilibrium of a given metagame interaction in the Snowdrift quadrant is *absolute* with no sensitivity to the initial conditions. This means the sign of the change in the frequency of the mutant trait is relative. By contrast, in the other quadrants the outcome of the games is relative so the sign of the change in the frequency of the mutant type is absolute.

In the Prisoner's Dilemma quadrant, no matter what the initial frequency m of the mutant type M is, it will always decline in frequency until the metagame interaction reaches strategy equilibrium and only the two defector strategies remain. The absolute outcome – the equilibrium frequencies of the four types – will vary based on the initial conditions as these will affect the length of time it takes for the metagame interaction to reach strategy equilibrium, but for all initial conditions that start with all four types present the sign of the change in the frequency of the mutant type will be negative.

In the Snowdrift quadrant, whatever the initial conditions (as long as all four types are present) the equilibrium frequencies will be the same, so the sign of the change is relative. If the equilibrium frequency of the M allele is 0.6, then this will represent an increase if the M allele is introduced with a frequency of 0.3, but a decrease if it is introduced at a frequency of 0.9. Therefore the unstable equilibrium in the Snowdrift quadrant moves from the border of the Prisoner's Dilemma ($\theta = 0$) when m is (asymptotically close to) 0. As m increases, the metagame equilibria moves in the direction of increasing θ round to the border of the Harmony Game ($\theta = \frac{\pi}{2}$) at $m = 1$.

We know that the equilibrium frequency of cooperators in a Snowdrift Game is $\frac{S_K}{S_K+T_K-1}$, and have $\theta_K = \theta_N + \frac{\epsilon}{2}$. Combining the two we have that the equilibrium frequency of the mutant trait in a metagame interaction in the Snowdrift quadrant of the circle between games at θ_N and θ_M is:

$$m_{\theta_N,\theta_M} = \frac{\sin\theta_K}{\sin\theta_K + \cos\theta_K} = \frac{\tan\theta_K}{1 + \tan\theta_K} \tag{5.6}$$

Therefore, the frequency of the mutant trait increases if $m < m_{\theta_N,\theta_M}$, as the equilibrium frequency of cooperators is greater than the initial frequency, and decreases if $m > m_{\theta_N,\theta_M}$.

To find the Snowdrift quadrant metagame equilibria divider for a given initial frequency of the mutant trait m we find the game where $m = m_{\theta_N,\theta_M}$:

$$\begin{aligned}
&\frac{\tan\theta_K}{1+\tan\theta_K} = m \\
&\qquad \implies \tan\theta_K = m(1+\tan\theta_K) \\
&\qquad \implies \tan\theta_K = \frac{m}{1-m} \\
&\qquad \implies \theta_K = \theta_N + \frac{\epsilon}{2} = \tan^{-1}\left(\frac{m}{1-m}\right) \\
&\qquad \implies \theta_N = \tan^{-1}\left(\frac{m}{1-m}\right) - \frac{\epsilon}{2}
\end{aligned} \tag{5.7}$$

What happens when m is near the bounds? There are two stable equilibria at $\theta = 0$ and $\theta = \pi$. All metagame interactions on the circle are between two points on the circle, though they may be arbitrarily close. Thus, if m is sufficiently large or sufficiently small, and the gap between games is sufficiently large, one or other of the stable equilibria may in fact disappear.

If G_N is the point at $\theta = 0$, then the mutant GCT trait increases in frequency if:

$$\epsilon > 2\theta_m = 2\tan^{-1}\left(\frac{m}{1-m}\right) \tag{5.8}$$

If, say, $m = 0.01$, then $\theta_m = 0.0101$ (4dp) and so the mutant GCT trait will always increase in frequency for every metagame in the Snowdrift quadrant if $\epsilon > 0.0202$, effectively removing the stable metagame equilibria at $\theta = 0$ on the boundary of the Snowdrift and Prisoner's Dilemma quadrants.

Through this game theoretic reasoning we can explain the results of the numerical solutions of the differential equations for the overall behaviour of the metagame around most of the circle. The exceptions are at those boundary points where the two GCT games cross dilemmas. However, we have implicit information about the behaviour of the metagame at these points by looking at the points surrounding them; for instance that the points where the Snowdrift Game meets the Prisoner's Dilemma and Harmony Games must be the stable equilibria of the metagame; any perturbation of the game from these points will be pushed back towards the boundary point.

This example case illustrates the principles at work in the metagame. Breaking down the behaviour of the metagame reveals the importance of relative differences in the S and T values of the different genotypes. We can better understand the overall results therefore by looking at a metagame over the whole of the ST-plane. Our expectation about what should happen in this case is primed by the face that when the local social strategy ESS is all-cooperators the metagame moves to the point on the circle with maximal S, likewise when the local social strategy attractor is all-defectors it moves towards the point with maximal T.

5.4.3 Dynamics Under Constraint of Constant Total Utility

We might wonder from the metagame on the circle if the total utility of the payoff matrix of the possible games biases the results. On the circle, the sum $(T-R)^2+(S-P)^2$ is constant, but the total utility of the payoff matrix varies; it is highest in the Snowdrift region and lowest in Stag-Hunt region. We investigate this using metagames in which the total utility of the games in the GCT allele space is constant — a kind of 'zero-sum' metagame. On the ST-plane this is equivalent to restricting the metagame to lines where $S+T=L$ for some constant L. This gives a family of metagames on the sets $\Gamma_L = \{\left(\begin{smallmatrix} 1 & S_L \\ T_L & 0 \end{smallmatrix}\right) : S_L + T_L = L\}$, choosing the encounter function to be the average of the two games.

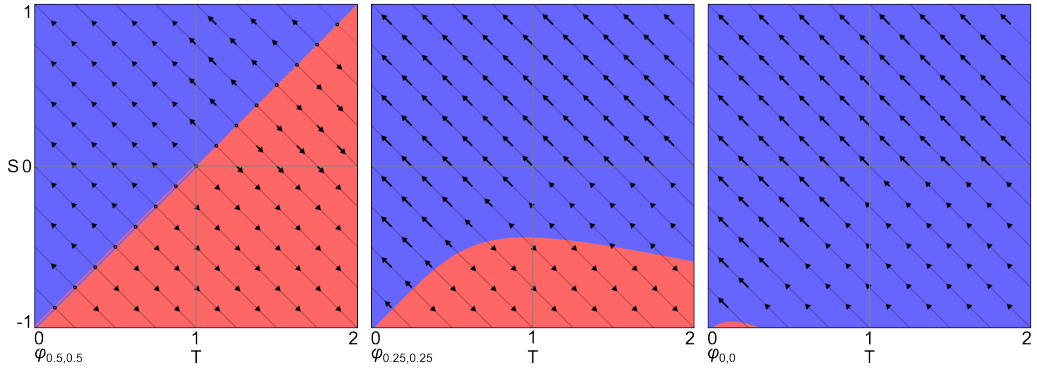


Figure 5.4: Metagames on lines where $S+T$ is a constant, with the initial conditions of all metagame interactions set to $c=0.5$, $m=0.5$, but three different interaction functions: $\phi_{0.5,0.5}$ (left), $\phi_{0.25,0.25}$ (centre) and $\phi_{0,0}$ (right). The lines along which the metagame can vary are marked. The selective pressure is in the direction of increasing S in the blue shaded area; in the red it is in the direction of increasing T. The length of the arrows indicates the relative magnitude of the change.

Two results are pictured using three different interaction functions, though in all cases the initial conditions are $c=0.5$, $m=0.5$. Unlike the metagame on the circle, we see here that there are no attractors in the different metagames. Instead, the metagames diverge. In the left figure, which uses the interaction function $\phi_{0.5,0.5}$ as in the rest of this paper (so the cross-GCT game is the mean of the two GCT-determined games), the metagames diverge depending on which side of the line $S=T-1$ the local game is at. This line has an interpretation on the ST-plane as the line where $S-P=T-R$ — the line where the difference between unilateral and mutual interactions is constant. In the right figure, where the interaction function is $\phi_{0.25,0.25}$ so cross-GCT interactions result in reduced payoff, the basin of attraction for the metagames changing in favour of increased cooperation is enlarged, with the result that there is a path from the Prisoner's Dilemma quadrant to the Harmony Game quadrant.

Changing the initial frequency m of the mutant GCT type has no effect on the direction of the flow along the lines, apart from the extreme cases when $m=0$ or $m=1$, though

it does affect the magnitude of the change. This is for the same reasons described in the detailed analysis of games on the circle in 5.4.2. Changing c changes the results for the Stag Hunt quadrant; as c increases, so does the portion of the Stag Hunt region in which the metagame favours movement that increases S and decreases T.

5.4.4 Unconstrained Metagames on the ST-Plane

Having considered metagames with different restrictions, we now extend the analysis to unconstrained metagames where the GCT allele space is all of ST-space, unlike the constrained constant total utility and constant selection strength metagames. As discussed earlier, one of the advantages of the ST-plane is the possibility of visualising the plane. This means we can plot the selective pressures on the metagame over the entirety of a space representative of the complete set of possible symmetric two-player games. A metagame interaction for games on the ST-plane has the form:

$$\begin{array}{c} \\ NC \\ ND \\ MC \\ MD \end{array} \begin{array}{c} \begin{array}{cccc} NC & ND & MC & MD \end{array} \\ \begin{pmatrix} 1 & S_N & 1 & S_K \\ T_N & 0 & T_K & 0 \\ 1 & S_K & 1 & S_M \\ T_K & 0 & T_M & 0 \end{pmatrix} \end{array} \tag{5.9}$$

The dynamics of the metagame over the whole of the ST-plane were systematically examined by constructing a pseudo-vector field showing the direction of movement through GCT allele space. This was done on a square lattice of points spaced 0.1 units apart in S and T. For each point representing a game G_N, 36 metagame interactions were computed between G_N and a mutant GCT trait playing the game G_{M_i}, where the G_{M_i} were spaced evenly around a small circle centred on G_N with radius r (here $r = 0.05$). By choosing a small radius here we make an assumption that in general mutations to a game-changing trait will have a small effect on the resulting game being played, such as creating a small increase in assortment and hence a small transformation to the effective game matrix. Where this assumption does not hold the radius could be increased.

For each of these metagame interactions, the change in the frequency of the mutant allele $\Delta m_i = x_{M_iC} + x_{M_iD} - x_{NC} - x_{ND}$ was recorded after the metagame interaction had reached equilibrium. The vector representing the general selection pressure from that point on the lattice was then computed as the vector sum $\sum_{i=0}^{36} \Delta m_i \begin{pmatrix} T_{M_i} - T_N \\ S_{N_i} - S_N \end{pmatrix} /36$. Again, in each metagame interaction the initial population state was determined by two parameters: c, the initial frequency of the cooperator allele, and m, the initial frequency of the mutant GCT allele. At the beginning of each metagame interaction we set the population set to be in linkage equilibrium, so for example the genotype MC had initial frequency mc.

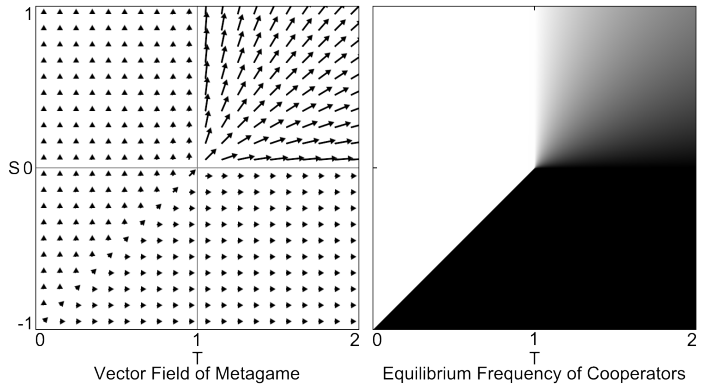


Figure 5.5: Vector field diagram mapping the metagame in ST-space for $c = 0.5$, $m = 0.5$ (left), compared with a plot of the equilibrium frequency of cooperators starting from $c = 0.5$ (right).

Figure 5.5 displays the behaviour of the metagame when $c = 0.5$, $m = 0.5$, compared with the social equilibrium under the replicator dynamics of the game with the same initial level of cooperation. There are two immediate conclusions. The first is that the behaviour of the metagame is closely linked to the dominant social strategy in that region. Where all-cooperators is the evolutionarily stable state in the region, the metagame moves in the direction of increasing S. Where all-defectors is the evolutionary stable strategy the metagame moves in the direction of increasing T. Second, the behaviour in the Snowdrift region is different from the rest of the plane — outside of the Snowdrift region, the magnitudes of all the vectors are small and they are parallel to one of the axes; inside the Snowdrift region, because of the polymorphic equilibria, the magnitudes are much larger and not parallel to either axis.

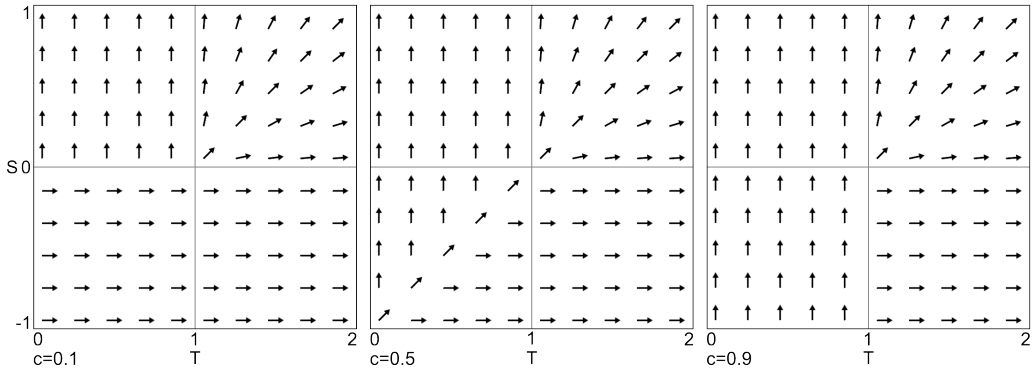


Figure 5.6: Vector field diagrams mapping the metagame in ST-space from a range of initial frequencies of cooperators: $c = 0.1$ (left), $c = 0.5$ (centre), $c = 0.9$ (right). In all cases $m = 0.5$. In this figure all arrows are shown the same size, indicating the direction of net change in the metagame (but not the relative magnitude).

Figure 5.6 shows the behaviour of the metagame as the initial frequency of cooperators c is varied. In these plots all the vectors are shown with the same magnitude to give a clearer indiction of the direction of change in the metagame. The comparison with

Figure 3.2 shows a strong connection with the equilibria of the underlying games on the ST-plane. Varying m was found not to have any effect on the direction of change in the metagame (apart from in the degenerate cases $m = 0$ and $m = 1$), though it could change the magnitude. The metagame functions in this way because each strategy type acts to maximise its own fitness (expected payoff). In the region where cooperation dominates, there is a selective pressure in favour of increasing S to benefit a cooperator on those (possibly rare) occasions when it encounters a defector, and likewise when defection is the dominant strategy there is selective pressure in the direction of increasing T. Where $S_M > S_N$ and cooperators dominate, linkage disequilibrium develops linking cooperation and the M GCT allele. Similarly when $T_M > T_N$ and defectors dominate, linkage develops between the D and M alleles.

The magnitude of the change in the metagame is small in the regions where there is a single social strategy present at equilibrium. When the metagame interaction reaches equilibrium in terms of the social strategy trait, so the population consists of just cooperators or just defectors, there is no further selective pressure on the game-changing trait. This is because R and P are constant for all games on the ST-plane, so one of the strategies MC and NC can at most weakly dominate the other, and likewise with MD and ND. However in the Snowdrift quadrant there is a polymorphic social strategy equilibrium so the metagame displays more complicated dynamics. The selective pressure is on increasing S and T proportionately to the ratio of cooperators to defectors at equilibrium, so the vector field lines flow radially outwards from the point $T = 1$, $S = 0$. The magnitude of the vectors for change in GCT frequencies in the Snowdrift region are the same regardless of the initial frequency of cooperators, whereas in the other regions it changes according to the changed basins of attraction for the 'all-cooperators' and 'all-defectors' evolutionary stable states of the social game. This is because the polymorphic population at equilibrium allows continual interaction between cooperator and defector types, so the metagame interactions reach the same equilibrium from any initial conditions.

5.5 Metagame Dynamics Under Constraint of Increasing/-Decreasing Assortment

The applications of the metagame model presented so far have been motivated primarily by abstract considerations like maintaining a constant strength of selection. We can also apply a metagame analysis where the biological motivation is more obvious. As we reviewed in Chapter 4, a particularly important example is where the game-changing trait is one that creates positive assortment on the social trait. This is because it is widely argued that the assortment of cooperative social interactions is a key to the evolution of cooperation (Hamilton, 1964b; Okasha, 2006; Godfrey-Smith, 2009). We

demonstrated that playing a game with increased assortment of interactions is mathematically equivalent to playing a different effective game in a population that is still freely mixed. This provides a natural way to use metagames to model whether or not traits that affect assortment will increase or decrease in frequency when they coevolve with social behaviours. Rather than impose assortment, metagames allow us to model its evolution.

As in Chapter 4, let $\alpha \in [0, 1]$ be a measure of positive assortment on the social trait. $\alpha = 0$ means no assortment, so encounters occur at the same frequency as types are present in the population. In a situation of full assortment, $\alpha = 1$, members of the population interact exclusively with those with the same behavioural trait. An intermediate level of assortment is taken to mean that a member of the population will interact with an individual with the same behavioural trait with probability α, and the rest of the time will interact with the different strategy types with probabilities proportionate to the population composition.

Recall that for any game $G = \left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$ there exists a corresponding fully assorted game $G^{assort} = \left(\begin{smallmatrix} R & R \\ P & P \end{smallmatrix}\right)$. No matter what social dilemma G represents, G^{assort} is a Harmony Game. Then any game G played with an assortment level of α is equivalent to a game G^{α} played in a freely mixed population:

$$G^{\alpha} = (1-\alpha)G + \alpha G^{assort} = \begin{pmatrix} R & S + \alpha(R-S) \\ T + \alpha(P-T) & P \end{pmatrix} \tag{5.10}$$

This equivalence allows us to translate questions about games played with different levels of assortment into questions about games with different payoff matrices, and hence to apply the tools of metagames developed so far. For any game G, we define the associated assortment metagame on the set $\Gamma = \{G^{\alpha} : \alpha \in [0, 1]\}$ with encounter function $\phi(G^{\alpha}, G^{\beta}) = G^{\frac{\alpha+\beta}{2}}$. This is a symmetric metagame. Metagame interactions are given by the matrix:

$$\begin{pmatrix} R & S + \alpha(R-S) & R & S + \frac{\alpha+\beta}{2}(R-S) \\ T + \alpha(P-T) & P & T + \frac{\alpha+\beta}{2}(P-T) & P \\ R & S + \frac{\alpha+\beta}{2}(R-S) & R & S + \beta(R-S) \\ T + \frac{\alpha+\beta}{2}(P-T) & P & T + \beta(P-T) & P \end{pmatrix} \tag{5.11}$$

To simplify calculations, it is always possible to let $G' = G^{\alpha}$, $\alpha' = 0$ and $\beta' = \frac{\beta-\alpha}{1-\alpha}$.

Figure 5.7 is a vector field diagram showing the behaviour of the metagame along lines of assortment for games on the ST-plane. Since $R = 1$ and $P = 0$ for every game on the ST-plane there is a unique fully assorted matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. The results are similar

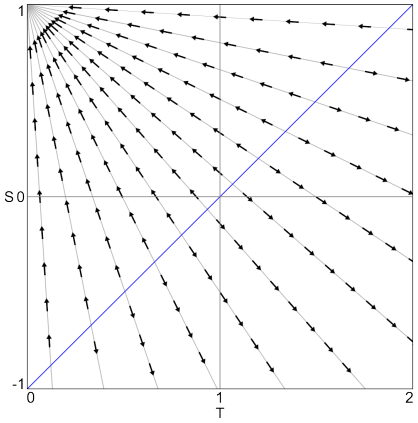


Figure 5.7: The dynamics of the metagame when the game-changing trait is social assortment. From these initial conditions ($c = 0.5$, $m = 0.5$) there is selection in favour of games with increased assortment when $S > T - 1$ (above the blue line).

to the metagames for games of constant total utility. Whether positive assortment is likely to increase or decrease depends on the equilibrium frequency of cooperators in that region of the ST-plane. An increase in assortment means that S will increase while T decreases and vice versa. Thus when cooperators dominate, the metagame will move in the direction of increased assortment, while when defectors dominate the metagame moves away from increased assortment.

From the perfectly balanced starting conditions of $c = 0.5$, $m = 0.5$m, assortment tends to increase when $S - P > T - 1 = T - R$ and decrease when $T - R > S - P$. The exception to this is when $S = 1 = R$ so selection is neutral with respect to cooperative strategies; here the metagame instead moves towards stable coexistence of strategies in the Snowdrift region. As with the other metagames we have studied, the initial conditions do not qualitatively affect the behaviour of the assortment metagame in the Harmony Game or Prisoner's Dilemma quadrants: assortment increased in the Harmony Game and decreases in the Prisoner's Dilemma quadrant. In the Stag Hunt quadrant, whether or not assortment increases depends on the basin of attraction of the local game: as the initial frequency of cooperators increases, so does the proportion of the Stag Hunt quadrant in which assortment spreads.

5.5.1 The Evolution of Assortment in the Snowdrift Quadrant

We can understand the evolution of assortment in the Snowdrift quadrant more precisely by returning to the results we first found for metagames on the circle. Our game-theoretic reasoning for the behaviour of metagame interactions on the circle in the Snowdrift quadrant only relies on the fact that as θ increases, the corresponding values of S increase and values of T decrease. This is also the case in the assortment metagame. If $G_M = {G_N}^{\alpha}$ is a Snowdrift Game G_N played with a higher level of assortment then

$S_M > S_K > S_N$ and $T_M < T_K < T_N$. Thus we know that for an assortment metagame where both G_M and G_N are in the Snowdrift region (as will be the case when increase in assortment is small) then the metagame interaction reduces to G_K as in the circle.

We know that:

$$G_K = \begin{pmatrix} 1 & (1-\frac{\alpha}{2})S_N + \frac{\alpha}{2} \\ (1-\frac{\alpha}{2})T_N & 0 \end{pmatrix} \tag{5.12}$$

We also know that this game G_K in the Snowdrift quadrant can be written in polar coordinates as $(S_K, T_K) = (r_K \sin\theta_K, r_K \cos\theta_K + 1)$. We know that the radius r_K does not affect the equilibrium of this game, so the same reasoning to understand games on the circle applies. The mutant assorting trait will increase in frequency in the metagame interaction if the equilibrium frequency of cooperators for G_{θ_K} is greater than the initial frequency of the mutant trait m.

Thus assortment evolves for a game in the Snowdrift region if the radial angle of the intermediate game G_K is greater than θ_m. As we know the coordinates on the ST-plane of G_K, we know that $\theta_K = \tan^{-1}\left(\frac{S_K}{T_K - 1}\right)$, so we have:

$$\begin{aligned} &\theta_K > \theta_m \\ &\quad\iff \tan^{-1}\left(\frac{S_K}{T_K - 1}\right) > \tan^{-1}\left(\frac{m}{1-m}\right) \\ &\quad\iff \frac{S_K}{T_K - 1} > \frac{m}{1-m} \\ &\quad\iff \frac{(1-\frac{\alpha}{2})S_N}{(1-\frac{\alpha}{2})T_N - 1} > \frac{m}{1-m} \end{aligned} \tag{5.13}$$

Therefore if m is small then assortment will increase over almost the entirety of the Snowdrift quadrant. Because the equilibrium frequency of bearers of the mutant assorting trait is absolute and unaffected by the initial conditions, when the initial frequency of the assorting trait is high the proportion of the Snowdrift region over which assortment will increase becomes smaller.

5.6 Discussion

A tempting, but incorrect, hypothesis to explain the behaviour of metagames would be that it is determined solely by the nature of the games in a given metagame interaction. After all, apart from the extreme cases of $c = 0$ or 1, changing the initial frequency of cooperators has no effect on the direction that the game changes in three-quarters of

the ST-plane (and changing the initial frequency of the GCT alleles has no effect on the direction of the vector field, only the magnitude of the vectors). It is the Stag-Hunt region that most obviously demonstrates this hypothesis is false; in fact the behaviour of the metagame is determined by the strategy frequencies in the metagame, which are in turn determined by the social equilibria of the games and the initial conditions. However, this effect is partially obscured because outside of the Stag Hunt region games have only one stable social equilibrium, so the initial social frequencies have no lasting effect.

The behaviour of the metagame in Stag Hunt games, though, shows that the initial conditions do matter — as there are two stable equilibria for the social strategy traits in the Stag Hunt game, the initial conditions determine which social equilibrium the population will reach and thus change the behaviour of the metagame. It is the local social equilibrium that determines the behaviour of the metagame, but only in the Stag Hunt region is this clearly distinguishable from the effects of the social game. The vector field follows the local social equilibrium; where there are no cooperators at equilibrium, the selective pressure on the GCT is in the direction of increasing T, while S increases when there are no defectors present in the social equilibrium. So depending on the initial frequency of cooperators and hence the social equilibrium in the Stag Hunt game, the selective pressure in the metagame will favour either increased S or increased T. We saw this behaviour replicated consistently across our different models, where game-changing traits that favoured increased cooperation (increasing the payoff for S or decreasing that for T) were favoured in regions where cooperation already dominated. This is an important initial result that we will return to.

It is worth considering whether using the ST-plane as a setting biases the results. The many benefits of using games on ST-plane have been covered, but the fact that the payoffs for mutual cooperation and mutual defection are fixed has two important consequences. First, the game can only change via changes in S and T, so strategy types can only increase their expected payoff by increasing the payoff for cross-strategy interactions. Were the payoffs for mutual cooperation and mutual defection subject to evolutionary change there would be more opportunity for the payoffs for coordinated actions to increase, keeping the metagame in a coordination game region. Nevertheless, simple geometric arguments reinforce the implication that a bistable region of game space will always be unstable in the metagame if the local social strategy equilibria are self-reinforcing as these will direct the metagame deeper and deeper into the basin of attraction of one of the equilibria.

The second consequence is that as the payoffs for mutual cooperation and mutual defection are fixed, strategies with the same social trait can at best only weakly dominate each other. This means that they are selectively neutral with respect to each other, so if the social game between cooperators and defectors reaches an equilibrium of all-cooperators or all-defectors, no further change is possible in the frequency of the GCT

alleles. Because of this, the absolute magnitude of the change in the GCT is often small. It also means that looking at multiple changes in game space requires the frequencies of the social strategies be perturbed, through a mechanism such as mutation. We will go on to develop a model that does this in Chapter 8.

It is interesting to consider the metagame in light of the breakdown of the ST-plane into radial coordinates, where the radius of the circle is equivalent to the strength of selection. We saw the technical benefits of reframing problems in ST-space in terms of the circle when we modelled the assortment metagame, but it can also aid conceptual understanding. Generally a game moves in the metagame in a direction that also increases its distance from the central point at $T - R = S - P$, agreeing with the intuition that when GCTs for similar games interact the one with increased strength of selection will prevail. This can also be seen in parallel with making and breaking links in games played on dynamic networks; when the game played is one suitable for the social equilibrium of the population this connection is strengthened, or at least does not break. It is only when the social game is unsuited to the equilibrium population state — for instance in a Stag Hunt game that is very close to the Harmony Game but where the initial frequency of cooperators is sufficiently low that the social equilibrium is all-defectors — that the movement in game space reduces the distance from the central point, effectively weakening the current social dynamics.

The metagame shows that social dilemmas that take the form of a Stag Hunt are unstable when the game is subject to evolutionary control. In this region the metagame dynamics change the social game so that the basin of attraction of the dominant social strategy is larger. This process, where time spent at a particular equilibrium enlarges the basin of attraction of that equilibrium, parallels Hebbian learning and the formation of associative memory in neural networks (Watson et al., 2011b), an example of the deep analogies between learning and evolution (Watson et al., 2015). Where there are two stable attractors for the social game, there is always a selective pressure in favour of enlarging one or other of the basins of attraction, and so through positive feedback the metagame is likely to move to either the Harmony Game or Prisoner's Dilemma regions. Worden and Levin (2007) and Akçay and Roughgarden (2011) also both found that a Prisoner's Dilemma could transition into a Snowdrift Game.

However, this is the opposite conclusion to Fort (2008), where the outcome is always a Stag Hunt coordination game. Where does this discrepancy come from? There are a few possibilities. First, in Fort's model, strategy-cells have complete control over their own payoff, but not the strategy of the opponent they'll face. In the metagame model, an individual's payoff depends some form of compromise between the game that it wants to play and the game its opponent wants to play. Second, Fort's model is a spatial one, and so is likely to feature a heterogenous distribution of strategies over the cellular automaton, and hence clusters of users of the same strategy due to the 'best neighbour takes over' update rule. There is a potential parallel to the work on effective game

transformations in games on networks in Pinheiro et al. (2012), which suggests that Prisoner's Dilemma games played on heterogenous networks transform to Stag Hunt games, while Prisoner's Dilemma games played on homogenous networks transform to Snowdrift Games.

How might we apply metagames? As well as being a technical model, metagames can be used conceptually when looking at examples of social evolution. Consider the *Anthopleura* genus of sea anemones, in particular *Anthopleura elegantissima* and *Anthopleura sola*, two species believed to be very closely related — indeed, considered a single species until recently (Pearse and Francis, 2000). *A. sola* is a solitary-living sea anemone; they are antagonistic toward one-another and compete for space on the rocky shore. *A. elegantissima* is a colonial anemone that clones itself by fission, sharing resources within the colony but is antagonistic towards other clones (Ayre and Grosberg, 2005). It is unclear whether *A. elegantissima* arose from a solitary ancestor that evolved clonality or a clonal ancestor that invaded the present habitat of the solitary form, but it is believed that there have been repeated shifts between clonality and solitariness (Geller and Walton, 2001). Here the life history of the anemones is a large scale game-changing trait, or more likely the product of many game-changing traits acting as conflict modifiers between individual and colonial interests, with *A. sola* and *A. elegantissima* occupying different points in metagame space — solitary and colonial living. Though it may not be feasible to perform a technical metagames analysis on the evolution of life history traits in *Anthopleura*, experience with the metagames model informs what is required. In particular, it is important to understand the constraints on the way that the life history can evolve. In this case, evidence suggests that the life history model is adaptive, but cannot change in a completely unconstrained manner — genetically based variation for fission rate, or phenotypic plasticity, is insufficient to explain intraspecific differences (Geller and Walton, 2001).

5.7 Conclusions

There already exist approaches in the literature (such as Traulsen et al., 2009) that explicitly model the way a game will change in particular circumstances; for example, due to changes in population structure determined by evolving the structure of an interaction network. The level of abstraction in our evolutionary formulation of metagames gives the model a complementary role. By modelling the way a population might change the game it is playing in a simple and explicit way, the results can be applied to a wide range of situations. Specific scenarios can then be modelled by imposing appropriate constraints on the metagame. By remaining ultimately agnostic to the mechanisms that induce game change, whether they be changing social ties, greenbeard signalling traits or the evolution of a propagule-based reproductive method, the metagame model

can illustrate general principles and thus sit alongside approaches that provide greater specificity.

Metagames provide a simple model that demonstrates the feedback mechanisms between a social game and its equilibria when both are subject to evolutionary pressures. When the game can change in an unconstrained way, the metagame changes to increase the fitness of the dominant social strategy. In the case of a metagame on the ST-plane this results in an increase in S, or T, or both when the metagame is in the Snowdrift region that supports a polymorphic population. The extent to which S and T are maximised depends on the frequency of cooperators and defectors in the population, which in turn depends on the equilibria of the social game. While the social game is important in influencing the dynamics of the metagame, the Stag Hunt region, where it is possible to make a distinction between the effects of the game and the effects of the social equilibrium, shows that it is the current equilibrium that determines the direction of change of the metagame and not the current game *per se*.

We have demonstrated, however, that when populations can change the game under different constraints it leads to metagames with very different properties. In some metagames there are attractors, in others the metagame completely diverges. When individual traits can alter payoffs but cannot alter the strength of selection (so it can be represented on the circle), the metagame has two attractors — one that only supports cooperators, and one that only supports defectors. Depending on the initial social frequencies in this metagame, a Stag Hunt game where both all-cooperators and all-defectors are possible evolutionarily stable states will move into the basin of attraction of only one of the two. When the initial conditions lead to the all-cooperator ESS, the game-changing trait will move towards a Harmony Game, when they lead to the all-defectors ESS the game will move towards a Prisoner's Dilemma. Because of the constraints in the metagame, Harmony Games and Prisoner's Dilemmas approach Snowdrift Games without leaving the basins of attraction for their single-strategy equilibria.

By contrast, in the constant total utility metagame where individuals have traits that can alter their payoffs but cannot alter the sum of all payoffs in the game, the metagame is divergent. Stag Hunt and Snowdrift Games become either Prisoner's Dilemma or Harmony Games depending on the initial game and initial social frequencies (Figure 5.4). The important metagame where the game-changing trait is assortment lies somewhat between the two — there is a single attractor for games that are fully assorted, but the basin of attraction does not extend to cover the entire ST-plane. Indeed, we find consistently across almost all the different metagames that the pro-cooperation game-changing traits are favoured in conditions that already support some level of cooperation.

In particular, this means that assortment does not evolve when the underlying conditions are a Prisoner's Dilemma. There is an apparent contradiction between this result

and previous models of social niche construction, which has claimed increasingly cooperative population structures will always be favoured subject to individuals living in the population structure they are genetically predisposed towards (Powers et al., 2011). This contradiction is important since we are claiming that metagames can serve as a formal model of social niche construction. We will examine this contradiction in detail in this next chapter, but our investigation is guided by the one investigated situation in which the game-changing trait did move from a Prisoner's Dilemma to a Harmony Game: in the constant total utility metagame where we changed the encounter function. We proposed that this can be interpreted as individuals meeting other individuals with the same GCT more often than the frequency of that GCT allele in the population – in essence, introducing assortment on the game-changing trait.

Here then, in a series of different applications of metagames we have shown that the constraints on how individual traits can alter the game result in metagames with very different outcomes. This ability to apply different constraints to the space of games that the metagame ranges over is one of the key strengths of the model. Metagames can have attractors or diverge: feedback between social behavioural and social context traits pushing games that favour cooperators into ones that more strongly favour cooperators, and games that favour defectors into games that more strongly favour defectors. The evolution of game-changing traits can transform Stag Hunt or Snowdrift Games into either Prisoner's Dilemmas or Harmony Games, depending not just on the initial game, but also the initial state of the population and the different mechanisms facilitating game change.

Why then do individuals play the game they play? At this point we have shown that there are conditions where individuals change the game to support increased cooperation despite being driven by the maximisation of individual utility, and we have provided a framework for analysing game changing behaviour in general cases. Even when the initial game is a Prisoner's Dilemma, we have shown that if the metagame has suitable constraints it is possible for the game to change to any of the other fundamental two-player games: the Snowdrift, Stag Hunt or Harmony Games. However, for game-changing traits that support increased cooperation to evolve in a Prisoner's Dilemma, we require special conditions on the encounter function that mediates interactions between the bearers of different GCT alleles.

