Modeling human-coupled common pool resource systems with techniques in evolutionary game theory and reinforcement learning

by

Isaiah Farahbakhsh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Chair:	Matthew Scott Professor, Dept. of Applied Mathematics, University of Waterloo
Supervisor(s):	Chris T. Bauch Professor, Dept. of Applied Mathematics, University of Waterloo Madhur Anand Professor, Dept. of Environmental Sciences, University of Guelph
Other Member(s):	Chrystopher L. Nehaniv Professor, Dept. of Systems Design Engineering, University of Waterloo Jesse Hoey Professor, Cheriton School of Computer Science, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Isaiah Farahbakhsh was the sole author for Chapters 1, 3, and 4 which were written under the supervision of Dr. Chris Bauch and Dr. Madhur Anand and were not written for publication. This thesis consists in part of one manuscript written for publication.

Exceptions to sole authorship of material are as follows:

Research presented in Chapter 2:

This research was conducted at the University of Waterloo by Isaiah Farahbakhsh under the supervision of Dr. Chris Bauch and Dr. Madhur Anand. Isaiah Farahbakhsh, Dr. Chris Bauch and Dr. Madhur Anand contributed to the model conception and formulation. Isaiah Farahbakhsh was solely responsible for coding. Isaiah Farahbakhsh conducted the analysis with advice from Dr. Chris Bauch. Isaiah Farahbakhsh wrote the draft manuscripts, which all co-authors contributed intellectual input on.

Citations:

Chapter 2: Farahbakhsh, I., Bauch, C. T., & Anand, M. (2021). Best response dynamics improve sustainability and equity outcomes in common-pool resources problems, compared to imitation dynamics. Journal of Theoretical Biology, 509, 110476.
https://doi.org/10.1016/j.jtbi.2020.110476

Abstract

Shared resource extraction among profit-seeking individuals involves a tension between individual benefit and the collective well-being represented by the persistence of the resource. In these systems, the decisions of rational agents have been modeled from a game theoretic, and more recently, a reinforcement learning approach. Within game theoretic models, the mechanisms used for learning dynamics are often assumed, and the influence of the type of learning dynamics are not systematically compared under identical models. Models using reinforcement learning techniques are a relatively recent addition to this field, and the literature on multi-agent systems with spatial structure is very sparse. This thesis presents two common pool resource models, each using one of these two different approaches.

In the second chapter, an evolutionary common pool resource game is simulated on a social network with payoff functions that depend on the state of the resource. Model predictions under two types of learning, best response and imitation dynamics are compared and it is shown that best response dynamics lead to an increase in sustainability of the system, the persistence of cooperation while decreasing inequality and debt. Given the strikingly different outcomes for best response versus imitation dynamics for common-pool resource systems, our results suggest that modellers should choose strategy update rules that best represent decision-making in their study systems.

In the third chapter, an analogous model to the one above is presented, however it uses reinforcement learning techniques to inform the agents' harvesting decisions. Here, the harvesting strategies of the agents are learned, rather than prescribed a priori, and the payoff function is the weighted sum of a profit goal and a social conforming goal. Preliminary results show that an increased cost of harvesting has a positive effect on the resource level and sustainability of the system, however, a high cost parameter brings the system to an unprofitable state where agents harvest above the analytically derived optimal level. Additionally, the effect of the weight of the conforming goal shows contradictory outcomes, which are highly dependent on the profitability of the system. These different outcomes are posited to be due to strong social conformity amplifying existing trends in the social dynamics.

Results from both chapters demonstrate the profound effect human learning models can have on common-pool resource systems, as well as the potential for sustainable outcomes to emerge among a non-hierarchical system of self-interested agents.

Acknowledgements

I would like to thank my family; Mona Farahbakhsh, Carmel Farahbakhsh, Rachel Farahbakhsh and Khosrow Farahbakhsh for their boundless love, encouragement and support.

I would like to thank my partner Charity Cruz for their infinite care, tenderness and sweet sweet love. I am forever grateful to have spent quarantine with such an inspirational, creative and kind soul.

I would like to thank my supervisors Chris Bauch and Madhur Anand for their insight, advice and understanding throughout much of my academic career.

Dedication

This is dedicated to my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Common-pool resources

Common-pool resources are resources such as forests and fisheries which are both available for public extraction and finite, therefore being very susceptible to overuse by profit-seeking individuals [1–5]. These resources also play a role across much larger temporal and spatial scales, and current global problems, such as climate change can be explored through this lens [6]. The ubiquity of common-pool resources as well as their fragility in the face of individual self-interest has led to their study in many diverse fields such as economics, sociology, applied mathematics, and ecology [1, 2, 7–15]. A pervasive idea regarding the outcome of common-pool resources is known as *the tragedy of the commons* [16]. The underlying argument is that given a resource shared among rational individuals, each individual can increase their personal profit by increasing their level of resource extraction. There is an associated cost to the health of the resource; however, this cost is shared by all individuals accessing the commons and is consequently less than the expected profit of increased extraction. The conclusion of the tragedy of the commons is that any common-pool resource is doomed to depletion in the absence of control by a central government or private ownership [17]. This paradigm, however, was informed by Hardin's white nationalist agenda which used misguided fears of scarcity to justify the rhetoric of the racist and anti-immigration organizations in which he enacted leadership roles [18, 19]. Since the dissemination of Hardin's theory, scholars investigating the validity of his claims have found that many human communities are in fact very capable of sustainably harvesting common-pool resources without a centralized governing body [1, 2, 20–22]. One important reason for these successes are the value systems that these communities follow regarding

proper resource use. These value systems, known as social norms, can range from a limit to how much an individual should harvest to the importance of telling the truth. They can be enforced by sanctioning from community members, or self-enforced through internal feelings of guilt when an individual violates a social norm. These social norms are dynamic and can evolve among individuals and communities due to both external and internal pressures [1, 2, 20, 21, 23, 24]. There are a number of techniques used to model human decision-making in common-pool resource systems and the theory behind some of these techniques is presented in the following sections.

1.2 Game theory

Game theory studies mathematical models which represent the strategic decision making of two or more rational agents. These agents are often referred to as players, playing a game comprised of the rules by which the players interact. These games often represent situations of social conflict in which each player's decisions are rewarded by some payoff function which represents the utility of a given player choosing a given strategy, dependent on the choices of other players in that game. The assumption that players are rational specifically refers to the fact that players are self-interested and their objective when choosing a strategy is to maximize the expected value of their payoff [25].

In its simplest form, a game theoretic model consists of two players (player 1 and player 2) who each choose between one of two strategies (A and B) and then receive a payoff determined both by their own strategy and the strategy chosen by their opponent. This scenario is often represented by a payoff matrix in the form,

		Player 2	
		A	B
Player 1	A	(a, b)	(c, d)
	B	(e, f)	(g, h)

In the above payoff matrix, the payoff of both players is located using the row corresponding to the strategy of player 1 and the column corresponding to the strategy of player 2. The first and second element of each tuple represent the payoffs of player 1 and player 2, respectively. This can be generalized to games with N players as well as allowing for more that two strategies. For these generalized games, the payoff of player i can be written as

$\pi_i(\boldsymbol{x})$, where the strategy profile, $\boldsymbol{x} = (x_1, x_2, ..., x_N)$, is an N-tuple with the j^{th} element, e_j, being the strategy utilized by player j. The set of strategies available to each player is denoted $\mathcal{X}_i$, where $x_i \in \mathcal{X}_i$ [25].

An important concept in game theory is the Nash equilibrium. In N-player deterministic games, a Nash equilibrium would be the strategy profile $\boldsymbol{x}^*$ that satisfies $\pi_i(\boldsymbol{x}^*) \geq \pi_i(\boldsymbol{x}^*_{-\boldsymbol{i}})$, where $\boldsymbol{x}^*_{-\boldsymbol{i}}$ is the strategy profile $\boldsymbol{x}^*$, with the strategy of player i replaced by a different strategy. In other words, *the Nash equilibrium is a strategy profile for which no single player can change their strategy to increase their expected payoff.* If the inequality is made strict such that, $\pi_i(\boldsymbol{x}^*) > \pi_i(\boldsymbol{x}^*_{-\boldsymbol{i}})$, then $\boldsymbol{x}^*$ is referred to as a strict Nash equilibrium.

Game theoretic models can be organized into a number of general categories and some of the most common are briefly addressed below.

Symmetric games satisfy the requirement that the payoff a player receives depends only on the strategies that each player uses and not the player who uses them. This means that for any permutation denoted α, we have $\pi_{\alpha(i)}(\boldsymbol{x}) = \pi_i(\alpha(\boldsymbol{x}))$, where $\alpha(\boldsymbol{x}) = (x_{\alpha(1)}, ..., x_{\alpha(i)}, ..., x_{\alpha(N)})$. For *asymmetric games*, this condition does not hold for all players and strategies [26].

Pure strategy games are composed of deterministic strategies, that define a player's decision for any given state of the game. In contrast, *mixed strategy games* are composed of probability distributions over the set of available strategies, such that a player using a mixed strategy would be choosing between two or more pure strategies with probabilities defined by that mixed strategy.

1.2.1 The Prisoner's Dilemma

Perhaps the most well-known example of a symmetric two-strategy game is the Prisoner's Dilemma. This game describes a situation where two partners-in-crime are imprisoned and unable to communicate. The prosecutors lack evidence and can only convict each prisoner for a lesser charge, so they offer the prisoners a deal. This deal comes as a dilemma to the prisoners as they need to choose between remaining silent or betraying their partner which would grant them a sentence lighter than that of the lesser charge, only if their partner remains silent. These two options can be represented as strategies in a game where remaining silent is referred to as cooperation and betraying the partner-in-crime is referred to as defection.

		Player 2	
		C	D
Player 1	C	(R, R)	(S, T)
	D	(T, S)	(P, P)

Table 1.1: The payoff matrix for the Prisoner's Dilemma game

In order for this game to be a Prisoner's Dilemma, the payoff matrix (Table 1.1) must follow the condition that $T > R > P > S$. An additional constraint often imposed is that $2R > S + T$. The parameter T is known as the temptation to defect and will be further discussed throughout this chapter. The implications of these constraints are discussed below.

i. Each agent can maximize their individual payoff by defecting so long as the other player cooperates

This is equivalent to stating that mutual defection is the Nash equilibrium. This means that $\pi(D, D) \geq \pi(C, D)$, or equivalently, if both players choose to defect, no single player can change their strategy to increase their expected payoff. Because of this, any rational player will always choose to defect in a one-shot game.

ii. Mutual cooperation is favoured over mutual defection

This is captured by the condition that $R > P$ and creates a tension when choosing between strategies as cooperation has the potential to offer a higher payoff if the other player also chooses to cooperate.

iii. The total payoff of all players is maximized when both choose to cooperate

This is captured by the condition that $2R > S + T$ and creates a global incentive for all players to cooperate and maximize their collective payoff.

With these implications, it is clear that the Prisoner's Dilemma models how individuals often are faced with the choice of acting in their self-interest at the expense of their

broader community, or acting in the collective interest of the community at the expense of maximizing their own personal profit. This brings us back to common-pool resources and the tragedy of the commons discussed in Section 1.1.

In many real-world social dilemmas analogous to the Prisoner's Dilemma, altruism has been observed, where parties have chosen to cooperate rather than defect, appearing to contradict the assertion of defection being the logical strategic choice. However, through iterated games in the sub-field of evolutionary game theory, researchers have been able to show that in many cases, self-interested players will chose cooperation over defection.

1.2.2 Evolutionary game theory

Evolutionary game theory emerged as a sub-field in the 1970s combining the framework of classical game theory with biological models of evolution. In evolutionary game theoretic models, there is always a temporal aspect to the game, allowing it to be played repeatedly. Rather than focusing on the strategy a rational player should choose, there is more emphasis on the dynamics of strategies, or how the frequency of strategies in a given population of players changes throughout time. These dynamics are governed by the current strategy composition of the population as well as the relative payoff of a given strategy, which can be thought of as being analogous to the biological fitness of a given phenotype. This framework offers a very natural way to model the propagation of human decisions such as harvesting strategies in which individuals are represented by players of a game, and an agent's harvesting decisions are the game strategies. There have been many types of models exploring the persistence of common-pool resources as well as the tragedy of the commons in general. These models range from systems of ordinary differential equations assuming a well-mixed population to discrete spatial models where the agents interact on a network or a lattice [10, 13, 25, 27–44].

Replicator dynamics

In spatially homogeneous evolutionary game theory models, the fitness of a given strategy is simply proportional to both the current frequency of the strategy in the population and the relative utility of that strategy. For a pure strategy symmetric game, let the frequency of the population playing strategy i be denoted x_i. To represent the rate of change for strategy i's frequency, the replicator equation is often used, defined as,

$$\dot{x_i} = x_i(\pi_i - \bar{\pi}). \tag{1.1}$$

Here, the rate of change is equal to the relative utility of a given strategy multiplied by it's current frequency in the population, where $\bar{\pi}$ is the mean payoff over the entire population [44–46]. Note that the payoff for a given strategy, π_i is generally a function of the strategy frequencies, and can be written as $\pi_i(\boldsymbol{x})$.

From the replicator equation we can introduce the concept of an evolutionary equilibrium, or stable state. Here, the state of the system refers to it's current strategy composition, which we can represent by the vector $\boldsymbol{x}$. From the theory of ordinary differential equations, we can say that a state, $\boldsymbol{x}^*$ is at equilibrium if $\dot{\boldsymbol{x}}(\boldsymbol{x}^*) = \mathbf{0}$. An equilibrium is asymptotically stable if it is the case that when the system is in a state close enough to the equilibrium, it will eventually approach that equilibrium. From this, it follows that an asymptotically stable state is resilient to small perturbations. This is equivalent to,

$$\bar{\pi}(\boldsymbol{x}^*) > \bar{\pi}(\boldsymbol{x}^* + \boldsymbol{\epsilon}) \tag{1.2}$$

for some small perturbation $\boldsymbol{\epsilon}$. The values of $\boldsymbol{\epsilon}$ for which this expression holds represents the size of the *basin of attraction* for state $\boldsymbol{x}^*$. A basin of attraction is simply the set of states surrounding $\boldsymbol{x}^*$ such that if the current state of the system, $\boldsymbol{x}(t)$ is a member of this set, the system will approach $\boldsymbol{x}^*$ as time progresses. The existence and size of basins of attraction are dependent on model parameters, and as these parameters change, a state that was once in the basin of attraction for a given steady state may find itself in the basin of attraction for an alternative stable state. The concept of alternative stable states are well studied in many human-environment systems and in many cases, when the system shifts towards an alternative stable state, the original equilibrium can be difficult to recover.

An influential common-pool resource model using replicator dynamics was published in 1996 by Sethi & Somanathan [33]. In this model, a well mixed population harvested from a generalized resource function using three available strategies: defectors, who harvested at a high level, cooperators, who harvested at a low level, and enforcers, who harvested at the same level as cooperators and punished defectors at a small cost to themselves. This punishment and cost was represented by a decreased payoff to the respective strategies when enforcers were present in the system. The study found that although defective harvesting was the Nash equilibrium for a static game, introducing evolutionary dynamics allowed for a mixed equilibria of cooperators and enforcers. If the punishment to defectors was sufficiently high, these equilibria were asymptotically stable, however if the cost of punishment was decreased, the size of the basin of attraction for these mixed equilibria could be decreased to the point where adding defectors to the population brought it to a defective state where the sustainable equlibria could no longer be recovered. These results

presented theory supporting ideas of the emergence and persistence of cooperative social norms and the challenge in recovering these social norms if external forces bring the system into an unsustainable alternative stable state.

Spatial dynamics

In agent-based evolutionary games with spatial structure, the dynamics of strategies is often informed by the frequency of strategies in an agent's spatial neighbourhood. For example, if players are represented as cells in a rectangular lattice, it is common for a player to sample from agents in their Moore or von Neumann neighbourhood to determine what strategy to utilize in the next iteration of the game. For choosing the strategy to adapt in the next time step, a common practice is for a given player to imitate the strategy of their highest earning neighbour. This is referred to as *imitation dynamics* and is defined as the strategy of agent i at time t, $x_{i,t}$ is that of their neighbour who received the highest payoff in the previous time step,

$$x_{i,t} = \underset{x_{j,t-1}}{\arg\max}\{\pi_j(x_{j,t-1})\}_{j\in\mathcal{N}_i} \tag{1.3}$$

where $\mathcal{N}_i$ is the set comprised of agent i and it's neighbours. With imitation dynamics, the strategies available to a given agent are limited to what is currently utilized by their neighbours, and can change throughout the course of the game. Imitation dynamics can also be stochastic such that the probability that a player copies the strategy of a given neighbour is positively correlated with the payoff that neighbour received in the previous time step. For mixed strategies, an agent would imitate the probability distribution of actions rather than the action chosen by the agent they are imitating [47].

An alternative to imitation dynamics, is what we call *best response dynamics*. Here, each agent chooses their strategy in the subsequent time step using the concept of the best response. In game theory, a player's best response is choosing the strategy that will give them the highest payoff, taking into account other player's strategies. In a spatial iterated game this can be formulated as,

$$x_{i,t} = \underset{x\,\in\,\mathcal{X}_i}{\arg\max}\{\pi_i(\text{with } x \text{ replacing } x_i \text{ in } \boldsymbol{y_{t-1}})\} \tag{1.4}$$

where $\boldsymbol{y}_{t-1}$ is the state of the system at the previous time step, including, but not limited to the set of strategies of each neighbour. With best response dynamics, the strategies available to a given agent do not change from the outset of the game.

In spatial models, cooperative strategies, which are not favoured in well-mixed models, have been shown to be much more resilient, challenging the base assumptions of the tragedy of the commons. This can be explained through the formation of cooperative spatial clusters, where defection becomes disadvantageous at the boundaries and cannot invade [30, 31, 42]. In an iterated Prisoner's Dilemma played on a lattice, Nowak & May found that so long as the inequality $2R > S+T$ (1.1) was satisfied, spatial clusters of cooperators were able to grow as the game was repeated [30].