This means that to fully understand the evolution of a game-changing trait, we must understand the kind of metagame that it corresponds to: in particular, the constraints on the way that the game can change (given the mechanics of the specific situation), and the way the game-changing trait affects interactions between the bearers of its different alleles. We now have a hypothesis that assortment on the game-changing trait itself plays an important role in the evolution of game-changing traits that support increased cooperation, the crucial mechanism for social niche construction. We spend the rest of this thesis understanding how assortment on game-changing traits affects their evolution.

Chapter 6

Assortment on Game-Changing Traits

So far in this thesis, we have identified in the literature why cooperation is so important, such as for its essential role in evolutionary transitions (Chapter 2. We have reviewed the arguments that assortment is the ultimate explanation behind many instances of the successful evolution of cooperative behaviours, and how we can mathematically represent assortment as a game-changing trait that transforms the payoff matrix of an evolutionary game (Chapter 4).

In Chapter 5 we introduced metagames to model the coevolution of social traits and game-changing traits such as assortment. In our initial modelling work we found that assortment and other game-changing traits that support increased cooperation evolve in conditions where cooperation is already favoured, such as a Harmony Game, or a Stag Hunt game where the initial conditions lead to a cooperative equilibrium. In no models with our standard encounter function did assortment spread in conditions where cooperation was not favoured. In particular, it never evolved in a Prisoner's Dilemma. Indeed, under Prisoner's Dilemma conditions, game-changing traits unfavourable to cooperation spread instead.

The results of this first investigation into the evolution of assortment pose a problem because they are in apparent conflict with those of social niche construction (Powers et al., 2011). We claim that metagames can serve as a formal model for social niche construction, yet the logical argument for social niche construction claims that when there are two different population structures, one of which supports a higher level of cooperation than the other, then as long as individuals on average live in the population structure they have a genetic preference for the more cooperative structure will always spread.

In this chapter we attempt to resolve this conflict. We saw that the only metagame model in which the game-changing trait evolved from a Prisoner's Dilemma to a Harmony Game was when we used a different encounter function that reduced the strength of interactions between bearers of different GCT traits. In the extreme case where such an encounter function zeroes cross-GCT games we can see how this would be equivalent to having no interactions between bearers of different GCT. Thus we proposed that encounter functions could approximate the effects of what is essentially assortment on a game-changing trait.

This has led us to the hypothesis that assortment on the game-changing trait itself plays an important role in the evolution of game-changing traits that support increased cooperation. In this chapter we pursue this by investigating how to model assortment on game-changing traits. We begin by developing a simulation model of an agent-based population that is subject to different kinds of assortment (Jackson and Watson, 2013). This demonstrates how the assortment on social and game-changing traits are not orthogonal, but can have complex interactions that reduce the strength of cooperation while aiding the spread of the underlying social conditions that could support more cooperation in the future. We then demonstrate how we can use interaction functions (Chapter 4) to transform game space to represent the interaction frequency data obtained from the simulation model.

This leaves us with two candidate techniques for modelling assortment on game-changing traits: encounter functions and interaction functions. We compare the two methods and find that the encounter functions method approximates the results found using interaction functions, but also introduces mathematical artefacts that the interaction functions technique avoids. With this technique to model GCT assortment we can determine how GCT assortment affects the evolution of game-changing traits.

6.1 The Logical Argument for Social Niche Construction

We have highlighted the contradiction between our initial modelling results on the evolution of cooperation-supporting game-changing traits such as assortment and the claims of social niche construction. Now we examine this discrepancy in more detail, considering the logical argument supporting social niche construction.

Social niche construction claims to provide a higher-order explanation of the evolution of cooperation, driven not just by adaptation to particular social conditions but in turn driving the evolution of those social conditions (Powers, 2010; Powers et al., 2011; Ryan et al., 2016). In this sense, social niche construction is a circular process where organisms modify their own social niche to change the conditions of their own social evolution. Social niche construction theory predicts that cooperative behaviours will be accompanied

by coevolved traits that support cooperation enabling it to be stable. We have represented the evolution of such traits in our metagames formalism as game-changing traits that transform the effective social game, such as by altering population structure, creating life-history bottlenecks, segregating slowly dividing germlines or policing eusocial insect workers.

But social niche construction is more than just the conceptual application of the idea of niche construction to social evolution. Powers et al. (2011) provides a logical argument that social niche construction is a process that leads to positive feedback loops between the evolution of cooperation and cooperation supporting traits. Other authors have previously suggested cooperation can drive the evolution of social conditions, such as the mating system (Breden and Wade, 1991), creating such 'runaway' increases in cooperation(Santos et al., 2006b; Rosas, 2010; Van Dyken and Wade, 2012b; Clarke, 2014). The logical argument for social niche construction is advanced as a general argument that can apply to any heritable trait affecting the population structure of its bearers (which in our analysis we broaden to include all types of game-changing trait). It is this logical argument that we appear to contradict in finding that cooperation supporting traits did not evolve in the important case of the Prisoner's Dilemma region of game space.

The problem for social niche construction is the same as that for other evolutionary accounts of cooperation: individual selection occurs according to relative fitness, and defectors will generally have a relative fitness advantage over cooperators as they do not incur the costs of cooperation, even though on a population level a higher mean fitness would be achieved were all individuals to cooperate. As such, we might expect that selfish individuals will take advantage of their greater relative fitness and change the social conditions to best favour them. Indeed, that is exactly what we found in our modelling work in Chapter 5: in Prisoner's Dilemma regions, game-changing traits evolved to increase the payoff for defection (T).

As in the metagames model, we consider the interaction of two different types of trait: a social trait (C or D) and a game-changing trait. Here we specify that the game-changing trait corresponds to a particular population structuring trait, such as one that creates increased assortment. Take two game-changing traits that lead to such structures, M and N, where the mutant trait M supports an increased level of cooperative behaviour. Though the two traits may start off at linkage equilibrium (as we have modelled in our metagame scenarios), the logical argument for social niche construction claims that linkage disequilibrium will develop through the following process:

- Cooperators living in structure M will increase in frequency more than those living in structure N.
- Therefore there will be positive linkage disequilibrium between the M game-changing trait and the cooperative trait C.

- Therefore there will be positive linkage disequilibrium between the N game-changing trait and the non-cooperative trait D.
- Since cooperation increases mean fitness, and cooperation is linked to the M game-changing trait, individuals living in structure M (possessing the M game-changing trait) have a higher component of fitness due to social behaviour than those living in structure N.
- Therefore the M game-changing trait will increase in frequency.

What are the potential resolutions to the discrepancy? In the metagame modelling we did in Chapter 5, the logical argument falls victim to Simpson's Paradox. Positive linkage disequilibrium does develop between the cooperative trait and the more cooperative game-changing trait M, and likewise between the non-cooperative trait and N. However, when the effective games are in the Prisoner's Dilemma quadrant the total frequency of cooperators decreases to zero. Yet the logical argument does hold for the model in Powers et al. (2011) under Prisoner's Dilemma conditions, so it is not the case that social niche construction can never occur in a Prisoner's Dilemma.

We can be confident too in the formal treatment of metagames, since it is based on the mathematics of two-player four-strategy symmetric games and the replicator equation. If both metagames and the logical argument are valid, then it must be the case that our interpretation of metagames models is incomplete, or does not account for additional features of the logical argument.

In particular, we come to the assumption of the logical argument that individuals possessing a game-changing trait for population structure M have to live in structure M on average. Individuals in a metagame that possess the game-changing trait M all live in a social niche modified by their possession of the GCT M. In their interactions with other individuals with the GCT trait M they play the effective game G_M, while in their interactions with individuals possessing other game-changing traits they play an effective game determined by the encounter function applied to M and the other game-changing trait.

So under the metagames model, although individuals do live in social conditions created by their own trait, they also interact with others bearing game-changing traits that create different social conditions. If M and N are both game-changing traits that create different levels of assortment of the social trait, with M creating the social conditions for a higher level of assortment than N, then the bearers of M will experience greater social assortment than the bearers of N despite the interactions between the two.

However, when we use the encounter function of $\phi_{0.5,0.5}$ that averages the payoff matrices determined by the two game-changing traits this is not enough in our model for pro-cooperation traits to spread in a Prisoner's Dilemma. We raised our interpretation that

this is akin to the two game-changing traits being well-mixed in the overall population. But this is not necessarily the case if the effect of the game-changing trait were to segregate its bearers, for instance by distributing them between different demes. This use of different encounter functions suggests a route to the resolution. We saw that when we changed the encounter function to $\phi_{0,0}$, more games move towards the Harmony Game – including some games in the Prisoner's Dilemma.

We propose a modification: what is required is that individuals assort on their game-changing traits. It is not just that individuals need to live, on average, in the population structure their GCT specifies, and so interact with other individuals possessing the same game-changing trait, but that they do not interact as frequently with individuals possessing other game-changing traits.

6.2 Directly Modelling the Effects of GCT assortment

We have repeatedly highlighted the particular importance of game-changing traits that create assortment on the social trait. When the game-changing trait is one that creates assortment then any assortment on that trait itself is a kind of 'second-order' assortment. Not all game-changing traits do result in second-order assortment. We can imagine how assortment on social and game-changing traits can vary across biological and modelling-inspired scenarios:

- When individuals have a group size preference (Powers and Watson, 2011), assortment on the social is generated by the increased variance from sampling small groups in a large population coupled with the differential success of more cooperative groups. If the group size preference leads to a greater chance of living in a group of the desired size then it also implicitly creates groups composed of individuals with similar group size preferences, so there is also assortment on the game-changing trait.

- In a network model where the game-changing trait determines an individual's preferred number of social connections, these preferences can be satisfied without creating assortment on the game-changing trait since an individual with a preferred degree of four can 'live in' the structure it desires by having four social connections with individuals with completely different degree-preference traits.

- When individuals possess a game-changing trait for dispersal radius, densely dispersed relatives will be assorted primarily on the game-changing trait, with the consequent possibility for assortment on the social trait.

- Assortment generated by signalling (greenbeard traits) actually generates assortment on the game-changing trait, the green beard. The green beard is useful to

the extent that the assortment it creates on this game-changing trait leads to assortment on the social trait, with parasitism and intragenomic conflict (suppressor mutations arising at other *loci*) potentially acting to break this link (Okasha, 2002).

- In general, when individuals are genetically related by common descent, the relatedness measures the chance of similarity between individuals at both the social trait and game-changing trait. Thus we see relatedness has the special property of creating assortment on both game-changing and social traits.

Our ultimate aim in this chapter is to be able to represent assortment on game-changing traits in metagame models.

First though, we develop our understanding of the interaction between these two types of assortment by presenting an agent-based model where we can directly vary the levels of assortment between different traits. In this model we investigate the effects of 'second-order' GCT assortment on the evolution of game-changing traits that support cooperation and compare different types of GCT assortment. We do this through an artificial life model in which different levels of assortment are expressed literally — assorters are physically more likely to interact with each other.

We then show how we can reproduce these results mathematically by using interaction functions to modify the replicator equation. We show that when GCT assortment is random or absent then the spread of the game-changing traits that support increased cooperation is strongly linked to the success of cooperators. When game-changing traits affect both assortment on social behaviours and on themselves, then the conditions under which populations can evolve game-changing traits beneficial to cooperation are enlarged.

6.2.1 Simulation Model Details

To investigate the effects of game-changing trait assortment on the evolution of population structures favourable to cooperation we created an evolutionary algorithm based on a fixed size population of asexual agents living in non-overlapping generations. A range of potential assorting behaviours are abstractly represented by having agents physically cluster to varying extents determined by their game-changing traits. This physical assortment influences fitness by affecting who each agent plays with, specifically as each agent's fitness is determined by playing a specified evolutionary game G against its nearest neighbour. The game-changing traits do not alter the payoffs of the base game or the rule to play with the nearest agent, but by changing the way that agents cluster together they alter the likelihood of encountering similar strategies and so change the effective game, though here the change is done directly rather than through payoff matrix transformation.

The model consists of a population of $P = 100$ agents living in a continuous space that is topologically toroidal. Each agent is represented as a circle $r = 4$ units in radius. The world is $100r \times 100r$ units in size. Each agent's genotype is haploid with two *loci* — a social trait gene that determines its social behaviour (with either 'cooperator' C or 'defector' D alleles) and a game-changing trait with alleles A for assorting behaviours and M for freely mixing behaviour. This gives four genotypes: CA, DA, CM and DM. The model is initially populated with these four genotypes present at equal frequency with agents placed randomly in the world.

There are two stages to the evolutionary algorithm. First the agents selectively aggregate for $T = 10000$ timesteps. Then each member of the population plays a game with its nearest neighbour to determine its fitness. 'Nearest neighbour' is not a symmetric relationship, so while all agents play at least one game (against their own nearest neighbour), some may be played against multiple other times by other agents that have it as their nearest neighbour. As all the games here are symmetric this is equivalent to the agent playing multiple times with different opponents.

For this model, an agent's fitness is defined to be its average payoff received per interaction rather than cumulative payoff; we choose not to reward or penalise an agent for being involved in multiple games. Recording the payoff of the focal player would also achieve this result but would mean a cooperator closest to another cooperator but surrounded by defectors received the same payoff as a cooperator with no defectors near it. However, there are arguments in favour of other mechanisms and the choice of average payoff represented a trade-off.

After the fitnesses are calculated the population reproduces using tournament selection up to the fixed size P: for P repetitions two agents are drawn from the population and the agent with the highest fitness (or a random agent if they had equal fitness) reproduces clonally with a small chance of mutation. To represent the intuition that population structure evolves more slowly than social behaviours the mutation rate for the social behaviour allele is set to a probability of $m_{SB} = 0.01$, while the probability of a mutation of the game-changing trait is $m_{GCT} = m_{SB}/2$.

The agents' movement is modelled by a variation on gravitational attraction. To simplify calculations, each agent is defined to have a mass of 1. The 'gravitational' force between agents i and j is then calculated as $\frac{G_{i,j}}{d^2}$. Here d is the distance between the two agents' centres taking into account the fact that the world is toroidal. $G_{i,j}$ is the attractive constant between the two agents, determined by their genotypes and three parameters - α, β and γ. These parameters influence the levels of first and second-order assortment in the model:

- α is the attractive force that agents with the assorting allele A feel towards agents with the same social trait as themselves. So agents with the genotype CA feel an attractive force of strength α towards other CA and CM-typed agents.
- β is the attractive force between agents with the assorting GCT allele A.
- γ is the attractive force between agents with the mixing GCT allele M.

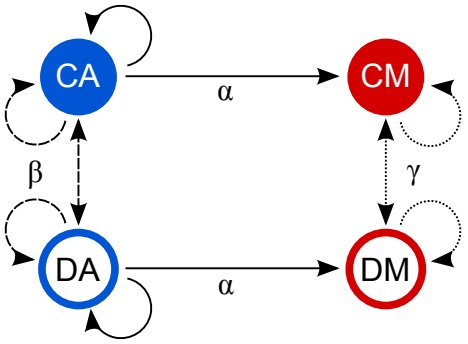


Figure 6.1: The attraction between the different genotypes is determined by the three parameters α, β and γ

.

Figure 6.1 shows diagrammatically the different forces that exist between each genotype. The combined forces can be tabulated to give the strength of attraction between genotypes (Table 6.1). Note that the attractive forces are not symmetrical, unlike real models of gravitation — agent i may be attracted to agent j more than j is to i. $G_{i,j}$ is then computed as rg where g is the attractive force from agent i to agent j as given in the table.

	CA	DA	CM	DM
CA	$\alpha+\beta$	β	α	0
DA	β	α	0	α
CM	0	0	γ	γ
DM	0	0	γ	γ

Table 6.1: The attractive force from an agent of one genotype (rows) to an agent of another genotype (columns).

There is also a repulsive force with magnitude -1 that affects the agents when they are closer than $2r$ apart and prevents them from overlapping. At each timestep the forces between all agents are calculated and the net force F_i for each agent calculated. Friction is then accounted for using the equation $F_i = F_i - 0.2v_i$, where v_i is the agent's previous velocity. Each agent's acceleration, velocity and position are then calculated and updated using numerical integration and the standard equations of motion in a plane, using an integration timestep of 0.1.

6.2.2 Model Scenarios

The model was run using four different sets of parameters, the relative strengths of α, β and γ defining four different scenarios with respect to the levels of first and second-order assortment and the nature of the second-order assortment:

6.2.2.1 No Assortment

($\alpha = 0$, $\beta = 0$, $\gamma = 0$). In this control scenario there is no attraction between agents. Consequently each agent's nearest neighbour is randomly determined by the initial placement of agents in the world and frequencies of the social strategies are expected to reach the equilibrium of the game being played.

6.2.2.2 Social Trait Assortment Only

($\alpha = 1$, $\beta = 0$, $\gamma = 0$). In this scenario, the only assortment is between social behaviours, where agents with the assorting game-changing trait are attracted to agents with the same social strategy traits, regardless of whether or not the other agent possesses the assorting trait.

6.2.2.3 Emergent GCT Assortment

($\alpha = 1$, $\beta = 1$, $\gamma = 0$). In this model agents with the assorting trait are attracted to others with the same strategy, but there is also attraction between agents with the assorting trait. This produces a scenario in which second-order GCT assortment is tied to the mechanism that generates first-order social trait assortment. This could be the case where strategy-assorters possess greenbeard traits.

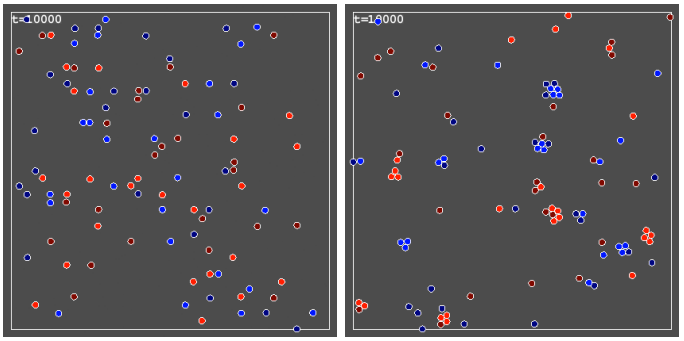
6.2.2.4 Enforced GCT Assortment

($\alpha = 1$, $\beta = 1$, $\gamma = 1$). In this model as well as strategy assortment there is also uniform assortment on game-changing traits. This is intended to produce a situation in which levels of assortment are uniform across different traits such as may be the case with relatedness.

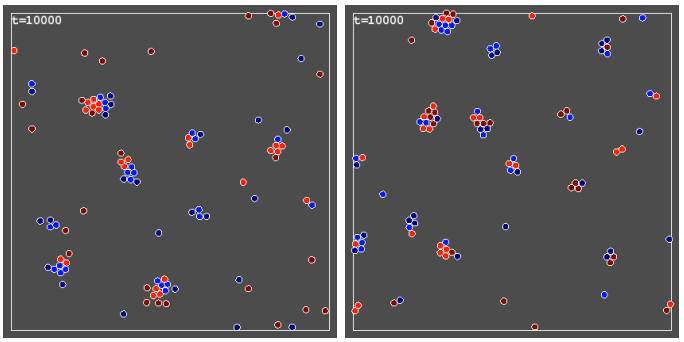
6.2.3 Simulation Model Results

Figure 6.2 shows the clustering that occurs in the model as the agents aggregate. In the control scenario, no clustering takes place. In the scenario with only social-trait

assortment, cooperators cluster with cooperators and defectors with defectors. This is predominantly agents with the CA genotype grouping with other CAs and DAs grouping with DAs as these are the agents attracted to others with the same strategy allele. When GCT assortment is emergent, cooperators and defectors cluster together, with CAs and DAs at the heart of the clusters and CMs and DMs at the edges attracted to other cooperators and defectors respectively. When GCT assortment is enforced, the clusters are more mixed.



(a) No Assortment (b) Social Only

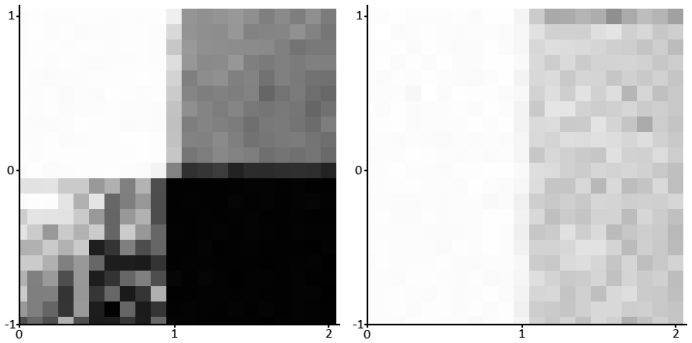


(c) GCT Emergent (d) GCT Enforced

Figure 6.2: Visualisations of the model at the end of the first generation ($T = 10000$) showing the way that agents cluster. Agents are coloured according to their genotype as in Figure 6.1.

To investigate the effects of GCT assortment over a wide range of social dilemmas we took a 21×21 lattice of points across ST-space 0.1 units apart and ran each scenario for every game on the lattice. Each scenario was repeated 5 times for every game and the results averaged. One of the dynamics in the model is that there is no selective difference between individuals with the same social strategy allele in the absence of individuals with the other social strategy allele (the same as with the metagame models in Chapter 5). The difference in fitness between, for instance, the CA and CM genotypes comes from their different interactions with the DA and DM genotypes. So when the model reaches a state in which only one of the social strategy alleles is present then there is no selective pressure between them and their frequencies begin to take a random walk. Experimental testing indicated 20 generations provided a balance between letting the

model reach equilibrium and mitigating the effects of the random walk, so this was the length of each run of the model.



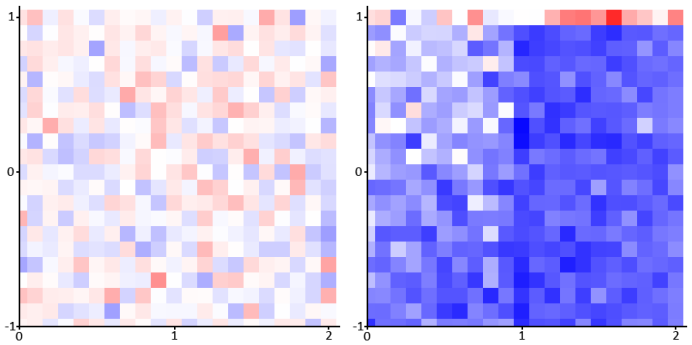
(a) No Assortment (b) Social Only

(c) GCT Emergent (d) GCT Enforced

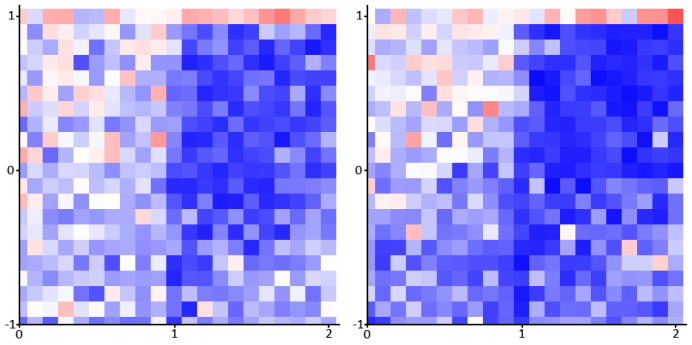
Figure 6.3: The mean absolute frequency of the C allele over the ST-plane in all four models on a scale where white indicates 100% cooperators, black 100% defectors.

Our hypothesis was that increasing GCT assortment would increase the spread of the assorting game-changing trait. We found some support for this view but with complications, some of which were obvious in retrospect. Figure 6.3 plots the mean absolute frequency of the C allele over ST-space in all four models. In all scenarios in which there is assortment, cooperators perform better than in the control. However, counter to our initial expectations, cooperation is more successful when there is just assortment on social strategies.

Figure 6.4 plots the mean frequencies of the A allele over ST-space. In the control model the frequency of the A allele is essentially random. In the other models, as expected, increasing levels of assortment increases the spread of the A allele. These results are summarised in Table 6.2 which records the mean frequencies of the C and A alleles over all games and all runs. As the table shows, there is essentially no net change in frequencies over the whole of ST-space in the control model. In the three scenarios with assortment, the frequencies of the C and A alleles increase, but with a trade-off between increased levels of cooperation and assortment on the game-changing trait.



(a) No Assortment (b) Social Only



(c) GCT Emergent (d) GCT Enforced

Figure 6.4: The mean frequencies of the A allele over the ST-plane on a red-white-blue scale. Red indicates the A allele decreases in frequency (< 0.5), blue that the A allele increases in frequency (> 0.5)

Scenario	Mn C	Var C	Mn A	Var A
No Assortment	0.505	0.140	0.501	0.005
Social Only	0.814	0.019	0.669	0.017
GCT Emergent	0.748	0.088	0.684	0.025
GCT Enforced	0.729	0.103	0.731	0.028

Table 6.2: The mean final frequencies of the C and A alleles over the ST-plane in each scenario.

6.2.4 Applying Interaction Functions

Agent-based simulation models like the one presented here are subject to noise, and it can be difficult to tune the desired behaviours precisely, so we can benefit from reproducing the results in a mathematical model. We can do this by using the interaction functions that we introduced in Chapter 4. Recall that these work by changing the replicator equation to match the actual frequencies with which genotypes interact with others due to the effect of game-changing traits, as opposed to the freely mixed case where the interaction probabilities match the frequencies at which the genotypes occur in the population.

These interaction functions are a set of n functions $e_i : S_n \to S_n$ that map the population state vector to the actual frequencies at which the i-th genotype encounters other genotypes. The fitness of the type is the composition of the fitness and interaction functions $f_i \circ e_i(x)$, and the mean fitness is $\sum_j^n x_j f_j \circ e_j(x)$. This gives the new replicator equation:

$$x_i = \dot{x_i}\Big(f_i \circ e_i(x) - \sum_j^n x_j f_j \circ e_j(x)\Big) \tag{6.1}$$

The set of modified fitness functions $g_i : g_i = f_i \circ e_i$ then defines a new game that when played in a well-mixed population is equivalent to the original game played with the given population structure. We saw that when the encounter functions are affine then they define a transformation of the payoff matrix. However, in this case the interaction functions are non-linear (so as to map map valid population states to valid population states where the entries sum to 1). This means their effect cannot be represented as a payoff matrix transformation, but they can still be used to modify the replicator equation.

We use interaction functions to mathematically model the results of the simulated scenarios, constructing the functions based on data from the simulations. First we define the four-strategy game in the absence of interaction functions. If we consider the strategies to be $x_1 = CA$, $x_2 = DA$, $x_3 = CM$, $x_4 = DM$ then for an arbitrary social game $G = \left(\begin{smallmatrix} R & S \\ T & P \end{smallmatrix}\right)$ the matrix of the full four strategy game is:

$$\begin{array}{c} \\ CA \\ DA \\ CM \\ DM \end{array} \begin{array}{c} \begin{array}{cccc} CA & DA & CM & DM \end{array} \\ \begin{pmatrix} R & S & R & S \\ T & P & T & P \\ R & S & R & S \\ T & P & T & P \end{pmatrix} \end{array} \tag{6.2}$$

This is then modified using interaction functions. To represent the changed number of interactions due to the population structure we use a simple non-linear interaction function — multiplying each entry in the population state vector by a scalar representing an increased chance of encountering that genotype and then normalising the resulting vector so the entries sum to 1. The scalars were calculated by running the model until the first reproduction event (at $T = 10000$) 100 times starting from evenly distributed population frequencies and recording the total number of games played between each pair of genotypes. Dividing this by the total number of interactions gave the mean frequencies at which a given genotype would encounter each other genotype when the actual frequency of each genotype was 0.25, so dividing again by 0.25 gives the actual encounter

rate between different genotypes as a multiple of what would have been expected in a well-mixed population.

We used these scalars to define the interaction functions for the four types in the model. This was a basic way of determining the interaction functions — a more complex way would have been to generate scalars for different actual population frequencies and interpolate between them to create more complex interaction functions. However, this simple method was sufficient to capture the behaviour of the three non-control models; the match between the results is illustrated in Figure 6.5.

	CA	DA	CM	DM
CA	2.32	0.14	1.19	0.36
DA	0.13	2.35	0.33	1.18
CM	1.54	0.44	0.93	1.09
DM	0.46	1.53	1.09	0.92
CA	2.14	0.93	0.72	0.20
DA	0.93	2.15	0.21	0.70
CM	0.96	0.28	1.22	1.54
DM	0.26	0.94	1.55	1.25
CA	2.22	1.07	0.49	0.22
DA	1.06	2.21	0.22	0.50
CM	0.51	0.23	1.50	1.76
DM	0.24	0.54	1.77	1.45

Table 6.3: The rows define the interaction functions giving the coefficients that modifying how likely it is for the row genotype to encounter the column genotype for the three scenarios with assortment, listed in order.

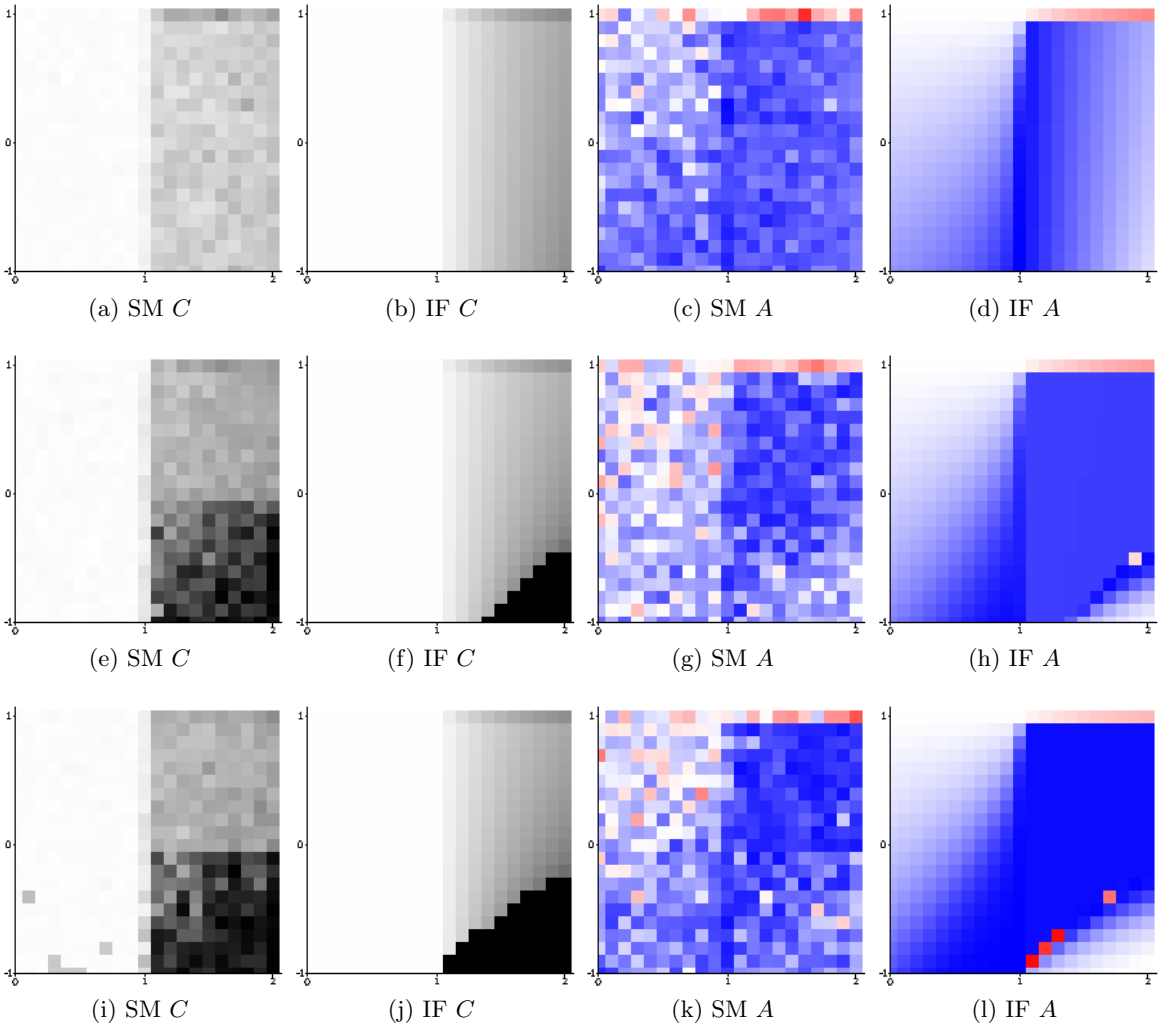
Table 6.3 gives the interaction coefficients that were used to define the interaction functions for the three scenarios in which there was assortment in the model. For example the first row defines the interaction function e_1, describing the transformation in the interaction frequencies for the genotype CA in the social strategy assortment-only model:

$$e_1\begin{pmatrix} x_{CA} \\ x_{DA} \\ x_{CM} \\ x_{DM} \end{pmatrix} = \frac{1}{2.32x_{CA}+0.14x_{DA}+1.19x_{CM}+0.36x_{DM}} \begin{pmatrix} 2.32x_{CA} \\ 0.14x_{DA} \\ 1.19x_{CM} \\ 0.36x_{DM} \end{pmatrix} \tag{6.3}$$

6.2.5 Simulation Model Discussion

Our expectation for this model was that increasing the level of assortment on the game-changing traits would lead to the increased prevalence of the GCT that supported cooperative behaviours — in this model represented by a GCT that directly increased correlated interactions between individuals with the same social strategy. This was true, but we did not anticipate that there would be a degree of trade-off between increased levels of the cooperative (C) and assorting (A) alleles. The comparison of the different scenarios reveals that cooperative strategies are more successful when there is just social

strategy assortment (with a mean of frequency of cooperators of 0.814 over the whole of ST-space), though the frequency of cooperators when there is GCT assortment (mean frequencies of 0.748 and 0.729) is still closer to the level obtained with only strategy assortment than to the case of no assortment at all (0.505).



(a) SM C (b) IF C (c) SM A (d) IF A

(e) SM C (f) IF C (g) SM A (h) IF A

(i) SM C (j) IF C (k) SM A (l) IF A

Figure 6.5: Visual comparison of simulation model (SM) results and the interaction function recreation (IF) for the three scenarios with assortment. The simulation model results are as in Figures 6.3 and 6.4, the interaction function graphs reproduce these scenarios.

The reason for this trade off is that assortment on social strategy and population structuring traits are not orthogonal processes. When there is just social strategy assortment, cooperators interact preferentially with cooperators and defectors with defectors, greatly reducing the number of cross-strategy interactions. The inclusion of assortment on the GCT alleles reduces this effect by bringing together cooperators and defectors with the same GCT allele. While increasing GCT assortment decreases the frequency of cooperators (relative to strategy assortment only), it increases the spread of the GCTs that ultimately promote cooperation. GCT assortment decreases the dependency between

cooperation and cooperation-promoting GCTs; when there is GCT assortment the A allele is able to spread even when the local social game is a Prisoner's Dilemma dominated by defectors.

An alternate way of looking at this is that it demonstrates that the dominant social behavioural trait does not necessarily have to control the evolution of the game-changing trait. This is an important result for social niche construction – social niche modifying GCTs that support enhanced cooperation must be able to evolve even in conditions unfavourable to cooperative behaviours or social niche construction would be a mechanism that just accelerates the evolution of cooperation rather than enabling it where defection would otherwise be favoured. If we imagine that there is a separation of timescales where traits that change the game evolve more slowly than social behaviours then when there is assortment on GCTs, these GCTs can evolve to become more supportive to cooperation even when the current social dilemmas are dominated by defectors. This would establish the social conditions for cooperative traits to spread more easily when cooperation becomes more favourable.

The model could be extended in a number of ways, such by allowing for repeated interactions and hence iterated strategies, or examining a wider range of model parameters. The successful realisation of the simulation results in a mathematical model using interaction functions to change the replicator equation also opens up avenues for future work. In particular it is possible to more precisely model different levels of social and game-changing trait assortment using interaction functions. Because the assortment in the simulation model is generated by the gravitational attraction it is difficult to tune and potentially presents an issue in comparing the results across different scenarios. This work also demonstrates that interaction functions can be applied to empirical or simulation-derived data to model the results mathematically.

6.3 Modelling Assortment in Multi-Trait Models

Our simulation model showed that although assortment on game-changing traits can partially disrupt assortment on social behaviours, it increases the range of behaviours in which population structures that support increased cooperation can evolve. This supports our hypothesis that GCT assortment plays an essential role in the spread of pro-cooperative GCTs. We also found that we can model GCT assortment using interaction functions. This gives us a second candidate method for representing GCT assortment in a metagame to go along with encounter functions. In the rest of this chapter we compare these two methods to model assortment on a game-changing trait.

First, we take a step back and think general terms about how to represent assortment on single traits in a multiple-trait model. When there are mechanisms promoting assortment on multiple different traits at all the same time, the different forces encouraging

assortment on the different traits will interact. The results of this interaction could increase or decrease the amount of assortment on any single trait as we saw in the simulation model where assortment on the game-changing trait could break up highly assorted strategy traits. If a cooperator genotype is almost exclusively encountering other cooperators then increasing the assortment on the game-changing trait could reduce strategy-assortment by promoting interactions between cooperators and defectors possessing the same game-changing trait. In the extreme case, it is clear that a population cannot be fully assorted on two traits unless the two are coextensive. Our analysis of metagame models so far and the logical argument for social niche construction suggests that cooperator and assorter traits will indeed often become fully (or mostly) coextensive – but this is as an outcome of social niche construction creating linkage disequilibrium between the traits, not the initial conditions. Our modelling methodology needs to be sufficiently robust to handle a wide range of multiply assorted scenarios.

When modelling assortment on a single trait (social strategy) we have defined an assortment index α. When there is no assortment on this trait ($\alpha = 0$) interactions occur at the population frequency. Full assortment on this trait ($\alpha = 1$) means bearers of the assorted trait interact with each other all the time. For intermediate values, bearers interact with each other at frequency α; the rest of the time $(1 - \alpha)$ they play the population. This is the standard understanding of single trait assortment, going back to Hamilton (1964b), but it can result in the occasional perversity in interpretation – for example, in the case where there is full assortment but one type has extremely low frequency. We have seen that this definition can be represented as a transformation to the payoff matrix of a two-player game (Chapter 4).

In the multi-trait case, modelling assortment becomes more challenging. For the rest of this section we use $\beta \in [0, 1]$ for a secondary trait assortment index (namely, GCT assortment) equivalent to the single-trait assortment index α. If one has an assortment level of α for one trait, and of β for another, then it may not be mathematically possible to reconcile the two by simply extending the definition of assortment we described for α to β.

We simplify the problem of balancing the representation of strategy and GCT assortment in metagames by invoking the separation of timescales between the evolution of social frequencies and the evolution of game-changing traits. Social strategy assortment occurs on the faster timescale of the effective games, GCT assortment the slower timescale of metagame interactions that move the population through the GCT trait space. This principle suggests where in the modelling process we need to apply assortment on the game-changing trait: the metagame interaction. As we have done throughout this thesis, we represent strategy assortment by transformations to the payoff matrix of an individual game determining the *effective game* that is being played. GCT assortment then applies to the metagame interaction matrix composed of the (possibly transformed) effective games. We will see that depending on the scheme chosen for GCT assortment, it may

or may not be possible to represent it in turn as a change to the payoff matrix of the metagame interaction.

- Assortment on a strategy trait is a transformation to a particular evolutionary game.
- Assortment on a game-changing trait is a transformation of a metagame interaction.

Though we are focusing here on metagames where the game-changing trait is assortment, this modelling decision gives us flexibility: we can look at the level of assortment on a game-changing trait regardless of whether if the game-changing trait is itself assortment.

6.4 Modelling GCT assortment in Metagames

In Chapter 5 we introduced our definition of a metagame. We consider a population with two traits: a social strategy trait and a game-changing trait. There are two possible strategy types of strategy trait: cooperation (C) and defection (D). Though there may be many possible values for the game-changing trait, for any given metagame interaction we also consider two: one mutant (M) and one non-mutant (N). Of particular interest are mutant GCTs that create a higher level of assortment on the strategy trait. This gives us four types in the population: MC, MD, NC and ND.

The traits M and N correspond to effective games G_M and G_N, drawn from a set of possible games. Each metagame is defined by the set of possible games, and by an *encounter function* ϕ that determines the games played when individuals with the M and N GCTs play each other. The encounter function determines two games: $G_K = \phi(G_M, G_N)$ and $G_L = \phi(G_N, G_M)$ – though typically we consider a symmetric metagame in which $G_K = G_L$. Our general metagame interaction matrix then is composed of these four games:

$$\begin{pmatrix} G_M & G_K \\ G_K & G_N \end{pmatrix} = \begin{pmatrix} R_M & S_M & R_K & S_K \\ T_M & P_M & T_K & P_K \\ R_L & S_L & R_N & S_N \\ T_L & P_L & T_N & P_N \end{pmatrix} \tag{6.4}$$

Our separation of timescales gives us the modelling intuition that GCT assortment occurs on the level of these metagame interaction matrices. How might we transform this matrix to reflect the presence of assortment on the game-changing trait?

6.4.1 Scale the Component Games in the ST-plane

For most of this thesis, we have narrowed our focus not just to two-player symmetric games, but to games on the ST-plane where $R = 1$ and $P = 0$. Our default encounter function has been the average of the games G_M and G_N. This is just the point on the ST-plane between M and N.

There is an intuitive principle behind this. M bearers 'want' to play G_M and N bearers 'want' to play G_N, so when an M meets an M or an N meets an N this can happen without discord. But when an M meets an N, there must be a compromise – and the game between the two is the most obvious compromise. However, this will no longer be the case when we are looking to transform the game to reflect the fact that there is assortment on the GCT. The more assortment on the GCT, the less likely it will be that M bearers will play against N bearers and vice versa. When we mathematically modelled assortment on the strategy trait, we saw that reducing the frequency of a particular interaction was equivalent to playing in a well-mixed population a game in which the payoff for that interaction was scaled to a lower value (Chapter 4).

This would suggest one naive method to represent assortment on the game-changing trait: scale the component games in the ST-plane. When a social trait assortment of α meant that S and T were less likely payoffs, we scaled the value of these two coefficients by $(1 - \alpha)$. We could repeat this idea but this time for $(1 - \beta)$. Geometrically, we are taking the point that lies between M and N on the ST plane and then moving it towards the counterpoint of the equilibrium circle that lies in the plane. As we know from our analysis of circles of constant selection strength in ST-space (Section 5.4.1), this preserves the equilibria of the cross-GCT game but reduces the strength of selection. The metagame interaction matrix would then look like:

$$\begin{pmatrix} 1 & S_M & 1 & (1-\beta)(S_M+S_N) \\ T_M & 0 & (1-\beta)(T_M+T_N) & 0 \\ 1 & (1-\beta)(S_M+S_N) & 1 & S_N \\ (1-\beta)(T_M+T_N) & 0 & T_N & 0 \end{pmatrix} \tag{6.5}$$

This seems a tempting option. We are saying that assortment means that selection will be weaker in the cross-GCT game though we keep all of the component games in the ST-plane. On reflection, however, this is not actually a benefit: it is overly focused on the ST-plane. Though there are good reasons why we have restricted most of our modelling to the ST-plane, our mathematical methods have not been limited to it. Clearly this approach makes little sense if R_M and R_N are different.

Worse, this *ad hoc* approach is deficient even when G_M and G_N are games on the ST-plane because there is no attempt to consider the effect that assortment on the game-changing trait would have on the R_K and P_K terms. Just because the component games are in the ST-plane, there is no *a priori* reason why the cross-GCT encounter games would be too, especially after a transformation. We draw out this point because it shows that while leveraging the ST-plane to take advantage of our geometric intuitions is a powerful tool that we have used throughout this thesis, it is not applicable in all situations.

Instead, we have to consider the two methods we have already discussed which provide consistent and justifiable procedures to affect the metagame interaction matrix: the *encounter functions* that determine the cross-GCT interactions and *interaction functions*.

6.4.2 Modelling GCT assortment with Encounter Functions

The idea of cross-GCT interactions are built into the very idea of an encounter function since it is the encounter functions of a metagame that determine the effective game played when individuals with different GCT traits encounter each other. This makes them a natural candidate for a mechanism to represent GCT assortment. Indeed, the naive method we just dismissed was actually already using encounter functions – it was just doing so poorly.

First though, we need to ask why might we want to use encounter functions to model GCT assortment when we know that there is a powerful tool – interaction functions – that we have used successfully to derive a payoff matrix transformation for single-trait assortment.

There are two main reasons: one on principle and one practical. The principled reason is that choosing an encounter function is a definitional requirement when constructing a metagame. Therefore it would be more parsimonious if encounter functions were sufficient in themselves to model the effects of GCT assortment. The practical reason is that the way a game changes due to the use of interaction functions cannot always be represented as a transformation of that game's payoff matrix. Since a metagame's encounter function is by definition part of what determines the payoff matrix of interactions within that metagame, any method to represent GCT assortment using encounter functions will ensure that representation is ultimately in payoff matrix form.

So far we have almost exclusively used encounter functions that average the effective games corresponding to our GCT traits (or take the radial average when we considered games on the circle). There is a natural interpretation of this as a population that is well-mixed with respect to the game-changing trait. It does not take a substantial shift from our naive method to define a more rigorous way to use encounter functions. If we drop the requirement that all the component matrices of the metagame interaction

matrix lie in the ST-plane, then we can define an encounter function for a level of GCT assortment β as:

$$\phi^{\beta}(G_M, G_N) = (1-\beta)(G_M + G_N) \tag{6.6}$$

This produces a constant transformation that can be solved using the standard replicator dynamics on the metagame interaction payoff matrix:

$$\begin{pmatrix} G_M & (1-\beta)(G_M+G_N) \\ (1-\beta)(G_M+G_N) & G_N \end{pmatrix} \tag{6.7}$$

Using this method, no GCT assortment ($\beta = 0$) produces the default well-mixed case that we have studied so far in this thesis. Under full GCT assortment ($\beta = 1$), the encounter functions will produce zeroed matrices. This means that types will obtain no fitness increase or decrease from cross-GCT interactions – it will be as if there were no cross-GCT encounters at all, exactly what we would expect if the population is fully assorted by GCT.

This is not, however, the same as there being two distinct subpopulations, or no connection between the two different GCT trait-groups at all. Because the evolution of type frequencies in the population under the replicator equation is determined by the relative difference between the type's fitness and the mean population fitness, the two GCTs will still indirectly affect each other by their effects on the average fitness. For example, if both MC and MD types have much higher fitness than NC and ND types then even though the two types will not experience direct fitness consequences from interactions with each other, the NC and ND types will diminish in relative frequency because of the higher mean population fitness.

6.4.3 Modelling GCT assortment with Interaction Functions

The other method we consider is using interaction functions, which we saw earlier in this chapter could be used to mathematically model the results of a direct simulation model of increased assortment. To use this method, we need the original fitness functions:

$$f(x_{MC}) = x_{MC}R_M + x_{MD}S_M + x_{NC}R_K + x_{ND}S_K \tag{6.8}$$

$$f(x_{MD}) = x_{MC}T_M + x_{MD}P_M + x_{NC}T_K + x_{ND}P_K \tag{6.9}$$

$$f(x_{NC}) = x_{MC}R_L + x_{MD}S_L + x_{NC}R_N + x_{ND}S_N \tag{6.10}$$

$$f(x_{ND}) = x_{MC}T_L + x_{MD}P_L + x_{NC}T_N + x_{ND}P_N \tag{6.11}$$

Note that we still need to decide what encounter function the metagame will use in this method to determine the games G_K and G_L. We also need to determine the family of interaction functions that map the actual frequencies the types are present in the population to the actual rate a type encounters the other types. We do this by thinking about what assortment on a GCT means from first principles.

Take a GCT assortment level of β. Again, when $\beta = 1$ it means there is full assortment on the GCT trait, so individuals play bearers of the same trait with probability β, or with $(1 - \beta)$ play the population. However, as bearers of the GCT traits will also have one of the two types of social trait, C and D, we have to take this into account when rewriting the replicator equation. For instance, the type x_{MC} will play bearers of the M allele with probability β as a result of GCT assortment, meaning that as a result of GCT assortment it will play against other x_{MC} types with probability $\frac{x_{MC}}{x_{MC}+x_{MD}}\beta$ and against x_{MD} types with probability $\frac{x_{MD}}{x_{MC}+x_{MD}}\beta$. The remaining $(1-\beta)$ of the time x_{MC} will play the population.

This gives our interaction functions, and lets us write the modified fitness functions ${f^\beta}_i = f_i \circ {e^\beta}_i(x))$:

$$f^\beta(x_{MC}) = \beta\Big(\frac{x_{MC}}{x_{MC}+x_{MD}}R_M + \frac{x_{MD}}{x_{MC}+x_{MD}}S_M\Big) + (1-\beta)f(x_{MC}) \tag{6.12}$$

$$f^\beta(x_{MD}) = \beta\Big(\frac{x_{MC}}{x_{MC}+x_{MD}}T_M + \frac{x_{MD}}{x_{MC}+x_{MD}}P_M\Big) + (1-\beta)f(x_{MD}) \tag{6.13}$$

$$f^\beta(x_{NC}) = \beta\Big(\frac{x_{NC}}{x_{NC}+x_{ND}}R_N + \frac{x_{ND}}{x_{NC}+x_{ND}}S_N\Big) + (1-\beta)f(x_{NC}) \tag{6.14}$$

$$f^\beta(x_{ND}) = \beta\Big(\frac{x_{NC}}{x_{NC}+x_{ND}}T_N + \frac{x_{ND}}{x_{NC}+x_{ND}}P_N\Big) + (1-\beta)f(x_{ND}) \tag{6.15}$$

Rewriting $f^\beta(x_{MC})$ in the same form as the original fitness functions we see it is:

$$f^\beta(x_{MC}) = \Big(\frac{\beta}{x_{MC}+x_{MD}} + (1-\beta)\Big)x_{MC}R_M \tag{6.16}$$

$$+ \Big(\frac{\beta}{x_{MC}+x_{MD}} + (1-\beta)\Big)x_{MD}S_M \tag{6.17}$$

$$+ (1-\beta)x_{NC}R_K \tag{6.18}$$

$$+ (1-\beta)x_{ND}S_K \tag{6.19}$$

The other modified fitness functions are similar. When $\beta = 0$, the fitness functions are the same as the originals ($f^0 = f$). When $\beta = 1$ we have:

$$f^1(x_{MC}) = \frac{1}{x_{MC} + x_{MD}}(x_{MC}R_M + x_{MD}S_M) \tag{6.20}$$

$$f^1(x_{MD}) = \frac{1}{x_{MC} + x_{MD}}(x_{MD}T_M + x_{MD}P_M) \tag{6.21}$$

$$f^1(x_{NC}) = \frac{1}{x_{NC} + x_{ND}}(x_{NC}R_N + x_{ND}S_N) \tag{6.22}$$

$$f^1(x_{ND}) = \frac{1}{x_{NC} + x_{ND}}(x_{ND}T_N + x_{ND}P_N) \tag{6.23}$$

This differs from the encounter functions method in the presence of the fractional terms. The equivalent fitness function for x_{MC} using encounter functions is $f^1(x_{MC}) = x_{MC}R_M + x_{MD}S_M$.

We can simplify how we write these equations by introducing a *GCT-scaling function*. Let $x_n = x_{NC} + x_{ND}$ and $x_m = x_{MC} + x_{MD}$ be the frequencies of the two GCT types. Then we define the GCT-scaling function as:

$$s_\beta(x) = \frac{\beta}{x} + (1 - \beta) \tag{6.24}$$

Of course this function is *not* a constant, but a shorthand — any transformations that depend on some value of $s_\beta(x)$ will only be constant if x is. Also, as $x_n + x_m = 1$, we have $s_\beta(x_m) = s_\beta(1 - x_n)$ and vice versa.

Rewriting the equations to include the scaling function, we can simplify the metagame interaction matrix to be:

$$\begin{pmatrix} s_\beta(x_m)G_M & (1-\beta)G_K \\ (1-\beta)G_K & s_\beta(x_n)G_N \end{pmatrix} \tag{6.25}$$

Interestingly, the interaction functions method which proceeds from first principles produces a very similar result to the encounter functions method. The only difference is the presence of the scaling function that modifies the payoffs for same-GCT interactions. Because the scaling function is a dynamic product of the population state, the metagame interaction matrix does not represent a fixed game but a dynamic game dependent on the population state. This means we cannot use all the techniques for finding Nash equilibria that apply to static game matrices and have to use the replicator equation. This has no bearing on this thesis which would use the replicator equation method regardless, but is a substantive difference between the two methods.

The scaling function takes a single argument x that is the combined frequency of the M or N alleles, so $0 \leq x \leq 1$. At the extreme values of x, $s_\beta(1) = 1$ while $s_\beta(0)$ is undefined. This is a limitation of the method, but if all the possible types are present in the initial

state of the population at non-zero frequency it will not matter, since we know that under the replicator dynamics a type that is initially present will never go extinct (though it may asymptotically approach 0 and hence can only ever asymptotically approach 1) (Weibull, 1997).

If the M and N alleles are present at non-zero frequency at the start, $0 < x < 1$ is a strict equality, and we have:

$$0 < x < 1 \implies \frac{1}{x} > 1 \implies \frac{\beta}{x} > \beta \implies \frac{\beta}{x} - \beta + 1 > 1 \tag{6.26}$$

So the scaling function always increases the relative payoff for same-GCT interactions. We also have for any $x_1, x_2 \in [0, 1]$ that $x_1 < x_2 \implies s_\beta(x_1) > s_\beta(x_2)$. This means that as the fraction of the population bearing a particular GCT shrinks, the scaling function term gets larger to represent the greater number of within-GCT interactions still occurring between those bearers than in the case of a well-mixed population.

The behaviour of the scaling function when one GCT is small does lead to a practical difficulty using interaction functions, because whenever one GCT trait asymptotically tends to 0 the output of the scaling function becomes very large. This source of numerical instability forces us to take great care when using numerical methods to evaluate the replicator equation.

6.5 Method Comparison

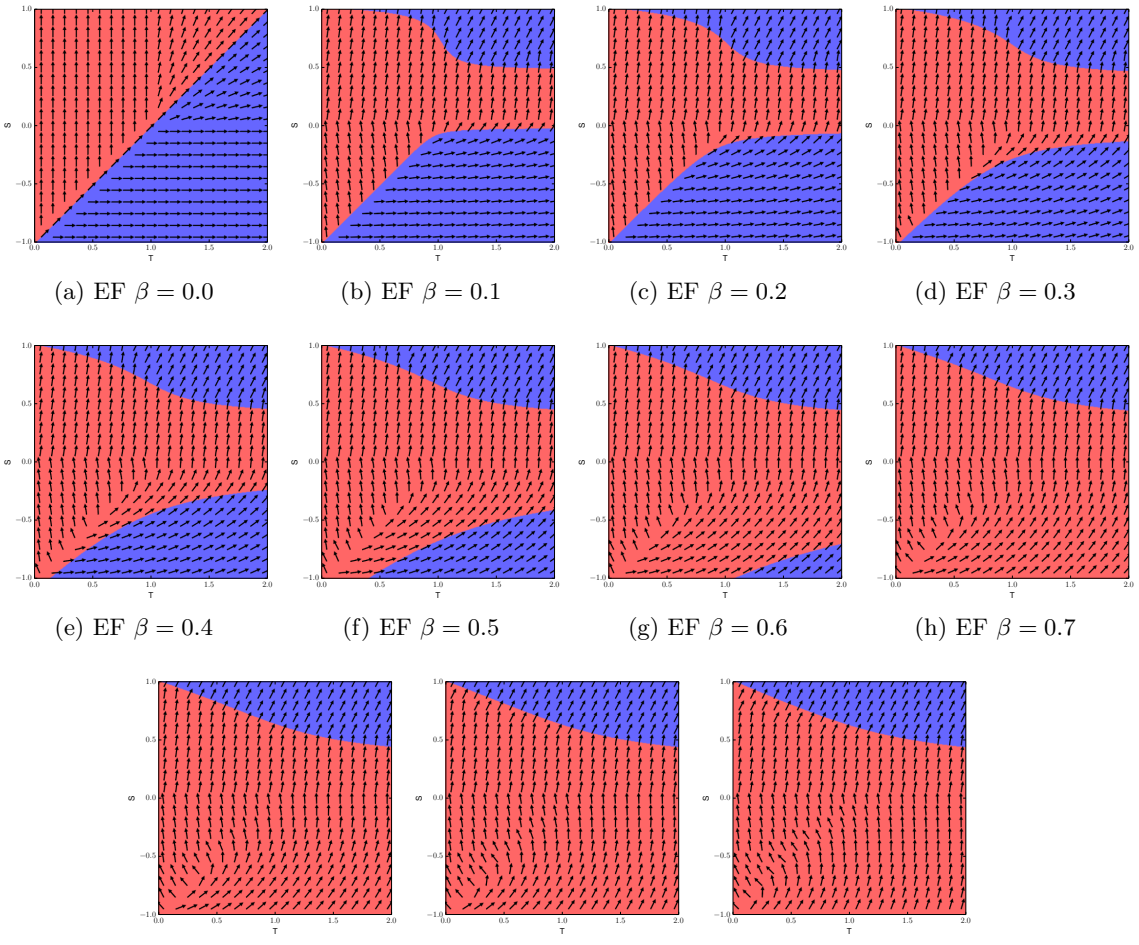
We have seen that the two methods produce metagame interaction matrices that are similar apart from the scaling function. Now we compare the results obtained using these two methods on two of the most important models used in this thesis.

The first is a vector field over the whole of ST-space (as in Figure 5.6). We take a lattice of points over the plane and a small circle of points around each point on the lattice. Each point on the lattice is a game G_N, each point on the circle around that lattice is a G_M. We evaluate a metagame interaction between the centre point and each point around the little circle. We then multiply the change in frequency of the M allele by the vector between the centre and that point M on the circle. The weighted sum of these vectors for each M around the little circle gives a mean game-change vector for each N on the lattice, producing a pseudo-vector field over the whole space. For a complete explanation of this model see Section 5.4.4. Note that in all the diagrams we picture here, we only show the direction of the resulting vectors, not the magnitude.

The second model is the assortment metagame (Section 5.5). Again we take a lattice of points on the ST-plane. We then perform a metagame interaction between this original game G_N and a mutant game assorted by some δ_α, $G_M = {G_N}^{\delta_\alpha}$. Here we are using

$\delta_\alpha = 0.01$. We visualise the results of this over ST-space by colouring the region in which the more-assorted M trait increases in frequency red, and the region where it decreases in frequency blue. Note that for the purposes of all these comparisons, we are taking the balanced initial population state where $c = 0.5$ and $m = 0.5$ (so all four types start at frequency 0.25).

We superimpose the vector field on the results of the assortment metagame for the results derived using encounter functions (Figure 6.6) and interaction functions (Figure 6.7) to compare the behaviours as GCT assortment (β) increases.



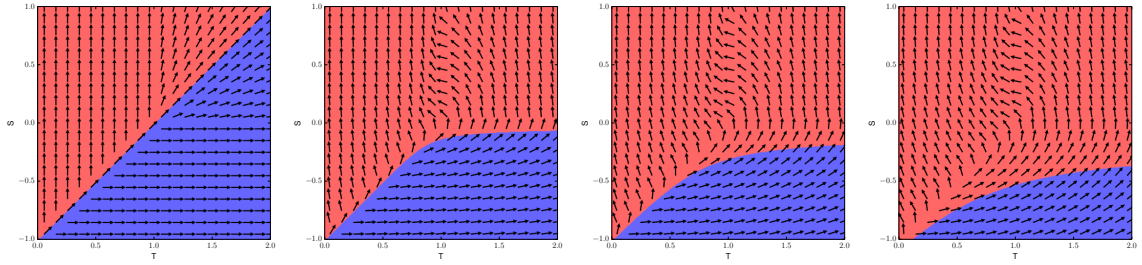
(a) EF $\beta = 0.0$ (b) EF $\beta = 0.1$ (c) EF $\beta = 0.2$ (d) EF $\beta = 0.3$

(e) EF $\beta = 0.4$ (f) EF $\beta = 0.5$ (g) EF $\beta = 0.6$ (h) EF $\beta = 0.7$

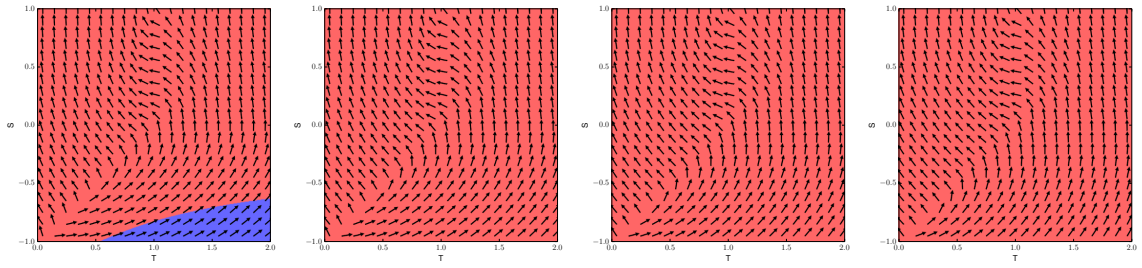
(i) EF $\beta = 0.8$ (j) EF $\beta = 0.9$ (k) EF $\beta = 1.0$

Figure 6.6: Encounter function calculated vector field superimposed on the assortment metagame (red means assortment increases, blue assortment decreases) for the ST-plane for increasing GCT assortment (β) (Initial conditions $c = 0.5$, $m = 0.5$).

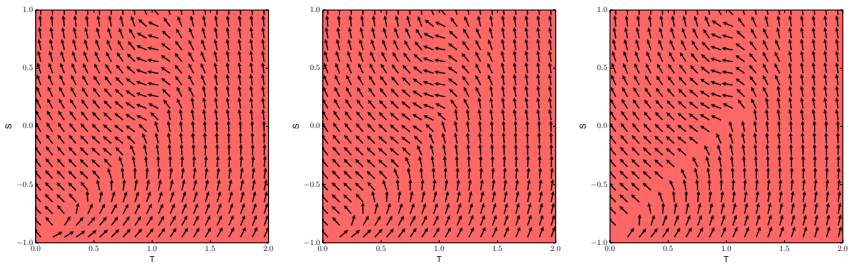
When there is no GCT assortment ($\beta = 0$) we see the behaviours discussed in Chapter 5, namely that the direction of selection pressure in the unconstrained metagame is in the direction of increasing S in the Harmony Game region, increasing T in the Prisoner's Dilemma, either S or T in the Stag Hunt depending on the initial conditions, and a combination of S and T according to their equilibrium frequency in the Snowdrift



(a) IF $\beta = 0.0$ (b) IF $\beta = 0.1$ (c) IF $\beta = 0.2$ (d) IF $\beta = 0.3$



(e) IF $\beta = 0.4$ (f) IF $\beta = 0.5$ (g) IF $\beta = 0.6$ (h) IF $\beta = 0.7$



(i) IF $\beta = 0.8$ (j) IF $\beta = 0.9$ (k) IF $\beta = 1.0$

Figure 6.7: Interaction function calculated vector field superimposed on the assortment metagame (red means assortment increases, blue assortment decreases) for the ST-plane for increasing GCT assortment (β) (Initial conditions $c = 0.5$, $m = 0.5$).

region. The assortment metagame model displays similar behaviour, with assortment increasing when $S > T + 1$ and decreasing otherwise – though this would change if the initial frequencies were different.

In this chapter we are primarily concerned with comparing the two methods of modelling GCT assortment so we will postpone a full analysis of the effects of GCT assortment to Chapter 8, but the overall results are clear. Using both methods, when there is GCT assortment the vectors show a selective pressure in favour of increasing S – good for cooperators as they will gain more (or lose less) payoff in games played against defectors. GCT assortment also greatly increases the region of ST-space over which a game-changing trait for increased assortment increases in frequency.

The methods produce similar results over the Prisoner's Dilemma quadrant, though the interaction functions method produces results more favourable to cooperation: it takes

a lower amount of GCT assortment for a point in the Prisoner's Dilemma to go from one where assortment decreases to assortment increasing. Indeed, using the interaction function method assortment increases over the whole of ST-space for GCT assortment above $\beta = 0.5$. This is almost the case using the encounter functions method above $\beta = 0.7$.

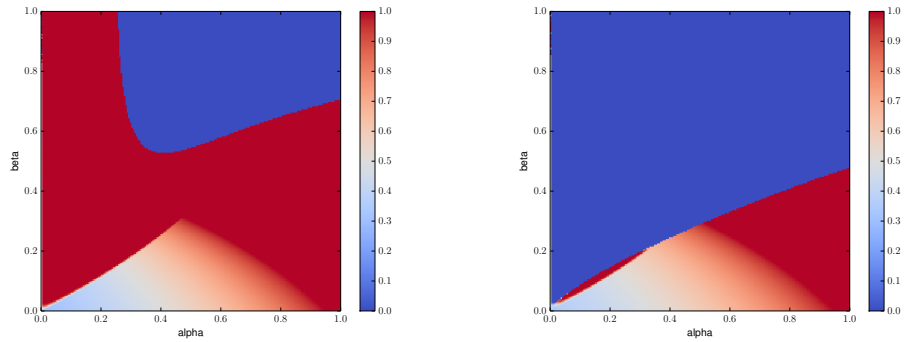
6.5.1 The Region where Assortment Never Increases under the Encounter Functions Method

The big difference between the two methods is the high-S region of the Harmony Game and Snowdrift quadrants, where assortment never increases when modelling increased GCT assortment using encounter functions. Comparing the vector fields in these quadrants, we see that the direction of game-change favours increased S across both methods. But when using interaction functions, higher levels of GCT-assortment result in a decrease in T being favoured, further benefiting cooperators. Using the encounter function method, the vectors instead change to show that there is a selective pressure in favour of increased T, which benefits defectors, even in the regions where assortment does still increase.

We look into this further by considering points in the Snowdrift quadrant. Fixing $T = 1.9$, we look at the points $S = 0.4$ and $S = 0.5$ (so these points are in the mid-right of the Snowdrift quadrant). Letting the GCT be social trait assortment, we examine the effect of varying social trait (α) and GCT assortment (β) at these points and record the equilibrium frequency of the assorting type M. This means that G_N is the game at the point $T = 1.9$, $S = 0.4$ or 0.5 and $G_M = {G_N}^{\alpha}$, with the metagame interaction performed between the two with the encounter function using a GCT assortment level of β.

In our previous assortment metagame models we have only looked at a fixed change in social trait assortment of $\delta_\alpha = 0.01$, but here we let α vary from 0 (where $G_M = G_N$) to 1 (where G_M is the fully assorted game). Figure 6.8 shows the $\alpha - \beta$ plots for these two points in the Snowdrift quadrant, with the equilibrium frequency of the increased assortment type M shown on a blue (no-assorters) to red (all-assorters) scale.

We see that in the Snowdrift region, increasing GCT assortment changes the metagame. With no GCT assortment ($\beta = 0$), the metagame reduces to an equilibrium with the two types MC and ND present, as we discussed in Section 5.5. As GCT assortment increases, we see two additional attractors for the metagame interaction: one in which there are all assorters (possessing the M trait), shown in red, and one in which there are none of the increased assorters (the blue region). We essentially have three states: MC-ND, all-M and all-N, with the metagame reducing to one of the games $\phi(G_M, G_N)$, G_M or G_N.



(a) $S = 0.4$, $T = 1.9$ (b) $S = 0.5$, $T = 1.9$

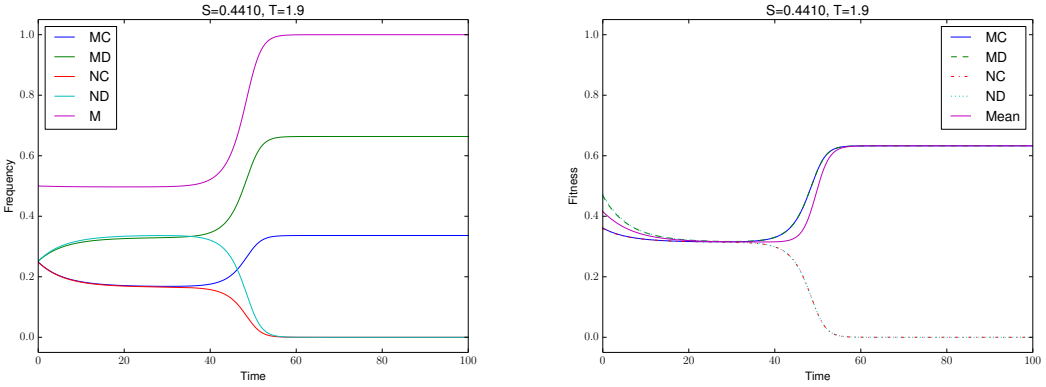
Figure 6.8: The equilibrium frequency of the assorting trait for two games in the Snowdrift quadrant as social trait assortment (α) and GCT assortment (β) vary. The parameter space is split into three different states: all-assorters (red), no-assorters (blue) and mixed (intermediate colours).

The shift between the two games, both in the Snowdrift region spaced only 0.1 units apart on the S-axis demonstrates that the basin of attraction for the no-assorter state has expanded greatly. In both figures, the mixed-equilibrium region is almost the same, but when $S = 0.5$ the basin of attraction of no-assorters expands so significantly it cuts into this region. In general, the non-assorters are benefitting from when there is higher-levels of GCT assortment under the encounter functions modelling method.

The switch between the metagame basins is sudden and dramatic. Taking the case of full GCT assortment ($\beta = 1$), but limiting the competition in the metagame interaction to game-changing traits corresponding to only a small increase in social trait assortment ($\delta_\alpha = 0.01$) as we have been doing in the assortment metagame model, then for $T = 1.9$ the boundary between the all-assorter and no-assorter regions lies between $S = 0.4410$ and $S = 0.4411$.