Additionally, many studies have extended these spatial model such that players have a limited memory of their past interactions with other players, such that instead of simply choosing to cooperate or defect, each player would utilize a strategy that informed them whether to defect or cooperate given their past interaction with any player [48]. One evolutionary robust strategy with a memory of one time step is called tit for tat [48, 49]. A player utilizing this strategy would cooperate during its first game with any given player, and from then on, copy the strategy that their opponent played in the last time step [48, 49]. The spatial structure of these models can vary in any given context, however any framework for local interactions can be represented by graphs which are discussed below.

1.3 Graph theory

Graph theory is the field of discrete mathematics that studies properties of graphs, also known as networks. A graph is a mathematical structure composed of N vertices or nodes and edges or links that connect them. A graph can be defined as the ordered pair $G = (V, E)$, where V is the set of nodes such that $|V| = N$, and E is the set of edges. Each edge in a graph connects two nodes, and edges can be ordered pairs in directed graphs (where each edge has a given orientation) or unordered pairs in undirected graphs. A simple graph has the restrictions that any two nodes cannot have more than one edge linking them and a node cannot have an edge connecting it to itself, known as a loop. For the rest of this thesis we will only be discussing undirected simple graphs. With these types of graphs, the set of edges is a subset of all distinct pairs of vertices such that $E \subseteq \{(v_i, v_j) \ : \ v_i, v_j \in V \text{ and } v_i \neq v_j\}$. Using more compact notation, e_{ij} denotes an edge connecting nodes v_i and v_j. Each node in a graph has an associated value called the *degree* denoted k_i for v_i. The degree of a node is equal to the number of other nodes v_i shares an edge with, or alternatively, the number of edges connected to it. If each node shares an edge with every other node, we call this type of graph a complete graph, and each $k_i = N - 1$. A well-studied attribute of graphs is the degree distribution, $P(k)$, which is the probability that any given vertex has degree k.

An important idea in graph theory is the concept of a path. In an undirected simple graph, a finite path is a sequence of nodes, $P_{i,j} = (v_i, v_{i+1}, ..., v_j)$, such that v_n and v_{n+1} share an edge. A path is often referred to in relation to the initial and final node in the sequence, where $P_{i,j}$ would be a path from v_i to v_j. A graph for which there is a path between any two nodes is called a connected graph. The shortest path between any two nodes is the sequence with the least elements of all paths connecting v_i and v_j. Note that the shortest path does not have to be unique. A property of connected graphs is the average path length, which is defined as the mean of all shortest paths between each pair of vertices. This can be calculated as,

$$l_G = \frac{2}{N(N-1)} \sum_{i<j} d(v_i, v_j) \tag{1.5}$$

where $d(v_i, v_j)$ is the distance of the shortest path between vertices v_i and v_j, measured by the number of edges between them or $|P_{i,j}| - 1$.

Another important concept is graph clustering, which measures the extent which subsets of nodes share more edges than what is expected from the average probability of any two nodes sharing an edge. The metric most commonly associated with clustering is the clustering coefficient, which has two formulations: the local clustering coefficient and the global clustering coefficient. The local clustering coefficient is calculated for each node in a graph and measures how close the neighbours of a given node are to forming a complete graph. Since the number of distinct edges between neighbours of any given node is $\frac{k_i(k_i-1)}{2}$, the local clustering coefficient for node v_i is defined as,

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in \mathcal{N}_i,\ e_{jk} \in E\}|}{k_i(k_i-1)} \tag{1.6}$$

where $|\cdot|$ denotes the cardinality of a given set. This can be used to measure the clustering in the entire network by taking the mean over all vertices, which we call the average local clustering coefficient,

$$\bar{C} = \frac{1}{N} \sum_i C_i \tag{1.7}$$

Another metric used to quantify clustering in a graph is the global clustering coefficient or network transitivity. To define this, we need to introduce the concept of a triplet. A triplet in graph theory is three vertices that share either two or three edges with each other. A triplet with three edges is called a closed triplet, where a triplet with two edges is

called an open triplet. The transitivity of a graph measures the proportion of triplets that are closed in a graph. If we represent the number of open triplets as T_o and the number of closed triplets as T_c, the transitivity of a graph is simply,

$$C = \frac{T_c}{T_o + T_c} \tag{1.8}$$

Although these metrics both measure clustering, they differ from each other as nodes with a higher degree have a greater influence on the transitivity of a graph compared with the average local clustering coefficient where each node contributes equally.

Often when graphs are used in mathematical models, nodes are given attributes, which is a set of state descriptors for each node. These can range from simply the colour of the node to a representation of how an agent corresponding to a node makes decisions in a social model. The extent by which nodes with similar or identical attributes share edges (or, alternatively, how like nodes tend to be connected to each other) is called homophily, and this can be measured using modularity and assortativity. Modularity, denoted Q, is used for attributes that are not rank ordered such as colours or social descriptors. It is defined as,

$$Q = \frac{\sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2|E|} \right) \delta_{ij}}{2|E|} \tag{1.9}$$

$$A_{ij} = \begin{cases} 1 & e_{ij} \in E \\ 0 & e_{ij} \notin E \end{cases} \tag{1.10}$$

$$\delta_{ij} = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \tag{1.11}$$

$$\tag{1.12}$$

where A_{ij} is index (i, j) in the graph's $N \times N$ adjacency matrix, representing whether v_i and v_j share an edge, $\frac{k_i k_j}{2|E|}$ is equal to the probability that v_i and v_j share an edge and δ_{ij} represents whether v_i and v_j share the same attribute x. The modularity of a graph falls in the range $[-\frac{1}{2}, 1]$, where $Q > 0$ if nodes that have the same attribute share more edges that what would be expected by chance alone.

When node attributes are rank ordered, such as the age or cumulative payoff of a node represented by a scalar, homophily is measured by the assortativity coefficient, r. This coefficient is an instance of the Pearson correlation coefficient between nodes that share

edges and is calculated using,

$$r = \frac{\sum_{ij} A_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i k_i(x_i - \bar{x})^2} \tag{1.13}$$

where x_i is the scalar attribute of v_i and $\bar{x} = \frac{1}{2m}\sum_i k_i x_i$ is the mean value of x weighted by the degree of the node [50].

1.3.1 Graph topologies

Graphs are often categorized in terms how nodes are connected to one another, which we will refer to as the graph topology. Graphs with desired course-grained characteristics can be generated, often using probabilistic models. Three common graph topologies and their generative models are introduced below.

Random graph

A random graph is characterized by having a degree distribution similar to the Poisson distribution, centered at the average degree of connectivity, $\bar{k}$. A common generative model for random graphs is the Erdős–Rényi model. This model is initialized with the set of nodes, V and a desired average degree, $\bar{k}$. Then, for any two unique nodes, an edge linking them is created with probability $\frac{\bar{k}}{N-1}$ [51].

Small-world graph

A small-world graph is characterized by the fact that any pair of nodes are unlikely to be directly connected, yet have a small shortest path length as well as higher clustering than what would be expected in a random graph. More specifically, the average shortest path length is proportional to the logarithm of the number of nodes,

$$l_G \propto \log(N) \tag{1.14}$$

In social networks, this can be interpreted as the phenomenon that any two strangers are often connected through a small chain of mutual acquaintances. A common generative model for small-world graphs is the Watts-Strogatz model. This model is initialized with the set of nodes, V, a desired average degree, $\bar{k}$ restricted to be an even integer, and the parameter $\beta \in [0, 1]$, known as the probability of re-wiring. First, a regular ring lattice is

constructed, for which $e_{ij} \in E$ if and only if $0 < d(i,j) < \frac{\bar{k}}{2}$. Here, the distance between v_i and v_j is defined as $\min(|i-j|, N-|i-j|$ to capture that V is cyclic ordered. Then, for all v_i, each $e_{ij} \in \{(i, j \mod N) : i < j < i + \frac{\bar{k}}{2}\}$ is changed to $e_{ij'}$ with probability β where $v_{j'} \in V$ is chosen at random and $j' \neq i$. Graphs generated using this model have the property that $C(\beta) \approx C(0)(1-\beta)^3$, where $C(\beta)$ is the transitivity of the small-world graph, parameterized by β, and $C(0)$ is the transitivity of the initial ring lattice [52–54].

Scale-free graph

A scale-free graph is characterized as having a degree distribution fitting a power-law such that $P(k) \propto k^{-\alpha}$. These graphs are often found in online social networks and contain a small number of nodes with a very high degree. The generative model often used in creating a scale-free graph is known as the Barabási-Albert model. It is initialized with a connected graph of m_0 nodes such that there is a path between any two nodes. Then a single node is added with the potential to have edges connecting it to $m \leq m_0$ existing nodes. These edges are added, connecting the new node to v_i with probability $p_i = \frac{k_i}{\sum_j k_j}$. This process is repeated until the N^{th} node is added. From the expression p_i, we can see that a newly added node has a much higher likelyhood of sharing an edge with another node of a high degree and as this process is iterated, we have a positive feedback loop increasing the amount of connections to existing high-degree nodes [55].

1.4 Self-organized criticality

As mentioned when introducing evolutionary stable states, a concept that is important in the literature regarding evolutionary game-theoretic models is the system's sensitivity to external perturbations [27, 56]. In discrete systems with spatial structure, stability analysis from ordinary differential equations often cannot be applied. Instead, this concept is often investigated by changing the strategy of an agent regardless of its perceived benefit when the system is at equilibrium, or introducing a mutant strategy. If the strategy composition of the system is robust to these kinds of perturbations, the system is at an evolutionary stable state. Additionally the time to return to equilibrium is often studied, counting the resulting time steps that the system takes in order to reach equilibrium again. This number of time steps is referred to as a cascade. Previous studies have demonstrated that the cascade size can vary quite drastically covering many orders of magnitude and in many cases, it behaves in line with the criteria for self-organized criticality proposed by Bak et al. [56, 57]. Self-organized criticality posits that many complex systems with

local interactions tune themselves into a critical state that displays scale-invariant spatial or temporal characteristics. In this state, small perturbations can cascade throughout the system with the cascade size-distribution fitting a power law [57]. Power law distributions are scale-invariant, having the form,

$$p(x) \propto x^{-\alpha} \tag{1.15}$$

which forms a straight line when graphed on a log-log plot. These distributions are found in many physical systems and are most prominent when the system is in a critical state and extremely sensitive to perturbations [58].

A study bridging this phenomenon with concepts discussed in Sections 1.2.1, 1.2.2, and 1.3 presented an iterated Prisoner's Dilemma game on a random graph with best response dynamics [27]. Here, the strategy profile available to all players was all possible strategies with memory of a single game. The authors found that for a critical range of the temptation to defect, T, the probability distribution of cascade sizes closely fit a power-law suggesting self-organized criticality within their model [27].

1.5 Reinforcement learning fundamentals

Reinforcement learning is a sub-field of machine learning concerned with the problem of an agent or group of agents learning a goal in an environment through experience. More formally, reinforcement learning systems can be represented by a Markov decision process where,

- the state of the environment and agent is given by $s \in \mathcal{S}$
- the action of an agent or group of agents available in state s is denoted $a \in \mathcal{A}(s)$
- the probability of transitioning to state s' from state s under action a is $P_a[s, s'] = P[S_{t+1} = s' \mid S_t = s, A_t = a]$
- the reward an agent receives immediately after the above transition is R_{t+1}

Often, in a reinforcement learning environment, the agent will not have access to the full state of the system, s, and instead will take actions based on their observations, giving them only partial information of the current state of the environment. In these cases, the reinforcement learning process would be represented as a partially observable Markov

decision process where observations can be denoted $o \in \mathcal{O}$. To correspond with the model presented, for the rest of this thesis we will only be addressing systems that are partially observable Markov decision processes and have a deterministic state transition such that $P_a[s, s'] \in \{0, 1\}$. Since s' is a function of (s, a), the reward an agent receives can be written as $R_{t+1} = r(s, a)$. In an N-agent system, the action, a and reward $r(s, a)$ would be N-dimensional vectors with the i^{th} element corresponding to the reward of agent i. In the formulation of most reinforcement learning algorithms, instead of directly focusing on the probability of transitioning to the next state, the emphasis is shifted to the agent's decision-making - more specifically, the probability of an agent to take action a in state s,

$$\pi(a \mid s) = P[A_t = a \mid S_t = s] \tag{1.16}$$

This is called the *policy* of an agent and iteratively optimizing this policy to maximize the agents expected reward is the primary goal of reinforcement learning.

1.5.1 Value functions and the Bellman equation

Perhaps the most useful tool utilized in improving an agents' policy are the value functions. The first value function we will look at is the state-value function for policy π, denoted $v_\pi(s)$. This function represents the utility of state s, under an agents' policy π, and is formulated as,

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \quad \forall s \in S \tag{1.17}$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1} \tag{1.18}$$

Note that instead of only focusing on the immediate reward, R_{t+1}, the agent is taking into account all future rewards, G_t. This is called the return, with a time discounting factor of γ.

Similar to the state-value function, we have the action-value function for policy π, denoted $q_\pi(s, a)$. The action-value function represents the utility of taking action a in state s and then following policy π where,

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \tag{1.19}$$

Now we can define an optimal policy, π^* as one that maximizes the value functions such

that,

$$v_{\pi^*}(s) = \max_{\pi} v_{\pi}(s) \tag{1.20}$$

$$q_{\pi^*}(s,a) = \max_{\pi} q_{\pi}(s,a) \tag{1.21}$$

It follows that the state-value function and the action-value function are related by,

$$v_{\pi}(s) = \sum_{a} \pi(a \mid s) q_{\pi}(s,a) \tag{1.22}$$

$$v^*_{\pi}(s) = \max_{a} q_{\pi^*}(s,a) \tag{1.23}$$

In most cases, value functions cannot be computed exactly and are instead estimated recursively using the Bellman equation that reformulates the value functions as,

$$\begin{aligned} v_{\pi}(s) &= \mathbb{E}_{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] && (1.24)\\ &= \sum_{a} \pi(a|s) \sum_{s',r} P(s',r|s,a) \left[r + \gamma \mathbb{E}_{\pi}[G_{t+1} \mid S_{t+1} = s']\right] && (1.25)\\ &= \sum_{a} \pi(a|s) \sum_{s',r} P(s',r|s,a) \left[r + \gamma v_{\pi}(s')\right] && (1.26)\\ q_{\pi}(s,a) &= \sum_{s',r} P(s',r|s,a) \left[r + \gamma v_{\pi}(s')\right] && (1.27) \end{aligned}$$

Applying the Bellman equation to the optimal value functions, we have,

$$v_{\pi^*}(s) = \max_{a} \sum_{s',r} P(s',r|s,a) \left[r + \gamma v_{\pi^*}(s')\right] \tag{1.28}$$

$$q_{\pi^*}(s,a) = \sum_{s',r} P(s',r|s,a) \left[r + \gamma \max_{a'} q_{\pi^*}(s',a')\right], \quad A_{t+1} = a' \tag{1.29}$$

By learning an estimation of v_{π^*}, the Bellman optimality equation allows reinforcement learning agents to approximate expected rewards over an infinite time horizon by only looking one time step ahead [59, 60].

1.5.2 Deep reinforcement learning

There are many different methods for learning these value functions, however we will only be discussing algorithms that use neural networks as function approximators as these algorithms are most commonly used and serve to model the harvesting decisions of agents in the model presented in Chapter 3.

The structure of neural networks

Neural networks, more formally referred to as artificial neural networks are a broad class of algorithms modelled after biological neural networks of human brains. These algorithms are often represented as graphs, with their nodes and edges analogous to neurons and synapses. In a feed-forward neural network, an input vector, $\boldsymbol{x} \in \mathbb{R}^n$ is classified into meaningful categories with the combination of linear functions and non-linear activation functions. Each neuron in a neural network has an associated weight vector $\boldsymbol{w} \in \mathbb{R}^n$, where each element can be mapped to an edge, a bias, $b \in \mathbb{R}$, and non-linear activation function, $\varphi : \mathbb{R} \to \mathbb{R}$. With these, we can compactly denote the output of any given neuron as,

$$a = \varphi(\boldsymbol{w}^T \boldsymbol{x} + b) \tag{1.30}$$

In deep neural networks, neurons are grouped into ordered layers such that the output of layer $(l-1)$ is the input to layer l. The outputs of a given layer with k neurons form the k-dimensional vector represented by,

$$\boldsymbol{a}_l = \varphi(\boldsymbol{w}_l^T \boldsymbol{a}_{l-1} + b_l) \tag{1.31}$$

The activation function, φ must be differentiable (except possibly at a single point) and have a finite range. Commonly used activation functions are sigmoid functions such as the logistic function or the hyperbolic tangent function. Another common type of activation function is the rectified linear unit or ReLU which is defined as simply $f(x) = \max(x, 0)$. Once the data has passed through all hidden layers in the neural network, it is then fed into an output layer that classifies the data into predefined categories or describes the data as a continuous vector or scalar quantity [61–64].

Training neural networks

When training a neural network, the goal is to learn weights and biases that can accurately classify data outside the training set. This is often done using a process called supervised

learning, where the network is given a training data set which includes the desired labels. The training process aims to simply minimize the error between the output of the model and the provided labels. The error is referred to as the loss function, often defined as the mean squared error,

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.32}$$

where $\hat{\boldsymbol{y}}$ is the output of the neural network and $\boldsymbol{y}$ is the desired output. To minimize this loss function, the partial derivatives with respect to the weights $\partial L/\partial w$, and the bias, $\partial L/\partial b$ are calculated starting from the final layer of the network and working backwards. This process, called *backpropagation*, is essentially an application of the chain rule for computing the derivatives for compositions of functions.

The importance of these derivatives is that they are needed to calculate the gradient of the cost function with respect to all the parameters of the neural network, which we will refer to as $\nabla_\theta L$, where θ is compact notation for all the weights and biases of the neural network. Since this gradient points in the direction of greatest increase for the loss function, updating the weights and biases in the direction opposite to that of the gradient will decrease this loss. This iterative process to minimize the loss function is called *gradient descent*, and using this process the parameters of the neural network are updated at each time step, t, where,

$$\theta_{t+1} = \theta_t - \alpha\nabla_\theta L \tag{1.33}$$

Here, α is the step size and determines the magnitude of updates to θ.

1.6 Reinforcement learning algorithms

Although many different types of reinforcement learning algorithms exist, they are often categorized in two ways. One is whether the algorithm is model-based or model-free. In model-based algorithms, agents attempt to learn a model made up of the functions $P_a[s, s']$ and $r_a(s, s')$ for the environment they are interacting with. In contrast, model-free algorithms attempt to learn the policy directly without being able to predict the dynamics of the environment. Another important category is whether the algorithm is on-policy or off-policy. An agent in an on-policy algorithm calculates the expected return with the current policy it is using. For an off-policy algorithm, an agent will calculate the expected return of the learned optimal policy, but choose its actions based on a different policy [59].

This can be better understood by discussing the trade-off of exploration and exploitation fundamental to reinforcement learning. In order to maximize its reward, an agent needs to choose actions that led to high rewards in previous time steps, or *exploit its previous experience*. However, in order to find these actions (and discover actions that provide an even higher reward), the agent must choose actions that it had not previously used in a given state, or *explore the system*. Many reinforcement learning algorithms are off-policy so that an agent can learn a deterministic optimal policy by choosing actions based on a different stochastic policy that allows for exploration.

An additional challenge in reinforcement learning is that the data an agent is learning from is significantly autocorrelated. This is a problem in the learning process as the theory behind training a neural network assumes that the data used to train the network is independent and identically distributed. To remedy this, instead of learning from each (s, a, r, s') transition as they are observed by an agent, these transitions are stored in a buffer which is then uniformly sampled from in the learning process. This process is called *experience replay* and is used in all algorithms discussed below.

1.6.1 Deep Q-networks

Deep Q-networks (DQN) are an off-policy, model-free algorithm that approximate the action-value function $q_\pi(s, a)$ with a Q-value function learned through a neural network, parametrized by θ. In order to learn this Q-function, the loss is defined as the mean squared error, similar to (eq. 1.32). In this formulation, the output of the neural network is the Q-value, $Q(s, a|\theta)$, and the desired output is the target Q-value, y, derived from the Bellman optimality equations (see Section 1.5.1). Since the algorithm learns from batches of size n sampled from the replay buffer, the loss function is computed using,

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - Q(s_i, a_i|\theta)^2 \right) \tag{1.34}$$

$$y_i = \begin{cases} r_i & \text{episode terminates at step } i+1 \\ r_i + \gamma \max\limits_{a'} Q'(s_{i+1}, a'; \theta') & \text{otherwise} \end{cases} \tag{1.35}$$

This loss is minimized by taking its gradient with respect to θ and using backpropagation on the neural network. Note, the target, y_t uses Q' instead of Q. This is because most applications of DQN use two Q-networks, where Q' is referred to as the *target network* and is essentially a copy of the primary Q-network that is updated at a much lower frequency. Using this target network adds stability to the training process since the target is not

changing as frequently or with the same magnitude as the primary Q-network. The target Q-network is initialized to be identical to the original Q-network and can be updated in one of two ways. If a hard update is used, the weights are updated to that of the original Q-network every m time steps, and if a soft update is used, the weights are updated such that $\theta' := \tau\theta + (1-\tau)\theta'$, where $\tau \ll 1$ [59, 65].

Since this network is only learning the Q-function and not a policy directly, it uses an ϵ-greedy policy such that with probability $1-\epsilon$, the action $\arg\max_a Q(s,a)$ is chosen and otherwise the action is randomly chosen with all actions having equal probability. As the network is trained, the value of ϵ decreases. A limitation to DQNs is that since they are approximating $q(s,a)$ for each $a \in \mathcal{A}(s)$, the action space must be discrete and if a continuous action space is discretized to a high-dimensional action space, this can add a large computational overhead to running the algorithm.

1.6.2 Deep deterministic policy gradient

Deep deterministic policy gradient (DDPG) uses an off-policy, model-free actor-critic algorithm to extend DQNs into continuous action spaces. This algorithm consists of two neural networks, an actor network and a critic network. The actor network's role is to learn a deterministic policy $\mu : S \to A$, represented by the actor function $a = \mu(s|\theta^\mu)$, parametrized by the weights of the actor network, θ^μ. Similar to Q-learning, the critic network, parametrized by θ^Q, takes state-action pairs as input, and outputs a Q-value, estimating the utility of taking action, a given a state, s. To improve stability, a target network is used for both the actor and critic, denoted μ' and Q' respectively. These networks are updated using the soft update rule.

The critic network uses the same loss function as the DQN, however the target value, y, is calculated using the output of both the actor and critic target networks:

$$y_i = r(s_i, a_i) + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}) \tag{1.36}$$

The actor network is trained by maximizing the performance objective which is simply the expected return from the start state for policy μ,

$$J(\mu) = \mathbb{E}[G_1|\mu] \tag{1.37}$$

This is done by using the gradient of the policy's performance,

$$\nabla_{\theta^\mu} J(\mu) \approx \mathbb{E}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t|\theta^\mu)}] \tag{1.38}$$

$$= \mathbb{E}[\nabla_a Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta_\mu} \mu(s|\theta^\mu)|_{s=s_t}] \tag{1.39}$$

To allow for exploration, the action chosen by the deterministic policy is given additive noise sampled from a normal distribution, such that the action implemented by the agent is $\mu + \mathcal{N}(0, \sigma_t^2)$, where the variance is a function that decreases with t [59, 66, 67].

1.7 Reinforcement learning and evolutionary game theory

Although reinforcement learning and evolutionary game theory originate from different fields, they both share many similarities as they address the problem of training autonomous agents to make meaningful decisions. In game theory, we discussed the player playing a game, which is analogous to an agent interacting with the environment in reinforcement learning. In many cases, including the models presented in the following chapters, these systems can be formulated as Markov decision processes where the agents are learning an optimal policy. In game theory the strategy a player utilizes, given the state of the system can be thought of as a policy, albeit one that is often much less complex than what can be learned with a neural network. Additionally, there is a parallel between the payoff of a player and the reward for an agent as these both represent the immediate utility of choosing a given strategy or action. A common difference in its application is that in evolutionary game theory, players generally aim to maximize their immediate payoff, whereas in most reinforcement learning models, an agent is maximizing their expected return over a discounted time frame.

1.8 Objectives & direction

So far this thesis has discussed: the history of common-pool resource systems (Section 1.1), spatial models in evolutionary game theory (Section 1.2.2), common types of networks and their generative models (Section 1.3.1), self-organized criticality (Section 1.4), as well as reinforcement learning algorithms (Section 1.6).

These concepts provide a framework for the following two chapters, the first of which will explore a common-pool resource model formulated as an iterated game where players interact on a social network, and the second of which will explore a similar system where agents learn policies to harvest a common-pool resource through reinforcement learning models.

The game theoretic model (Chapter 2) is inspired by the model published by Sethi & Somanathan [33] formulating a human-environment system as an evolutionary game, and the model published by Ebel & Bernholdt [27] investigating self-organized criticality on a spatial iterated Prisoner's Dilemma played on a network. In this new model, two mechanisms for strategy propagation: best response and imitation dynamics (see Section 1.2.2) are compared under identical parameters. This model offers novel insight for a couple reasons. One, being that the literature on common-pool resource models with a social network structure is sparse. Models that explore this interaction can give significant insight into common-pool resources since most human interactions are determined by social networks rather than physical location within a community [68, 69]. One study that has explored this analyzed an empirical network common-pool resource model where cooperators maximize their payoffs over a longer time horizon than defectors. Using a single strategy update rule, a higher number of links between cooperators, as well as smaller networks, were found to promote cooperation and efficient resource management [36]. A second study investigated agents on a lattice informed by both social norms and organizational rules when harvesting a forest; however, the long term dynamics were governed by sampling strategies from a normal distribution [70]. A third study modelling agents on a network found that in most cases increasing the probability of rewiring cooperator-defector links increased cooperation in the system [71].

In the previous studies, as well as the majority of game theoretic models, one mechanism for the evolution of strategies among agents is assumed. Rarely in the literature are the dependence of model predictions on this choice of mechanism compared. Two of the most common mechanisms, imitation and best response dynamics, are supported by the literature; however, their direct comparison in a single model could shed light on the qualitative differences resulting from these contrasting psychological inclinations [72, 73]. The closest we have seen to this is a study that systematically compared imitation dynamics and strategy evolution using genetic algorithms with an N-player Prisoner's Dilemma game on a lattice. The authors found evolution to significantly promote cooperation and increase strategy convergence rates in their system. The genetic algorithm incorporated aspects of imitation and the model did not include payoffs coupled to a common-pool resource [38]. In our model, we are using best-response dynamics instead of genetic modification as well as an explicit common-pool resource, allowing us to further separate social learning from

independent prediction in a human-environment system.

The reinforcement learning model (Chapter 3) is presented as a work in progress with some preliminary results with the goal of finalizing the analysis and publishing the final results in the near future. This model is a variation of the model published by von der Osten et al. [13] as well as an extension of the previous model published by Zhu et al. [74]. These studies investigate the sustainability of a common-pool resource in a spatially homogeneous system with up to 15 agents using both the DQN algorithm and DDPG algorithm discussed in sections 1.6.1 and 1.6.2, respectively. The model presented here differs from these previous studies in the definition of the reward function as well as by using a minimum of 24 agents whose interactions are structured by a social network. Additionally, the actions available to each agent is significantly increased, allowing agents to learn their own preferred actions rather than having them assigned a priori. These changes offer a higher degree of realism and present a novel framework for modeling human-environment systems, where many of the social parameters are learned instead of prescribed by the modeller.

Chapter 2

Best response dynamics improve sustainability and equity outcomes compared to imitation dynamics in common-pool resources problems[1]

Abstract

Shared resource extraction among profit-seeking individuals involves a tension between individual benefit and the collective well-being represented by the persistence of the resource. Many game theoretic models explore this scenario, but these models tend to assume either best response dynamics (where individuals instantly switch to better paying strategies) or imitation dynamics (where individuals copy successful strategies from neighbours), and do not systematically compare predictions under the two assumptions. Here we propose an iterated game on a social network with payoff functions that depend on the state of the resource. Agents harvest the resource, and the strategy composition of the population evolves until an equilibrium is reached. The system is then repeatedly perturbed and allowed to re-equilibrate. We compare model predictions under best response and imitation dynamics. Compared to imitation dynamics, best response dynamics increase sustainability of the system, the persistence of cooperation while decreasing inequality and

[1]Material from this chapter was published as Farahbakhsh et al. (2021). Best response dynamics improve sustainability and equity outcomes in common-pool resources problems, compared to imitation dynamics. *Journal of Theoretical Biology 509*(110476) https://doi.org/10.1016/j.jtbi.2020.110476.

debt corresponding to the Gini index in the agents' cumulative payoffs. Additionally, for best response dynamics, the number of strategy switches before equilibrium fits a power-law distribution under a subset of the parameter space, suggesting the system is in a state of self-organized criticality. We find little variation in most mean results over different network topologies; however, there is significant variation in the distributions of the raw data, equality of payoff, clustering of like strategies and power-law fit. We suggest the primary mechanisms driving the difference in sustainability between the two strategy update rules to be the clustering of like strategies as well as the time delay imposed by an imitation processes. Given the strikingly different outcomes for best response versus imitation dynamics for common-pool resource systems, our results suggest that modellers should choose strategy update rules that best represent decision-making in their study systems.

2.1 Introduction

The model presented in this chapter explores a human population arranged on various network topologies that harvest an ecological resource. All agents have equal access to this resource and it's growth is modeled by a logistic difference equation. Both strategy evolving mechanisms are compared over identical model parameters. Topologies are also compared since spatial structure has often been found to influence outcomes in tragedy of the commons scenarios [11, 28, 29, 75]. Through running this model across a large parameter space, insight will be gained regarding the mechanisms which lead to cooperation and the persistence of resources in common-pool resource systems.

2.2 Harvesting model

The model simulated a network of N individuals or nodes harvesting from a generalized common-pool resource. At initialization, each node is randomly given one of two harvesting strategies; cooperation or defection, with the probability of either strategy being 0.5. At each time step, every node simultaneously harvests the resource with cooperators harvesting less than their 'maximal equal share' and defectors harvesting more than their 'maximal equal share'. The 'maximal equal share' is defined by $\frac{R_t}{N}$ where R_t is the resource at time t. When harvesting, cooperators inflict a punishment proportional to the depletion of the resource to any defective nodes to which they are directly linked. This proportionality is justified because enforcement by social norms is often more severe when the resource is close to depletion [76–78]. This punishment also incurs a small cost to the cooperators' own harvest. Each nodes respective net harvest is given by the following payoff functions,

$$\pi_c = \frac{1}{N}(cR_t - a \cdot p \cdot n_d(1 - R_t)), \quad c, a < 1 \tag{2.1}$$

$$\pi_d = \frac{1}{N}(dR_t - p \cdot n_c(1 - R_t)), \quad d \geq 1 \tag{2.2}$$

where π_c and π_d are the cooperators and defectors payoffs respectively, c and d are the proportions of the 'maximal equal share' that the cooperative and defective nodes harvest respectively, p is the magnitude of the punishment inflicted by cooperative nodes on defective nodes, a is the relative cost of that punishment for the cooperative nodes, n_c and n_d are the amount of cooperative and defective nodes directly connected to any given node in the network.

The resource is updated at each time step using a logistic difference equation from which the net harvest of all the nodes in the network is subtracted. It is modeled by,

$$R_{t+1} = R_t(1 + F(1 - R_t)) - \frac{R_t}{N}(dN_d + cN_c) \tag{2.3}$$

where F is the maximal growth rate, or fecundity of the resource, and N_d and N_c are the total number of defectors and cooperators in the network respectively.

After all nodes harvest and update their payoffs, one randomly selected node changes its strategy if it is perceived to increase its individual payoff. Only one node is selected per time step as this is common for simulations of evolutionary games [27, 31, 79]. Additionally, in many cases, the time scale over which social norms change is often longer than that over which a single resource extraction occurs [80–82]. This process is repeated until a steady state is reached where the frequency of strategies either reaches equilibrium or a periodic steady state. As no periodic steady states were detected, this steady state is simply referred to as equilibrium throughout the rest of this chapter. Once equilibrium is reached, the system is then perturbed by randomly choosing a node and changing its harvesting strategy regardless of any perceived profit increase. The system is then left to re-equilibrate. This process continues until the resource is depleted or a predetermined number of perturbations is reached. In this simulation a single node is chosen to be perturbed for similarity with other evolutionary games on networks that explored the concept of self-organized criticality [27, 56]. Although the number of nodes that have their strategy changed during a perturbation must be small, the selection of a single node is arbitrary and any number of nodes could be perturbed so long as that number is significantly smaller than the size of the network.

If at any point during the simulation, the resource drops below a critical level, ϵ, the resource is extinct and the simulation ends. This represents the resource reaching a depleted level from which it cannot recover. The value for ϵ scales with the network size and is given by $\frac{1}{N \cdot 10000}$. A flowchart of this model is shown in Figure 2.1.

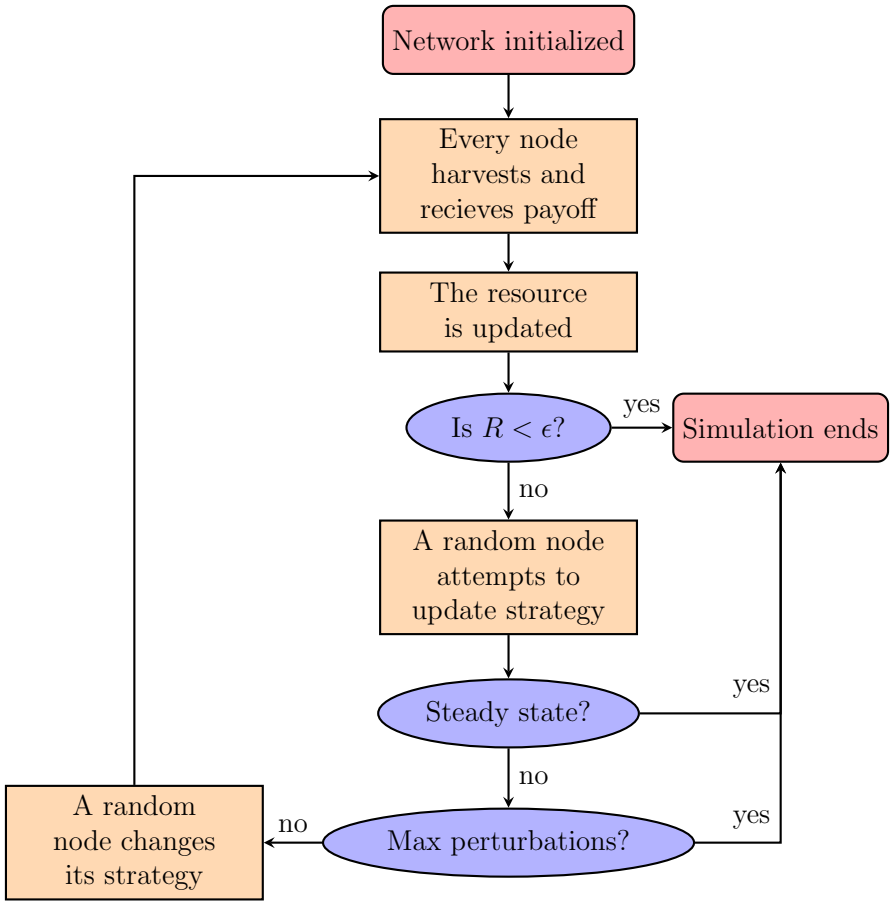


Figure 2.1: Flowchart of the model

2.3 Strategy propagation

Two different mechanisms for the propagation of harvesting strategies among nodes were compared. One is best response dynamics and the other we call imitation. In the case of imitation, the node that is selected for strategy reassignment compares its cumulative payoff with that of its neighbouring nodes. If any neighbouring node has a higher cumulative payoff, the selected node will change its strategy to that of its highest earning neighbour. If there are multiple highest earning neighbours, the selected node will randomly choose between the highest earning neighbours with equal probability. In the case of best response dynamics, the selected node will change its strategy to that which is most profitable for

the next time step. It does so by simulating harvesting from the updated resource with the strategy it wasn't previously using. If the expected payoff from that simulation is greater than the selected nodes' mean payoff over all time steps, the node will change its' harvesting strategy to the new one.