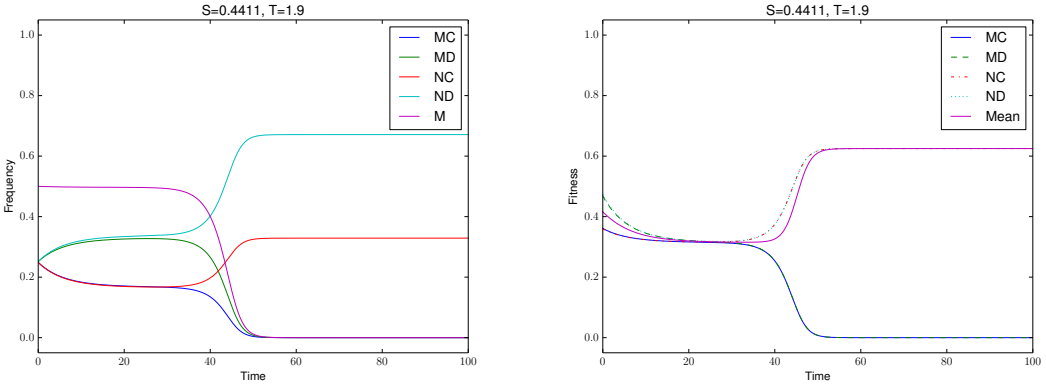
Figure 6.9 shows the evolution of the type frequencies (left) and fitnesses (right) under the replicator dynamics in the assortment metagame ($\delta_\alpha = 0.01$) for the game G_N with $S = 0.4410$, $T = 1.9$ modelled using encounter functions (balanced initial conditions $c = 0.5$, $m = 0.5$). Here the game reduces to $G_M = {G_N}^{0.01}$. We see that MC and NC both decrease in frequency while MD and ND increase in frequency, with ND increasing the most. But then ND and NC collapse while MC and MD increase to equilibrium frequency. The population state at equilibrium has $x_{MC} = 0.336$ and $x_{MD} = 0.664$ with the other two asymptotically close to 0. These equilibrium frequencies of the assorting types are asymptotically close to the ESS frequencies obtained for G_M.

By contrast, Figure 6.10 shows the evolution of the type frequencies (left) and fitnesses (right) for G_N with $S = 0.4411$, $T = 1.9$. This tiny increment of 0.0001 in S is enough for the game to instead reduce to G_N. The type frequencies behave the same for the



(a) Frequencies (b) Fitnesses

Figure 6.9: The evolution of the type frequencies (left) and fitnesses (right) of the assortment metagame ($\delta_\alpha = 0.01$) at the point $S = 0.4410$, $T = 1.9$ modelled using encounter functions (balanced initial conditions $c = 0.5$, $m = 0.5$)



(a) Frequencies (b) Fitnesses

Figure 6.10: The evolution of the type frequencies (left) and fitnesses (right) of the assortment metagame ($\delta_\alpha = 0.01$) at the point $S = 0.4411$, $T = 1.9$ modelled using encounter functions (balanced initial conditions $c = 0.5$, $m = 0.5$)

initial timesteps, but then the N types dominate and the N types collapse. The population state at equilibrium has $x_{NC} = 0.329$ and $x_{ND} = 0.671$. So even though $S = 0.4411$ means this game should be more favourable to cooperators, the outcome of the assortment metagame is less favourable to cooperators.

The same behaviour occurs in the Harmony Game quadrant, as an $\alpha - \beta$ plot from the game at $S = 0.75$, $T = 0.75$) shows (Figure 6.11), though here there is no polymorphic equilibrium as all the games are in the Harmony Game quadrant. We see that higher GCT assortment (β) means the metagame interaction is more likely to be in the no-assorters basin, while higher social trait assortment (α) means the metagame is more likely to be in the all-assorters basin.

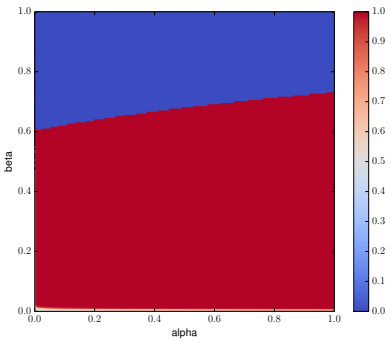


Figure 6.11: The equilibrium frequency of the assorting trait for the Harmony Game at $S = 0.75$, $T = 0.75$ as social trait assortment (α) and GCT assortment (β) vary. The parameter space divides into two regions all-assorters (red) and no-assorters (blue).

As we have seen, which basin of attraction the metagame interaction falls in is complex, with a small change transitioning between the two regions though it does not alter the rank ordering of the payoffs, the mean fitness of the game or other obvious criteria. In general, higher S and higher GCT assortment means a metagame interaction is more likely to be in the no-assorters basin, while higher social trait assortment means it is more likely to fall in the all-assorters basin. This is mathematically interesting, but the vital question is whether this is an actual property we expect to see in reality or an artefact of the modelling process.

6.6 Conclusions

Modelling GCT assortment with interaction functions consistently produces results more favourable to the spread of assortment, and does not result in the surprising behaviour in the high-S regions of the Harmony and Snowdrift Games.

On the other hand, modelling GCT assortment with encounter functions has the advantage of parsimony: GCT assortment can be represented by the encounter function alone. Modelling with interaction functions still requires us to pick an encounter function to define the metagame, which we have done by taking the 'well-mixed' encounter function and then applying our interaction functions to this matrix.

However, it could be argued that this is only apparent parsimony, since our general encounter function $\phi^{\beta}(G_N, G_M) = (1 - \beta)(\lambda G_N + \mu G_M)$ still requires two separate choices - the choice of β, and the separate choice of λ and μ. Wrapping both into one function only seems more parsimonious — but there are still two different duties being performed by the encounter function.

Seen this way, we can argue that interaction functions are preferable because they make these two different tasks distinct. While encounter functions have the unambiguous practical advantage that they result in a static transformation of the matrix, interactions functions are a mathematical method based on clear assumptions that we developed and validated for the case of social trait assortment (Chapter 4). We have also seen the utility of interaction functions in how they can replicate the results of data obtained through a simulation model.

This gives us the confidence in interaction functions to believe that the results of the encounter functions method, which approximates them, is indeed producing aberrant results in the high-S regions of the Harmony and Snowdrift games. Interaction functions are more rigorous and at the same time provide continuous results. They are a more powerful tool that we have used in other situations, and for the rest of this thesis are the method we will use to model GCT assortment as we continue to investigate the evolution of game-changing traits.

Chapter 7

The Coevolution of Assortment on Social and Game-Changing Traits

We now have a method to model assortment on game-changing traits in a metagame. In this chapter we extensively model the effects of game-changing trait assortment when that game-changing trait is one that creates assortment on the social trait. This means we have 'first-order' assortment on social traits governed by the game-changing traits, and 'second-order' assortment on the game-changing traits themselves. We might think that assortment on a game-changing trait will just amplify the effects of assortment on strategy. However, we know it is more complex than this. As we found in the results of the simulation model (Chapter 6), assortment on game-changing traits can actually reduce strategy assortment by inducing more interactions between individuals with different strategy traits but the same game-changing trait.

This forces us to analyse many different variables at different points in ST-space. This is a high dimensional-space: the initial S, T, frequency of cooperators c, frequency of game-changing trait mutants m, step change in social trait assortment δ_α and level of game-changing trait assortment β can all vary at every point, so it is at least a six-dimensional space. Our goal in this chapter is to identify the key relationships between these parameters, constructing a series of models to identify and investigate causally significant cross-sections of this space, and attempt to develop summary measures to present this data in an intelligible way.

This chapter serves as a stepping stone where we examine a lot of different scenarios but focus on understanding the behaviour of the model. This lays the groundwork for Chapter 8 where we apply our understanding to explain the biological implications of these mathematical models. How does assortment on game-changing traits link with assortment on social strategy traits? Does it merely amplify the existing power of the

assortment mechanism or is there a more complex interaction between the two? What then are the implications of GCT assortment for the evolution of assortment and the spread of cooperation?

7.1 The Continuous Increase in Social Assortment Model

The *Continuous Model* examines whether greater assortment (and, ultimately, increased cooperation) can evolve through the evolution of game-changing traits corresponding to a series of small increases in the level of assortment. This is the same setup we have used in our previous analysis of the assortment metagame (see Section 5.5 and Figure 6.7).

Recall that on the ST-plane, there is a single unique fully assorted game $G^{assort} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. No matter what social dilemma the game G represents, G^{assort} is a Harmony Game. Any game G played with an assortment level of α is equivalent to a game G^{α} played in a freely mixed population:

$$G^{\alpha} = (1-\alpha)G + \alpha G^{assort} = \begin{pmatrix} R & S + \alpha(R-S) \\ T + \alpha(P-T) & P \end{pmatrix} \tag{7.1}$$

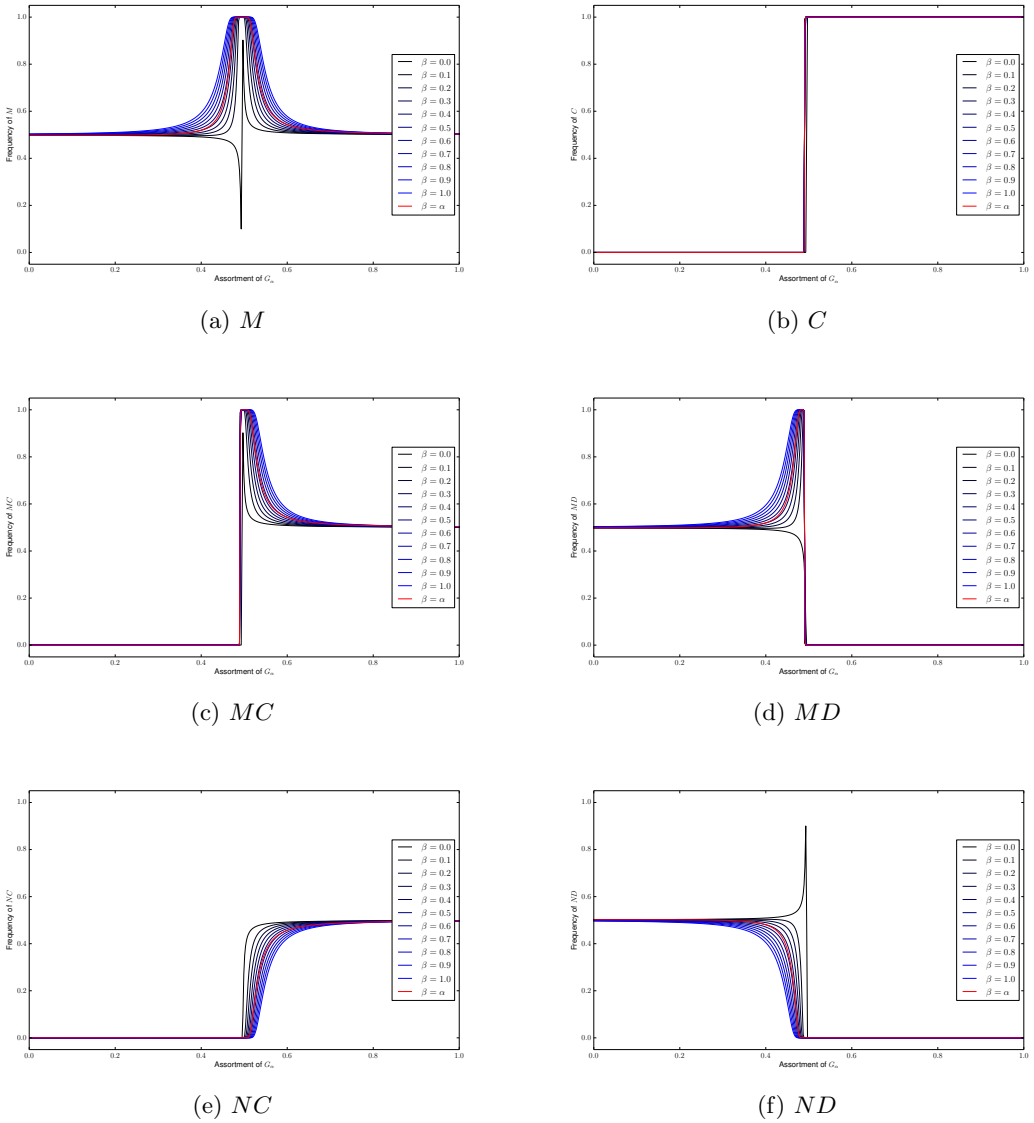
We take a point on the ST-plane corresponding to a game G. Increasing the assortment of this game creates an *assortment line* – a line segment in the ST-plane from the point G to the fully assorted game. The set of games on this assortment line, $\Gamma = \{G^{\alpha} : \alpha \in [0,1]\}$, creates a metagame with encounter function $\phi(G^{\alpha_1}, G^{\alpha_2}) = G^{\frac{\alpha_1+\alpha_2}{2}}$.

We perform a series of metagame interactions between the effective game $G_N = G^{\alpha}$ and a slightly more assorted effective game $G_M = G^{\alpha+\delta_{\alpha}}$ for some small increment d_{α}. Here we are going to take $\delta_{\alpha} = 0.01$. The initial game G played in a population where the game-changing traits are causing respective assortment levels of α and $\alpha + \delta_{\alpha}$ is equivalent to G_N and G_M played in a well-mixed population.

We record the results of the continuous model as α increases along the assortment line for metagame interactions transformed by different levels of GCT assortment (β), modelled by transforming the metagame interaction payoff matrix using interaction functions as described in Section 6.4.3. We also consider the situation where the level of GCT assortment is linked to the level of assortment on the social trait, by performing metagame interactions with β set to the average of α and $\alpha + \delta_{\alpha}$. Though the equivalency is not direct, we can imagine this equating to a scenario where a mechanism like relatedness through common descent links the levels of assortment on multiple traits.

7.1.1 The Prisoner's Dilemma

We look at the results of the continuous model starting at the game in ST-space where we imagine it would be most difficult for cooperation or increased assortment to evolve in. This is the game in the bottom right corner of ST-space as we depict it, the Prisoner's Dilemma game with $S = -1$, $T = 2$. We denote this game PD. For this game, if $\alpha < 0.5$, the game PD^{α} is a Prisoner's Dilemma, if $\alpha > 0.5$ then PD^{α} is a Harmony Game.



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.1: The frequency of the alleles M and C and the different genotypes in the Continuous Model starting from the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) for increasing levels of GCT assortment (β) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

Figure 7.1 shows the results of the continuous model for the frequencies of all the different types in the population, and the combined frequencies of the M assorting trait and C cooperative trait. Note that these graphs do not plot the behaviour of the metagame over time. Rather they plot the equilibrium frequency of the relevant type or trait for a series of metagames as the assortment index α of $G_N = G^\alpha$ ranges from 0 to 1.

Each figure plots the results for a series of lines corresponding to increasing levels of GCT assortment (β), from no GCT assortment (black line) to full GCT assortment (blue line). We also plot the case where the level of GCT assortment varies, with $\beta = \frac{2\alpha+\delta_\alpha}{2}$ (red line).

When there is no GCT assortment, this figures demonstrate the same results as the assortment metagame in Section 5.5. The assortment line from PD runs across a diagonal of ST-space, from the point PD in the bottom right to the fully assorted game in the top left. Where both G^α and $G^{\alpha+\delta_\alpha}$ are Prisoner's Dilemma games (below $\alpha = 0.5$), the frequency of the mutant type decreases and the equilibrium frequency of cooperators is zero; if $\alpha > 0.5$ then both effective games are Harmony Games, assortment increases and the equilibrium frequency of cooperators is one.

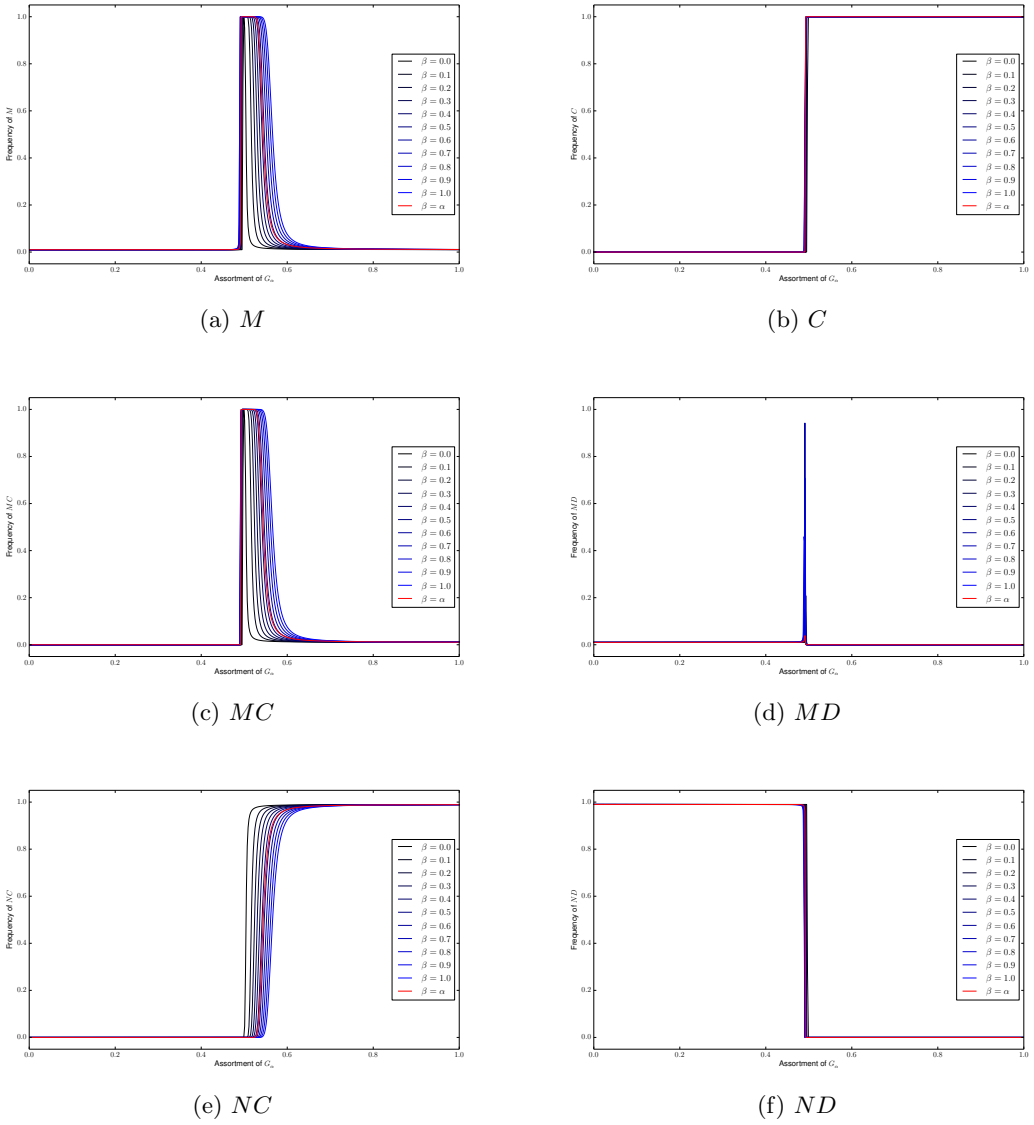
We see that increasing GCT assortment has little effect on the equilibrium frequency of cooperators, which still undergoes a sudden shift between no-cooperators and all-cooperators outcomes. However, it does affect the evolution of the more assorting game-changing trait M. Increased GCT assortment is beneficial for the spread of M, though the absolute change is small since the two games G^α and $G^{\alpha+\delta_\alpha}$ are usually both either Prisoner's Dilemmas or Harmony Games. As we found in Chapter 5, when the metagame interaction reduces down to a single social trait, the bearers of that trait are selectively neutral. And indeed, the least change in absolute frequency of the M type comes when the two games are both extreme Prisoner's Dilemmas or extreme Harmony Games and so reduce to a single strategy trait more quickly (as this assortment line passes through the origin of the circle of constant selection strength on the ST-plane, we can imagine it a line of fixed equilibria but varying radius – so varying the intensity of selection).

When the intensity of selection is lower, there is more opportunity for differentiation between the two bearers of the social trait that ultimately survives during the period in which all four types are present. As GCT assortment increases, this increasingly benefits the MD trait when the games are Prisoner's Dilemmas and the MC trait when the games are Harmony Games.

7.1.2 Low Frequency Invasion Scenario

As well investigating what happens when the initial population composition is balanced to understand the behaviour of the metagame all things being equal, it is important to look at what happens when the more assortive mutant type is introduced at a low

frequency, for instance as the result of a mutation. Figure 7.2 displays the results of the continuous model when the initial conditions are changed to $c = 0.01$ and $m = 0.01$.



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.2: The frequency of the alleles M and C and the different genotypes in the Continuous Model starting from the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) for increasing levels of GCT assortment (β) under low frequency invasion initial conditions ($c = 0.01$, $m = 0.01$).

Comparing the results, we see that the outcome is the same for the cooperative trait: GCT assortment has little effect on the equilibrium frequency of cooperators, which starts at zero, but increases to one when in the Harmony Game, even though cooperators only started as 1% of the population.

The general results are the same in that the M trait changes very little from the initial frequency when the local games are more extreme Harmony Games or Prisoner's Dilemmas, though the region where this is not the case is smaller and more sharply defined.

We also see different results on either side of $\alpha = 0.5$ divide. This is because we are keeping the initial frequencies constant across this assortment line – so the model has gone from one in which there is a very small difference between the initial frequency of 0.01 cooperators to the equilibrium frequency of no-cooperators to where there is a significant difference between the initial frequency of 0.01 and the equilibrium frequency of all-cooperators.

7.1.3 Results

These results are sections through a three dimensional surface, the different lines equalling cross-sections with fixed β (and the 'intrinsic line' a diagonal section where $\alpha = \beta$). Figures 7.3 and 7.4 plot the full surface with increased levels of assortment on the game-changing trait plotted against the social trait assortment of the effective game G_N and the equilibrium frequency of the M and C traits.

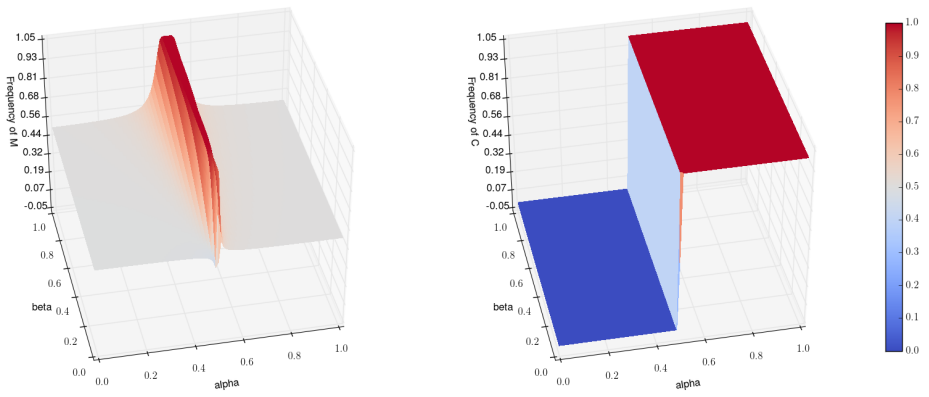


Figure 7.3: α-β plots showing the frequency of the alleles M and C in the Continuous Model starting from the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

In the low frequency invasion model, the equilibrium frequency of the more assorting type M does not reach a high absolute level. But what we are interested in is whether or not it increases in frequency from the initial level of 0.01, as these increases would make further increases incrementally more successful in the future. The same is true for the balanced initial conditions model, where the equilibrium frequency of the M trait is also close to the initial frequency when the game is on the extremes.

We can see this a 'critical value': the level of GCT assortment β above which the equilibrium frequency of the more assorted trait M will increase from the initial conditions. Using an interval bisection method, we find that a GCT assortment level of $\beta > 0.497487$ is required for assortment to increase above the initial frequency when G_N is the extreme Prisoner's Dilemma and $G_M = {G_N}^{\delta_\alpha}$ where $\delta_\alpha = 0.01$. At this particular point in the space, this critical value of β is almost identical across both sets of initial conditions (in

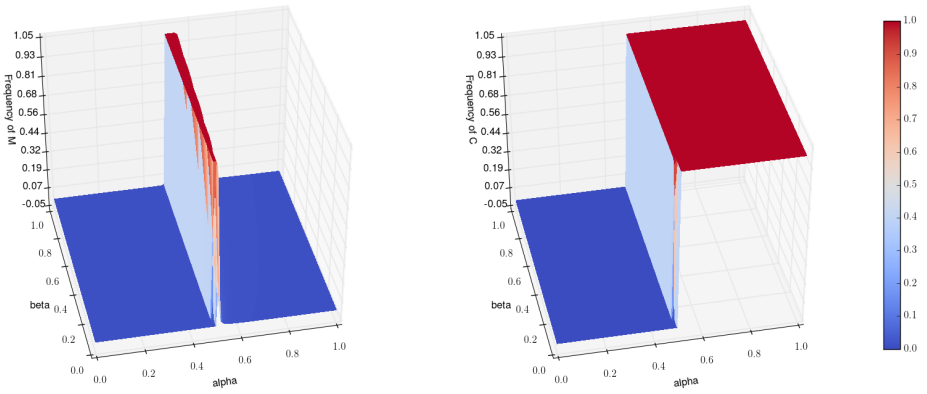


Figure 7.4: α-β plots showing the frequency of the alleles M and C in the Continuous Model starting from the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under low frequency invasion initial conditions ($c = 0.01$, $m = 0.01$).

full: 0.4974874371859137 and 0.4974874371832101, calculated with a search depth of 50 bisections).

This critical value gives us one summary statistic to measure over the entirely of ST-space, which we shall return to at the end of this chapter.

7.2 The Discrete Jump in Social Assortment Model

The continuous model represents the case where a game-changing trait causes a small, continuous shift in the level of social trait assortment – in our model, a small increase of $\delta_\alpha = 0.01$. It is not an unreasonable modelling assumption that mutations in assortment promoting game-changing traits will mean small changes in the level of assortment created on the social trait. However, it is also not the only possible relationship between the two. A game-changing could instead create a larger discrete jump in the level of social trait assortment.

In the discrete model we relax the assumption that the change in social trait assortment has to be small. This lets us ask how much social trait assortment is required to change the social outcome of a game, and how is this affected by second-order assortment on the game-changing trait. We can look more broadly at the parameter space mapped out by the step change in social trait assortment (α) and the level of game-changing trait assortment (β). We used the discrete model before, in Section 6.5.1 when we analysed the behaviour of the metagame in the Snowdrift region modelled using encounter functions.

As in the continuous model, we pick a game G_N. We then perform a series of metagame interactions between G_N and the game $G_M = {G_N}^\alpha$, as α ranges from 0 to 1. Essentially, performing a metagame interaction between the starting game, and a second game that

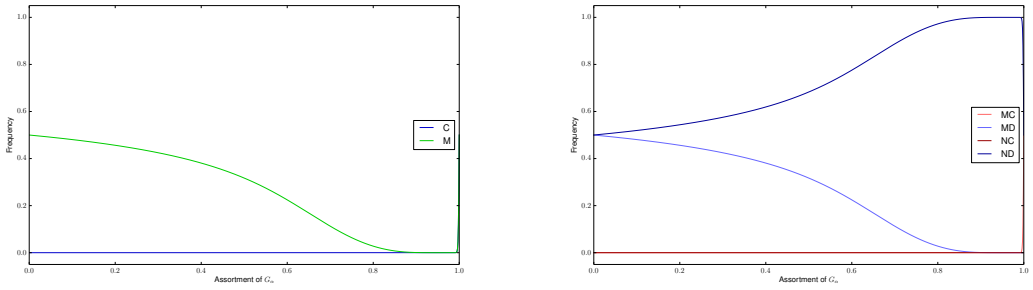
ranges along the assortment line segment, instead of two nearby games running along that line.

We easily can translate questions in the continuous model to those in the discrete model. If we are doing a metagame interaction between two points G^{α} and $G^{\alpha+\delta_\alpha}$ in the continuous model, it is the same as looking at the discrete model between $G' = G_\alpha$, ${\alpha_1}' = 0$ and $G^{\alpha'} = \frac{\delta_\alpha}{1-\alpha}$.

7.2.1 The Prisoner's Dilemma

7.2.1.1 No Assortment on the Game-Changing Trait

The first point we model is again the most extreme Prisoner's Dilemma in ST-space, the game with $S = -1$, $T = 2$, which we expect to be the most difficult point in ST-space for cooperation to evolve in. We start by plotting the discrete model when there is no assortment on the game-changing trait.



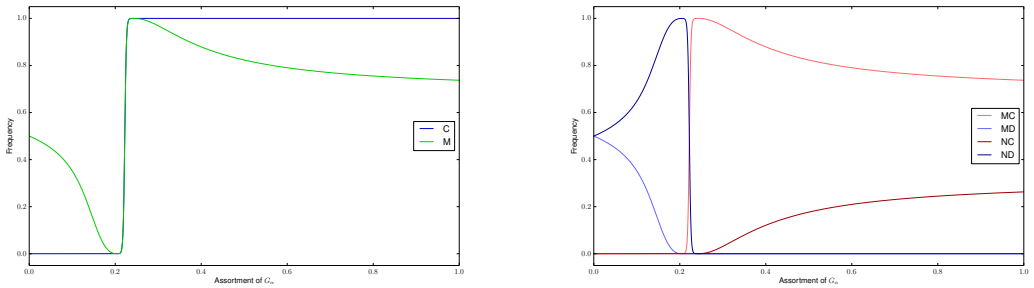
(a) Traits (b) Types

Figure 7.5: The frequency of the traits M and C and the different types for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$) with no GCT assortment ($\beta = 0$).

Figure 7.5 plots the frequencies of the M and C traits and the four types in the discrete model with no GCT assortment ($\beta = 0$). Here, any more assortive mutant game-changing trait declines below its initial frequency of introduction, unless the mutant game-changing trait induces full social trait assortment when the resulting frequency is $M = 0.5$. As the step change in social trait assortment (α) increases, the equilibrium frequency of the M trait decreases, following a sigmoidal curve. The frequency of C starts near 0, but jumps to 0.5 when $\alpha = 1.0$.

Breaking the results down into the four types, we see that MC and NC are 0% of the population at equilibrium (technically, asymptotically close to 0%) when the mutant GCT is anything short of full social trait assortment. When $\alpha = 0$, MD and ND are the same strategy, so both make up 50% of the population. As α increases, MD decreases and ND increases. This is because higher levels of social trait assortment

makes cooperation temporarily more successful – and so gives more opportunity for ND to receive its greater payoff for defecting against cooperators. This causes the sigmoidal decrease in the frequency of the M trait. In the extreme case when $\alpha = 1$, the metagame interaction is balanced, with all four types present at 25% of the population.



(a) Traits (b) Types

Figure 7.6: The frequency of the traits M and C and the different types for a weaker Prisoner's Dilemma ($S = -0.125$, $T = 1.125$) under balanced initial conditions ($c = 0.5$, $m = 0.5$) with no second-order assortment ($\beta = 0$).

What if instead we take a less extreme Prisoner's Dilemma for the starting point? We might be tempted to think that since the extreme Prisoner's Dilemma model ended in a point with all types balanced when the step change was to full social trait assortment, then for a weaker initial Prisoner's Dilemma the frequency of the M and C traits will continue to increase. Figure 7.6 shows equilibrium frequencies of the traits and the types for the discrete model starting at the game with $S = -0.125$, $T = 1.125$. This is the same as our original Prisoner's Dilemma game played with an initial assortment of $\alpha = 0.4375 = \frac{7}{16}$. This time G^α is a Harmony Game if $\alpha > \frac{1}{9}$. In general, we have the relationship that a game played in this Prisoner's Dilemma with an assortment of α is equivalent to the extreme Prisoner's Dilemma played with an assortment of $\frac{9\alpha+7}{16}$.

For $\alpha \leq \frac{2}{9}$, the results are a scaling of the results in Figure 7.5. The four types are balanced at $\alpha = \frac{2}{9}$. Then the M and C alleles increase in frequency one as we might expect. However, while the frequency of the C trait stays at 1, the equilibrium frequency of the M trait decreases. This is because the equilibrium frequency of MC decreases while NC increases. We can see this as a converse of the initial situation for low α – under these conditions, the metagame interaction is a Harmony Game with some Prisoner's Dilemma aspects, and as it becomes a more extreme Harmony Game there is not enough opportunity for MC to be differentially more successful than NC. Note that the frequency of M has still increased from the initial conditions.

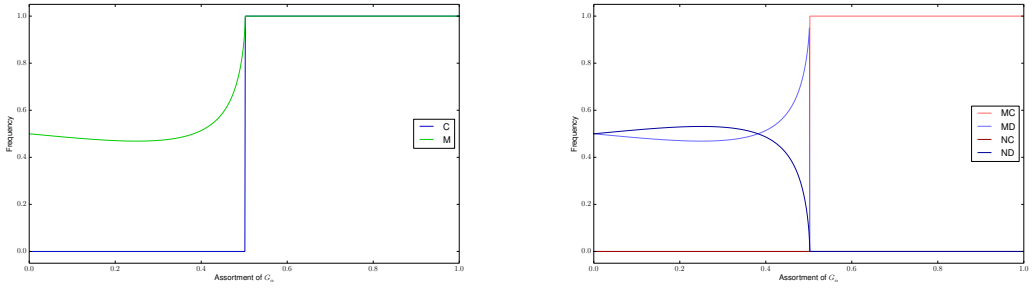
These results reflect the discrepancy we identified in Section 6.1 between metagame models with no assortment on the game-changing trait and the logical argument for social niche construction (Powers et al., 2011). It is a surprising result – increasing the additional social trait assortment that a new game-changing trait provides does

not necessarily benefit the bearers of that game-changing trait. This runs counter to our expectation that increasing assortment benefits assorters and cooperators. In some instances, increasing the step change in assortment actually decreases the equilibrium frequency of the assorters – up to a certain limit.

This means a large discrete jump in the intensity of assortment is required to escape an extreme Prisoner's Dilemma – when there is no assortment on the game-changing trait at least. We do see in the less extreme Prisoner's Dilemma that there is a is a threshold level of increased social assortment above which all-cooperators is the social equilibrium and the assorting trait benefits, but up to this value increasing the intensity of the assorting trait decreases its success.

7.2.1.2 Linked Social and GCT Assortment

We can begin by introducing game-changing trait assortment in the case where this second-order assortment is linked to the amount of first-order assortment on the social trait – that is, $\beta = \alpha$ as G^{α} ranges along the assortment line.