2.4 Network simulations

Four network topologies were tested in this model; lattice, random, scale-free and small-world networks. These were chosen as they are very common network topologies used in the literature, and both scale-free and small-world networks share similarities with social networks [52, 83]. In the square lattice, the von Neumann neighbourhood was used with periodic boundary conditions. Random networks were generated by the Erdös-Rényi model, scale-free networks were generated by the Barabási–Albert model [55], and small-world networks were generated by the Watts-Strogatz model with β, the probability of rewiring an edge in the ring equal to 0.08 [52]. For all networks the average degree of connectivity, $\bar{k} = 4$, so that they can be compared to the lattice.

2.5 Parameters

Each network was generated with $N = 15 \times 15 = 225$ nodes and average degree of connectivity, $k = 4.0$. These values were chosen to allow for direct comparison with the square lattice (as N is a perfect square) using a von Neumann neighbourhood while at the same time being within the range of those used in Ebel & Bornholdt's Prisoner's Dilemma network study [27]. For each network topology, all combinations of parameters shown in Table 2.1 were run. Cooperators harvest was limited to $c \leq 0.6$ since cooperators are conservationists, thus committing to harvest significantly less than their fair share of the resource. For values of $c > 0.6$, the resource was depleted very quickly while being initialized with half of the nodes harvesting at $d \geq 1.0$. Defectors harvest was limited to $d \leq 1.9$ since values larger than this result in immediate resource depletion. Fecundity was limited to $0.5 \leq F \leq 0.9$ to agree with values that allowed for the persistence of the resource within realistic constraints. For punishment, values larger than 0.3 were not included since they made punishment and its cost too severe to be realistic. For the cost of punishment, $a = 0.1$ such that this cost would not be high enough to disincentivize cooperation. This parameter space resulted in 450 parameter sets that were simulated for each network topology. Baseline parameter values were chosen based on mid-range parameter values which allowed

for the persistence of the resource over multi-parameter variations (see Table 2.1). This was done as resource persistence is a pre-condition for gathering much of the data that will be analysed. The resource was initialized at $R = 1.0$ to simulate a community harvesting a resource that has not been harvested previously. For each parameter set and topology, 50 networks were generated and each network was perturbed 100 times.

Parameter	**Minimum**	**Maximum**	**Step**	**Baseline**
Cooperators Harvest (c)	0.2	0.6	0.1	0.3
Defectors Harvest (d)	1.0	1.9	0.1	1.4
Fecundity (F)	0.5	0.9	0.2	0.9
Punishment (p)	0.1	0.3	0.2	0.2
Cost of Punishment (a)	–	–	–	0.1

Table 2.1: Parameter ranges used in model simulations: minimum and maximum of parameter range, increments sampled, and baseline values.

2.6 Outcome metrics

From the simulations, the cascade size, defined as the amount of strategy switches taken for the system re-equilibrate after a perturbation, was recorded. The perturbed node's degree and clustering coefficient were collected as well. For each cascade, the average number of strategy switches per time step was also logged. The rest of the data was collected every time the system reached equilibrium after a perturbation. This data consisted of the level of the resource and the number of cooperators at equilibrium. Average local clustering coefficient and network transitivity over the whole network and over the subnetworks of same strategy-types were recorded as well as network modularity with respect to the strategy of the nodes. The Gini index for total node payoff was also recorded over the whole network to investigate the role of the distribution of wealth in the system [32]. The Gini index was calculated using,

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|P_i - P_j|}{2N\sum_{i=1}^{N}P_i} \tag{2.4}$$

where P_m, $m = \{1, 2, ..., N\}$ is the total payoff of a given node, m. [84].

Finally, the sustainability level of each parameter combination was calculated. Here, sustainability is defined by whether the resource can persist and is quantified by the frequency of simulations in which the resource is not depleted to a value less than ϵ. Therefore, this sustainability level is a value between 0 and 1, with 1 being maximally sustainable and 0 being completely unsustainable.

2.7 Sustainability of the resource

Over the parameter space and all topologies, the systems with best response dynamics were more sustainable than those employing imitation. However, these differences were not statistically significant, due to many simulations having extreme values for sustainability. Two important variables related to sustainability are the number of cooperators and the level of the resource. These variables were recorded when the system was at equilibrium and the analysis of their means and coefficients of variation (CV) over all topologies were conducted as seen in Figure 2.2. Similar to the sustainability of the system, the mean cooperator frequency and resource level were both significantly higher for best response dynamics; however, there was a greater differential between the mean resource levels for each network topology. The CV of the cooperator frequency was low for both best response dynamics and imitation; however, the difference between them was statistically significant with imitation having a higher CV. Additionally, over all topologies imitation had a higher CV for the resource level. This difference can be explained by the underlying distributions of the data. Here, best response dynamics have single values with very high frequency surrounded by much lower frequency values whereas imitation dynamics result in distributions closer to a skewed Gaussian with many values having mid range frequencies over a larger range (Figure A.1). The mechanism driving these distributions is that best response dynamics converge to a global optimum rapidly as each node evolving its strategy at a given time step is always able to access the strategy it is not using. In contrast, imitation dynamics allow only for sampling strategies that are in use by connected nodes. Therefore, the system can drastically change between perturbations as seen in Figure 2.3.

2.8 Multi-parameter analysis

To gain insight into the different qualitative regimes within each network topology, the data was visualized in parameter planes visualized as heatmaps. The parameters varied in the heatmaps were c, the cooperators harvest and d, the defectors harvest. This is

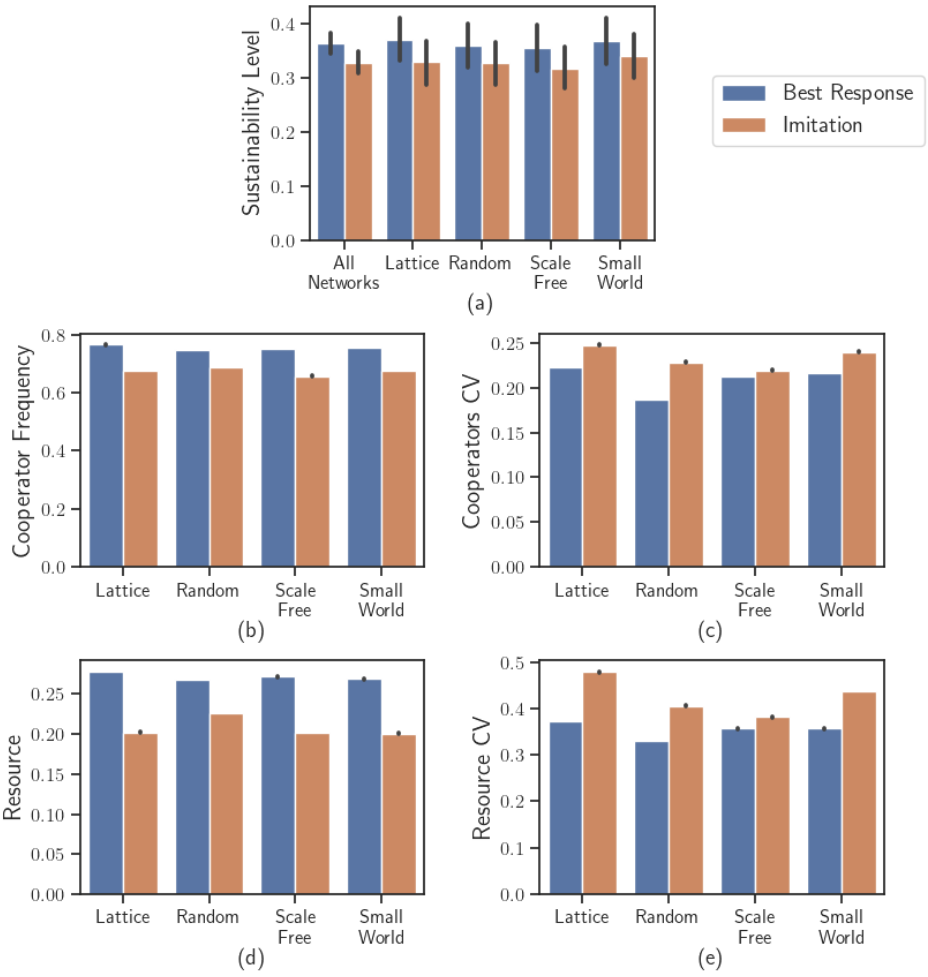


Figure 2.2: Mean metrics relating to the systems' sustainability level defined as the frequency of simulations in which the resource is not depleted below ϵ. Best response dynamics promote higher sustainability (a), cooperator frequency (b) and resource level (d) at equilibrium while also having lower coefficients of variation (CV) (c,e). Error bars represent the 95% confidence interval.

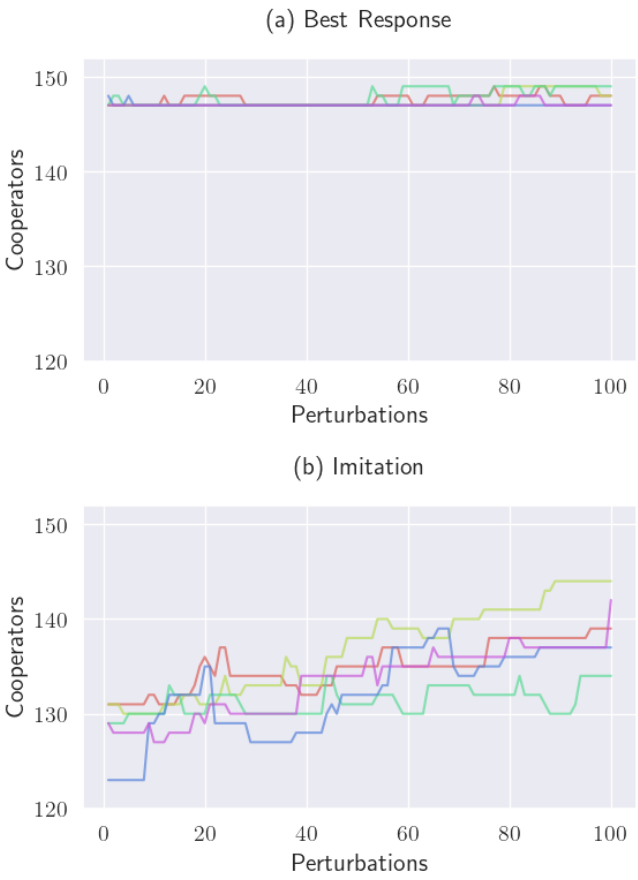


Figure 2.3: Cooperators at equilibrium after perturbations for 5 runs on a scale-free network with identical parameters. Each colour represents a different stochastic realization of the model. In the best response system (a), the number of cooperators does not vary significantly after the first perturbation; however, in the imitation system (b) the there is much more variance.

because these two parameters cover a much larger range than the others as they can vary greatly among human systems and have much more relaxed realism constraints [85, 86]. Heatmaps for the scale-free network are shown here as network topology did not have a significant effect on these results. Examining the heatmaps for the mean sustainability over each topology, the plots show that as both harvesting rates are increased, the system becomes less sustainable to the point where not a single simulation can persist (Figure 2.4a,b). Intuitively, an increase in sustainability would be correlated with an increase in cooperation. However, the heatmaps for the number of cooperators at equilibrium on all network topologies in the imitation system show that the number of cooperators increase with both c and d such that they are maximal at the point in which the system cannot persist due to over-extraction of resources (Figure 2.4d). In this case, stress in the system encourages cooperation. For increasing values of c, an increase in cooperation is expected because the payoff, and therefore the incentive to cooperate, has increased. With high values of d, the increase in number cooperators is less obvious. When d increases, the incentive to defect is strong, which could cause a large-scale propagation of defection in the network. However, with a large defectors' harvest, the resource decreases very quickly and once it reaches a very low level, the cooperators' punishment becomes very large, to the point that even with few cooperators in the system, defectors adjacent to them cannot harvest profitably and are likely to switch to the cooperator strategy.

For best response dynamics, there is a difference to this narrative as c is increased. At high c levels ($c > 0.4$), there is a drastic increase in cooperators and the average number of cooperators does in fact decrease with d. This could be caused by having a large enough incentive to cooperate at initialization such that there are never enough defectors to bring the resource to a level in which the cooperators' punishment causes defective harvesting to be unprofitable. This agrees with the resource having its maximal value at $c = 0.5$ as well as the system showing a near saturation of cooperators at $c = 0.6$ (Figure 2.4c,e). At this saturation of cooperators, the resource level decreases for low d, demonstrating that above $c = 0.5$, cooperators have taken over the system and any higher payoff for them results in greater resource extraction as defectors are not influencing the system. With imitation dynamics, the system does not experience the same regime change to a saturation of cooperators, most likely due to clusters of defective nodes which are immune to invasion by cooperators.

Regarding modularity in the best response system, for $c \leq 0.4$, modularity is maximized when both d and c are minimized (Figure 2.4g). This happens in order for both cooperators and defectors to maximize payoff as now punishment and cost of punishment play a much more significant role in reducing payoffs for nodes connected to others with opposing strategies. For $c > 0.4$, modularity approaches 0 since the network is approach-

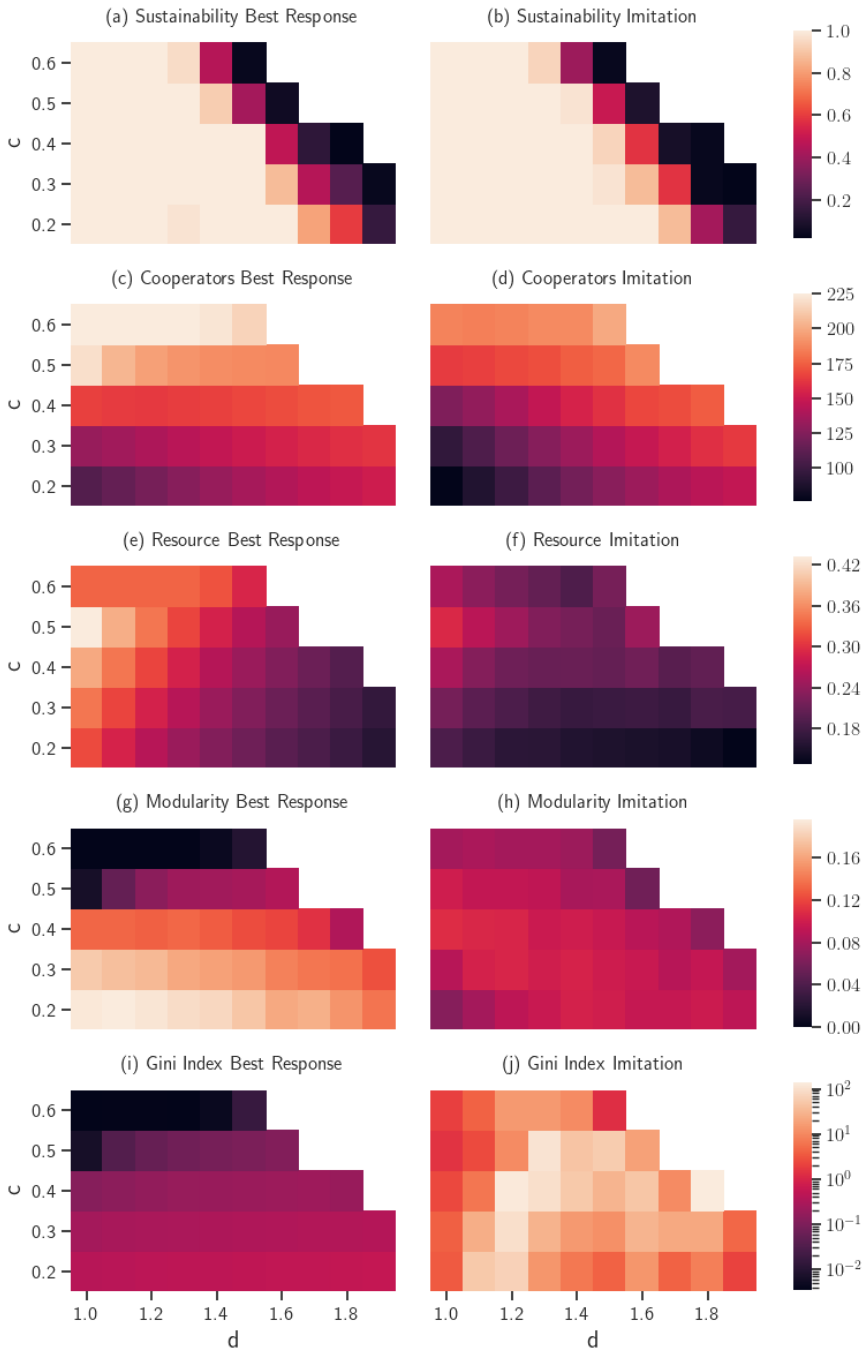


Figure 2.4: Comparing mean values for best response (left) and imitation dynamics (right) on the scale-free network for baseline parameters. The axes are d, corresponding to the defectors' harvest and c, corresponding to the cooperators' harvest. The values shown are sustainability (a, b), cooperators at equilibrium (c, d), resource at equilibrium (e, f), modularity at equilibrium (g, h) and the absolute value of the Gini index at equilibrium (i, j).

ing a globally homogeneous population. In the imitation system, these phenomena are not observed (Figure 2.4h). Instead, modularity is maximized for intermediate values of d and c. The reason it is not maximized for low values of d and c could be because the number of cooperators is low enough that the system is approaching global homogeneity towards defection.