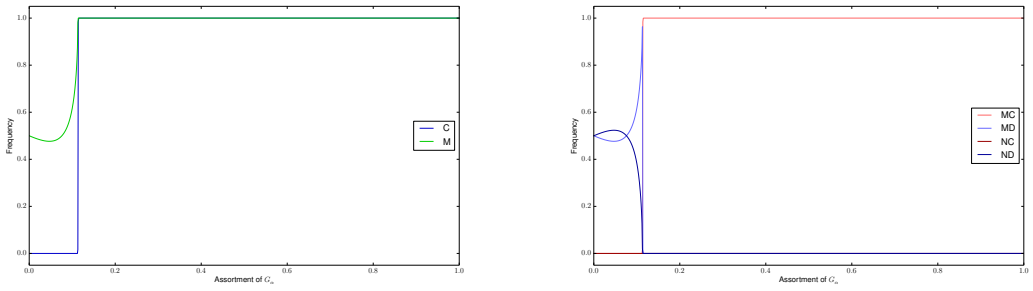


(a) Traits (b) Types

Figure 7.7: The frequency of the traits M and C and the different types for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$) with intrinsic second-order assortment ($\alpha = \beta$).

Figure 7.7 plots the results of the discrete model in this case. We see the equilibrium frequency of the cooperative trait (C) starts at 0, then sharply increases to 1 after a critical value κ. The frequency of the assorters (M) starts at 0.5, decreases, then increases to 1 in line with the increase of cooperators at κ. Looking at the result by type, up until the step change value of κ, the equilibrium is composed of the MD and ND types. Though ND is favoured by small increments of α, as the discrete jump becomes larger it is actually the assorting defector type MD that increases in equilibrium frequency, benefitting from the greater assortment on the game-changing trait. When $\alpha > \kappa$ both types collapse asymptotically towards 0.

Looking at the same model from the weaker Prisoner's Dilemma (Figure 7.8), we see that again this is a scaled version of Figure 7.7. The behaviour is much the same, with



(a) Traits (b) Types

Figure 7.8: The frequency of the traits M and C and the different types for a weaker Prisoner's Dilemma ($S = -0.125$, $T = 1.125$) under balanced initial conditions ($c = 0.5$, $m = 0.5$) with intrinsic second-order assortment ($\alpha = \beta$).

again a critical value κ, lower this time, above which MC is the dominant type. Unlike when there is no assortment on the game-changing trait, the assorting trait remains at 100% of the population above the critical value κ.

However, in both models we find that there is initially a negative gradient to the frequency of the assorting trait, so even with this linked game-changing trait assortment a discrete jump in the intensity of social assortment the new assorting GCT provides would be necessary for assorters to increase in frequency.

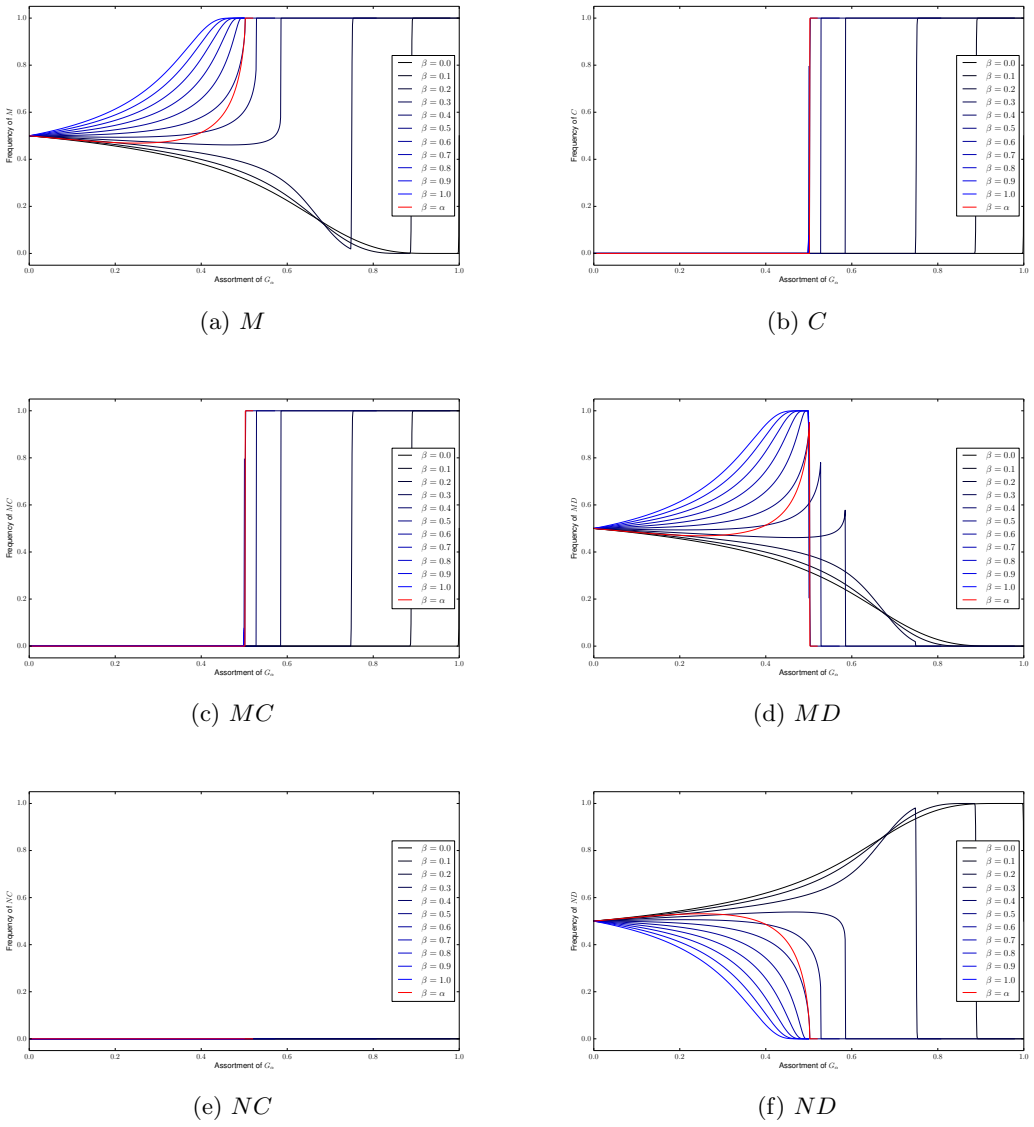
7.2.1.3 Increasing GCT Assortment

Figure 7.9 shows the results of the discrete model starting at the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) as we increase the assortment on the game-changing trait.

There is a critical threshold value of discrete jump in α above which the metagame interaction is in the basin of attraction of an all-cooperators equilibrium. This threshold lowers as the level of GCT assortment increases.

The cooperators present at this equilibrium are all of the assorting type MC. The defecting type NC is asymptotically close to extinction at all points save the situation when $\alpha = 1$ and $\beta = 0$ that we described earlier, where all types at equilibrium are balanced at 0.25 of the population. Therefore, from this extreme Prisoner's Dilemma the all-cooperators metagame equilibrium is also an all-assorters metagame equilibrium.

As with the continuous model, the curves in Figure 7.9 provide cross sections of a surface in the α-β parameter space that we can also show in full (Figure 7.10). Plotting the surface shows distinctly the sharp boundary of the all-cooperators social trait equilibrium in the metagame. The equilibrium frequencies of the assorting trait M are more complex, since assorting defectors are present at equilibrium across the region where all-defectors is the social trait equilibrium.



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.9: The frequency of the traits M and C and the different types for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

For a given α, the frequency of the M trait s always equal or greater as β increases, showing that game-changing trait assortment can allow for the evolution of pro-cooperative assorting traits, even when cooperation itself is not a successful strategy. From this extreme Prisoner's Dilemma though, the assorting trait must still cause a discrete jump in the intensity of social trait assortment to increase in frequency.

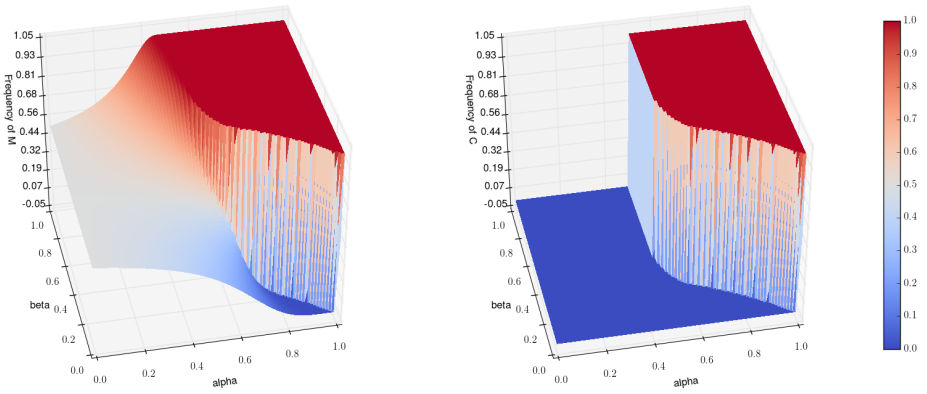


Figure 7.10: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

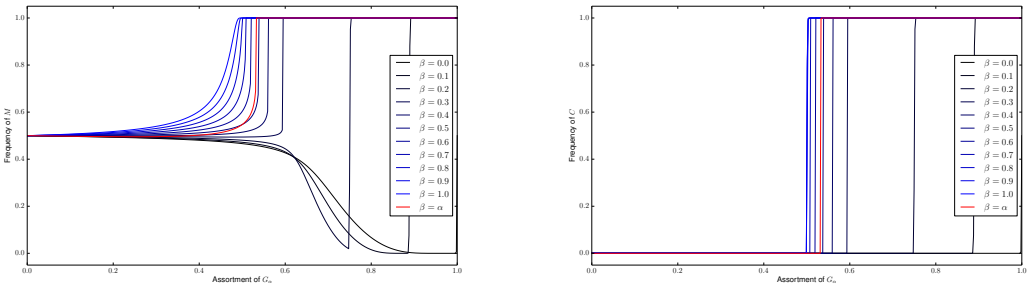
7.2.1.4 Varying the Initial Frequency of Cooperators

As we stated at the outset, in this chapter we will be showing the results of the discrete model for a wide range of parameters. We start by changing the initial frequency of cooperators.

Lowering the initial frequency of cooperators to $c = 0.1$ produces the curves in Figure 7.11 and the surface in Figure 7.12. The lower initial frequency of cooperators has a barely perceptible effect on the boundary of the all-cooperators social equilibrium basin. Fewer cooperators does mean that the metagame interaction is faster to reach social equilibrium, so there is a smaller relative change in frequency between MD and ND, so the surface of the M trait starts flatter then steeply increases or decreases as α increases.

The sign of the change is the same however: if we look at the critical β value above which the equilibrium frequency of the more assorted trait M will increase from the initial conditions, as we did for the continuous model, with the same interval of $\delta_\alpha = 0.01$ to look at the initial gradients of the M surface, then again a GCT assortment of $\beta > 0.497487$ is required for the slope of the M surface to be positive. This is the same value as we found for the continuous model.

If we raise the initial frequency of cooperators to $c = 0.9$ we find the curves in Figure 7.13 and the surface in Figure 7.14. The results are what we would expect from the case of low initial cooperators: again the boundary of the basin of attraction of the all-cooperators all-assorters metagame equilibrium barely changes. This time, since there are more cooperators initially present, there is more opportunity for ND and MD to differentiate themselves in the all-defectors social equilbrium regions of parameter space. The critical β for a small change in α of 0.01 is once again 0.497487.



(a) M (b) C

Figure 7.11: The frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under low cooperation initial conditions ($c = 0.1$, $m = 0.5$).

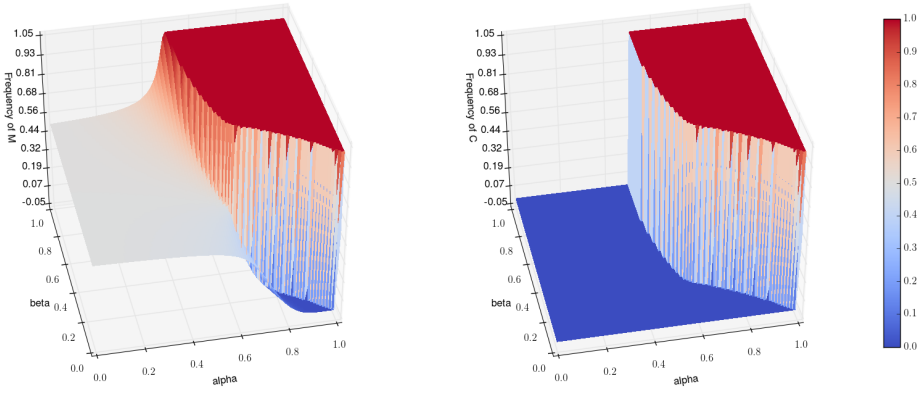
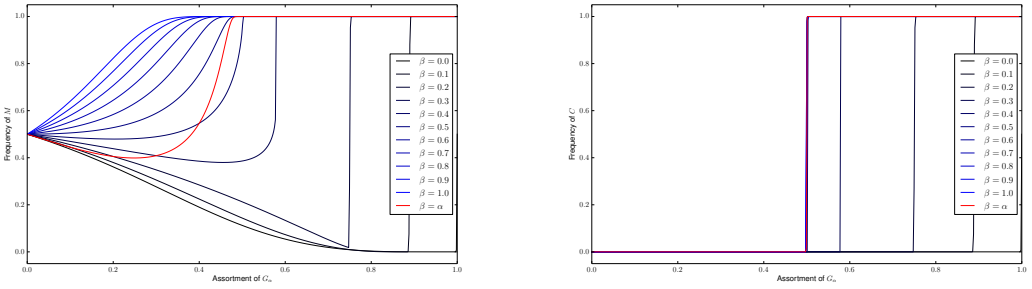


Figure 7.12: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under low cooperation initial conditions ($c = 0.1$, $m = 0.5$).



(a) M (b) C

Figure 7.13: The frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under high cooperation initial conditions ($c = 0.9$, $m = 0.5$).

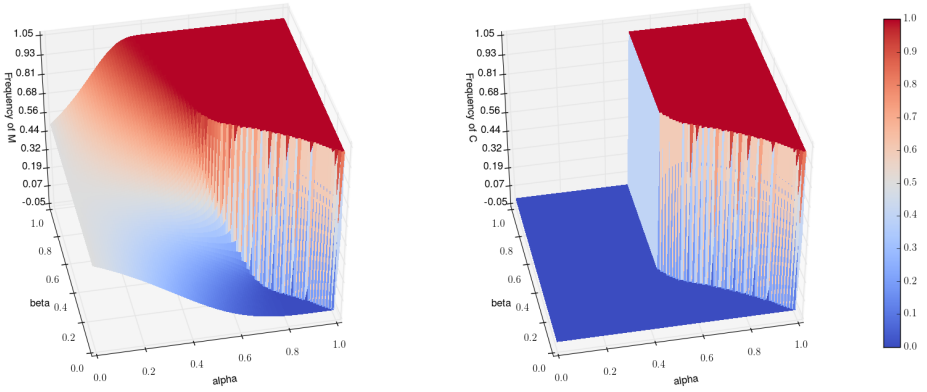
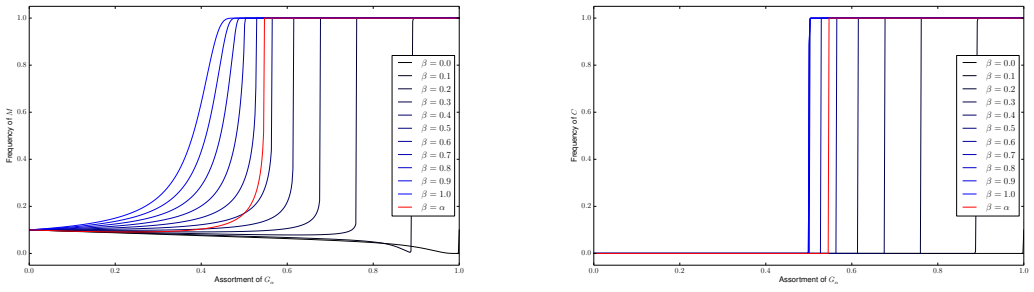


Figure 7.14: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under high cooperation initial conditions ($c = 0.9$, $m = 0.5$).

7.2.1.5 Varying the Initial Frequency of Social Assorters

Changing the initial frequency of the cooperating trait has little effect on the overall results of the discrete model, though it can flatten or steepen the gradient of change in the equilibrium frequency of the M trait. The other population composition parameter to investigate is the initial frequency of the GCT that creates social trait assorters.



(a) M (b) C

Figure 7.15: The frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under low assorter initial conditions ($c = 0.5$, $m = 0.1$).

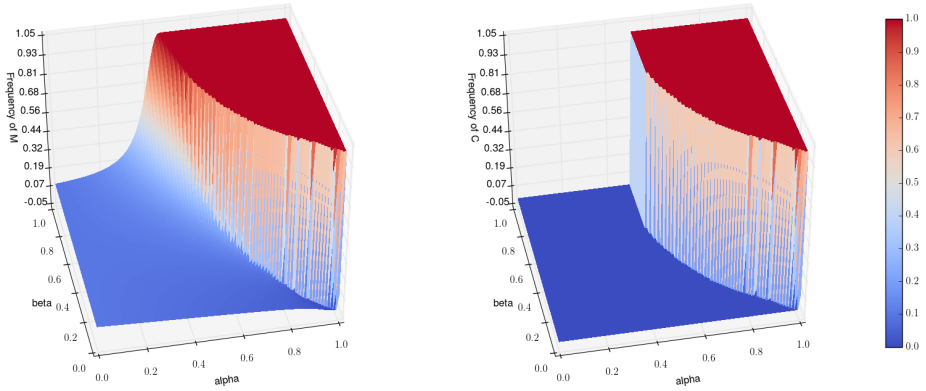
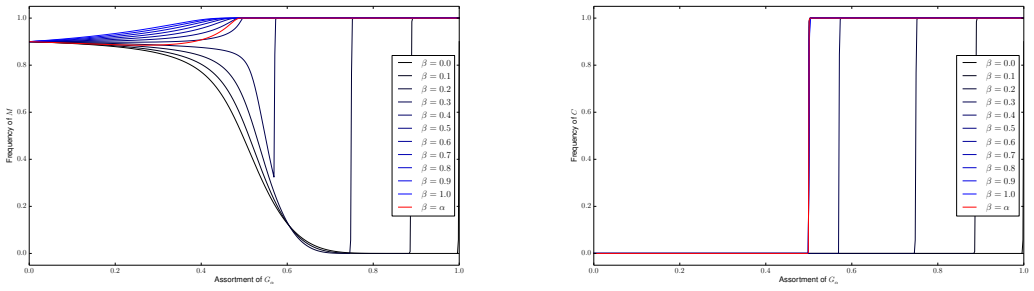


Figure 7.16: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under low assorter initial conditions ($c = 0.5$, $m = 0.1$).

Lowering the initial frequency of the social trait assorters to $m = 0.1$ produces the curves in Figure 7.15 and the surface in Figure 7.16. The basin of attraction of the MC metagame equilibrium is slightly smaller in this instance. This change has a significant impact on the equilibrium frequencies of the M trait, since these often shift by a small amount from the initial conditions. Indeed, when $\alpha = 0$ the metagame interaction is between identical games, so the initial frequency of the assorter trait is unchanged. The critical β for a small change in α of 0.01 remains 0.497487.



(a) M (b) C

Figure 7.17: The frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under high assorter initial conditions ($c = 0.5$, $m = 0.9$).

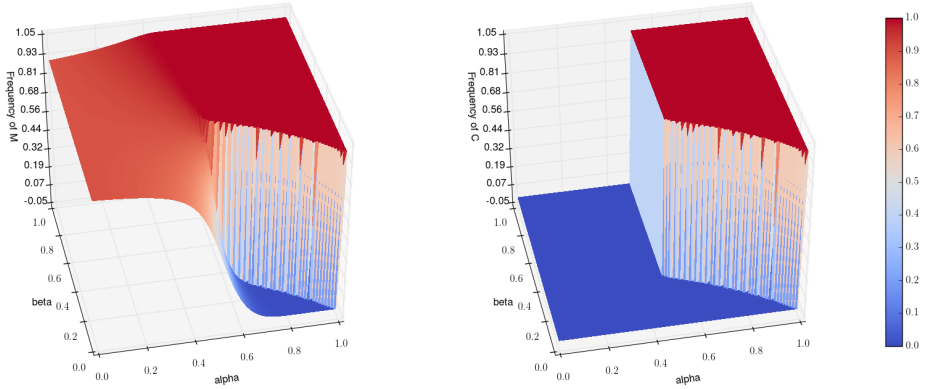


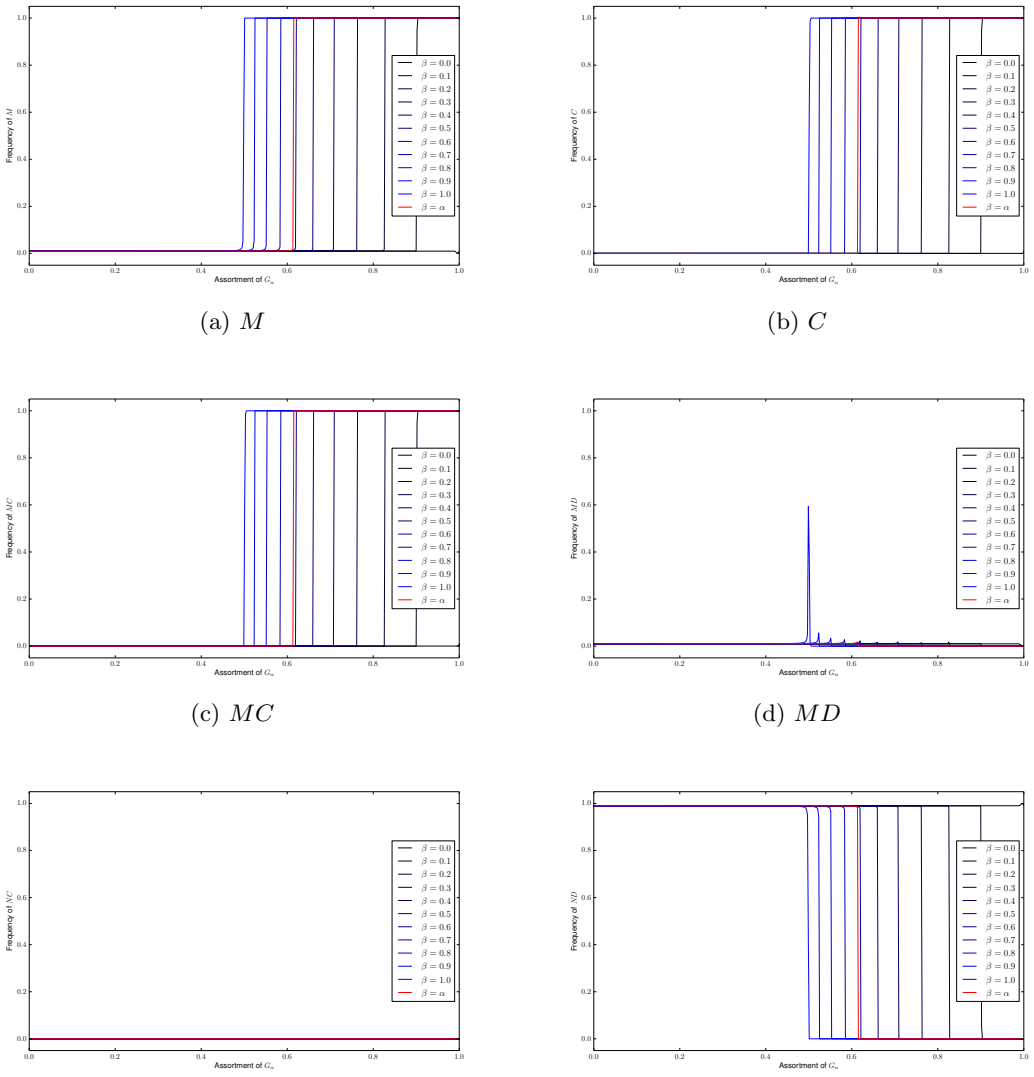
Figure 7.18: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under high assorter initial conditions ($c = 0.5$, $m = 0.9$).

If we raise the initial frequency of the social trait assorter type to $m = 0.9$, we get Figures 7.17 and 7.18. Comparing the discrete model under these two initial conditions, we see that the absolute outcome for the M trait is relative to the initial conditions, though the relative outcome is not – where the M trait increases relative to the initial conditions, it generally does so whatever those initial conditions may be.

This is also seen in the consistent value for the critical β above which assortment increases, which for a small change in α of 0.01 remains at 0.497487. There is a a high-α, high-β region of parameter space over which the all-cooperators all-assorters basin is persists across the initial conditions, but its boundary does change, growing larger when there is a higher initial frequency of assorters or cooperators.

7.2.1.6 Low Frequency Invasion Scenario

As with the continuous model, we also want to see what assortive mutant type is introduced at a low frequency as the result of a mutation. Figure 7.2 displays the results of the continuous model when the initial conditions are changed to $c = 0.01$ and $m = 0.01$. This means that the type MC is introduced at an initial frequency of just 0.001.



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.19: The frequency for the traits M and C and the different genotypes for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under invasion initial conditions ($c = 0.01$, $m = 0.01$).

Breaking the equilibrium frequencies down by type, we see a near complete alignment of the success of assorters and cooperators. The type NC never rises above its introduction frequency of 0.01, while aside from some spikes at the boundary where G^α crosses into the Harmony Game quadrant, MD stays similar to its introduction frequency or decreases.

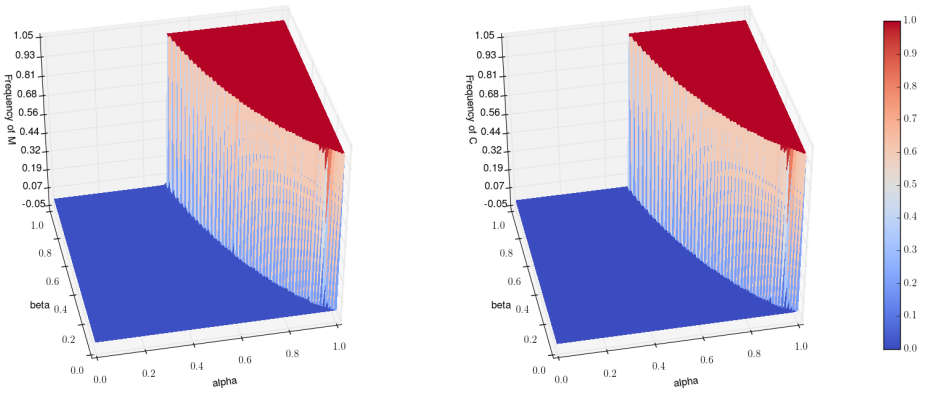


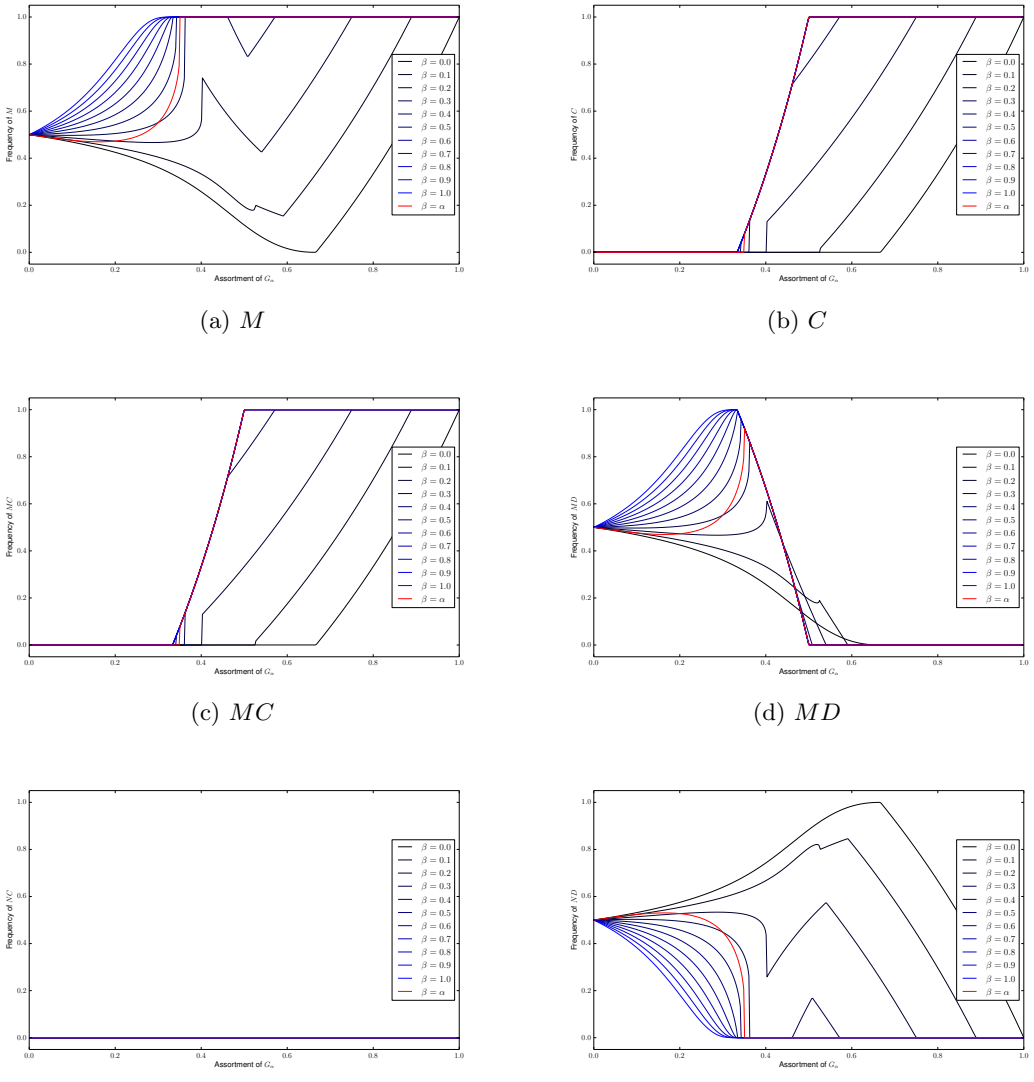
Figure 7.20: α-β plots showing the frequency of the traits M and C for the extreme Prisoner's Dilemma ($S = -1$, $T = 2$) under invasion initial conditions ($c = 0.01$, $m = 0.01$).

Instead, it is the success of MC that causes the increase in the frequency of the M and C traits, though the region of parameter space dominated by MC is smaller than the other models run, reflecting that this low frequency invasion scenario is more challenging for both assorters and cooperators.

7.2.1.7 Different Points in the Prisoner's Dilemma

We take another point in the Prisoner's Dilemma quadrant at $S = -0.5$, $T = 2$. The behaviour of the discrete model from this Prisoner's Dilemma is difference, since unlike the extreme Prisoner's Dilemma, its assortment line crosses into the Snowdrift quadrant. If $\alpha < \frac{1}{3}$ then G^α is a Prisoner's Dilemma, if $\frac{1}{3} < \alpha < \frac{1}{2}$ it is a Snowdrift game and otherwise it is a Harmony Game.

The results (Figures 7.21 and 7.22) reflect the different nature of the games along the assortment line. There is still the region dominated by the type MC, but the boundary to it is less steep as the assortment line passes through Snowdrift games. Examining the type frequency sections (Figure 7.21), we see that up to approximately $G^\alpha = 0.4$, the curves are very similar to those of the extreme Prisoner's Dilemma (Figure 7.9). Above approximately $G^\alpha = 0.66$, the curves look like scaled version of the curves seen in the Snowdrift quadrant (Figure 7.23).



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.21: The frequency of the traits M and C and the different types for a Prisoner's Dilemma with high temptation to defect and lower penalty for being the sucker ($S = -0.5$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

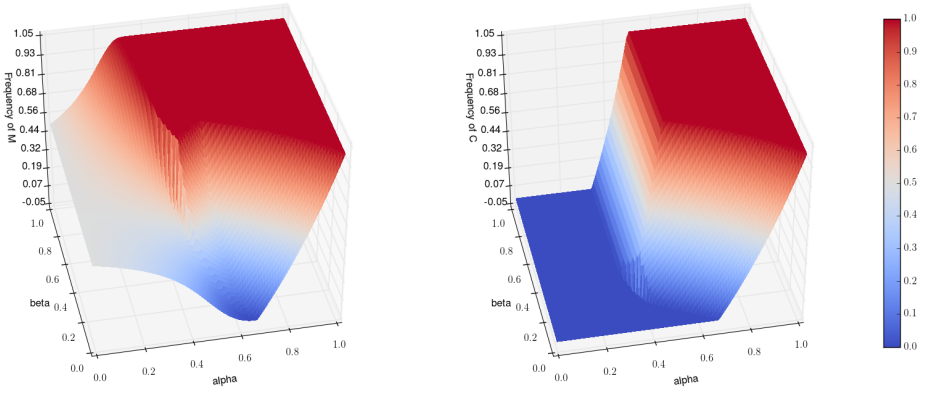
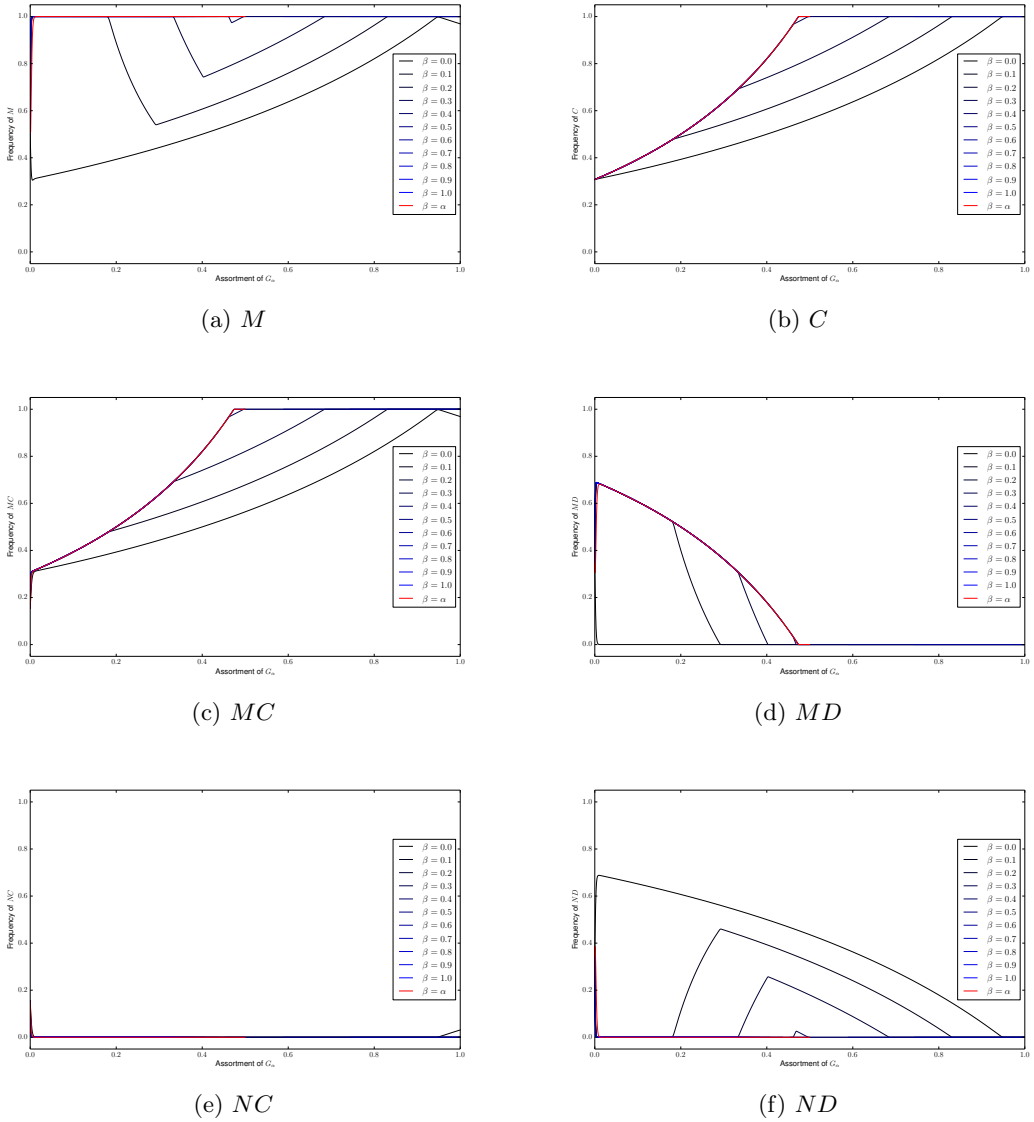


Figure 7.22: α-β plots showing the frequency of the traits M and C for a Prisoner's Dilemma with high temptation to defect and lower penalty for being the sucker ($S = -0.5$, $T = 2$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

7.2.2 The Snowdrift Game

We know the Snowdrift quadrant is a unique region of ST-Space, because the base two-player game has a polymorphic equilibrium containing both cooperators and defectors. When there is no GCT assortment, we have seen that since neither of the social traits tend towards extinction, the metagame equilibrium reaches a fixed equilibrium regardless of the initial conditions, containing the types MC and ND (Section 5.5.1).



(a) M (b) C

(c) MC (d) MD

(e) NC (f) ND

Figure 7.23: The frequency for the traits M and C and the different genotypes for a Snowdrift Game ($S = 0.4$, $T = 1.9$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

Figures 7.23 and 7.24 show the type sections and surfaces for the discrete model in a Snowdrift Game from the point $S = 0.4$, $T = 1.9$. Out of interest we note how similar the results of the $\alpha - \beta$ plot are to those obtained using the encounter functions method

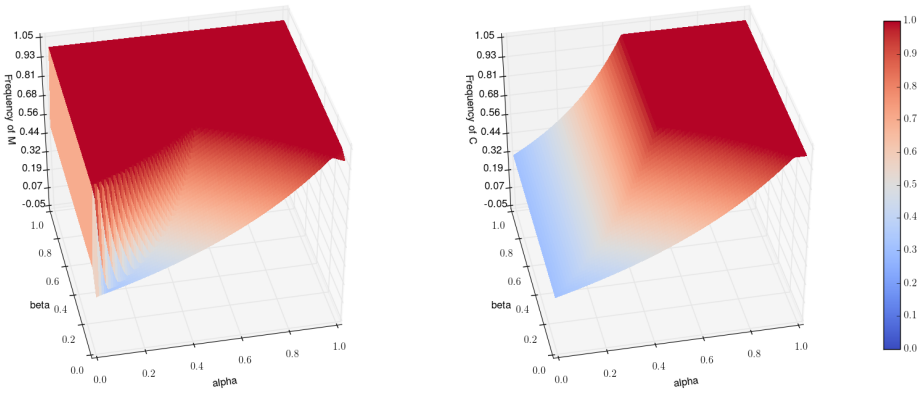
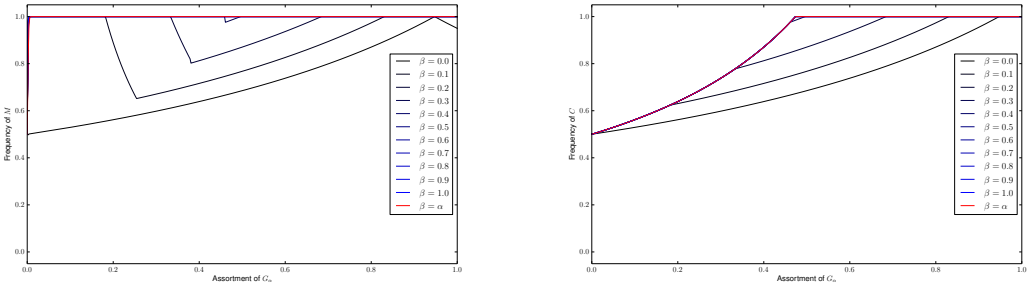


Figure 7.24: α-β plots showing The frequency of the traits M and C for a Snowdrift Game ($S = 0.4$, $T = 1.9$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

(Figure 6.8a) – except here there is no equilibrium region with none of the assorting trait.

As we saw there, for low GCT assortment, a moderate discrete jump in the level of social trait assortment actually decreases the success of the assorting trait because it increases the success of the ND type. In general, however, introducing assortment onto the game-changing trait is extremely beneficial to the assorting trait, which dominates the metagame over almost all of parameter space.



(a) M (b) C

Figure 7.25: The frequency for the traits M and C and the different genotypes for a Snowdrift Game ($S = 0.9$, $T = 1.9$) under balanced initial conditions ($c = 0.5$, $m = 0.5$).

if we compare the sections with another Snowdrift Game at $S = 0.9$, $T = 1.9$, we see that the two are like stretched versions of each other, with the initial frequency of cooperators and assorters are higher in the game at $S = 0.9$, $T = 1.9$ due to the different social equilibrium of the local game.

7.3 Results

Looking at the results of the discrete model, we see there are significant similarities. In particular, a consistent region in parameter space where MC is the dominant type, leading to a metagame interaction equilibrium with all-cooperators and all-assorters. This is at high α and gets larger as the amount of GCT assortment increases. When MC is introduced at low frequency this may be the only only region in parameter space favourable to cooperators. Since this basin enlarges with increased assortment on the game-changing trait, this shows how game-changing trait assortment can allow for the evolution of pro-cooperative assorting traits. From this extreme Prisoner's Dilemma though, the assorting trait must still cause a discrete jump in the intensity of social trait assortment to increase in frequency.

We have used another 'critical value' here, the β above which the assorting trait increased in frequency, and found that it was the same across multiple initial conditions for the extreme Prisoner's Dilemma. This gives us another summary statistic to look at over ST-space.

7.4 Critical Values

One of the stated aims of this chapter was to develop summary measures for the effects of assortment on social and game-changing traits on the evolution of those traits. We have seen how many different parameters we can vary for just a few points in ST-space. Therefore we want to develop summary statistics that compress information about the higher dimensional parameters into measures we can look at across ST-space.

We are asking two questions: when does cooperation evolve, and when does assortment evolve.

From looking at the models so far, we have already identified two interesting properties – two 'critical values'. A critical α value for the evolution of cooperation, κ_α, which is the step increase in social assortment required for a game α for a game in the ST-plane such that if the increase in social assortment created by the assorting type is above this threshold, the metagame interaction will have an all-cooperators equilibrium.

The critical β value for the evolution of social assortment, κ_β, is the level of game-changing trait assortment (β) such that above this threshold the frequency of the assorting game-changing trait will always increase from the given initial conditions.

The summary measure for each type of assortment requires we fix the value of the other (so we find the critical α for a given level of GCT assortment β), but we can also look at the case where the two types of assortment are linked ($\alpha = \beta$).

7.4.1 Critical Values for the Evolution of Cooperation

Here we are interested in what the summary measures tell us about what levels of assortment are required for cooperation to become the dominant social strategy at equilibrium – that is, for the equilibrium of the metagame interaction to only include bearers of the cooperator trait. For each of α, β and the combination of $\alpha = \beta$, at every point on the ST-plane we find the lowest value (α, β or $\alpha = \beta$) above which all values lead to an all-cooperators equilibrium.

To create this measure, we fix a level of assortment on the game-changing trait β, and then for every game G in ST-space we find the value κ_α such that if $\alpha > \kappa_\alpha$, a metagame interaction between G and G^α will have an all-cooperators equilibrium. We calculate these points by interval bisection. We create this measure for two different sets of initial conditions: balanced ($c = 0.5$, $m = 0.5$) and low frequency invasion ($c = 0.01$, $m = 0.01$).

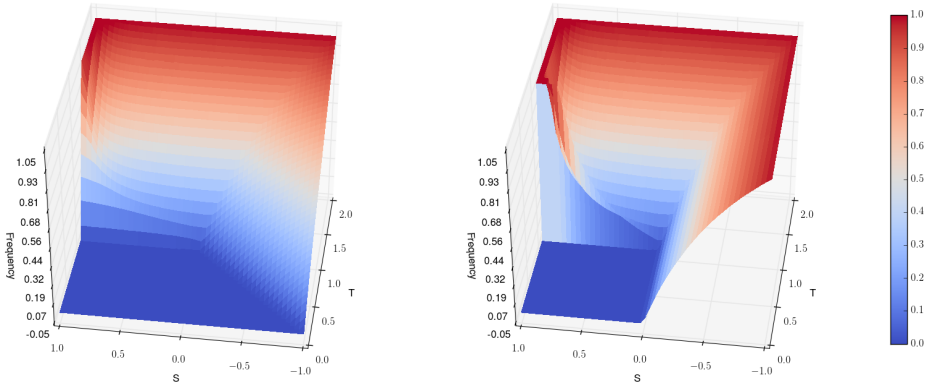


Figure 7.26: The critical α value for the evolution of cooperation over the whole of ST-space when there is no GCT assortment ($\beta = 0$) for a balanced initial population (left: $c = 0.5$, $m = 0.5$) and invasion frequency population (right: $c = 0.01$, $m = 0.01$).

Figure 7.26 shows the critical α for the evolution of cooperation. The height of the surface is the lowest α above which all values of α lead to 100% cooperators. When the metagame is already in this basin at $\beta = 0$, this will be the threshold - as is the case for the whole Harmony Game quadrant. When there is no assortment on the game-changing trait, the threshold is the α value such that $G^{\frac{\alpha}{2}}$ is in the basin of attraction of cooperation. This is different in the low frequency invasion scenario since the low initial frequency of cooperators makes the whole Stag Hunt quadrant essentially an extension of the Prisoner's Dilemma.

Introducing assortment on the game-changing trait linked to the intensity of social assortment (Figure 7.27), reduces the threshold value required for the metagame interaction at each point in the space for cooperation to evolve, but does not alter the relative

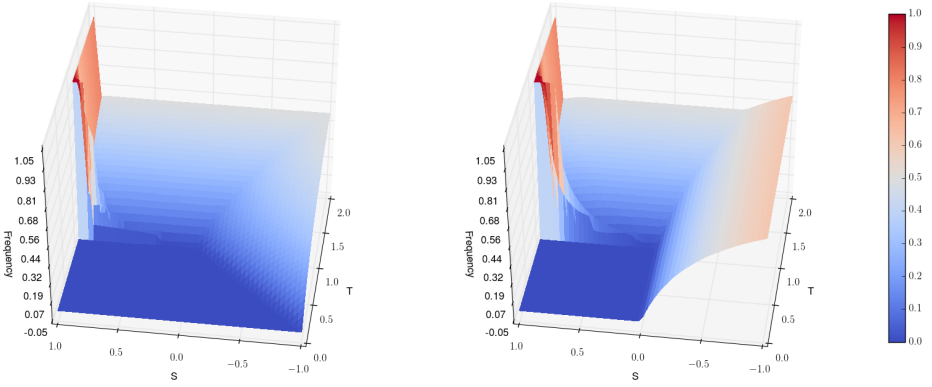


Figure 7.27: The critical $\alpha = \beta$ value for the evolution of cooperation over the whole of ST-space for a balanced initial population (left: $c = 0.5$, $m = 0.5$) and invasion frequency population (right: $c = 0.01$, $m = 0.01$).

levels of social trait assortment required between two points save for a region at the upper end of the Snowdrift Game where $S = 1 = R$. It remains the case that when there is linked social trait and game-changing trait assortment, discrete jumps in the social trait (and hence the game-changing trait too) for cooperation to become the equilibrium state of the metagame interaction when it is not the equilibrium state of the initial game.

7.4.2 Critical Values for the Evolution of Social Assortment

The critical β value for the evolution of social assortment, κ_β, is the level of game-changing trait assortment (β) such that above the threshold the frequency of the assorting game-changing trait will always increase from the given initial conditions. To calculate it we fix some step change in α (we use $\delta_\alpha = 0.01$) and use the interval bisection technique for points across ST-space.

In the case of the extreme Prisoner's Dilemma, we found that this value was consistently $\beta = 0.4978$ across a range of different initial conditions. Figure 7.28 plots the critical β for the evolution of social assortment over the whole of ST-space. We see that, as expected, the critical β in the extreme Prisoner's Dilemma is the highest such β in ST-space (save for the unusual line on the boundary of ST-space where $S = R = 1$, along which assortment is not favoured). The critical β decreases smoothly across the basin of attraction of all-defectors in the social trait game – that is the Prisoner's Dilemma quadrant and either the all-defect half of the Stag Hunt quadrant when the initial conditions are balanced or the entirety of the Stag Hunt quadrant when the low initial frequency of the cooperator type makes the entire quadrant essentially an extension of the Prisoner's Dilemma. We also see that for small δ_α, any GCT assortment will allow the assorting type to increase across the Harmony and Snowdrift quadrants.

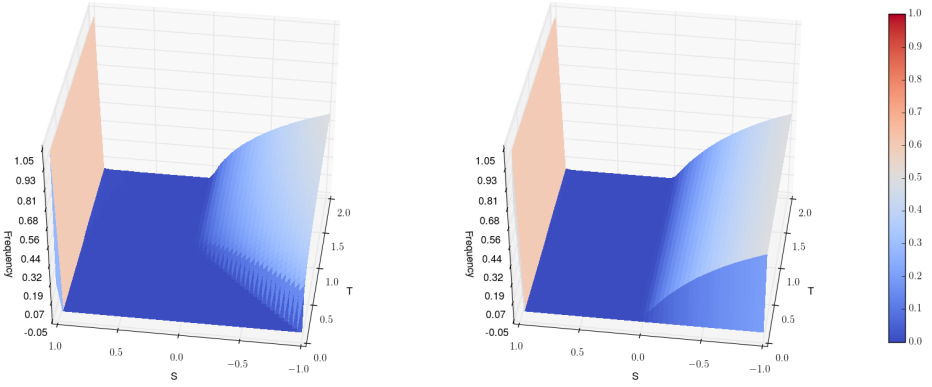


Figure 7.28: The critical β value for the evolution of social assortment over the whole of ST-space for a balanced initial population (left: $c = 0.5$, $m = 0.5$) and invasion frequency population (left: $c = 0.01$, $m = 0.01$).

Much the same as for the critical α, making the step change in social trait assortment instead linked to the level of game-changing trait assortment (so $\delta_\alpha = \beta$) acts to lower the critical value across the space (Figure 7.29). The greatest value, the critical linked $\alpha = \beta$ in the extreme Prisoner's Dilemma, is 0.3820.

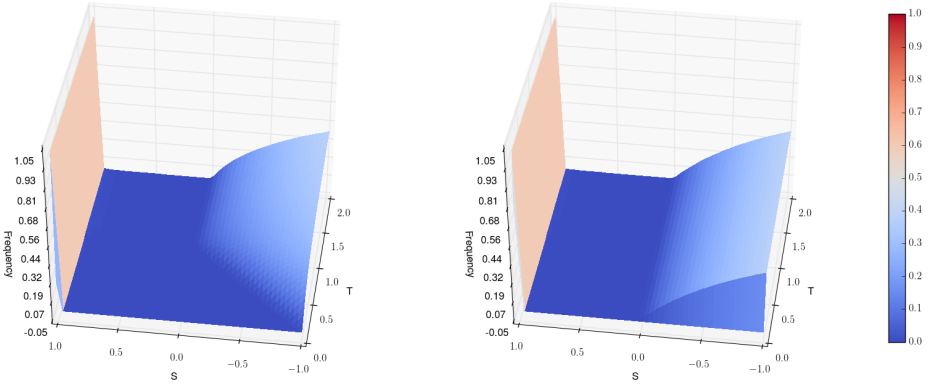


Figure 7.29: The critical $\alpha = \beta$ value for M at the whole of the ST-plane for a balanced initial population (left: $c = 0.5$, $m = 0.5$) and invasion frequency population (left: $c = 0.01$, $m = 0.01$).

7.5 Conclusions

This chapter has not presented substantial new results, but has given us greater understanding of the behaviour of the model across a range of parameters, and has let us represent our understanding of existing behaviours of the metagame in new ways, such as through the two critical values we have identified. We have seen that the intensity of additional social assortment created by a more assorting game-changing trait and the

level of assortment on that game-changing trait do both, in general, increase the success of bearers of the cooperating or assorting trait.

However, this relationship is not strictly linear. We have seen that in Prisoner's Dilemma games, increasing the intensity of the social assorting trait actually decreases the success of that trait up to a certain threshold, since the increased assortment allows cooperators to persist for long enough for non-assorting defectors to gain more from predating on those cooperators and thus further increase their fitness over the assorting defectors.

The initial conditions affect the critical values by changing the social equilibrium of the local games. Comparing the scenarios with balanced and low frequency invasion initial conditions for both the critical α for the evolution of cooperation (Figure 7.26) and the critical β for the evolution of social assortment (Figure 7.28), we see the consequences of a low initial frequency of cooperators causing almost the entire Stag Hunt quadrant to be inside the basin of attraction of the no-cooperations social equilibrium. As a result, the critical values in the Stag Hunt are the same as those in the Prisoner's Dilemma quadrant. When the initial conditions are balanced, the critical α and β both reduce as T reduces from the extreme Prisoner's Dilemma to the corner point of the Stag Hunt in ST-space at $S = -1$, $T = 0$. This sensitivity to the initial conditions is an issue for our modelling work as by choosing initial conditions we affect the outcome. In the next chapter we will mitigate this by presenting a natural set of initial conditions to use in our concluding models.

We have found that social trait assortment in the Prisoner's Dilemma can evolve if there is sufficient assortment on the game-changing trait, but that the assortment on the game-changing trait or the intensity of social assortment the game-changing trait creates must increase by a discrete amount to start the process. Social assortment will not increase in the Prisoner's Dilemma for small continuous changes in α and β when linked. In other words, a game-changing trait that intrinsically controls its own assortment cannot evolve over the whole plane – and cannot start to evolve from any point where the intrinsic critical value is non-zero (Figure 7.29). However, if the nature of the game-changing trait means that it is inherently assorted above the right threshold then the assorting trait can begin to spread.

We have concentrated on analysing the behaviour of metagames in the abstract here. In the next chapter we will address the issue of sensitivity to the initial conditions and then apply our understanding to develop two final models that address the evolution of game-changing traits that cause assortment, and the paths a population's effective game can take through the space of social dilemmas.

Chapter 8

The Evolution of Game-Changing Traits

In this chapter, we apply our results and techniques to study three key models of the evolution of game-changing traits. We tie together the techniques we have developed so far: transformations to games induced by game-changing traits (Chapter 4), the modelling formalism of metagames (Chapter 5) and the means to represent assortment on the game-changing trait (Chapters 6 and 7).

Because we found that the assortment metagame is sensitive to the initial conditions, we first model a natural scenario where the initial conditions for metagames at each point in ST-space arise as a mutation to the equilibrium frequency of the local social game. Given the separation of timescales in our model, where we assume that social trait equilibration is faster than that of game-changing traits, this is a reasonable scenario to privilege for our final models. We find that there must be a significant level of GCT assortment for social trait assortment to spread in the Prisoner's Dilemma quadrant – the region of game space where cooperation is equivalent to strong altruism.

With this method to create initial conditions, we return to modelling the evolution of game-changing traits that create social assortment, and how it is affected by assortment on the game-changing trait ('second order assortment'). Game-changing traits that create social assortment are particularly important because social trait assortment is given as an ultimate explanation for the evolution of cooperation (Chapter 4). Here we conclude the results of the extensive study in Chapter 7.

The final series of models look at the paths the effective social dilemma a population plays can take through the space of games. We have repeatedly discussed the idea that as a population's game-changing trait evolves over time, it will create a path through game space. In in our modelling work so far though, we have been restricted to creating vector fields showing the mean direction of selection pressure on the game-changing

trait at any point in game space, or to determining the regions where some constrained game-changing trait will increase or decrease in frequency. In this final model, small stochastic mutations to the game-changing trait create mutants that compete with the original, either displacing it or failing. The success or failure of a series of these mutations creates a path through game space. We take advantage of the invasion at mutated social equilibrium frequencies modelling scenario to provide a consistent basis for determining the initial conditions as the games change.

We plot the trajectories of these stochastic paths through game space and find that they do indeed follow the vector fields we constructed. We see the effect of increasing GCT assortment by calculating the mean paths from a single point in ST-space: populations engaged in a Prisoner's Dilemma that would evolve game-changing traits further favouring defectors will instead take paths through to the Snowdrift and even the Harmony Game when there is sufficient assortment on the game-changing trait. Finally, we assess the significance of GCT assortment for the evolution of cooperation by showing how populations taking paths through game space across the whole of ST-space changes the social equilibria of the final games.

8.1 The Invasion at Mutated Equilibrium Frequencies Scenario

Though we examined the behaviour of metagame models under a range of initial conditions in Chapter 7, in general we have considered the behaviour of metagame interactions where the initial conditions are balanced and in linkage equilibrium — that is, the initial frequency of each type at all points in the ST-space under consideration is 0.25. This scenario is useful to clearly demonstrate the metagame dynamics from a position of complete balance. However, biologically this is not a particularly plausible scenario.

We have also modelled a scenario in which the cooperative social trait C and mutant game-changing trait M both arose as mutations to a homogenous initial population, creating a scenario in which cooperators and mutant GCTs attempt to 'invade' the extant population from an initial low frequency. This is a more plausible biological scenario, though not in all instances: if the social game is a Snowdrift or Harmony Game then it is unlikely that the initial population would all be defectors, since any cooperative strategy would be favoured to reach a non-zero equilibrium frequency.

The other problem is that we have been imposing one set of fixed initial conditions to games over the whole of ST-space. When we create 'vector fields' of the selection pressure on the game-changing trait over the ST-plane, we want these vector fields to indicate the possible trajectories of a population through game space given that metagame. For this to be the case, the results of one metagame interaction should plausibly form the initial

conditions for the next. Under these conditions, it is implausible to assume that after a metagame interaction has reached equilibrium, the initial conditions for the subsequent metagame interaction will have been reset to the completely balanced all-0.25 case. While the low frequency invasion scenario would make sense in the Prisoner's Dilemma, if the metagame interaction were between games in a different quadrant then we would expect a greater initial frequency of cooperators.

An alternative scenario presents both a plausible narrative for game change over the long timescale of GCT evolution and a compelling initial scenario for investigating metagame interactions in its own right. That is the scenario of *invasion at mutated equilibrium frequencies scenario*: we assume the initial game G_N has reached its social strategy equilibrium before low-frequency mutations perturb the social trait and introduce a new mutant game-changing trait M.

So we can imagine then the following narrative for the model over the long timescale of GCT evolution:

- There is a population all possessing the game-changing trait N, so the population as a whole is playing the effective game G_N.
- By playing this game, the social traits in the population end up at the equilibrium frequency of the game G_N – with x_C cooperators and x_D defectors.
- A mutation at some low probability m_μ introduces a new game-changing trait M. Simultaneously, there is also a small probability (c_μ) of a mutation to the frequencies of the social traits.
- A metagame interaction occurs between G_N and G_M starting from these mutated equilibrium conditions.
- The metagame interaction equilibrium will be reached.
- We then assume that the GCT type which has increased in frequency will, over the long timescale, supplant the other – even if it is still in the minority in absolute terms. Thus the population as a whole will eventually be playing G_N again or have changed to play G_M.
- The cycle repeats.

When we are talking about the vector field model, we modify this narrative so instead say that instead of a single mutation G_M, we imagine that over the long timescale of GCT evolution, multiple possible changes in GCT might occur within a certain small radius in game space. Though the vector field model does not compare all these possible metagame interactions simultaneously, it indicates the net likely change in the GCT given the combined effects of all these possible GCT mutations.

Mathematically, we let the chance of a mutation to the game-changing trait be m_μ and to the social trait c_μ. Then the four-dimensional population state vector of initial conditions for a metagame interaction at each point in the space can be found as a linear transformation of the two-dimensional population state vector of equilibrium frequencies for the two-strategy game at the same point given a single initial condition, the initial frequency of cooperators (which only affects the Stag-Hunt game when both cooperators and defectors are initially present). This transformation is done with the matrix:

$$\begin{pmatrix} m_\mu(1-c_\mu) & m_\mu c_\mu \\ m_\mu c_\mu & m_\mu(1-c_\mu) \\ (1-m_\mu)(1-c_\mu) & (1-m_\mu)c_\mu \\ (1-m_\mu)c_\mu & (1-m_\mu)(1-c_\mu) \end{pmatrix} \tag{8.1}$$

So given an initial population state of $(C, D) = (x_C, x_D)$ we have an invasion at equilibrium frequency population state of:

$$\begin{aligned} x_{MC} &= m_\mu(1-c_\mu)x_C + m_\mu c_\mu x_D \\ x_{MD} &= m_\mu c_\mu x_C + m_\mu(1-c_\mu)x_D \\ x_{NC} &= (1-m_\mu)(1-c_\mu)x_C + (1-m_\mu)c_\mu x_D \\ x_{ND} &= (1-m_\mu)c_\mu x_C + (1-m_\mu)(1-c_\mu)x_D \end{aligned} \tag{8.2}$$

8.1.1 Vector Fields for the Invasion at Mutated Equilibrium Frequencies Scenario

As in Section 5.4.4, we can construct a 'vector field' over ST-space showing the net selective pressure on the game-changing trait at different points across the space. One of our goals with the invasion at mutated equilibrium frequencies scenario is to 'join the tails' of the metagame interactions depicted in the vector field model and move from talking about the selective pressure at points in game space to talking about trajectories through game space. First though we do want to construct the vector field to understand the expected behaviour of those paths through game space.

Recall, we construct the vector field by taking a square lattice of points spaced 0.1 units apart in S and T. For each point representing a game G_N, we compute 36 metagame interactions between G_N and a mutant GCT trait G_{M_i}, with each of the G_{M_i} spaced evenly around a small circle centred on G_N with radius $r = 0.05$. For each of these metagame interactions, we record the change in the frequency of the mutant trait $\Delta m_i = x_{M_iC} + x_{M_iD} - x_{NC} - x_{ND}$ after the metagame interaction reaches equilibrium. The vector demonstrating the selection pressure on the GCT is the mean of the vector offset of G_{M_i} from G_N scaled by the change in the frequency of the M_i trait.

The generated vector fields for increasing levels of GCT assortment are shown in Figure 8.1, superimposed on the results of the assortment metagame for the same β. We quantify these results by taking the metagame vector field over ST-space and calculating the mean vector at each point in the lattice over the whole of the space, and also for every point in each fundamental game quadrant (games lying on the boundary between two quadrants are not counted). Table 8.1 shows the results.

	Overall		HG		SD		PD		SH	
	S	T	S	T	S	T	S	T	S	T
$\beta = 0.0$	0.0397	0.0397	0.0000	0.0000	0.1586	0.1586	0.0000	0.0000	0.0003	0.0003
$\beta = 0.1$	0.0702	-0.0248	0.0000	-0.0000	0.2854	-0.1004	0.0000	0.0000	0.0019	-0.0010
$\beta = 0.2$	0.0704	-0.0241	0.0000	-0.0000	0.2891	-0.0973	0.0000	0.0000	0.0033	-0.0028
$\beta = 0.3$	0.0707	-0.0240	0.0000	-0.0000	0.2909	-0.0957	0.0000	0.0000	0.0050	-0.0047
$\beta = 0.4$	0.0714	-0.0239	0.0000	-0.0000	0.2935	-0.0933	0.0001	0.0000	0.0074	-0.0073
$\beta = 0.5$	0.0722	-0.0241	0.0000	-0.0000	0.2957	-0.0910	0.0001	0.0000	0.0107	-0.0107
$\beta = 0.6$	0.0732	-0.0251	0.0000	-0.0000	0.2957	-0.0910	0.0002	0.0000	0.0146	-0.0148
$\beta = 0.7$	0.0741	-0.0254	0.0000	-0.0001	0.2979	-0.0887	0.0003	0.0000	0.0184	-0.0185
$\beta = 0.8$	0.0748	-0.0260	0.0000	-0.0001	0.2979	-0.0887	0.0005	0.0000	0.0210	-0.0210
$\beta = 0.9$	0.0754	-0.0264	0.0000	-0.0001	0.2979	-0.0887	0.0014	0.0000	0.0223	-0.0222
$\beta = 1.0$	0.0769	-0.0265	0.0000	-0.0002	0.2979	-0.0887	0.0069	0.0000	0.0226	-0.0226

Table 8.1: The mean vector of GCT change over all games in ST-space broken down by game quadrant where the initial population state of each metagame interaction is the equilibrium state of the game at that point subjected to small mutations. Initial conditions: $c = 0.5$, $c_\mu = 0.01$, $m_\mu = 0.01$.

The table shows the general effect of increasing assortment on the game-changing trait. When there is no GCT assortment, the mean vector over the whole of ST-space is a perfectly balanced small increase in S and T, both of 0.0397. As GCT assortment increases, the mean change in S increases, while the mean change in T decreases. This shows the general effect of GCT assortment is to create conditions more favourable to cooperators (by increasing the sucker's payoff S) and less favourable to defectors (by decreasing the temptation to defect T). The biggest change comes from the introduction of any GCT assortment at all.

Breaking the results down by region, we can see the reason for this is due to the Snowdrift Game. We already know that the magnitude of game-change is much larger in the Snowdrift Game than in other games, because the polymorphic social equilibrium of the Snowdrift game means that both cooperators and defectors are present in the population long enough for the metagame equilibrium to fully equilibrate. In the Prisoner's Dilemma quadrant, net change when there is no GCT assortment is less than the four decimal places we restrict numbers to in the table. We see that increasing GCT assortment leads to a small increase in the sucker's payoff, benefitting cooperators in the Prisoner's Dilemma. The Harmony Game is the mirror image of this: as GCT assortment increases there are small absolute decreases in T, to the detriment of defectors. But in the Snowdrift Game, the absolute game change is orders of magnitude larger, affecting the

overall results. Here, the mean game change starts at a balanced increase in S and T both of 0.1586, indicating that the metagame selection pressure is in favour of increases in mean fitness. The introduction of a low GCT assortment of $\beta = 0.1$ is sufficient to shift the mean vector to one in which S increases but T decreases, so defectors are hindered while cooperators benefit.

Since the magnitude of game change in a quadrant can vary significantly and we are as interested in the direction of game change as the absolute values, we repeat this tabulation but first normalise the magnitude of all summed vectors so one game in a quadrant where the effect is particularly strong does not dominate the results. This quantifies the way we have plotted the vector fields with each vector shown as having the same magnitude (as in Figure 8.1). These results are given in Table 8.2).

	Overall		HG		SD		PD		SH	
	S	T	S	T	S	T	S	T	S	T
$\beta = 0.0$	0.7071	0.7071	1.0000	0.0000	0.7071	0.7071	0.0000	1.0000	0.7071	0.7071
$\beta = 0.1$	0.9688	0.2479	0.9833	-0.1822	0.9431	-0.3326	0.4134	0.9105	0.8657	0.5006
$\beta = 0.2$	0.9918	0.1278	0.9544	-0.2985	0.9476	-0.3196	0.6112	0.7915	0.9438	0.3305
$\beta = 0.3$	0.9992	0.0390	0.9215	-0.3885	0.9497	-0.3133	0.7294	0.6841	0.9811	0.1936
$\beta = 0.4$	0.9996	-0.0283	0.8871	-0.4615	0.9527	-0.3038	0.8040	0.5946	0.9966	0.0820
$\beta = 0.5$	0.9967	-0.0817	0.8529	-0.5221	0.9555	-0.2951	0.8533	0.5214	1.0000	-0.0089
$\beta = 0.6$	0.9918	-0.1277	0.8196	-0.5730	0.9555	-0.2951	0.8871	0.4616	0.9966	-0.0829
$\beta = 0.7$	0.9866	-0.1631	0.7876	-0.6162	0.9582	-0.2862	0.9110	0.4124	0.9897	-0.1434
$\beta = 0.8$	0.9807	-0.1955	0.7571	-0.6533	0.9582	-0.2862	0.9284	0.3715	0.9810	-0.1939
$\beta = 0.9$	0.9747	-0.2235	0.7281	-0.6855	0.9582	-0.2862	0.9415	0.3370	0.9716	-0.2367
$\beta = 1.0$	0.9688	-0.2480	0.6994	-0.7147	0.9582	-0.2862	0.9514	0.3080	0.9619	-0.2735

Table 8.2: Table 8.1 (invasion at mutated equilibrium frequencies scenario: $c = 0.5$, $c_\mu = 0.01$, $m_\mu = 0.01$) with all summed vectors normalised to have unit magnitude to examine the direction of game change.

We see that as GCT assortment increases, the mean normalised vector shifts from one of balanced increase in S and T to a strong increase in S and a decrease in T. The mean vector is most strongly in favour of increasing S for medium levels of GCT assortment; at even higher levels of GCT assortment as well as increasing the success of cooperators, the success of defectors decreases. The mean normalised vector lets us see more clearly the effect of GCT assortment in the Harmony and Prisoner's Dilemma quadrants, where the absolute change in frequencies is small due to the metagame interaction starting so close to equilibrium social frequencies. In the Harmony Game quadrant, the mean vector is initially exclusively in favour of increases in S that benefit cooperators. But as GCT assortment increases, we see the decreases in T that make the game less favourable to defectors.

In the Prisoner's Dilemma, the mean vector shows purely an increase in T, benefitting defectors. But as GCT assortment rises to $\beta = 1$, the mean vector turns to one where, while there is still some benefit being accrued to defectors, most of the selective pressure

on the game-changing trait favours game-changing traits that are advantageous to co-operators. This demonstrates how assortment on the game-changing trait is beneficial to future cooperation in a Prisoner's Dilemma, even though it might not be sufficient for the social trait equilibrium to change to all-cooperators.