Finally, the distribution of payoffs across the network was quantified using the Gini index. This metric can be interpreted as zero being complete equality with the distance from zero measuring the extent of inequality among agents. Usually the index is in the range $[0, 1]$; however, including negative individual payoffs in its calculation can take it out of this range. Additionally, a negative Gini index means that the net total payoff of the system is negative. When calculating the Gini index as well as debt, the cumulative payoff of each agent was used. When comparing systems with heatmaps, the absolute value of the Gini index is used in order to have a suitable resolution to compare both systems. Under the baseline parameters, the best response system has much lower Gini index values (Figure 2.4i,j). The inequality in the system increases with d and decreases with c. This is seen across all parameters for best response dynamics; however, the negative correlation with c is much stronger for the complete parameter space. This correlation with d and c follows from the fact that as the two parameters approach each other so do π_d and π_c, increasing equality in the system. There is also a sudden decrease in Gini index in the same high c region where the number of cooperators and network modularity also experience a drastic transition. This regime experiences the highest equality because the system is saturated with cooperators and therefore all payoffs are equal. For imitation dynamics, the Gini index is much higher and does not show a strong correlation with any parameters or variables. The reason for the drastic inequalities in payoff could be due to defective nodes which only share edges with like strategy nodes. These nodes will not receive any punishment and will harvest every round at the highest possible level. Unless the neighbouring nodes are perturbed, they will continue to imitate the high earning defector which could result in accumulation of debt if they share edges with cooperators, especially when the resource is at a low level. As the modularity is still relatively low for these high magnitude Gini index values, these groups of like strategies must not reach a significantly large size. To differentiate these from clusters around multiple nodes large enough to significantly influence the modularity, we will call this phenomenon 'micro-clustering'.

Consistently across all network topologies for best response dynamics, the extrema for Gini index values are positive and less than 1, meaning that there is no excessive debt among individual nodes and all systems have a net positive cumulative payoff (Figure 2.5c). For imitation dynamics, the extrema of the Gini index surpass those of best response

dynamics by several orders of magnitude and in fact vary significantly across network topologies. Regarding the frequency of debt in the system, global and individual debt is most prevalent in scale-free networks and rarest in small-world networks (Figure 2.5d). This could be explained by separately examining the clustering of individual strategies in each network quantified by the relative transitivity for each strategy type (Figure 2.5a,b). Transitivity offers a metric for clustering through the presence of triadic closure in the network. Due to the impossibility of triads in lattice networks, this topology was excluded from the transitivity analysis. When looking at same-strategy clustering, transitivity was calculated over the whole network as well as a subset of the network made up of nodes with a single strategy and the edges they share. Since the transitivity over the entire network varied significantly across topologies, to compare topologies, the transitivity of a given strategy type was scaled by the transitivity of the entire network. This gives us the relative transitivity, which measures whether or not the presence of triadic closure in a given strategy type is greater than that found in the entire network. For scale-free networks with imitation dynamics, the transitivity of cooperators is significantly lower than that of the entire network (giving a relative transitivity less than 1) and the transitivity of defectors is significantly higher (giving a relative transitivity greater than 1). This is due to the scale-free degree distribution which allows a small number of very high-degree nodes which are not found in other the other network topologies. If these high-degree nodes are initialized as defectors, they can be central in much larger 'micro-clusters' than are possible on other network topologies. The central defecting node of this 'micro-cluster' is then able to influence a greater number of nodes to remain defectors when it is unprofitable at low resource levels, resulting in a larger number of nodes accumulating debt. Interestingly, in best response systems, these 'micro-clusters' found in scale-free networks would more likely be surrounding a cooperative node. In small-world networks, the transitivity of each strategy is very close to the transitivity of the whole network. This is because having nodes with degrees as high as those found in scale-free networks has a low probability in this system since the re-wiring probability is low ($\beta = 0.08$). Additionally, these networks have a low degree of separation between any given pair of nodes, making it difficult for 'micro-clusters' to form, which are strong against invasion from the opposite strategy as there is a greater probability of high-earning cooperators to share edges with defectors.

2.9 Power-law analysis

In this section, all power-law fitting was conducted using the *powerlaw* Python package [87]. As smaller values in empirical data usually do not follow power laws and power-

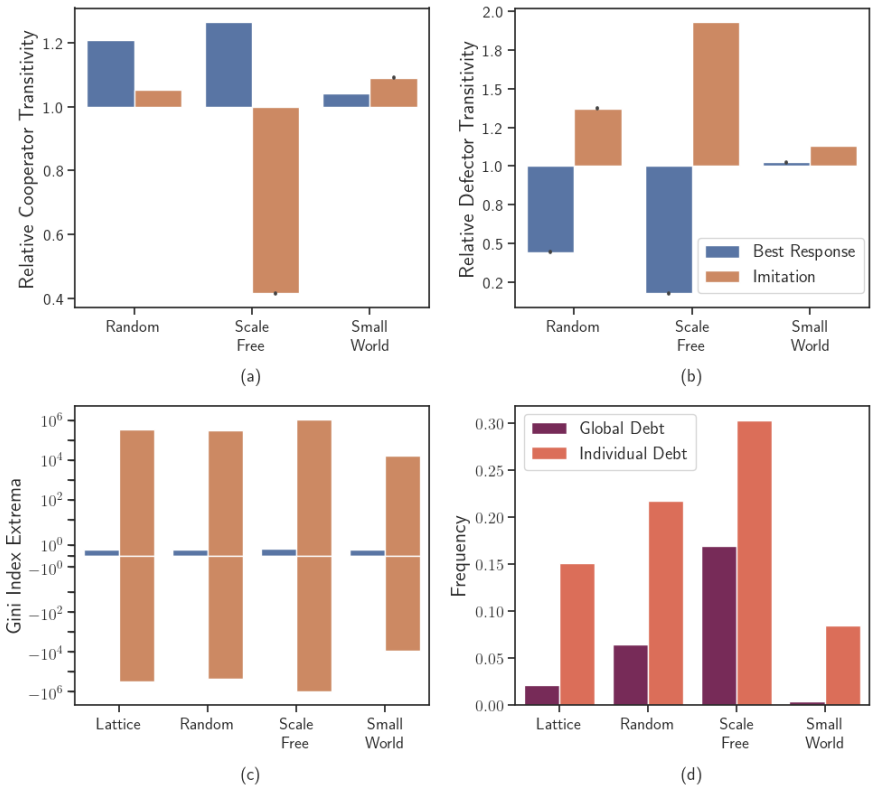


Figure 2.5: The clustering of like strategies as quantified by relative transitivity varies significantly across network topologies (top). Wealth distribution across network topologies (bottom). In imitation systems there is much more inequality than best response systems as seen in the Gini index extrema (c). There are also systems with debt in imitation dynamics (d) which is not seen in best response systems.

law fitting is often performed on the tails of probability distributions, the data was fit on cascades larger than the size of the system, $N = 225$ [88].

To quantify which cascade size distributions fit a power-law, the Kolmogorov– Smirnov goodness-of-fit test was performed on the data. As the Kolmogorov– Smirnov test is extremely sensitive to noise, instead of testing whether the distribution is a power law, the fit of a power law is compared to that of an exponential distribution. If the distribution is closer to that of an exponential, we can rule out that it is a power law fit, and if it is closer to a power law we can at least be certain that it is a heavy-tailed distribution [87, 88].

From the results of the Kolmogorov– Smirnov test, all of the cascade size distributions for imitation dynamics fit an exponential distribution more closely than a power law distribution, immediately ruling out the potential for self-organized criticality in these systems. However, for best response dynamics, about 21% of the parameter space had distributions that were closer to a power law for lattice (20.89%), random (22.44%) and small-world (21.77%) networks and therefore confirm the existence of heavy-tailed distributions suggesting the occurrence of self-organized criticality in these systems. For scale-free networks, only 6.67% of the parameter space had distributions closer to a power law. This could be due in part to the existence of larger 'micro-clusters' surrounding high degree nodes as these have a very low probability of invasion by the opposite strategy and could decrease the probability of very large cascade sizes leading to much fewer heavy-tailed cascade size distributions.

Along with displaying a power law distribution, self-organized critical systems must display system-wide effects that are triggered by small perturbations [57]. In terms of this model, the extent of system-wide cascades can be examined through the maximum cascade size as well as the proportion of cascades larger than the number of nodes in the system (see Figure 2.6b). For the simulation presented in this paper, it would be the proportion of cascades larger than $N = 225$. For system wide cascades, systems with best response dynamics had a significantly larger maximum cascade size (by 6-7 orders of magnitude) as well as a significantly greater proportion of cascades of size greater than N (Figure 2.6a). Interestingly, scale-free networks have a much higher proportion of cascade sizes larger than N suggesting that the high-degree nodes unique to this system can in fact promote strategy switching up to a point (Figure 2.6b). This, along with the results from the Kolmogorov– Smirnov test strongly suggests that systems with best response dynamics exhibit self-organized criticality much more than those with imitation; however, the existence of high-degree nodes in scale-free networks acts contrary to this phenomenon.

Finally, there are similarities across all network topologies regarding fitting the cascade size distribution to a power-law. The mean standard error, which determines the closeness

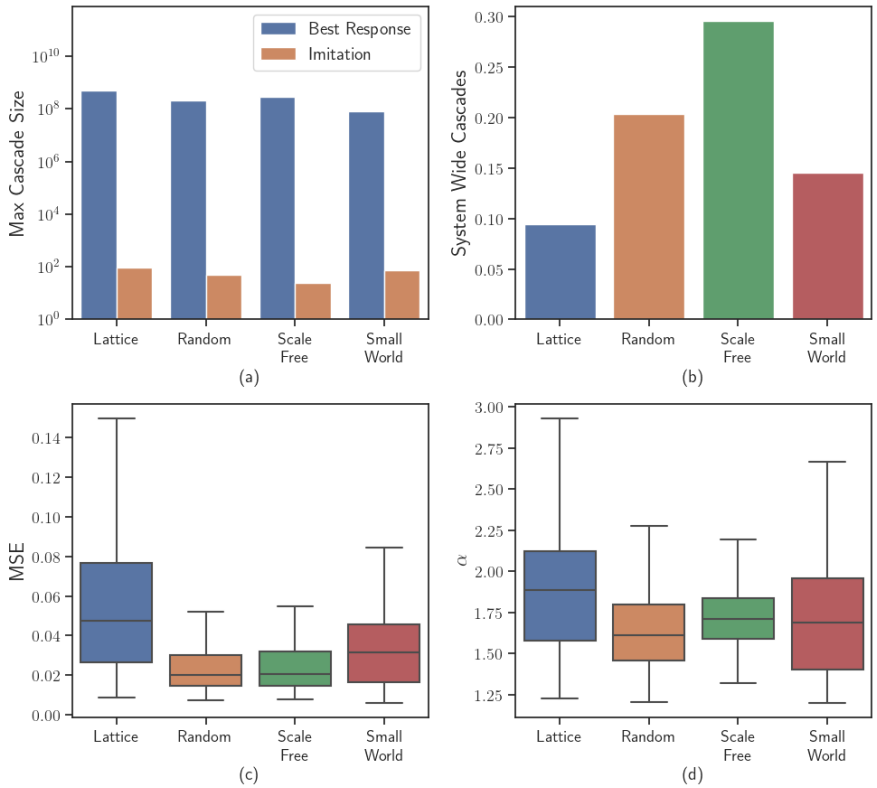


Figure 2.6: Metrics of self-organized criticality. Best response dynamics show evidence of self-organized criticality due to large maximum cascade sizes and a significant frequency of system wide cascades (top). All topologies in the best response systems show similarities in the mean standard error (MSE) and slope (α) of the power-law fit (bottom).

of the power-law fit over the entire parameter space is similar across all topologies despite scale-free networks scoring much lower as a power-law distribution in the Kolmogorov–Smirnov test (Figure 2.6c). Additionally, the slope of the power-law fit given by α does not significantly differ across topologies (Figure 2.6d). This implies an element of universality in the shape of the cascade distributions over all network types, even as the scale-free network distributions are less heavy-tailed. Regarding properties of the node being perturbed before each cascade, neither the local clustering coefficient nor the degree of the perturbed node were correlated with the cascade sizes.

2.10 Discussion

In this model, a tragedy of the commons was averted in many cases, with best response dynamics increasing the amount of sustainable runs compared to imitation dynamics. Just as importantly, outcomes were more equitable under best response dynamics, as measured by individual debt and the population's Gini index. Additionally, the cascade size distribution in systems with best response dynamics had a closer fit to a power-law and larger cascade sizes in general, demonstrating evidence of self-organized criticality.

Spatial structure, however, did not have a drastic effect on the persistence of the resource or number of cooperators at equilibrium. This agrees with the 2012 empirical study by Gracia-Lázarao et al. in which humans played a spatial Prisoner's Dilemma on a lattice and a scale-free network. In this study, there was no significant difference in cooperation among the two network topologies and in fact, the level of cooperation was comparable to that which arose in unstructured populations [29]. This spatial independence is surprising given the importance of local interaction in this model. However for other metrics, the network topology did have a significant effect, such as for the Gini index, the clustering of strategies and the goodness-of-fit of the cascade size distribution to a power-law. Most of these differences can be attributed to the existence of high-degree nodes in scale-free networks which can be central in 'micro-clusters' of the same strategy type.

The strategy propagation mechanism proved to be extremely important in determining system sustainability. In all cases, best response dynamics were much more successful than imitation at averting the tragedy of the commons. This, along with the rapid convergence of best response dynamics, complements similar findings when comparing strategy propagation in the N-player iterated Prisoner's Dilemma game [38]. These results hint at the importance of context-dependent foresight (modeled by best response dynamics) as opposed to pure imitation of peers regardless of context for the sustainability of common-pool resource-dependent human systems. They also suggest that in systems where imitation is important, more sustainable outcomes might be achievable by increasing the rate of social learning. This finding is consistent with other recent research on social-climate systems [89]. In both strategy update rules, each agent is acting in its own self interest; however, the agent who makes choices based on its current state and relation to others benefits itself and the overall community more than a self-interested agent who defaults to following another individual's path to success.

Imitation can be seen as a representation of social learning through copying others. History has shown social learning to be highly adaptive, primarily through the propagation of information such as technology and survival skills. However, in the past, this social

learning has been practised with a combined knowledge of an individual's context and social norms, as modelled with best response dynamics [90]. In this model, pure social learning is less sustainable in comparison to independent prediction modelled by best response dynamics. This could be due to the fact that best response dynamics take into account the current state of the resource while imitation dynamics do not. Consequently, agents using a best response strategy update rule are aware of the consequences of over-harvesting a resource near depletion. Therefore, they explicitly take into account the social norms of the community when making a decision on how to harvest. Imitating agents however, are only trying to increase their profitability using a strategy that was advantageous for others in the past, thus only indirectly accounting for social norms. This implicit time delay in decision making can have significant detriment to the wealth of the resource as well as the individual. With this time delay, it is possible that if the model was run for a longer period of time, the frequency of cooperators may converge to that of best response dynamics, warranting future work where the maximum number of perturbations to the system is increased.

The hypothesis of the benefit of best response dynamics versus imitation dynamics for common-pool resource problems should be further tested with existing models in the literature that employ either update rule. An identical model with the other update rule can then be analysed and the results of both models can be compared to investigate whether best response dynamics do in fact prove more beneficial to the community than imitation over a diverse array of model assumptions. Models with more complex harvesting strategies such as mixed strategies or strategies with memory of previous interactions should be included in this comparison.

Additionally, the socio-ecological coupling in this model could be modified such that the fecundity of the resource (F, from eq. 4) could be dependent on the number of cooperators. This has been examined in a number of studies with well-mixed populations and has shown to support the co-existence of cooperators and defectors as well as sustainable resource levels [91–93]. A further extension to this model could be to include a small number of cooperating nodes whose strategy remains fixed for the entirety of the simulation. This begs the question as to which nodes would be selected for maximizing sustainability, which has been explored in numerous studies [94–96]. This method of system control could also demonstrate additional qualitative differences between the topologies of social networks simulated. Furthermore, a model with global punishment could be compared to the model presented in this paper to investigate how including network structure in punishment affects the sustainability of the system. Finally, this model could be tested with empirical networks and case studies to verify whether the assumptions and conclusions gleaned from this study remain true in practice.

Chapter 3

Reinforcement learning model

The model presented in this chapter is an extension of the reinforcement learning common-pool resource models developed by von der Osten et al. as well as by Zhu and Kirley [13, 74]. Similar to the previous chapter, this model represents a community of agents that harvest a common-pool resource at discrete time intervals. A fundamental difference, however is that rather than having agents choose from a predefined set of strategies, they are learning much more complex strategies by optimizing a reward through experience with the environment. In the following sections, the model is introduced and preliminary results are presented.

3.1 Resource dynamics

Similar to the model presented in Chapter 2, a generalized common-pool resource, R_t, is harvested synchronously by agents on a social network who each invest an effort, x_i into harvesting. The change in resource at time t is represented by a growth term and a harvest term that are dependent on the state of the resource as well as the cumulative effort, $X = \sum x_i$ at the previous time step,

$$R_t = R_{t-1} + \Delta R \tag{3.1}$$

$$R_t = R_{t-1} + G(R_{t-1}) - H(X_{t-1}, R_{t-1}) \tag{3.2}$$

Just as in the game-theoretic model, $G(R)$ is a logistic growth function,

$$G(R) = r_g R \left(1 - \frac{R}{K}\right) \tag{3.3}$$

having a maximal growth rate r_g, and carrying capacity K. The harvest, $H(X, R)$ is modelled differently than the previous model, using the Cobb-Douglas production function for consistency with previous reinforcement learning common-pool resource models,

$$H(X, R) = \beta X^{\alpha} R^{1-\alpha}, \quad \alpha \in (0, 1) \tag{3.4}$$

The Cobb-Douglas production function can be interpreted as follows. X is the labour invested to the system, which we will refer to as the cumulative effort, R is available capital, represented by the resource in this system, and β is scaling factor representing the efficiency of production. The exponents α and $1-\alpha$ represent what is known in economics literature as *output elasticities*. These values represent the extent to which the cumulative effort and the state of the resource affect the amount of production. Looking at the partial derivative with respect to X,

$$\frac{\partial H}{\partial X} = \frac{\alpha H}{X} \tag{3.5}$$

we can see that for a constant resource level, the response of H to changes in X is proportional to α. The same holds for how H responds to changes in R, with the proprotionality being $(1 - \alpha)$. As $H(X, R)$ is a homogeneous function of degree one, meaning that the exponents, α and $(1 - \alpha)$ sum to one, the production changes exactly in proportion to the same relative change of the cumulative effort and capital. For example, a 1% increase in both X and R result in a 1% increase in $H(X, R)$ and economists call this property *constant returns to scale* [97–99].