8.2 The Evolution of Social Assortment Under Increasing GCT Assortment

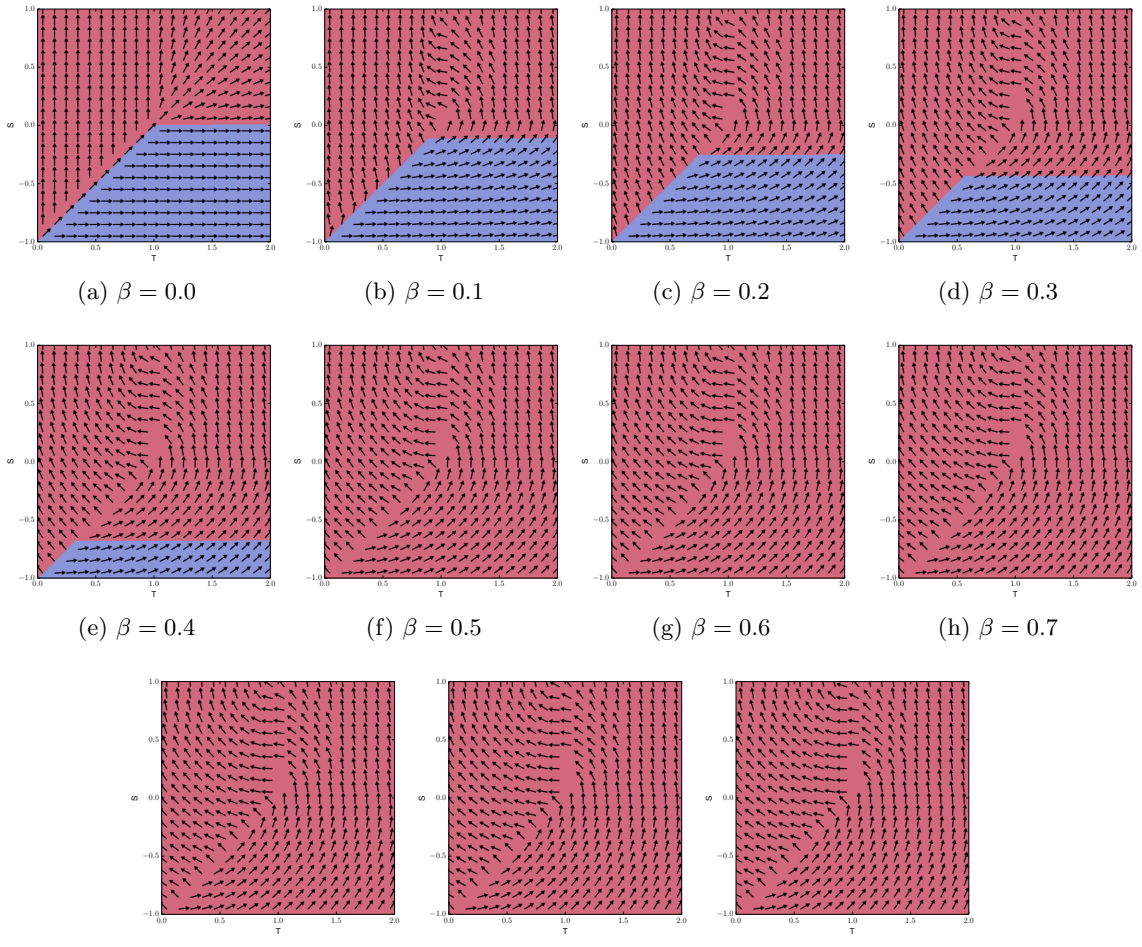
We have previously examined the evolution of game-changing traits that cause increased assortment on the social trait in the case where there is no assortment on the game-changing trait (Section 5.5), using different modelling methodologies for the game-changing trait (Chapter 6), and in the different runs explored with the continuous and discrete models in Chapter 7.

For the scenario with balanced initial conditions, we have already found that:

- When there is no assortment on the game-changing trait, assortment increases if $S > T - 1$. Assortment always decreases in the Prisoner's Dilemma, always increases in the Harmony Game, and splits the Snowdrift and Stag-Hunt regions. If the initial frequency of cooperators was changed, this would look the same, but for the Stag-Hunt region, which would change to match the changed equilibria of the base game.
- A low level of GCT assortment is sufficient for assortment to increase over the whole of the Snowdrift region.
- As GCT assortment increases, the region of the Prisoner's Dilemma for which assortment evolves increases. By $\beta = 0.5$ assortment increases over the whole of the Prisoner's Dilemma.
- This change is not linear – the contours are curved. Games where the strength of selection is lower (which would have a smaller radius if we switched to the circle centred in ST-space) require lower levels of GCT assortment for social trait assortment to always increase than games with the same strategy-equilibrium but larger radius.

Now we assess the behaviour of the assortment metagame over the whole of ST-space where the initial conditions at each point vary according to the equilibrium frequency of the social traits at that point in the space. We assume a low probability of mutations to the social and game-changing traits: $c_\mu = 0.01$, $m_\mu = 0.01$. We still have to decide on one initial condition: the initial frequency of cooperators for the social game at each

point. This affects the size of the all-cooperators and no-cooperators equilibria in the Stag Hunt region. We choose $c = 0.5$ for this initial condition. These conditions mean that across ST-space, the initial frequency of the more assortive M trait will be 0.01. However, the initial frequency of the cooperative trait will vary according to the location in the space: in the Harmony Game quadrant the initial conditions will be almost all cooperators (0.99), while in the Prisoner's Dilemma quadrant there will be almost no cooperators (0.01). The Stag Hunt quadrant is split between these two conditions, while the Snowdrift quadrant has a continuous range of initial frequency of cooperators ranging from low near the Prisoner's Dilemma to high near the Harmony Game quadrant.



(a) $\beta = 0.0$ (b) $\beta = 0.1$ (c) $\beta = 0.2$ (d) $\beta = 0.3$

(e) $\beta = 0.4$ (f) $\beta = 0.5$ (g) $\beta = 0.6$ (h) $\beta = 0.7$

(i) $\beta = 0.8$ (j) $\beta = 0.9$ (k) $\beta = 1.0$

Figure 8.1: ST-space showing the region in which assortment evolves (red) and the underlying vector field as GCT-assortment (β) increases in the invasion at mutated equilibrium frequencies scenario ($c = 0.5$, $c_\mu = 0.01$, $m_\mu = 0.01$).

Figure 8.1 plots the results of the assortment metagame over ST-space for increasing levels of GCT assortment, as well as the vector field. To calculate the assortment metagame we take a lattice of 400×400 points in the space and perform a metagame interaction between the game G at that point in the space and the slightly more socially assorted game G^{δ_α} where $\delta_\alpha = 0.01$. If the frequency of the assorting trait increases

from its initial frequency we colour that point in red, if it decreases we colour the point in blue.

We can compare these results with Figure 6.7 where we computed the assortment metagame from balanced initial conditions. We see that with no GCT assortment, social assortment now increases over the whole Snowdrift quadrant. This is due to the lower initial frequency of mutants 0.01. The region in which social assorting traits decrease coincides exactly with the region where all-defectors is the social equilibrium. As GCT assortment increases, the region where social assorting traits can spread increases parallel to the T-axis. This differs from when we had balanced initial conditions where the boundary between the regions was a smooth curve. However, as in the other models, a GCT assortment level of $\beta = 0.5$ is sufficient for social assorting traits to spread over the whole of ST-space.

We can also calculate the critical value of β required for social trait assortment to evolve (as in Section 7.4.2). For each point in ST-space, given our initial conditions, the critical β value κ_β is the lowest level of assortment on the game-changing trait required for the assorting game-changing trait to increase in frequency. This is plotted in Figure 8.2. We can see how the critical β surface here is a hybrid of those in the balanced and low frequency invasion scenarios (Figure 7.28): it displays the same increase parallel to the T-axis as S becomes lower, but cuts off at the diagonal in the Stag Hunt quadrant as in the balanced initial conditions because we are mutating at equilibrium frequencies obtained from an initial frequency of cooperators of 0.5. The critical β in the extreme point in the Prisoner's Dilemma is still $\kappa_\beta = -0.497487$. This is unsurprising as the initial conditions throughout the Prisoner's Dilemma quadrant in this scenario are the same as those in the low frequency invasion scenario.

We have already substantially analysed the social assortment metagame under other conditions. We see that social trait assortment can evolve easily when the initial conditions already support cooperators. On its own, however, this is a weak result since increased social assortment may be unnecessary when cooperation is already favoured. It is the evolution of assorting traits when the social conditions do not favour cooperation that is interesting, such as in the Prisoner's Dilemma. Here we find that increased assortment on the game-changing trait can lead to the spread of the increased assortment on the social trait, with the higher the assortment on the game-changing trait the greater the region of ST-space where this is possible.

8.3 The Evolution of Social Dilemmas Under GCT Assortment

Modelling the evolution of social trait assortment using the mutated equilibrium frequencies scenario is advantageous because it gives us the confidence that our results are valid

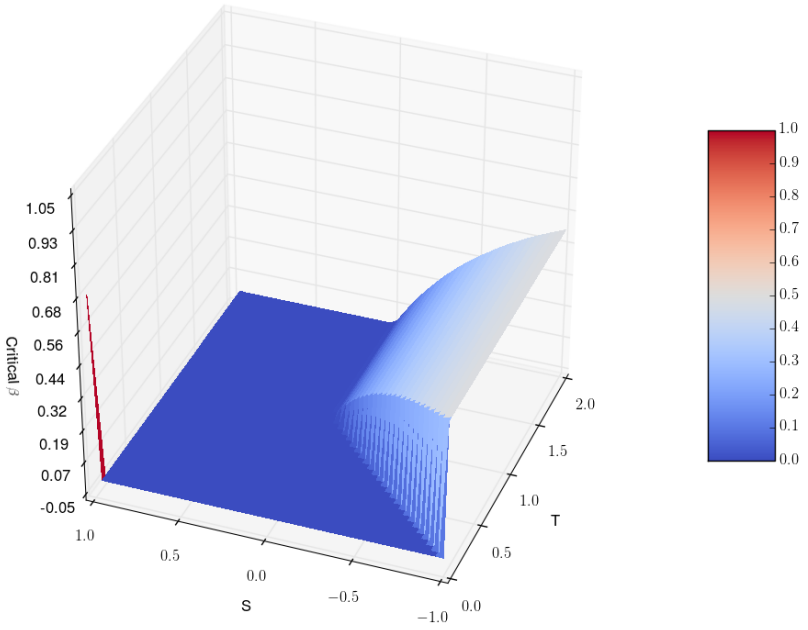


Figure 8.2: The critical β value for the evolution of social assortment over the whole of ST-space for the invasion at mutated equilibrium frequencies scenario ($c = 0.5$, $c_\mu = 0.01$, $m_\mu = 0.01$).

if we translate a series of metagame interactions into a narrative of movement through a space of possible games. We have now developed the techniques and understanding to go further and explicitly model populations taking paths through game space.

The scenario of mutant GCT traits arising in single-GCT populations (playing a single game) that have reached strategy equilibrium allows us to provide a coherent narrative description of the long-term dynamics of the metagames process over the slow timescale of GCT evolution with a minimum of assumptions. At first the population is engaged in a particular evolutionary game. Mutations to game-changing traits that affect the social context result in a subset of the population playing a different effective game. There is then competition, both on the behavioural level between the bearers of different social traits, and on the structural level between bearers of the GCT traits. The spread of the GCT traits will be influenced by the linkage that develops between social strategies and game-changing traits. As this process repeats, the successive fixation of game-changing traits traces the path the population's effective social dilemma takes through the space of possible games (ST-space). This is similar to the narrative of Worden and Levin (2007)'s escape from the Prisoner's Dilemma model, but there a single game is changed by the addition of new strategies; here there is a metagame of competition between different effective games corresponding to game-changing traits.

8.3.1 Model Description

Here we develop a model demonstrating this process of GCT evolution. Instead of mapping the metagame dynamics over ST-space, we have the GCT of a single population evolve by a continual process of small mutation and selection. We expect that the paths the populations take through game space should follow the vector fields that we have constructed. If not it would call into question our modelling work so far.

The model follows the same repeated procedure. We start with a population all possessing some GCT M_0 lying on the ST-plane, so all are playing the game G_{M_0}. Then, repeating this procedure for each GCT M_i:

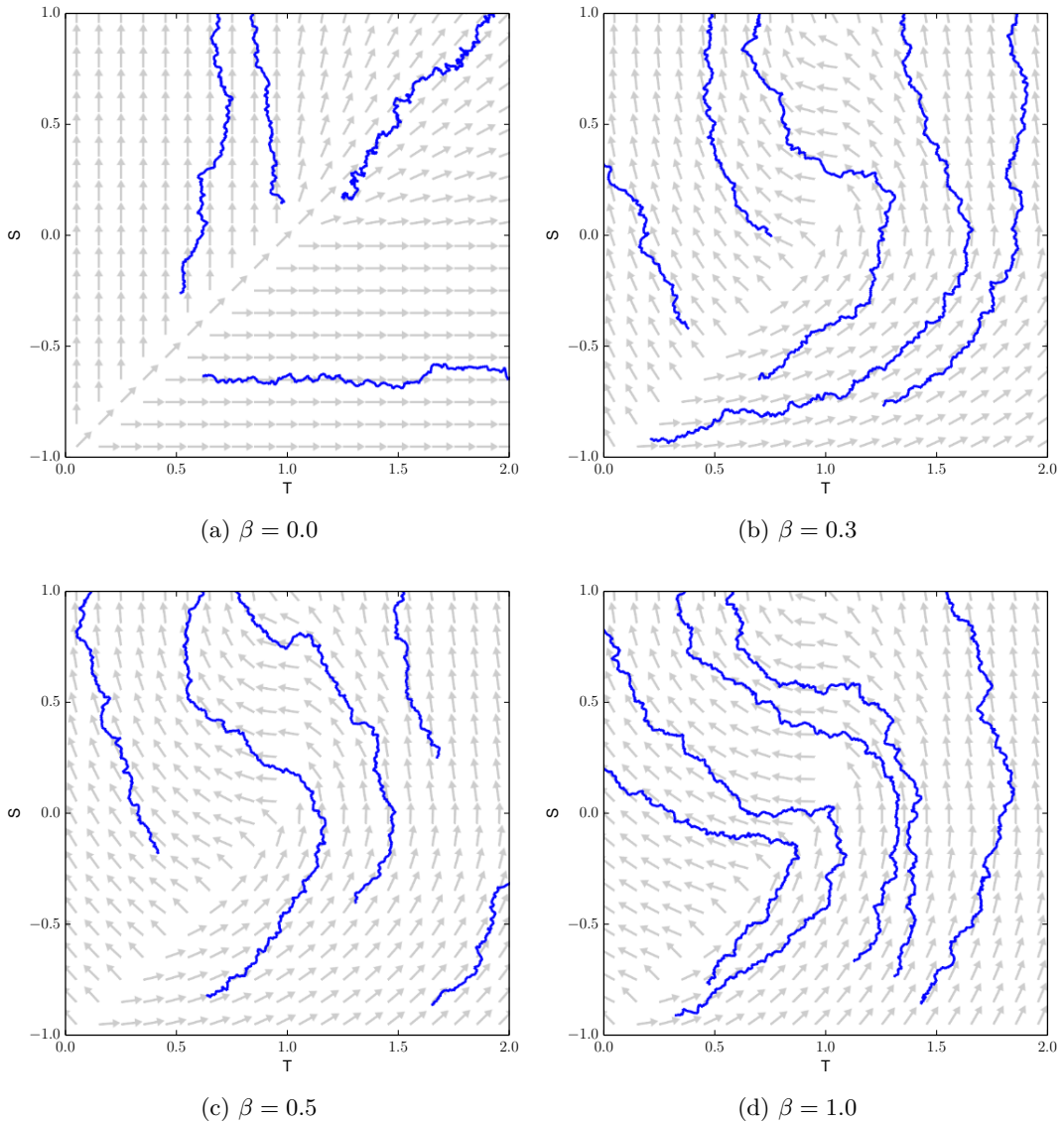
- The whole population possesses the GCT M_i, so they are playing the game G_{M_i}.
- We assume that the frequency of the social traits in the population has reached equilibrium – with x_C cooperators and x_D defectors.
- We pick a new GCT M_{i+1}, lying at a random angle on a small circle of radius r (here $r = 0.01$), centred on the point M_i.
- M_{i+1} is introduced by a low frequency mutation m_μ. Simultaneously, there is also a low frequency mutation in the social traits of c_μ.
- A metagame interaction occurs between G_{M_i} and $G_{M_{i+1}}$ starting from these mutated social equilibrium conditions.
- The metagame interaction is evaluated until it reaches equilibrium.
- We count the GCT trait which has increased in frequency as the winner of the metagame interaction. If the original trait M_i won, then we pick a new M_{i+1} and repeat the process. If M_{i+1} won then this trait displaces M_i, so we increment the index and repeat the cycle, this time with M_{i+1} as the original game.

We repeat this procedure until either the population leaves the bounded region of ST-space of the social dilemmas, or a population has remained at one point in the space for 100 iterations (so no mutant GCTs have been able to displace it). The series of points corresponding to the M_i show the path the population's social dilemma through the space of possible games.

8.3.2 Paths in ST-Space

We run the model multiple times for different values of GCT assortment (β). The amount of GCT assortment is fixed for a model run. We illustrate a selection of the results in Figure 8.3, showing the paths through ST-space from different starting points

and for different levels of GCT assortment. For comparison, the paths (in blue) are shown superimposed on the vector fields over ST-space for the relevant value of β (as in Figure 8.1).



(a) $\beta = 0.0$ (b) $\beta = 0.3$

(c) $\beta = 0.5$ (d) $\beta = 1.0$

Figure 8.3: The ST-plane showing the region in which assortment evolves (red) as GCT assortment (β) increases in the well-mixed scenario ($c = 0.5$, $m = 0.5$).

We see that the paths closely track the vector fields. This is the behaviour we were hoping for, though it was not guaranteed. Remember these vector fields were calculated by taking a lattice of points over ST-space and playing each game against a collection of points spaced equidistantly around each point on the lattice. We then multiply the change in the frequency of the mutant type for each of the mutants on the circle by the vector offset of that point. The result is a vector we interpret as the direction of selection. That these vector fields accurately indicate the behaviour of the paths populations take

through game space supports the modelling work we have done in this thesis to create them.

The close match between these paths and the vector fields means our basic analysis carries over. When there is no GCT assortment the game takes a path through ST-space that maximises the payoffs for the social trait that dominates the region of game space it is in, or in the case of the Snowdrift quadrant, that maximises the combination of both social traits that is the social equilibrium at that point in the Snowdrift. Paths starting in the Prisoner's Dilemma quadrant become more extreme Prisoner's Dilemmas; Harmony Games become more extreme Harmony Games and Snowdrift Games more extreme Snowdrift Games. The exception is the Stag Hunt quadrant: all paths starting in the Stag Hunt quadrant follow the local social equilibria out of the Stag Hunt to become Harmony or Prisoner's Dilemma or Snowdrift games.

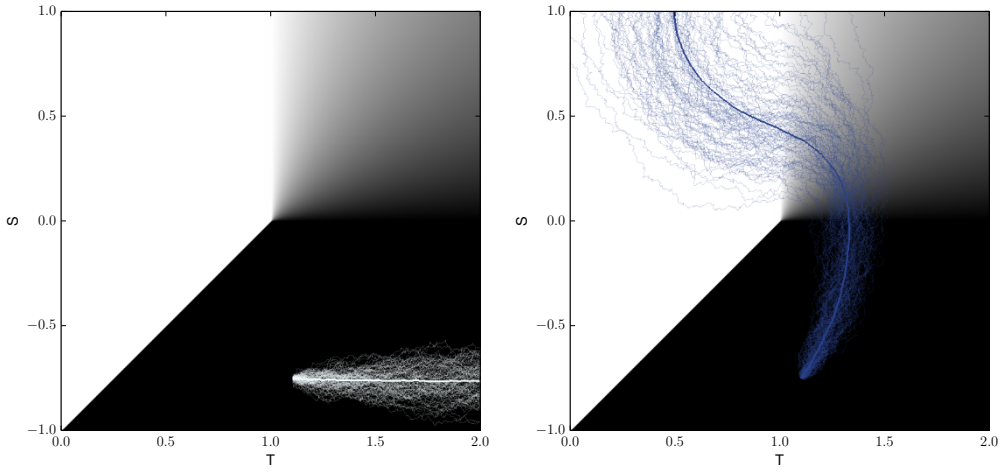
8.3.3 The Mean Path from the Prisoner's Dilemma

We have stressed the importance of the evolution of altruistic behaviours to processes as fundamental to biology as the extreme social integration required for a major transition in evolution(Michod, 2000; Bourke, 2011). Recall that an altruistic behaviour is one in which the social actor loses fitness while the recipient of the act gains fitness (2). In game theoretic terms, cooperating when the social dilemma is a Prisoner's Dilemma is akin to an altruistic act, since a recipient defector will gain in fitness ($T > 0$) while the actor loses fitness ($S < 0$).

From the start then, we have been motivated to show conditions under which a population can evolve game-changing traits that will change the incentive structure of its social interactions sufficiently that the effective social dilemma ceases to be a Prisoner's Dilemma but one in which cooperative acts can be sustained.

We can now model this. We pick a starting point in the Prisoner's Dilemma quadrant of ST-space with $S = -0.75$ and $T = 1.1$. This point is chosen artificially as it is a particularly clear example of what we are looking at, but as we will demonstrate later, this choice does not qualitatively bias our results. There is stochasticity in each path we calculate, since the mutant game-changing trait is a random point on a small circle around the original, so we calculate 100 paths from the starting point for increasing values of GCT assortment (β). We can then find the mean point at each step in the paths (every successful iteration) to plot the mean path. We illustrate this in Figure 8.4 for the case of $\beta = 0.0$ (shown in white) and $\beta = 1.0$ (shown in blue), which show the results of 100 paths from the starting point and the calculated mean path as a thicker line.

Figure 8.5 shows the mean paths for increasing GCT assortment. Note that some of the paths appear to terminate in the middle of the Snowdrift region. This is an artefact



(a) $\beta = 0.0$ (b) $\beta = 1.0$

Figure 8.4: 100 different paths starting at $S = -0.75$, $T = 1.1$, for no GCT assortment ($\beta = 0.0$, white) and full GCT assortment ($\beta = 1.0$, blue), showing the calculated mean path as a thicker line.

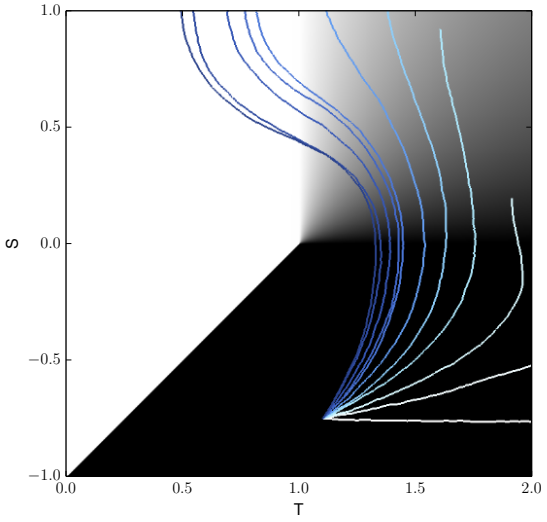


Figure 8.5: The mean paths starting at $S = -0.75$, $T = 1.1$ for increasing GCT assortment (β). When there is no GCT assortment, the mean path is purely an increase in T. As GCT assortment increases the mean path takes the population to increasingly cooperator-friendly regions of ST-space.

of the averaging process: as some paths terminate when they cross $S > 1$ and others terminate when they cross $T > 2$, taking the mean path results in a path that appears to terminate at some point with $S < 1$, $T < 2$. Table 8.3 lists the mean final position of the path, the normalised vector offset from the start to the mean final position and the equilibrium frequency of cooperators for the game at the mean final position.

We see that when there is no GCT assortment, the mean path is directly in the direction of increasing T, so the initial Prisoner's Dilemma evolves into a Prisoner's Dilemma even

β	S	T	S vector	T vector	N.E.
0.0	-0.763	2.004	-0.015	1.000	0.000
0.1	-0.520	2.004	0.247	0.969	0.000
0.2	0.192	1.910	0.758	0.652	0.174
0.3	0.917	1.602	0.958	0.288	0.604
0.4	1.004	1.376	0.988	0.155	0.728
0.5	1.004	1.114	1.000	0.008	0.898
0.6	1.004	0.814	0.987	-0.161	1.000
0.7	1.004	0.765	0.982	-0.188	1.000
0.8	1.004	0.694	0.974	-0.226	1.000
0.9	1.002	0.542	0.953	-0.303	1.000
1.0	1.000	0.492	0.945	-0.328	1.000

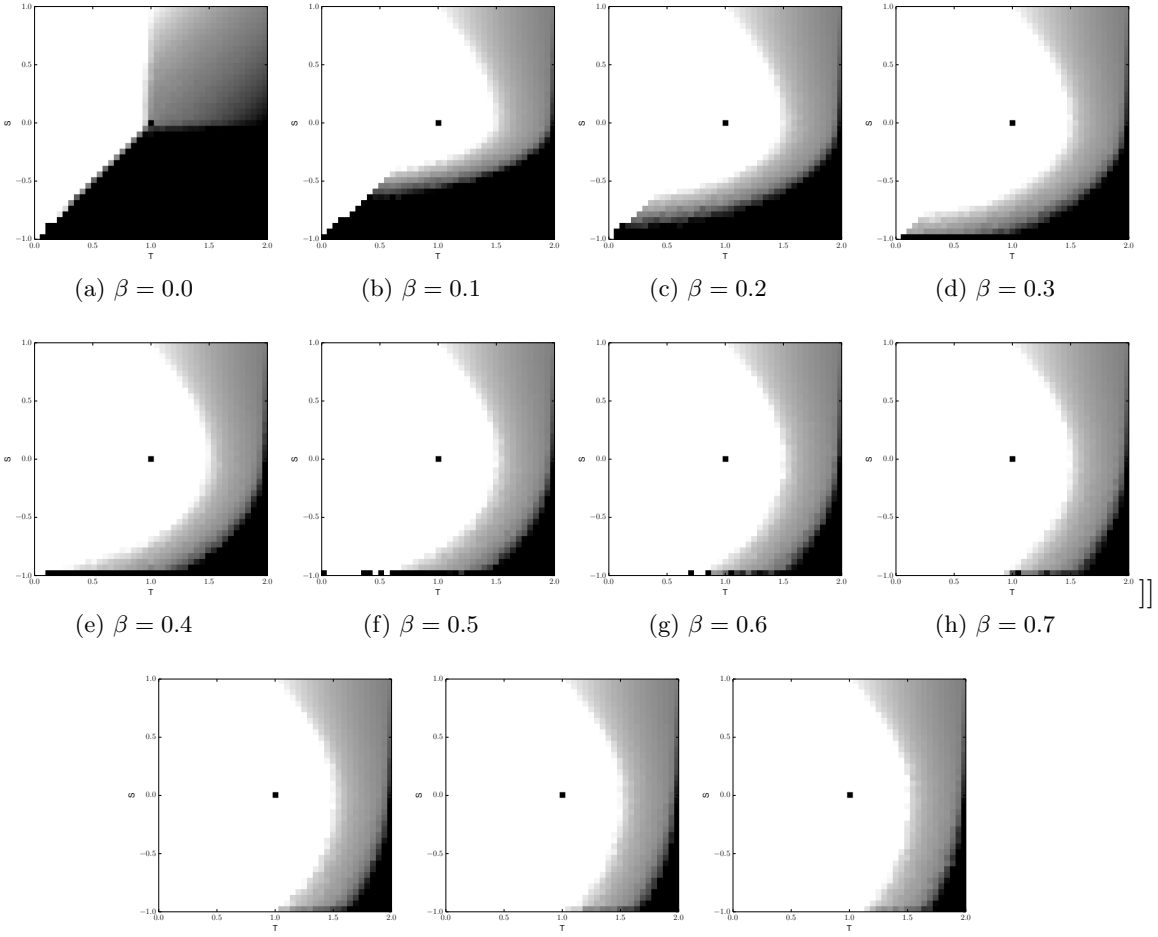
Table 8.3: From a starting point of $S = -0.75$, $T = 1.1$ in the Prisoner's Dilemma, the mean final position, normalised vector of the mean path and Nash Equilibrium frequency of cooperators at the mean final position.

more favourable to defectors. As GCT assortment increases, however, the mean path starts to curve around into the Snowdrift Game (for $\beta \geq 0.2$), and then into the Harmony Game (for $\beta \geq 0.6$) above which all-cooperators is the social equilibrium of the final point on the mean path. This is a dramatic change in the nature of the path the population takes through game space, from one of increasing defection to ultimate cooperation. If a game-changing trait is sufficiently assorted on itself, then even when it can evolve in an unconstrained manner it will do so to a state that supports cooperation. This recreates in metagames the result of the logical argument for social niche construction (Powers et al., 2011) that the initial metagames modelling contradicted (Section 6.1) – at least for this point in the Prisoner's Dilemma.

8.3.4 The Mean Path Over ST-Space

In Chapter 4 we looked at the 'first order' question: how does assortment on the social trait influence the evolution of cooperation (Figure 4.3)? We have investigated the 'second order' question: how does assortment on the game-changing trait influence the evolution of assortment on the social trait? Now we complete this investigation of the triangle of cooperation, assortment on social traits and assortment on game-changing traits by providing a general assessment of how assortment on the game-changing trait influences the evolution of cooperation when the game-changing trait can vary in an unconstrained way.

We do this by calculating paths for points on a lattice spread over ST-space in units 0.05 apart in S and T. For different levels of GCT assortment we calculate the outcome of the path model and the mean path. We can then calculate the resulting equilibrium frequency of cooperators. The results are shown in Figure 8.6.



(a) $\beta = 0.0$ (b) $\beta = 0.1$ (c) $\beta = 0.2$ (d) $\beta = 0.3$

(e) $\beta = 0.4$ (f) $\beta = 0.5$ (g) $\beta = 0.6$ (h) $\beta = 0.7$

(i) $\beta = 0.8$ (j) $\beta = 0.9$ (k) $\beta = 1.0$

Figure 8.6: The equilibrium frequency of cooperators at the end point of the final path for different starting points across ST-space, showing how the basin of attraction for the all-cooperators social equilibrium increases as the level of assortment on the game-changing trait increases.

Table 8.4 shows how increasing assortment on the game-changing trait promotes increased cooperation even when the game-changing trait can evolve in an arbitrary way instead of being constrained to represent a trait for increased assortment.

When there is no assortment on the game changing trait, the basin of attraction for both the all-cooperators and no-cooperators social equilibria are even, with 35.9% and 35.6% of paths over the whole of ST-space ending at the two social equilibria respectively. Unsurprisingly, almost all paths starting in the Prisoner's Dilemma end with no cooperators present at equilibrium. As assortment on the game-changing trait increases, the basin of attraction of all-cooperators grows and the basin of attraction of no-cooperators shrinks, with a corresponding increase in the mean level of cooperation at the end of both paths. In particular, in the Prisoner's Dilemma, even though cooperation is not completely successful – the proportion of games that reaches the all-cooperators equilibria is 36% – the mean equilibrium frequency of cooperators rises from 0.3% to 67.7%

as many Prisoner's Dilemma games become Snowdrift Games.

	ST-Space			Prisoner's Dilemma		
β	All-C	No-C	Mean C	All-C	No-C	Mean C
0.0	0.359	0.356	0.501	0.003	0.975	0.003
0.1	0.489	0.226	0.663	0.085	0.618	0.250
0.2	0.526	0.142	0.726	0.138	0.460	0.362
0.3	0.560	0.093	0.776	0.172	0.343	0.446
0.4	0.595	0.076	0.805	0.207	0.273	0.512
0.5	0.623	0.061	0.823	0.233	0.223	0.558
0.6	0.642	0.043	0.838	0.265	0.177	0.593
0.7	0.652	0.036	0.847	0.287	0.152	0.621
0.8	0.661	0.033	0.853	0.320	0.138	0.646
0.9	0.663	0.030	0.857	0.328	0.125	0.663
1.0	0.673	0.027	0.861	0.360	0.115	0.677

Table 8.4: The proportion of mean paths ending at different social equilibria (all-cooperators and no-cooperators, with the mixed equilibria not shown) and the mean equilibrium frequency of cooperators for paths starting over all of ST-space and just those starting in the Prisoner's Dilemma quadrant.

8.4 Conclusions

When the game-changing trait is restricted to one which affects the level of social assortment, GCT assortment is essential to the spread of social assortment in the Prisoner's Dilemma. Social trait assortment can easily evolve when the social conditions already support cooperators – but when the social conditions already support cooperators then the appearance of cooperation is by definition unsurprising. This is why GCT assortment is important: it enables cooperation-promoting traits for increased social assortment to spread in the Prisoner's Dilemma where the social conditions do not favour cooperation. Game-changing traits for increased social assortment spread across all of ST-space whenever the level of GCT assortment is above $\beta = 0.497$ (except for a few points where $R = S = 1$), so it is sufficient for the game-changing trait to only partially be assorted on itself.

This means that while we can talk about assortment on social traits in the abstract as the ultimate explanation for cooperation, the nature of the proximate mechanisms that create assortment on the social trait are still important because they must generate some assortment on themselves to evolve when the social conditions do not favour cooperation.

In Chapter 4 we saw that social trait assortment is a powerful benefit to the evolution of cooperation because it produces the most favourable continuous transformation of ST-space for cooperators, scaling the space towards the point in it where cooperation is most advantageous. When we could just impose assortment on the social trait, the all-cooperators equilibrium can encompass the whole of ST-space for a level of social

trait assortment above $\alpha = 0.5$ (Figure 4.3), and when the game-changing trait is social trait assortment it is possible for the entirety of ST-space to result in cooperation if the discrete jump is large enough.

When instead the social equilibria change as the result of the evolution of a game-changing trait through random small changes to the effective fitness benefits it provides, increasing GCT assortment increases the basin of attraction of all-cooperator social equilibria and allows all-cooperators equilibria to be reached from effective games that start as Prisoner's Dilemmas. However, it is no longer the case that the whole of ST-space will end at an all-cooperators equilibrium: even at the highest levels of GCT assortment some games in the Prisoner's Dilemma will remain Prisoner's Dilemmas. We see that there are limits to the increases in cooperation that assortment on the game-changing trait will allow.

This unconstrained model also lets us draw a subtler conclusion. When there are no constraints on the possible game-changing traits, so it can change to favour increased or decreased cooperation, or increased total payoff, or any other criteria, then the level of assortment on the game-changing trait changes the way that game-changing trait will evolve. Game-changing traits that would evolve to make the conditions even less favourable to cooperators instead evolve to support cooperation. This means that assortment on the social trait need not drive the evolution of cooperation: if the game-changing trait is assorted on itself then it will be more likely to evolve to favour the spread of increased cooperation.

Chapter 9

Discussion and Conclusions

Why do selfish individuals cooperate for the benefit of others and at a potential cost to themselves? This apparent paradox has been recognised since the beginnings of evolutionary theory (Darwin, 1871). The need to understand the evolution of cooperation has only grown as social behaviours have taken on an increasingly prominent role in evolutionary accounts of the biological world. They are now implicated as drivers of the major evolutionary transitions that created the hierarchical organisation of biological life (Maynard Smith and Szathmáry, 1997; Bourke, 2011; Szathmáry, 2015). Particularly problematic are instances of strong altruism where the altruist experiences a loss of fitness, since by definition altruistic traits are evolutionarily unstable. How can this be reconciled with the observed occurrence of cooperation in the natural world?

The generally accepted resolution to the apparent paradox of cooperation is that cooperative behaviours can be evolutionarily stable if the benefits of the cooperation are directed at other cooperators (Hamilton, 1964b; Eshel and Cavalli-Sforza, 1982; Michod and Sanderson, 1985; Lehmann and Keller, 2006; Fletcher and Doebeli, 2009). Many proximate mechanisms can create this positive assortment, such as relatedness due to common descent, signalling mechanisms ('greenbeards'), spatial structure (Nowak and May, 1992) or iterated interactions (Axelrod, 1987). These diverse mechanisms for creating assortment all explain why cooperation could emerge as an adaptive response to particular social conditions or population structures. Other resolutions to the apparent paradox modify the costs and benefits of the social interaction more directly, through mechanisms like side-payments (Jackson and Wilkie, 2005) or policing (Boyd et al., 2010).

Until recently, social evolution theory has not addressed the corollary to this resolution: how do social contexts that create positive assortment evolve? We have placed this thesis within the emerging paradigm that views the evolution of cooperation and the conditions that enable it as reciprocal: a process of social niche construction (Powers, 2010; Ryan et al., 2016). This approach recognises that social interactions do not occur

against the fixed backdrop of a predefined social context, but are influenced and in turn influence social conditions such as population structures that are themselves evolved features.

Where previous accounts of social evolution might take a particular population structure and determine whether cooperation was likely to evolve in that structure, a social niche construction account also seeks to explain the origins of that population structure, and why the population is living in that structure as opposed to another population structure. Evolved traits such as genetic bottlenecks in propagule based reproduction, limited dispersal and vertical transmission of interspecific mutualists are all ways that organisms can modify their own social environment to change the social context of their interactions.

We have argued why a complete account of the evolution of cooperation must include the evolution of assortment-promoting traits (Rosas, 2010), and how social niche construction aims to do exactly that. Furthermore, since organisms engaged in social niche construction alter the social context of their interactions and then adapt to those changed conditions, positive feedback between the evolution of the social behaviours of organisms in a population and the social context has the potential to create 'runaway' increases in cooperation which may be able to explain the upward shift in the units of evolution during major transitions (Ryan et al., 2016). These factors demonstrate the importance of understanding the linked evolution of social traits and traits that affect the social environment as social niche construction attempts to. We have proposed that metagames provide a mathematical model for social niche construction, and have developed the consequences in this thesis.

9.1 Two Levels of Explanation for the Evolution of Cooperation

We have argued that the evolution of cooperation can be broken down into two problems. The 'first-order' problem is how does cooperation can evolve, the generally accepted answer to which is 'in the presence of positive assortment.' The 'second-order' problem that we have been most concerned with asks how that positive assortment evolves.

The 'first-order' problem has been the traditional concern of social evolution theory. When there is positive assortment, cooperative behaviours (including strong altruism) can survive and spread across generations because the benefits fall on those with a tendency to pass the cooperative traits on (Godfrey-Smith, 2009). In Chapter 4 we analysed this question using interaction functions, a formal method to transform the replicator dynamics to reflect non-random interaction structures in the population. We proved that if the interaction function satisfied certain properties (if it was an affine transformation) then it induced a transformation of the payoff matrix of a game.

Using interaction functions to externally impose transformations to the payoff matrix of social dilemmas corresponding to playing those games with increased levels of assortment, we saw the strong effect of positive assortment on the evolution of cooperation. With an assortment level of $\alpha = 0.5$ on the social trait, even the most extreme Prisoner's Dilemma in ST-space — the game least favourable to cooperation — to switch to an all-cooperators equilibrium. Modelling the transformation due to positive assortment in this way revealed why it is so powerful: mathematically it is a scaling of the entire space of games to the single point in that space (the fully assorted game) most conducive to cooperation. There can be no continuous transformation of the ST-space of games more favourable to cooperation.

Yet in addressing the evolution of cooperation here, our account was incomplete. We imposed a change in the social context and watched cooperation either emerge or fail to emerge as a result. We could not give a full account of the origins of cooperation because we could not account for why such a transformation might occur.

If the answer to the first problem is assortment, the 'second-order' problem is the question of how that assortment evolves. In Chapter 5 we introduced the metagames model, a minimal formalism for the evolution of the payoff matrix of a game through competition between individuals possessing game-changing traits that change the incentives for a social interaction. Since game-changing traits can be represented as transformations to the payoff matrix of a game, this let us turn a model of competition between evolutionary games into a model that can investigate the evolution of game-changing traits.

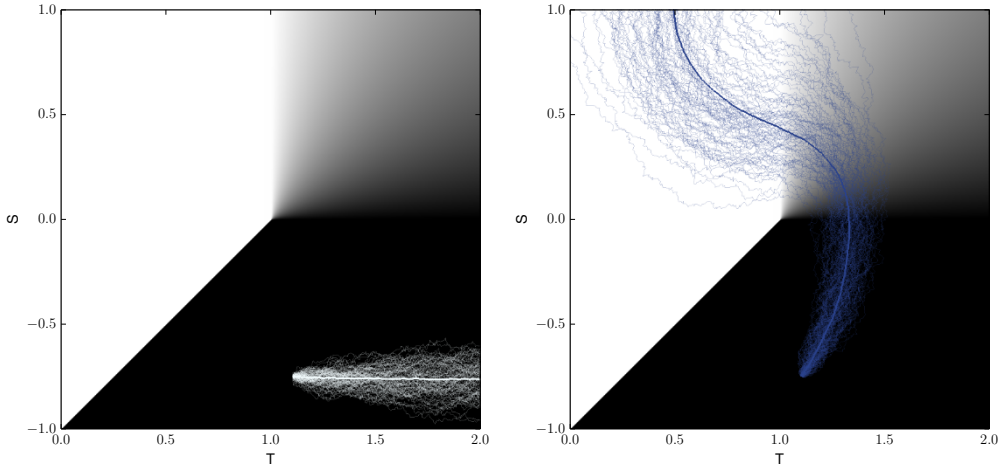
In particular, this allowed us to address the second-order question because the social context no longer has to be imposed. With this model we investigated the conditions under which assortment can evolve (Chapter 5). However, we found that our initial analysis disagreed with the results of other models of social niche construction (Powers et al., 2011): assortment only evolved under conditions already favourable to cooperation. Assortment would spread in the Harmony Game quadrant and in the region of the Stag Hunt quadrant that the initial conditions made favourable to cooperation. It would reach an absolute non-zero frequency in the Snowdrift quadrant, which would mean assortment increased in frequency if it was initially introduced at a low frequency. But assortment did not spread in the Prisoner's Dilemma or the no-cooperator social equilibrium region of the Stag Hunt.

Unpacking the concept of game-changing traits (Chapter 6) led to the realisation that a more sophisticated model of the coevolution of social and game-changing traits needs to include the fact that game-changing traits can induce assortment both on the social strategy trait (α) and on the game-changing trait itself (β). We therefore developed a mathematical representation of GCT assortment within the metagames formalism using interaction functions.

We have used this formulation to characterise the connections between the levels of social and game-changing trait assortment that a game-changing trait for assortment must have to increase in frequency (Chapter 7). We found that social trait assortment in the Prisoner's Dilemma can evolve if there is sufficient assortment on the game-changing trait, but that the assortment on the game-changing trait or the intensity of social assortment the game-changing trait creates must increase by more than a small 'continuous' increase in the social assortment level to start the process: the game-changing trait must be assorted above the critical β for the assorting trait to spread.

The most extreme form of cooperation is (strong) altruism, where the altruist experiences a loss in fitness. In game theoretic terms, being a strong altruist is equivalent to choosing to cooperate when engaged in a Prisoner's Dilemma. The question of when strong altruism can evolve can be translated in our formalism to asking when a population can evolve out of a Prisoner's Dilemma to a region of game-space where all-cooperation is the stable outcome (typically a Harmony Game).

In Chapter 8 we showed how critical the level of GCT assortment is to answering this question: populations starting in the same social niche in the Prisoner's Dilemma can take very different paths through the space of possible games depending on the level of assortment on the game-changing trait. They can end up playing a stronger Prisoner's Dilemma if there is no GCT assortment ($\beta = 0.0$) or a Harmony Game when there is full assortment on the GCT ($\beta = 1.0$) (Figure 9.1).



(a) $\beta = 0.0$ (b) $\beta = 1.0$

Figure 9.1: The trajectories of metagames from a single point in the Prisoner's Dilemma ($S = -0.75$, $T = 1.1$). The evolution of a populations GCT can take a path into an even stronger Prisoner's Dilemma if there is no GCT assortment ($\beta = 0.0$) or become a Harmony Game when there is full assortment on the GCT ($\beta = 1.0$).

This shows the power of assortment on the game-changing trait. When there are no constraints on the way the game-changing trait can evolve, so it could favour increased

cooperation, or increased selfishness, or increased total payoff, the level of assortment on the game-changing trait can determine how it will evolve. Game-changing traits that would evolve to make the conditions even less favourable to cooperators instead evolve to support cooperation if they are self-assorted. In this way, game-changing traits that generate assortment on themselves can drive the evolution of increased cooperation even in Prisoner's Dilemmas.

9.2 Metagames and the Evolution of Game-Changing Traits

The models and techniques we have developed over the course of this thesis allow us to support three claims. The first is a methodological claim: metagames are a model of social niche construction. We have supported this throughout the thesis with the results of our metagame analyses. We can say predictively what happens in evolutionary competitions between populations playing different games under a wide range of conditions in constrained and unconstrained metagames ranging over the classical social dilemmas. We have done this for metagames of mathematical interest, such as where there is constant selection strength (Chapter 5), and biological interest, investigating these conditions in detail for assortment (Chapter 7). Even when a population is initially engaged in a Prisoner's Dilemma, we have shown that if the social conditions vary under suitable constraints in the metagame, it is possible for the game to change to any of the other fundamental two-player games: the Snowdrift, Stag Hunt or Harmony Games. We have also shown that to fully understand the evolution of a game-changing trait, we must understand the kind of metagame that it corresponds to: in particular, the constraints on the way that the game can change (given the mechanics of the specific situation), and the way the game-changing trait affects interactions between the bearers of its different traits.

Second, we claim that while the positive assortment of cooperative behaviours enables a cooperative outcome to be stable, it is insufficient to allow game-changing traits that enable cooperation to evolve. We have identified assortment on a game-changing trait as the crucial factor that determines whether a population can evolve out of a Prisoner's Dilemma (as well as making it easier to evolve to more cooperative states from other games). Evolving out of a Prisoner's Dilemma is particularly significant though because of the connection between the Prisoner's Dilemma and strong altruism. Previous models that have considered the interplay of social traits and equivalents to game-changing traits (such as Karlin and McGregor (1974)'s use of modifier alleles) have not considered this issue. Earlier models of social niche construction also did not fully recognise the importance of assortment on the game-changing trait, instead requiring the condition that on average individuals live in the population structure they genetically prefer. Because assortment on a social-trait and game-changing trait are not fully orthogonal (as seen in the simulation model of Section 6.2), the effects of the two can easily be

conflated. In the abstract setting of a metagames model we have been able to distinguish clearly between the two and separate out their properties. This let us see that 'first-order' assortment was the answer to our first-order question – how cooperation can evolve, while 'second-order' assortment held the key to the second-order question.

If it is really the case that the evolution of assortment is driven by the evolution of assortment on assortment-promoting traits – 'second-order assortment', we might worry that we are merely shifting the problem up a level, threatening an inductive loop of ever-higher levels of assortment proposed to explain ever-higher level questions.

There are two responses to this. The first and most important is that assortment on the game-changing trait is a causally potent level. As we have shown, the presence or absence of GCT assortment can determine whether a population ends up playing a Prisoner's Dilemma or a Harmony Game – it can change the final equilibrium from one of all-defectors to all-cooperators. This is sufficient to demonstrate the importance of GCT assortment even if an infinite number of higher-order levels of assortment were to exist. The second is that assortment on social and game-changing traits are distinct, but not completely orthogonal. For example, a game-changing trait that creates high levels of relatedness in a population, such as by limiting dispersal, will create assortment on both the social and game-changing traits. A game-changing trait for a group size preference will create assortment on the game-changing trait because individuals will form groups with others with similar group size preferences, but only create assortment on the social trait in some small groups due to sampling error from a large meta-population.

Our third claim follows on from the second: we can characterise the conditions when a population can evolve out of a Prisoner's Dilemma. Indeed, our results are broader than this, letting us show how constrained and unconstrained games change in ST-space in a general model. By a general model we do not mean a completely unifying model that subsumes all other models within it, but in the sense that it can be applied to a general class of games. Therefore we have a theoretical framework that allows us to characterise the conditions necessary for game-changing traits that promote cooperation to evolve.

9.3 Modifying Previous Accounts

Before performing this modelling work, we could envisage different possible outcomes when trying to evolve cooperation-supporting traits in a Prisoner's Dilemma. There was the theory we explored in our first models of assortment metagames: that assortment could evolve in a Prisoner's Dilemma even when the equilibria of the social game is all-defectors because when the population is on the transient to the social equilibrium, the cooperating assorters would drag the population towards increased assortment. Another possible account would be that game-changing traits that decrease the rate at which an

all-defectors equilibrium is approached might be preferred, because even though cooperators are extinct when the social equilibrium is reached, cooperators obtain the differential benefits of cooperation on the transient. This would suggest that the reason a population might not be able to escape a Prisoner's Dilemma is that it loses all 'social mobility' when the frequency of the social traits reaches equilibrium – but cooperators will still benefit until this point.

However, we have shown that this is not the case because there is a crucial step that our metagame analyses force us to consider: the consequences within the population of bearers of a cooperative social trait having greater fitness. In Chapter 7 we saw that when the effective game is one in which cooperation is unfavoured, then up to a critical threshold traits that promote assortment and therefore aid cooperators actually have negative consequence for the assorting cooperators. In these conditions, a longer lasting pool of cooperators gives more time for defectors to 'prey' on the cooperators — obtaining the greater reward for unilateral defection against cooperators than for defecting against other defectors. More cooperators gives the defectors whose game-changing trait provides them with a higher temptation to defect more opportunity to differentiate themselves from those defectors with lower temptation to defect. So while cooperators possessing a mutant game-changing trait that promotes increased cooperation may see some benefit, it will be far outweighed by the differential growth in the less-assorting defectors. The result is a form of Simpson's paradox, with assorting cooperators coming to dominate the subpopulation of cooperators, but the total number of cooperators falling within the population as a whole meaning the assorting trait decreases in frequency. We have been able to find the critical thresholds that the levels of assortment on the social or game-changing traits must cross for this to no longer be the case, and so characterise the required levels of assortment on social and game-changing traits for 'runaway' social niche construction to occur.

As well as characterising when cooperation will evolve, we have shown when it will not. Given the apparent paradox of cooperation, there have been many attempt to demonstrate when cooperation might evolve, and far fewer seeking to place limits on this since defection was seen as the default state. The demonstration of so many proximate mechanisms for achieving cooperation might lead us to breathlessly assume cooperation is even inevitable, but in general, assortment does not evolve in the Prisoner's Dilemma unless there is sufficiently high assortment on the game-changing trait (Section 8.2). The same is true for a Stag Hunt game: assortment does not evolve and cooperation does not spread unless there is the requisite level of GCT assortment or a high initial frequency of cooperators. And when the game-changing trait can evolve in an unconstrained manner, though more cooperative games are favoured, games starting in the Prisoner's Dilemma do not always evolve to reach cooperation-supporting social equilibria.

9.4 Concluding Remarks

We have found consistently though not universally that increasing assortment on game-changing traits supports the evolution of game-changing traits that promote cooperation. Though it would be nice to have a simple conclusion, such as 'population structures that support increased cooperation will always spread,' exhaustively characterising the conditions of the metagames model has shown that this is not the case. From an initial Prisoner's Dilemma, we have shown that if the metagame has suitable constraints it is possible for the game to change to any of the other fundamental two-player games: the Snowdrift, Stag Hunt or Harmony Games. To fully understand the evolution of a game-changing trait, we must understand what these constraints are.

Our modelling work has let us identify the crucial and previously hidden role of assortment on the game-changing trait. While the positive assortment of cooperative behaviours can help a cooperative outcome to be stable, they are insufficient to allow the game-changing traits that enable cooperation to evolve. Previous authors have stated that strong altruism (cooperation in a Prisoner's Dilemma) requires assortment of social behaviours to evolve and have emphasised the importance of achieving this assortment over the particular mechanism for assorting, but we find that this is not the full picture. If assortment is imposed externally, as we did in Chapter 4, then it is true that no concept of GCT assortment is required. But assortment on the game-changing trait is required for social traits that promote assortment to evolve in the Prisoner's Dilemma.

When the assortment on the social trait is no longer externally imposed, it is this 'second-order' assortment on the game-changing traits that promote assortment that enables a population of selfish individuals engaged in a social dilemma requiring strong altruism to modify their social niche to create to a stable equilibrium of all-cooperators – but if these conditions are not met the social dilemma may become even less favourable to cooperation. This means that the nature of the particular mechanism by which a social assortment-promoting game-changing trait creates assortment does matter, since different game-changing traits for increased social assortment can create different levels of assortment on themselves.

Our analysis has suggested possible pathways for the evolution of altruism. For example, though greenbeard signalling mechanisms are unstable and subject to parasitism, such mechanisms could cause the population to reach sufficient levels of GCT assortment to reach regions of the α-β parameter space that facilitate the evolution of stable paths to social trait assortment such as relatedness. Weak assortment on the game-changing trait could therefore provide a path to strong social trait assortment.

There are many avenues for future work. We have confined our analysis to two-player two-strategy games on the ST-plane. These games include the canonical social dilemmas, so they have been a sufficiently rich substrate for our purposes, but the metagame model

could extend to more broader classes of evolutionary game. We have also limited the analysis to asexual populations. Introducing sexual reproduction would affect the model since recombination would reduce the effects of linked social and game-changing traits.

We have seen that the concept of social niche construction allows us to move beyond viewing the evolution of cooperation as an adaption to a particular social context to an active process that can create that context. By explaining how individual traits can align group fitness interests, social niche construction can give bottom up accounts for evolutionary transitions in individuality (Ryan et al., 2016). We have argued that the theory of metagames we have developed in this thesis serves as a mathematical model for social niche construction and supported this with the results of an extensive body of modelling work. We have shown how we can characterise the relationship between the 'first order' social assortment created by assorting game-changing traits and the 'second order' assortment on the game-changing traits, and found the critical thresholds required for social niche construction to take place.

But we have also shown for unconstrained game-changing traits that the assortment on the game-changing trait plays more than just a facilitating role. GCT assortment does not just allow cooperation-promoting game-changing traits to evolve, it makes game-changing traits that would potentially lead to greater defection instead evolve towards supporting greater cooperation. We have therefore shown that assortment on game-changing traits can drive the evolution of the conditions for the evolution of cooperation.

Bibliography

Akçay, E. and Roughgarden, J. (2011). The evolution of payoff matrices: providing incentives to cooperate. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715):2198–2206.

Archetti, M. and Scheuring, I. (2012). Review: Game theory of public goods in one-shot social dilemmas without assortment. *Journal of Theoretical Biology*, 299:9–20.

Axelrod, R. (1987). The evolution of strategies in the iterated prisoner's dilemma. *Genetic algorithms and simulated annealing*, pages 32–41.

Axelrod, R. and Hamilton, W. (1981). The evolution of cooperation. *Science*, 211(4489):1390.

Ayre, D. J. and Grosberg, R. K. (2005). Behind anemone lines: factors affecting division of labour in the social cnidarian anthopleura elegantissima. *Animal behaviour*, 70(1):97–110.

Berg, R. (1996). The indigenous gastrointestinal microflora. *Trends in microbiology*, 4(11):430–435.

Biernaskie, J., West, S., and Gardner, A. (2011). Are greenbeards intragenomic outlaws? *Evolution*, 65(10):2729–2742.

Binmore, K. (1992). *Fun and Games: A Text on Game Theory*. DC Heath and Company.

Birch, J. (2012a). Collective action in the fraternal transitions. *Biology & Philosophy*, 27(3):363–380.

Birch, J. (2012b). Social revolution. *Biology & Philosophy*, 27(4):571–581.

Bonner, J. (1988). *The evolution of complexity by means of natural selection*. Princeton Univ Pr.

Bourke, A. (2011). *Principles of Social Evolution*. Oxford University Press, USA.

Boyd, R., Gintis, H., and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328(5978):617–620.

Brand, S. (1974). *II cybernetic frontiers*. Random House.

Breden, F. and Wade, M. J. (1991). runaway social evolution: Reinforcing selection for inbreeding and altruism. *Journal of Theoretical Biology*, 153(3):323–337.

Buss, L. (1987). *The Evolution of Individuality*. Princeton University Press Princeton.

Cao, L., Ohtsuki, H., Wang, B., and Aihara, K. (2011). Evolution of cooperation on adaptively weighted networks. *Journal of Theoretical Biology*, 272(1):8–15.

Carter, M., Gibbs, M., and Harrop, M. (2012). Metagames, paragames and orthogames: A new vocabulary. In *Proceedings of the international conference on the foundations of digital games*, pages 11–17. ACM.

Clarke, E. (2014). Origins of evolutionary transitions. *Journal of Biosciences*, 39(2):303–317.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. John Murray, London.

Dawkins, R. and Krebs, J. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):489–511.

Dawklns, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Doncaster, C., Jackson, A., and Watson, R. (2013a). Manipulated into giving: when parasitism drives apparent or incidental altruism. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758).

Doncaster, C. P., Jackson, A., and Watson, R. A. (2013b). Competitive environments sustain costly altruism with negligible assortment of interactions. *Scientific reports*, 3.

Eshel, I. and Cavalli-Sforza, L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, 79(4):1331.

Fletcher, J. A. and Doebeli, M. (2009). A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society B: Biological Sciences*, 276(1654):13–19.

Fletcher, J. A. and Zwick, M. (2006). Unifying the theories of inclusive fitness and reciprocal altruism. *The American Naturalist*, 168(2):252–262.

Folse, H. and Roughgarden, J. (2010). What is an individual organism? a multilevel selection perspective. *The Quarterly Review of Biology*, 85(4):447–472.

Fort, H. (2008). A minimal model for the evolution of cooperation through evolving heterogeneous games. *EPL (Europhysics Letters)*, 81:48008.

Gardner, A. and West, S. (2006a). Demography, altruism, and the benefits of budding. *Journal of Evolutionary Biology*, 19(5):1707–1716.

Gardner, A. and West, S. (2006b). Spite. *Current biology: CB*, 16(17):R662.

Gardner, A. and West, S. (2010). Greenbeards. *Evolution*, 64(1):25–38.

Geller, J. B. and Walton, E. D. (2001). Breaking up and getting together: evolution of symbiosis and cloning by fission in sea anemones (genus anthopleura). *Evolution*, 55(9):1781–1794.

Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection*. Oxford University Press, USA.

Grafen, A. (1979). The hawk-dove game played between relatives. *Animal Behaviour*, 27:905–907.

Grafen, A. (2006). Optimization of inclusive fitness. *Journal of Theoretical Biology*, 238(3):541–563.

Hamilton, W. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16.

Hamilton, W. (1964b). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1):17–52.

Hamilton, W. (1975). Innate social aptitudes of man: an approach from evolutionary genetics. *Biosocial anthropology*, 133:155.

Hemelrijk, C. K. (1999). An individual–orientated model of the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1417):361–369.

Herron, M. and Michod, R. (2008). Evolution of complexity in the volvocine algae: transitions in individuality through darwin's eye. *Evolution*, 62(2):436–451.

Hines, W. G. S. and Smith, J. M. (1979). Games between relatives. *Journal of Theoretical Biology*, 79(1):19–30.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge Univ Press.

Hölldobler, B. and Wilson, E. (1990). *The Ants*. Belknap Press.

Howard, N. (1971). *Paradoxes of Rationality: Theory of Metagames and Political Behavior*. MIT Press.

Howard, N. (1974). 'General' metagames: An extension of the metagame concept. In *Game Theory as a Theory of a Conflict Resolution*, pages 261–283. Springer.

Howard, N. (1976). Prisoner's dilemma: The solution by general metagames. *Behavioral Science*, 21(6):524–531.

Jackson, A. and Watson, R. A. (2013). The effects of assortment on population structuring traits on the evolution of cooperation. In *Advances in Artificial Life, ECAL*, volume 12, pages 356–363.

Jackson, A. and Watson, R. A. (In Prep). Metagames: A formal framework for the evolution of game-changing behaviours.

Jackson, M. O. and Wilkie, S. (2005). Endogenous games and mechanisms: Side payments among players. *The Review of Economic Studies*, 72(2):543–566.

Johnson, M. W. and Leydesdorff, L. (2015). Beer's viable system model and luhmann's communication theory:organizations from the perspective of meta-games. *Systems Research and Behavioral Science*, 32(3):266–282.

Karlin, S. and McGregor, J. (1974). Towards a theory of the evolution of modifier genes. *Theoretical population biology*, 5(1):59–103.

Keller, L. and Ross, K. G. (1998). Selfish genes: a green beard in the red fire ant. *Nature*, 394(6693):573–575.

Laland, K. N. and Sterelny, K. (2006). Perspective: seven reasons (not) to neglect niche construction. *Evolution*, 60(9):1751–1762.

Lehmann, L. and Keller, L. (2006). The evolution of cooperation and altruism – a general framework and a classification of models. *Journal of Evolutionary Biology*, 19(5):1365–1376.

Leigh Jr, E. (1995). The major transitions of evolution. *Evolution*, 49(6):1302–1306.

Levin, S. A. (2014). Public goods in relation to competition, cooperation, and spite. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10838–10845.

Lieberman, E., Hauert, C., and Nowak, M. (2005). Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316.

Macy, M. and Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7229.

May, R. (1981). The evolution of cooperation. *Naturey*, 292:291–292.

Maynard Smith, J. (1978). Optimization theory in evolution. *Annual Review of Ecology and Systematics*, pages 31–56.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge Univ Pr.

Maynard Smith, J. (1998). The origin of altruism. *Nature*, 393:639–640.

Maynard Smith, J. and Price, G. (1973). The logic of animal conflict. *Nature*, 246(5427):15–18.

Maynard Smith, J. and Szathmáry, E. (1997). *The Major Transitions in Evolution*. Oxford University Press, USA.

Mesterton-Gibbons, M. (1991). An escape from 'the prisoner's dilemma'. *Journal of Mathematical Biology*, 29(3):251–269.

Michod, R. (2000). *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*. Princeton University Press.

Michod, R. (2007). Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104(Suppl 1):8613.

Michod, R. and Herron, M. (2006). Cooperation and conflict during evolutionary transitions in individuality. *Journal of Evolutionary Biology*, 19(5):1406–1409.

Michod, R. and Roze, D. (2001). Cooperation and conflict in the evolution of multicellularity. *Heredity*, 86(1):1–7.

Michod, R., Viossat, Y., Solari, C., Hurand, M., and Nedelcu, A. (2006). Life-history evolution and the origin of multicellularity. *Journal of Theoretical Biology*, 239(2):257–272.

Michod, R. E. and Sanderson, M. (1985). Behavioural structure and the evolution of cooperation. In Greenwood, P. J., Harvey, P. H., and Slatkin, M., editors, *Evolution: essays in honor of John Maynard Smith*, pages 95–104. Cambridge University Press.

Nash, J. (1951). Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295.

Nowak, M. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Belknap Press.

Nowak, M. and May, R. (1992). Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829.

Odling-Smee, F. J., Laland, K. N., and Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution*. Number 37 in Monographs in Population Biology. Princeton University Press.

Ohtsuki, H. and Nowak, M. A. (2006). The replicator equation on graphs. *Journal of Theoretical Biology*, 243(1):86–97.

Okasha, S. (2002). Genetic relatedness and the evolution of altruism. *Philosophy of Science*, 69(1):138–149.

Okasha, S. (2006). *Evolution and the Levels of Selection*. Oxford University Press.

Olinick, M. (1981). Mathematical models in the social and life sciences: a selected bibliography. *Mathematical Modelling*, 2(3):237–258.

Pacheco, J., Traulsen, A., and Nowak, M. (2006a). Active linking in evolutionary games. *Journal of Theoretical Biology*, 243(3):437–443.

Pacheco, J., Traulsen, A., and Nowak, M. (2006b). Coevolution of strategy and structure in complex networks with dynamical linking. *Physical Review Letters*, 97(25):258103.

Pearse, V. and Francis, L. (2000). *Anthopleura sola*, a new species, solitary sibling species to the aggregating sea anemone, *A. elegantissima* (Cnidaria: Anthozoa: Actiniaria: Actiniidae). *Proceedings of the Biological Society of Washington*, 113(3):596–608.

Pepper, J. and Smuts, B. (2002). A mechanism for the evolution of altruism among nonkin: Positive assortment through environmental feedback. *The American Naturalist*, 160(2):205–213.

Pepper, J. W. (2000). Relatedness in trait group models of social evolution. *Journal of Theoretical Biology*, 206(3):355–368.

Pinheiro, F. L., Pacheco, J. M., and Santos, F. C. (2012). From local to global dilemmas in social networks. *PLoS One*, 7(2):e32114.

Powers, S. (2010). *Social Niche Construction: Evolutionary Explanations for Cooperative Group Formation.* PhD thesis, University of Southampton.

Powers, S., Penn, A., and Watson, R. (2011). The concurrent evolution of cooperation and the population structures that support it. *Evolution*, 65(6):1527–1543.

Powers, S. and Watson, R. (2011). Evolution of individual group size preference can increase group-level selection and cooperation. *Advances in Artificial Life. Darwin Meets von Neumann*, pages 53–60.

Queller, D. and Strassmann, J. (2009). Beyond society: the evolution of organismality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3143–3155.

Queller, D. C. (1985). Kinship, reciprocity and synergism in the evolution of social behaviour. *Nature*, 318(6044):366–367.

Queller, D. C. (1997). Cooperators since life began. *The Quarterly Review of Biology*, 72(2):184–188.

Queller, D. C., Ponte, E., Bozzaro, S., and Strassmann, J. E. (2003). Single-gene greenbeard effects in the social amoeba *Dictyostelium discoideum*. *Science*, 299(5603):105–106.

Rapoport, A. (1985). Applications of game-theoretic concepts in biology. *Bulletin of Mathematical Biology*, 47(2):161–192.

Reeve, H. and Jeanne, R. (2003). From individual control to majority rule: extending transactional models of reproductive skew in animal societies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1519):1041–1045.

Ridley, M. and Grafen, A. (1981). Are green beard genes outlaws? *Animal Behaviour*, 29.

Rosas, A. (2010). Beyond inclusive fitness? on a simple and general explanation for the evolution of altruism. *Philosophy & Theory in Biology*, 2.

Roughgarden, J. (2009). *The Genial Gene: Deconstructing Darwinian Selfishness*. Univ of California Press.

Ryan, P., Powers, S. T., and Watson, R. A. (2016). Social niche construction and evolutionary transitions in individuality. *Biology & philosophy*, 31(1):59–79.

Santelices, B. (1999). How many kinds of individual are there? *Trends in ecology & evolution*, 14(4):152–155.

Santos, F., Pacheco, J., and Lenaerts, T. (2006a). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3490.

Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006b). Cooperation prevails when individuals adjust their social ties. *PLoS Comput Biol*, 2(10):e140.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Snowdon, J., Powers, S., and Watson, R. (2011). Moderate contact between subpopulations promotes evolved assortativity enabling group selection. *Advances in Artificial Life. Darwin Meets von Neumann*, pages 45–52.

Sober, E. and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Stewart, A. J. and Plotkin, J. B. (2014). Collapse of cooperation in evolving games. *Proceedings of the National Academy of Sciences*, 111(49):17558–17563.

Strogatz, S. (1994). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Pr.

Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. *Proceedings of the National Academy of Sciences*, page 201421398.

Tanimoto, J. and Sagara, H. (2007). Relationship between dilemma occurrence and the existence of a weakly dominant strategy in a two-player symmetric game. *BioSystems*, 90(1):105–114.

Taylor, C. and Nowak, M. (2007). Transforming the dilemma. *Evolution*, 61(10):2281–2292.

Taylor, P. and Jonker, L. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156.

Thompson, J. (2005). *The Geographic Mosaic of Coevolution*. University of Chicago Press.

Traulsen, A., Santos, F., and Pacheco, J. (2009). Evolutionary games in self-organizing populations. *Adaptive Networks*, pages 253–267.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, pages 35–57.

Tudge, S., Watson, R., and Brede, M. (2013). Cooperation and the division of labour. In *Advances in Artificial Life, ECAL 2013: Proceedings of the Twelfth European Conference on the Synthesis and Simulation of Living Systems*.

Turner, P. and Chao, L. (2003). Escape from prisoners dilemma in RNA phage Φ6. *The American Naturalist*, 161(3):497–505.

Van Dyken, J. and Wade, M. (2012a). Origins of altruism diversity I: the diverse ecological roles of altruistic strategies and their evolutionary responses to local competition. *Evolution*, 66(8):2484–2497.

Van Dyken, J. and Wade, M. (2012b). Origins of altruism diversity II: Runaway coevolution of altruistic strategies via"reciprocal niche constructio". *Evolution*, 66(8):2498–2513.

Van Schaik, C. (1983). Why are diurnal primates living in groups? *Behaviour*, pages 120–144.

Van Veelen, M. (2011). The replicator dynamics with n players and population structure. *Journal of Theoretical Biology*, 276(1):78–85.

Velicer, G. J. (2003). Social strife in the microbial world. *Trends in microbiology*, 11(7):330–337.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Watson, R., Palmius, N., Mills, R., Powers, S., and Penn, A. (2011a). Can selfish symbioses effect higher-level selection? *Advances in Artificial Life. Darwin Meets von Neumann*, pages 27–36.

Watson, R. A., Mills, R., Buckley, C., Kouvaris, K., Jackson, A., Powers, S. T., Cox, C., Tudge, S., Davies, A., Kounios, L., et al. (2015). Evolutionary connectionism:

algorithmic principles underlying the evolution of biological organisation in evo-devo, evo-eco and evolutionary transitions. *Evolutionary Biology*, pages 1–29.

Watson, R. A., Mills, R., and Buckley, C. L. (2011b). Global adaptation in networks of selfish components: Emergent associative memory at the system scale. *Artificial Life*, 17(3):147–166.

Weibull, J. (1997). *Evolutionary Game Theory*. MIT press.

Wenseleers, T., Helanterä, H., Hart, A., and Ratnieks, F. L. (2004). Worker reproduction and policing in insect societies: an ESS analysis. *Journal of Evolutionary Biology*, 17(5):1035–1047.

West, S., Griffin, A., and Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2):415–432.

West, S. A., El Mouden, C., and Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, 32(4):231–262.

Wilson, D. and Colwell, R. (1981). Evolution of sex ratio in structured demes. *Evolution*, pages 882–897.

Wilson, E. et al. (1971). *The Insect Societies*. Harvard University Press.

Worden, L. and Levin, S. (2007). Evolutionary escape from the prisoner's dilemma. *Journal of Theoretical Biology*, 245(3):411–422.

Zollman, K. J. (2008). Explaining fairness in complex environments. *Politics, philosophy & economics*, 7(1):81–97.