The harvest is distributed among agents proportional to the effort they invested and as labour incurs costs, each agents harvest is reduced proportional to the effort they invested. This proportionality is called the *cost per unit effort*, and is a global parameter, denoted c. From this, the payoff of each agent and the collective payoff is calculated as follows,

$$\pi_i = \frac{x_i}{X} H(X, R) - c x_i \tag{3.6}$$

$$\Pi = \sum_i \pi_i = H(X, R) - cX \tag{3.7}$$

Note that in equation 3.7, the cumulative payoff, Π is positive when $H(X, R) > cX$. Solving for X, we have a condition on the cumulative effort such that cumulative payoff is positive or in other words, the system is profitable. The cumulative efforts that allows for

a net profitable system is denoted,

$$X_{\text{profit}} < \left(\frac{\beta}{c}\right)^{\frac{1}{1-\alpha}} R \tag{3.8}$$

Similarly, solving for $\Delta R > 0$ gives us a condition for positive net resource growth,

$$X_{\text{sus}} < \left(\frac{r_g R(1-\frac{R}{K})}{\beta R^{1-\alpha}}\right)^{\frac{1}{\alpha}} \tag{3.9}$$

From these conditions, we have three distinct regions consisting of the disjoint sets $X_{\text{profit}} \setminus X_{\text{sus}}$, $X_{\text{sus}} \setminus X_{\text{profit}}$, and $X_{\text{profit}} \cap X_{\text{sus}}$, corresponding to regions that are purely profitable, purely sustainable and both profitable and sustainable (Figure 3.1).

By finding the intersection of upper the bounds of X_{sus} and X_{sus}, we can derive a condition for the optimal collective effort,

$$X_{\text{opt}} = K\left(\frac{\beta}{c}\right)^{\frac{1}{1-\alpha}}\left(1 - \frac{\beta}{r_g}\left(\frac{\beta}{c}\right)^{\frac{\alpha}{1-\alpha}}\right) \tag{3.10}$$

This expression informs the expected cumulative effort when the system is in a steady state, as it allows for the highest collective payoff without depleting the resource. We can see here, that $X_{\text{opt}} \geq 0$ when $c \geq \beta^{\frac{1}{\alpha}} / r_g^{\frac{1-\alpha}{\alpha}}$. If this condition is not satisfied, an optimal collective effort does not exist as effort is strictly non-negative.

Parameters for the common-pool resource system were chosen to be as similar as possible to those used by von der Osten et al. [13]. The main difference however is that there is no lower bound to the total effort possible in the system other than that imposed by realism constraints (i.e. $X \geq 0$). The motivation behind this was to give agents the option to refrain from harvesting if the system was unprofitable. Additionally the maximal growth rate, r_g was increased from 0.5 to 0.65 as the original growth rate led to repeated depletions over most parameters during the testing phase. All parameter values are shown in Table 3.1.

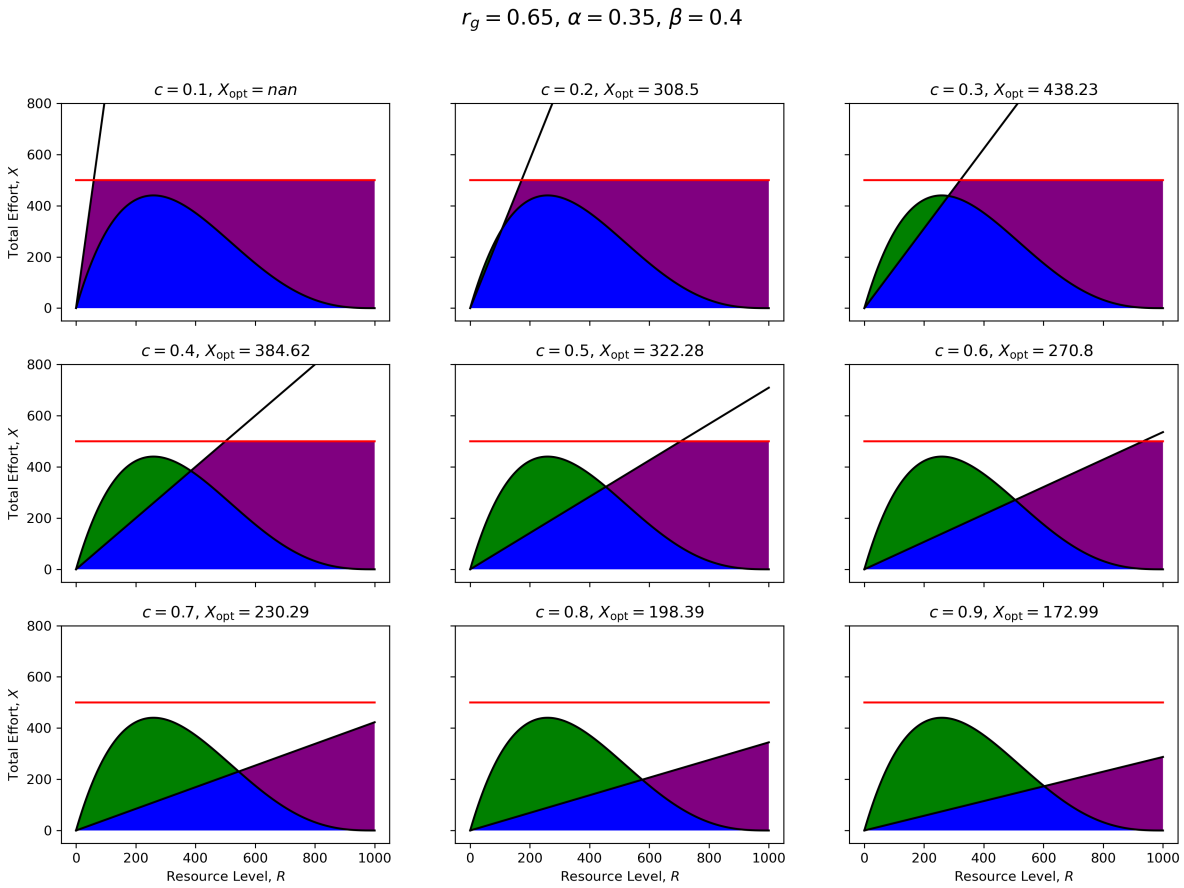


Figure 3.1: The three distinct harvesting regions, $X_{\text{profit}} \backslash X_{\text{sus}}$ (purple), $X_{\text{sus}} \backslash X_{\text{profit}}$ (green), and $X_{\text{profit}} \cap X_{\text{sus}}$ (blue) for varying values of the cost per unit effort, c.

Maximal growth rate (r_g)	0.65
Carrying capacity (K)	1000
Output elasticity (α)	0.35
Production efficiency (β)	0.4
Maximum cumulative effort ($X_{\max}$)	500

Table 3.1: Parameters that remained constant over all simulations and their values.

Before being fed to the deep learning model, all rewards were scaled to $[-1, 1]$. Additionally, during training, a penalty of -10 was incurred to all agents if the resource was depleted before the termination of an episode. This early episode termination penalty is common in reinforcement learning models and can be interpreted as either a fine or significant loss in profits once the resource being harvested can no longer persist.

3.2 Spatial structure

This model structures the social dynamics by arranging the agents on a social network, allowing for local interactions to influence the reward function. The network is comprised of N agents represented by nodes on a network, where the reward of agent i is dependent on the efforts of all agents for which they share an edge. Identical small-world graphs were used as social networks for all simulations. This topology was chosen as empirical offline social networks often demonstrate small-world characteristics [52, 100–103] In selecting the graph to use in all simulations, 10^6 small-world graphs were generated using the Watts-Strogatz algorithm (see Section 1.3.1), and the most characteristic small-world network was chosen. This was decided by the shortest average path length and the transitivity of the network.

3.3 Reward function

The reward function each agent uses to learn a policy is defined by two terms, a profit term, ξ_i, and a conforming term, λ_i. The profit term is defined using a logistic piecewise

function that reaches its maximal value when an agent receives the highest possible payoff during a given time step, $\pi_{\max}$. Since $\pi_i \propto \Pi$, we can define this maximum payoff as,

$$\pi_{\max} = \begin{cases} x_{\max}\left(\frac{H}{X} - c\right) & \frac{H}{X} - c > 0 \\ 0 & \frac{H}{X} - c \leq 0 \end{cases} \tag{3.11}$$

Here we can see that if the system is unprofitable, then the maximal payoff an agent can receive is 0 (as opposed to a negative payoff), and this is only possible when the agent has effort, $x_i = 0$, meaning they are not participating in the system during that round. The conforming term is defined by the local interactions of the agents and represents the social incentives to conform to existing norms in a social group. It is formulated as a Gaussian function centered at the mean local effort defined for agent i as,

$$\hat{x_i} = \frac{\sum_{j \in \mathcal{N}_i} x_j}{|\mathcal{N}_i|} \tag{3.12}$$

where $\mathcal{N}_i$ is the set of neighbours of agent i. A Gaussian function was chosen as it obtains its maximum when an agent harvests at $\hat{x}_i$ and is symmetric about that mean value. This function is defined in a piecewise form such that the tail of the function farthest from $\hat{x}_i$ reaches a value close to zero. Adding these two terms together, weighted by w, we have,

$$r_i = w\xi_i + (1 - w)\lambda_i \tag{3.13}$$

$$\xi_i = \begin{cases} \frac{1}{\exp(-k(\pi_i - (\pi_{\max} + s))) + 1} & \pi_{\max} \neq \pi_{\min} \\ 0 & \pi_{\max} = \pi_{\min} \end{cases} \tag{3.14}$$

$$\lambda_i = \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{3(x_i - \hat{x_i})}{x_{\max} - \hat{x_i}}\right)^2\right) & \hat{x_i} < \frac{x_{\max}}{2} \\ \exp\left(-\frac{1}{2}\left(\frac{3(x_i - \hat{x_i})}{\hat{x_i}}\right)^2\right) & \hat{x_i} \geq \frac{x_{\max}}{2} \end{cases} \tag{3.15}$$

$$\tag{3.16}$$

where weight, $w \in [0, 1]$ is a global parameter and $r_i, \xi_i, \lambda_i \in [0, 1]$.

As π_i is a function of x_i, we can write both rewards as a function of x_i,

$$r_i(x_i) = w\xi_i(x_i) + (1-w)\lambda_i(x_i) \tag{3.17}$$

$$\xi_i(x_i) = \begin{cases} \frac{1}{\exp\left(-\ln(\gamma)\left(\frac{x_i}{x_{\max}} - \left(1 + \frac{\ln\left(\frac{1}{0.99}-1\right)}{\ln(\gamma)}\right)\right)\right)+1} & X < \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R \\ \frac{1}{\exp\left(\ln(\gamma)\left(\frac{x_i}{x_{\max}} + \frac{\ln\left(\frac{1}{0.99}-1\right)}{\ln(\gamma)}\right)\right)+1} & X > \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R \\ 0 & X = \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R \end{cases} \tag{3.18}$$

$$\lambda_i(x_i) = \begin{cases} \exp\left(-\frac{1}{2}\left(\frac{3(x_i - \hat{x_i})}{x_{\max} - \hat{x_i}}\right)^2\right) & \hat{x_i} < \frac{x_{\max}}{2} \\ \exp\left(-\frac{1}{2}\left(\frac{3(x_i - \hat{x_i})}{\hat{x_i}}\right)^2\right) & \hat{x_i} \geq \frac{x_{\max}}{2} \end{cases} \tag{3.19}$$

For the steps taken to make this change of variables, an explanation for parameters chosen in the logistic and Gaussian function, and a graph of the reward function over various values for c and w, see Appendix B.

3.4 Two learning models

Here, the learning dynamics are informed by reinforcement learning techniques, where an agent's policy is learned through both deep Q-learning (DQN) for a discrete action space or deep deterministic policy gradient (DDPG) for a continuous action space [65, 66]. With a discrete action space, the set of actions available to an agent, $a \in \mathcal{A}(s)$ must be defined prior to initializing the model. In a generalized CPR system such as this, the choice of such actions can be difficult as there is little empirical work to inform values for these abstract actions. Often modellers choose social parameters to allow for the system to persist, within realism constraints. An arbitrary choice for $\mathcal{A}(s)$ could have unintended effects on the model outcome motivating a novel approach to deriving $\mathcal{A}$ from the dynamics of an identical model free of limitations to the agents' action space. Currently, only the DDPG model has been trained and run over all parameters, however future work will use the results from these experiments to inform the actions of a DQN model on a significantly larger population of agents.

3.5 Training the DDPG model

3.5.1 Exploration and exploitation

The agents in this system were trained over an episode length of $T = 1000$. This means that the system was run repeatedly for a maximum of 1000 time steps, before the agents' efforts and the state of the common-pool resource were reset. Training over a single episode constitutes one epoch, and the model is trained for a maximum number of epochs, $N_{\text{epoch}} = 400$. To allow for significant exploration at the beginning of training, all agents underwent a phase of pure exploration for 6000 time steps, in which the agents do not learn and are only collecting experience through a policy of random actions. After this phase of pure exploration, the agents begin to learn a policy, however they continue to explore the system as the action output contains an additive noise term, $\mathcal{N}_t$, for which the variance decays at the end of each epoch. This noise term is sampled from a Gaussian distribution, such that $\mathcal{N}_{t_e} \sim \mathcal{N}(0, \sigma_{t_e})$, where $\sigma_{t_e} = \left(1 - \frac{0.9}{(N_{\text{epoch}}-1)} t_e\right)^3$ and $t_e \in \{0, 1, 2, ..., (N_{\text{epoch}} - 1)\}$ is the current epoch. This insures that σ_{t_e} decays from a maximal value of $\sigma_0 = 1$ to a final value of $\sigma_{(N_{\text{epoch}}-1)} = 0.001$. This ensures that each agent is exploiting a successful policy by the end of training. A cubic decay rate was chosen as it was found to stabilize training most effectively.

3.5.2 Common-pool resource system initialization and parameters

The common-pool resource system has essentially two types of variables that require initialization, the level of the resource, R, and each agents' harvesting effort, x_i. The resource was always initialized at its carrying capacity, $K = 1000$, and each agents' initial effort was sampled from a uniform distribution, bounded around $X_0 = 300$ (Table 3.2).

Variable	Initial conditions
Resource (R)	$R_0 = 1000$
Mean initial cumulative effort (X_0)	$X_0 = 300$
Individual efforts (x_i)	$x_{i_0} \sim \mathcal{U}(\frac{2}{3}\frac{X_0}{N}, \frac{4}{3}\frac{X_0}{N})$

Table 3.2: Initial conditions used at the beginning of each episode: the resource level and each agents' individual effort.

3.5.3 DDPG formulation

The DDPG model was trained using an actor network with four hidden layers with 64, 128, 256, and 128 nodes, respectively. The critic network had three hidden layers with 128, 256, and 128 nodes, respectively. In the critic network, the first hidden layer received an agents' observations as input and the agents' actions were added to the output of the third layer before being passed to the activation function. All activation functions were ReLU, except for the output of the actor network, which used a hyperbolic tangent function to scale the action output to $[-1, 1]$. The training used a batch size of 256, and an Adam optimizer with learning rates of 10^{-4} and 10^{-3} for the actor and critic networks, respectively [104]. The future rewards were discounted with a gamma factor of $\gamma = 0.99$ and target networks for both the actor and critic were updated using the soft update rule with $\tau = 0.001$.

Each agent, represented by a unique DDPG, observed a 4-tuple from the environment as their input consisting of their own harvesting effort, x_i, the average effort of their neighbours, $\hat{x}_i$, their own payoff π_i and the cumulative effort of all agents, X, such that at each time step, $s_i = (x_i, \hat{x}_i, \pi_i, X)$. The actions taken by each agent were their change in harvesting effort Δx_i, such that at each time step, $a_i = \Delta x_i$ and for each agent, $x_{t+1} = x_t + \Delta x_i$. A schematic of the DDPG algorithm and it's interaction with the common-pool resource environment is shown in Figure 3.2.

All neural networks were coded using the TensorFlow python library and the common-pool resource environment interfaced with the DDPG through the gym environment. Code for the DDPG algorithm was adapted from the model presented by Zhu et al. [74, 105, 106].

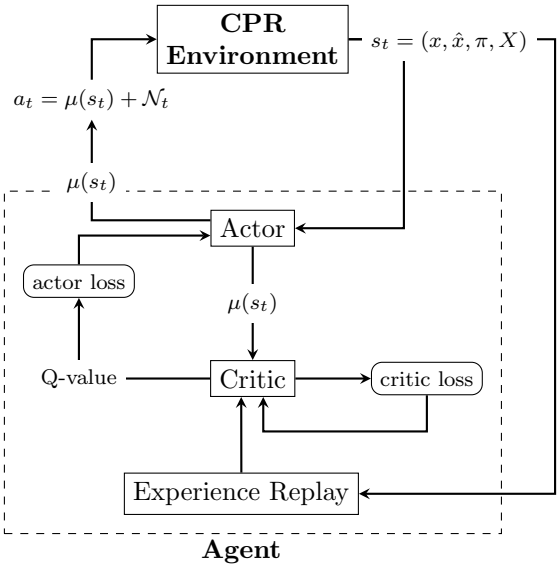


Figure 3.2: Simplified schematic of the DDPG structure for a single agent interacting with the common-pool resource environment. In the multi-agent system, the environment consists of both the resource dynamics and the efforts of a given agents' neighbours.

3.5.4 Training the agents

Due to limits in computation power, this model was trained for a system of $N = 24$ agents. Additionally, as opposed to training at each time step, common in most reinforcement learning models, the agents were trained simultaneously every 10 time steps. The motivation behind a slower time scale for the social dynamics relative to the resource dynamics was twofold. First, as each time step represents a consecutive instance of an agent harvesting a resource, it would be unrealistic to assume that this agent would change its harvesting strategy each time they harvest. This inertia of social dynamics has been discussed in various studies [80, 107, 108]. Second, having these two time scales greatly stabilized the system for both training and running the simulations.

For each distinct common-pool resource system (parametrized by the cost per unit effort, c), an individual optimal harvesting level can be predicted simply by equally distributing the global optimal harvesting effort between all agents such that $x_{\text{opt}} = X_{\text{opt}}/N$. Note that this assumes a well-mixed system which is not the case here, however it can serve as a useful approximation. Additionally, for $c = 0.1$, X_{opt} does not exist. From the training data (Figure 3.3), we can see that for intermediate cost values, $c \in [0.5, 0.7]$, the learned efforts closely approximate x_{opt}. In contrast, for low cost values, $c \in [0.1, 0.2]$ the efforts

show increased variability, often learning efforts much greater than x_{opt}, which approach x_{max}. In these regions, the short-term cost for an agent to increase their harvest is low, incentivizing a high harvesting effort to quickly maximize the profit term in their reward function. This is made more clear examining the effect of w on the learned effort. For $c \in [0.1, 0.5]$, a higher conform weight leads to a much slower increase in efforts as agents' are rewarded much more by conforming around their initial efforts rather than quickly maximizing their profit.

The variability in efforts for low c can be attributed to the fact that either X_{opt} is non-existent ($c = 0.1$), or unstable ($c = 0.2$). The instability is due to the fact that $X_{\text{profit}} > X_{\text{sus}}$ for most of the viable phase space, meaning that agents maximizing profit will rarely be harvesting in the sustainable region. For $c = 0.3$, this is not the case, allowing for agents to maximize their profit while still harvesting sustainably, so long as the resource is at a relatively low level. For high cost values, one can see that there is much less variability in efforts as there is a high risk for an agent to increase their harvesting effort. For these regions, we notice the weight of conformity having opposite effect on efforts, where higher weights correspond to learned efforts greater than x_{opt}. The reason behind this seemingly contradictory behaviour is that now, $X_{\text{opt}} < X_0$ so that the agents are learning to decrease their efforts over time, and process becomes much slower due to social inertia present when agents are primarily trying to conform, rather than maximize their payoffs.

3.6 Testing the DDPG model

Once trained, each model was tested with the same common-pool resource system it was trained on, with reduced noise where $\sigma = 0.05^3$ remained at a constant value for the entirety of the simulation. Each model was run under the same initial conditions as the system it was trained on with ten duplicates for each (c, w) pair. For the entirety of this chapter, all mean values for each (c, w) are averaged over these ten duplicate runs, beginning at $t = 1000$. In the testing phase, all agents learned every 100 time steps, rather than 10, as this greatly improved the sustainability of the system to allow for meaningful analysis.

3.7 Simulation results

For this model, as the two parameters that varied across simulations were the cost per unit effort, c and the weight of conformity, w, the following results focus on the effect that these parameters have on the dynamics of the common-pool resource system. Over all

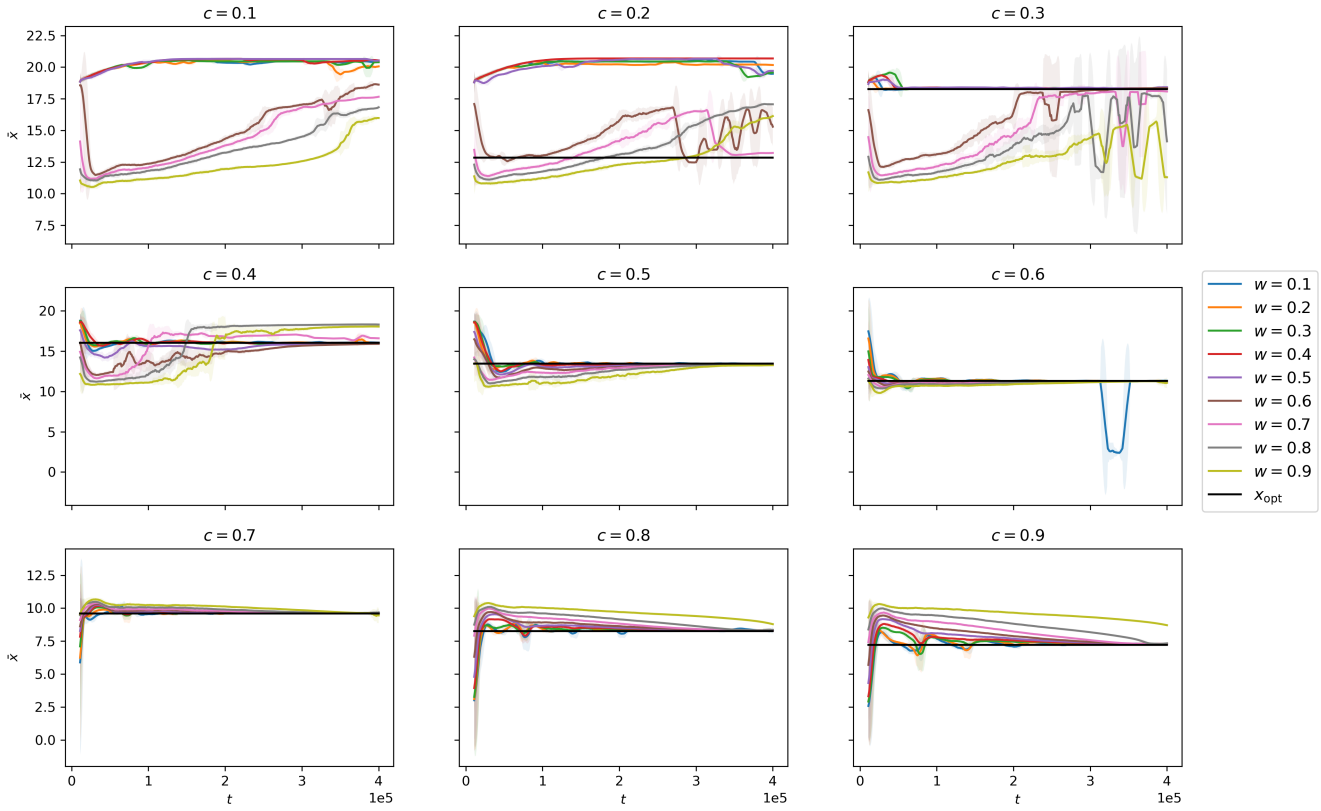


Figure 3.3: Mean efforts, $\bar{x}$ during training plotted with the optimal effort, x_{opt}, for each cost value, c. For $c > 0.3$, the efforts approach x_{opt}, however this rate of convergence is slower for a higher weight of conformity, w.

simulations of the trained model, the cost, c and weight of conformity, w had a significant effect on the mean resource level $\bar{R}$. This was tested pairwise with the Kruskal-Wallis test ($p < 0.01$). This also holds for the mean efforts $\bar{x}$ across all c and most w. One exception to this is that the mean effort for $w = 0.4$ and $w = 0.7$ did not differ significantly ($p \approx 0.831$).

For most values of the cost per unit effort, c, the mean resource level, $\bar{R}$ shows little correlation with the weight of conformity, w (Figure 3.4). The exceptions to this are for systems with high cost, $c > 0.6$ where there is a slight downwards trend in $\bar{R}$ with increased

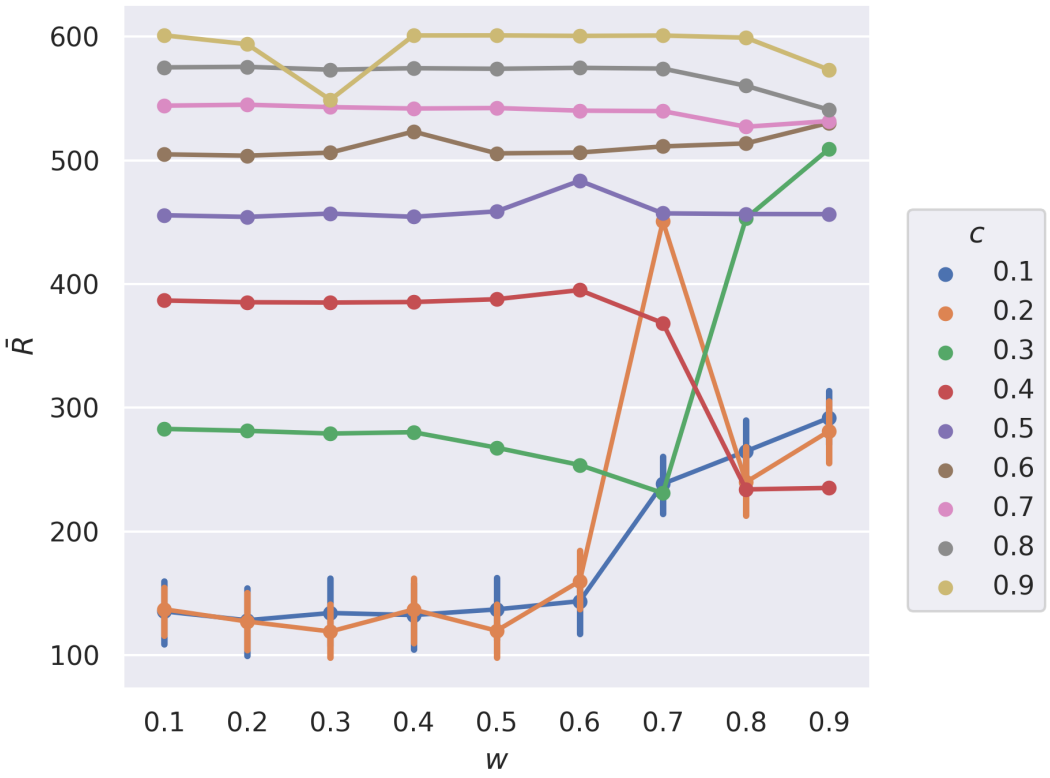


Figure 3.4: The effect of the conforming weight, w on the mean resource level, $\bar{R}$, for each cost per unit effort value, c. For high values of c, there is a slight downwards trend in $\bar{R}$ as w is increased.

w. This is due to inertia in the efforts caused by strong social conformity, discussed in Section 3.5.4. For c, the effect of increasing w shows contradictory results on the state of the resource. These upward spikes with increased w could be due to the fact that for low c, X_{opt} is either undefined or unstable, allowing for randomness in the system to influence the group harvesting dynamics in unpredictable ways.

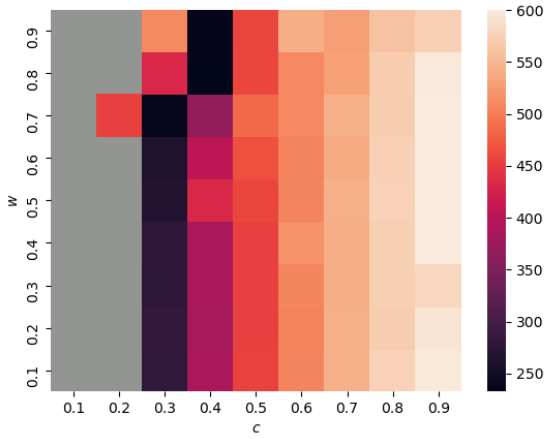
From the parameter planes, one can immediately observe that in most regions, the mean resource level, $\bar{R}$ is positively correlated with the cost per unit effort, c. This intuitively follows from the fact that if c is a high value, a greater harvesting effort comes with a greater risk to each agent and conforming to local efforts offers a much greater reward than diverging from social norms, even when w is low, as $H/X - c$ is much more likely to be negative making the system unprofitable. This can be seen from the mean asset level[1], $\bar{A}$, where the system drastically shifts from being profitable to unprofitable between $c = 0.6$ and $c = 0.7$ (Figure 3.5e). In the profitable regime, the resource is at a significantly higher level than R_{opt} while efforts are below x_{opt} and this difference increases with the weight of conformity, w. This trend could be in part, due to the fact that for $c < 0.6$ we have $x_0 < x_{\text{opt}}$. This means that a high weight of conformity will have a stronger effect of keeping efforts below x_{opt}, due to social inertia.

In the unprofitable regime, the resource is at a significantly lower level than its optimal level, and it follows that the mean efforts are above their optimal level. If an agent were to maximize its profit reward in this regime, it would need to refrain from participating in the system until it is profitable and then immediately harvest with effort $x_{\max}$. This policy is both difficult and risky to learn, as agents do not directly observe the state of the resource and incorrectly harvesting when the system is unprofitable incurs a large penalty. From the difference in mean effort to optimal effort, it is clear that agents were unable to learn this optimal policy as their efforts are consistently greater than the optimal effort, thus overharvesting and maintaining the system in an unprofitable state. This is supported by a decrease in the profit reward in the transition to profitable and unprofitable regimes, as well as an increase in the conform reward, showing that in the unprofitable regime, the agents' actions are informed by the conform goal, further removing environmental feedback from their decision making. Additionally, as the weight of conformity increases, the difference between the mean and optimal levels for both the resource and effort increase. This too, can be attributed to social inertia as for these c values, $x_0 > x_{\text{opt}}$.

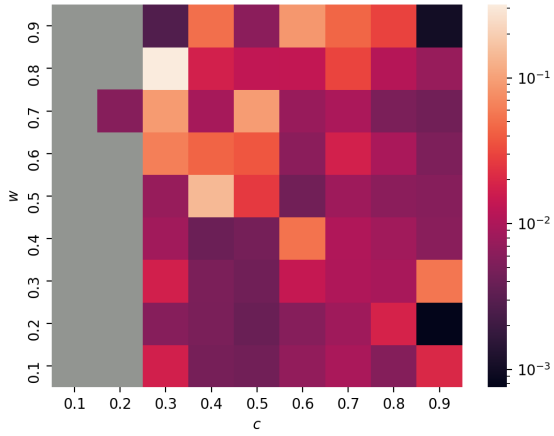
One area of interest is the stark transition from high to low resource levels, occurring at $c = 0.3$ for $w \in [0.1, 0.7]$ and $c = 0.4$ for $w \in [0.8, 0.9]$. At $c = 0.3$, this drastic decrease in resource level is expected, as it can be seen as a critical region between a sustainable

[1] An agent's assets is defined as their cumulative payoff where $A_i = \sum_t \pi_i$

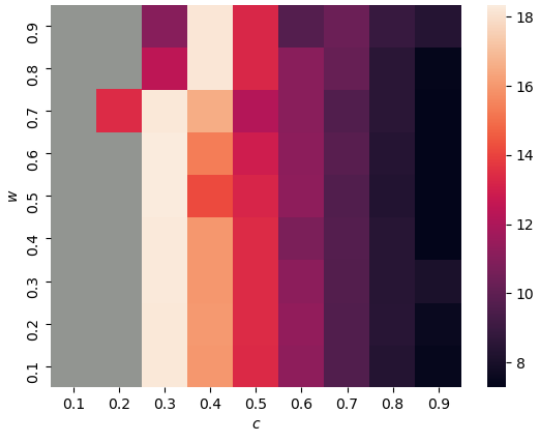
(a) Mean resource level, $\bar{R}$



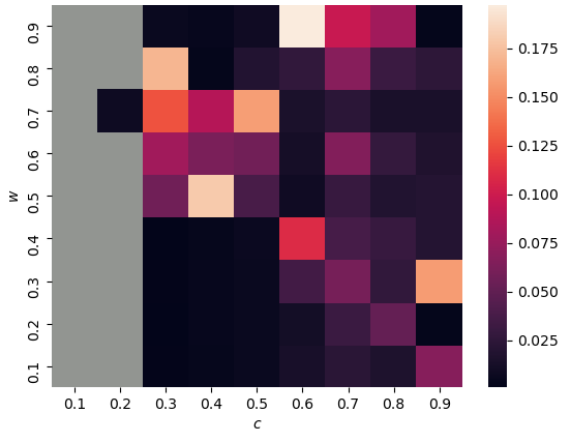
(b) Resource level coefficient of variation



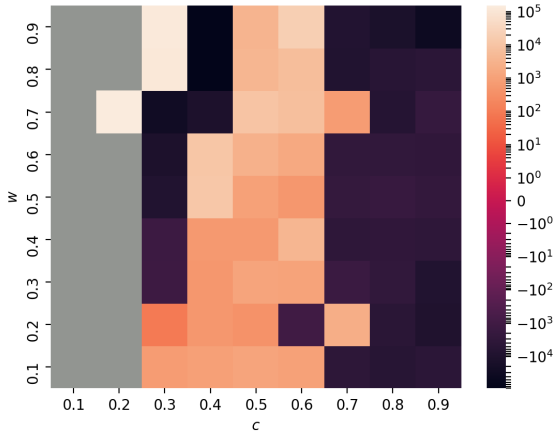
(c) Mean effort, $\bar{x}$



(d) Effort coefficient of variation



(e) Mean asset level, $\bar{A}$



(f) Asset level coefficient of variation

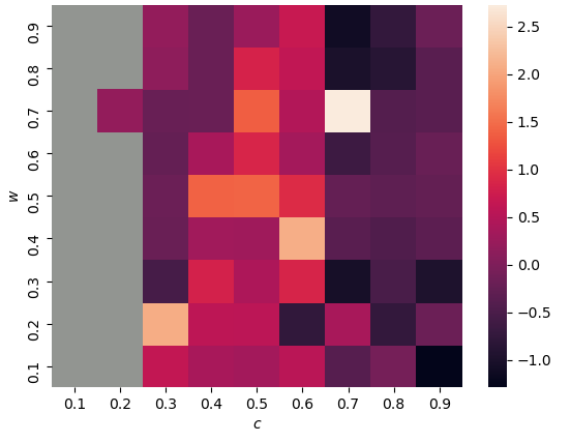
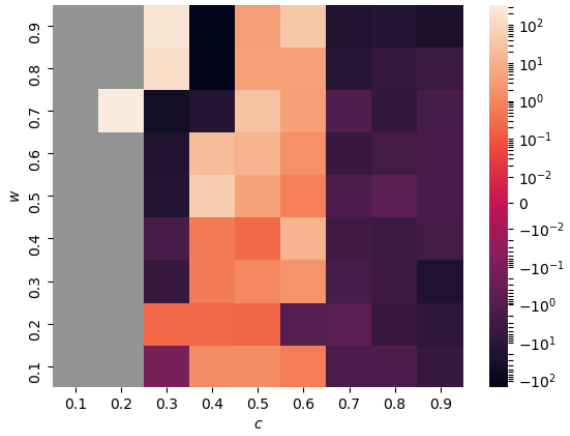
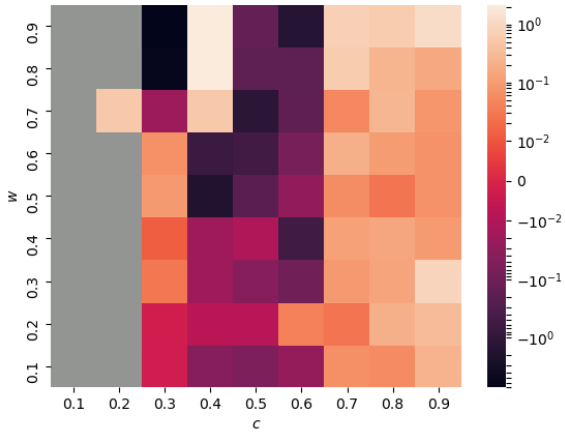


Figure 3.5: Heatmaps of mean results and coefficients variation over all combinations of the cost, c and weight of conformity, w. Mean values were taken after a transient of $t = 25000$ and grey areas denote systems that depleted in a short period of time, $t < 1000$. Note that some figures with extreme values have a log-scaled color bar for ease of viewing.



(a) The difference of the mean resource level and the optimal resource level $\bar{R} - R_{\text{opt}}$

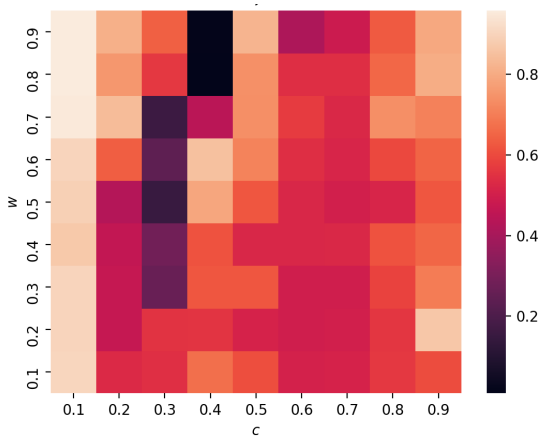


(b) The difference of the mean effort and the optimal effort $\bar{x} - x_{\text{opt}}$

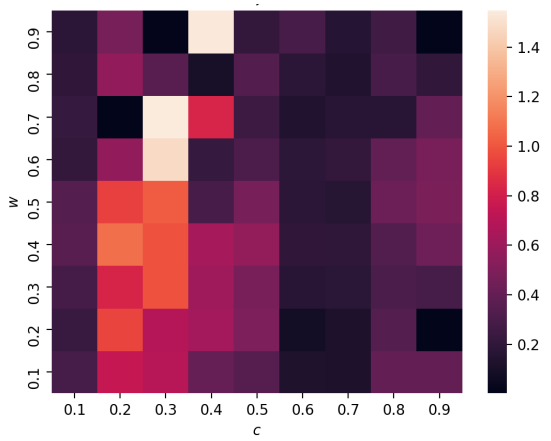
and depleted system. For $w \in [0.1, 0.7]$ however, the resource is at a much higher level than other w values, and this critical region appears to occur at $c = 0.4$ without leading to a depleted system for lower values of c. Looking at the time series for these parameters (Figure 3.8), we can see that for $c = 0.4$, as w is increased, the resource dynamics transition from oscillations about an equilibrium ($w = 0.7$), to a downwards trend in the resource ($w \in [0.8, 0.9]$). This stark difference in behaviour is due to the difficulty in learning a policy that maximizes the profit term in the reward such that when w is high enough, the agents essentially give up on the profit term and focus solely on conforming, a much easier goal for this system. When this occurs, the agents' actions cease to be informed by the state of the resource, allowing for these downward trends without consequence to the agents' rewards. For $w \in [0.8, 0.9]$, the profit reward remains close to its minimum value, however for $w = 0.7$, the profit reward oscillates between high and low values.

A notable exception to this is the single divergent trajectory at $w = 0.9$. Here, the agents receive a slightly higher profit reward near the beginning of the simulation and are incentivizing them to optimize the profit reward rather than discounting it. This leads to a positive feedback loop as once a sufficient number of agents learn that they can decrease their reward by decreasing their effort, others quickly follow suit as the weight of conformity is very high. On the contrary, for $c = 0.3$, the profit goal only begins to be taken into account by the agents for the later half of the time series, leading to an increase in the resource. As w is increased, contrary to intuition, high profit rewards become more frequent in the system, however this leads to depletions for $w = 0.8$ as these high profit rewards incentivize greedy agents to increase their harvesting effort, leading to depletion

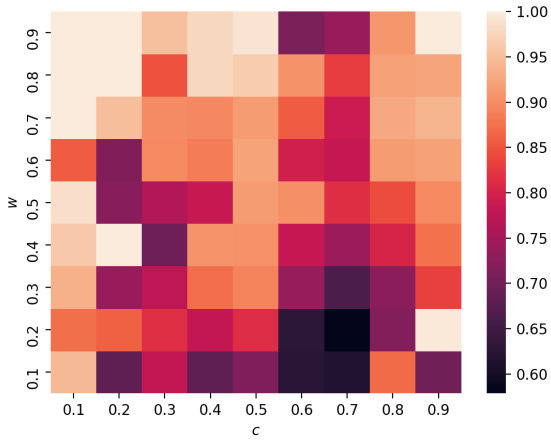
(a) Mean profit goal, $\bar{\xi}$



(b) Profit goal coefficient of variation



(c) Mean conform goal, $\bar{\lambda}$



(d) Conform goal coefficient of variation

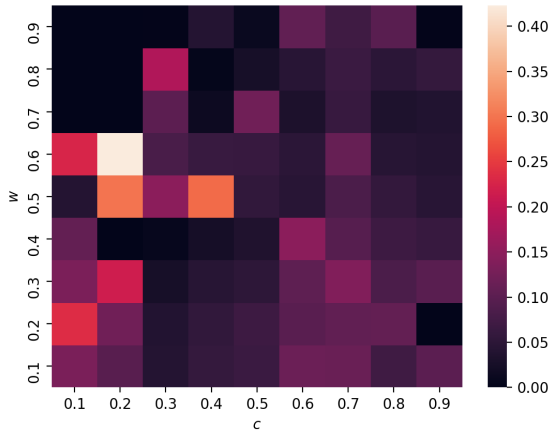


Figure 3.7: Heatmaps of mean values and coefficients variation for the two goals in the reward function. Mean values were taken after a transient of $T/2$, where T is the total length of the simulation.

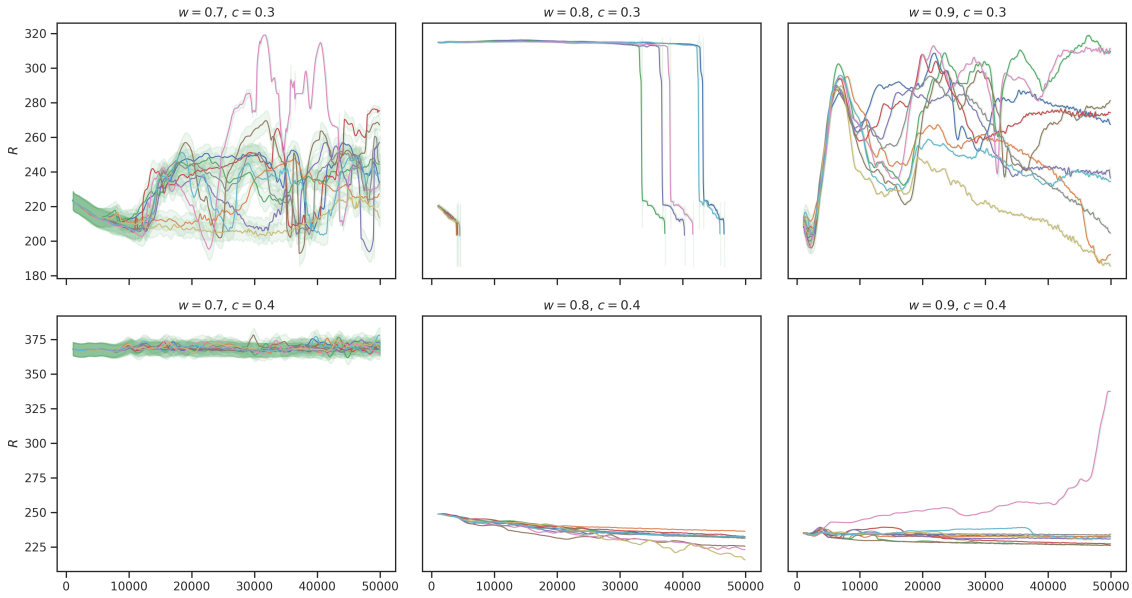


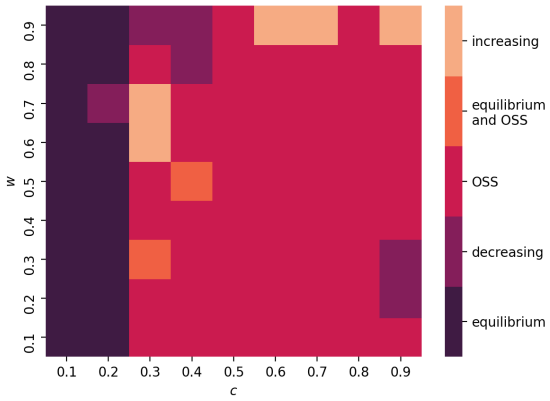
Figure 3.8: Resource trajectories at the transition between the sustainable and unsustainable regimes shown in Figure 3.5a. Each coloured trajectory represents an individual duplicate run with identical model parameters.

in all duplicate runs. For $w = 0.9$, the trajectories show both increasing and decreasing trends and this behaviour is likely due to a high conforming weight amplifying stochastic fluctuations in the system. This difference between dynamics for $c = 0.3$ and $c = 0.4$ warrants further investigation to tease out the effect on stochastic processes during both the training and testing phases as the initial actions that agents take leads to profoundly different outcomes.

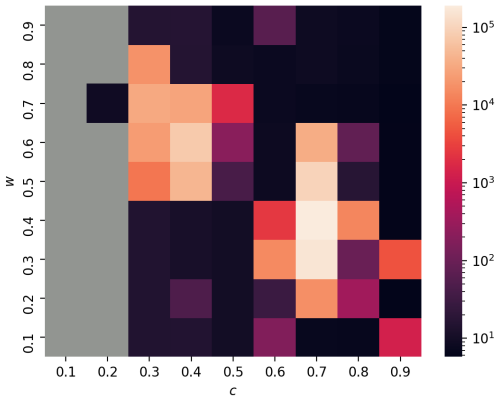
The majority of the resource and effort dynamics showed oscillations about some mean value, suggesting that agents learned to overcompensate their actions to approach x_{opt}, creating a boom-bust cycle with the resource. As these oscillatory dynamics were often centered about some mean, we will call these cases oscillatory steady states. Since the system showed various dynamical regimes, the testing simulation time series were classified by their qualitative behaviour, in this case, the resource dynamics. In Figure 3.9a, the system has been classified into five unique categories, representative of the richness of dynamics present in this system. Note that behaviour that differs from the dominant oscillatory steady state occurs for high cost per unit effort and weight of conformity, as well as at the transition between sustainable and unsustainable regimes. Additionally, many systems were dominated by regions of high frequency oscillations and these systems were identified by the magnitude of the fast Fourier transform for high frequencies in the time series data (Figure 3.9b). These regions of high frequency oscillations correspond to higher coefficients of variation in both the resource and efforts, which follows from the fact that with high frequency oscillations, there would be more extreme values, however, this may suggest that during these high frequency oscillations, the signal is amplified, as has been observed in many cases, but requires further analysis.

3.8 Discussion

This model demonstrated how self-interested agents were able to learn a sustainable harvesting policy in certain conditions without direct feedback for over-harvesting. This is significant as behaviours were learned without being prescribed by the model formulation and yet agents were still able to harvest sustainably over most parameters without having full knowledge of the system. The cost per unit effort of harvesting had a significant effect on the resource level and the efforts of the agents, increasing and decreasing them, respectively. Additionally, the system displayed a discrete transition from a profitable to unprofitable system as the cost per unit effort increased. This agrees with results from von der Osten et al, however their model showed a gradual decrease in assets, and did not have the discrete regime change observed in this system [13]. In the model presented in



(a) Classes of qualitative behaviour for the time series of the resource, R



(b) Presence of high frequency oscillations in the time series of the resource, R

this chapter, not only was the resource at a higher level in the profitable regime, it was also at a higher level than the optimal resource level expected with agents who are trying to maximize their profits without depleting the resource. Similar to von der Osten et al's model, the weight of conformity did not show as distinct of a trend with the mean values of resources and efforts, however when compared to the optimal resource and effort levels, a clear trend emerged where a higher weight of conformity amplified the differences from the optimal resource values. This amplification could be due to an increased discounting of the profit reward leading to a lack of feedback from the environment when choosing actions for high conforming systems. This suggests that strong social conformity can have significant effects in reinforcing and amplifying the status quo in common-pool resource systems. As the dynamics at this point ceased to be informed by the environment, this increased weight of conformity can lead to contradictory outcomes, dependent on the existing social norms of the system.

This model also displayed a diverse array of resource dynamics, characterized by oscillations on many different scales, usually about a mean value. This behaviour may be, in part due to overshoot dynamics and slow adaptation brought about by strong conformity within the system, decoupling the resource from the agents' actions to varying degrees. This phenomenon, however requires longer simulations and additional analysis and will be the focus of future work. These diverse dynamics demonstrate the value of modeling this system using an agent-based approach, as opposed to using top-down differential equations models. Trajectories that stray from analytical solutions show how agents with only partial information can bring the system to states that were not predicted from a top

down analysis and this can offer a more realistic understanding of real-world systems where decision-making over the management of common-pool resources can occur in a bottom-up manner. Additionally, the use of reinforcement learning for modelling common-pool resources shows great promise for future studies as many functions and parameters previously prescribed by the modeller for agent-based models can be learned from the given common-pool resource system. This allows for a higher diversity and complexity of actions and strategies that naturally arise from the socio-ecological dynamics.

Chapter 4

Conclusions and future work

In this thesis, common-pool resource models were compared using a variety of techniques in game theory and machine learning. In both models, the agents' actions are informed by maximizing an individual reward function, however under many conditions, this self-interest led to outcomes that benefited all agents collectively. An important concept in both models was the effect learning from peers had on the outcomes of the system. In Chapter 2, it was found that pure social learning was detrimental to the sustainability of the system compared with agents who only tried to maximize their payoff regardless of whether the strategy they were utilizing was successful among their peers. In Chapter 3, the dichotomy between these two learning strategies was replaced by a spectrum, where the weight of conformity, analogous to the extent of social learning was a global variable. In this system, since learning was performed through deep reinforcement learning algorithms, agents could learn much more complex strategies. In this model, the effect of social learning on the systems' behaviour was much more nuanced, amplifying existing trends in the profitable and unprofitable regimes. A stark difference between these two models was that in the game theoretic model, defectors with neighboring cooperators would receive an immediate penalty proportional to the degradation of the resource. For the reinforcement learning model, the penalty for depletion was shared equally and individual agents had much less immediate feedback on the effect their actions had on the state of the resource. Regardless of this difference, both models demonstrated an emergence of cooperation, even without direct or immediate incentives for sustainable actions. The mechanisms leading to these sustainable outcomes differed between models with some parameters, namely the extent of social learning, having contradictory effects. This demonstrates the extreme importance for modellers to choose learning dynamics informed by the system in study, and also emphasizes the dangers of prescribing broad top-down conservation policies on

diverse human-environment systems.

Relating these results to real-world common-pool resource systems, they show how changes to both the natural system, the human system and the way in which these two systems are coupled can have significant effects on how the social dynamics affect the sustainability of the resource. Although these models are formulated on assumptions that significantly decrease the degree of complexity observed in empirical systems, they can inform policy makers of the extreme importance to support governance initiatives that are specific to and informed by the existing social norms as well as the relationship between the community and the resource they are harvesting. Additionally, the manner in which agents learn proved to have a significant effect on the state of the resource, which can motivate policy makers and community leaders to devise systems of governance that are informed by the existing learning styles present in the community harvesting the common-pool resource. Creating policy through bottom-up initiatives informed by the local social norms and economy has led to sustainable outcomes in many real-world systems including Maine's lobster fisheries and the harvest of turtle eggs in the Turtle Islands of the Philippines [109, 110]. In the second example, a top-down policy banning the harvest of turtle eggs led to their subsequent depletion, further demonstrating how local systems of governance show much greater success when compared with sweeping rules that are not informed by the complex pre-existing human-environment dynamics.

As only preliminary results were presented for the reinforcement learning model, the most obvious next steps are continuing analysis on the data from the reinforcement learning model to gain a deeper understanding of when qualitative differences in the dynamics were caused by model parameters or randomness present in both the model initialization and the neural networks used in learning. Characteristics of the oscillatory dynamics, including the effect of high frequency oscillations in the system, the presence of chaotic dynamics, as well as the use of lag-1 autocorrelation as a tool for early warning signals of resource collapse could all offer critical insight to the dynamics and sustainability of the resource.

In terms of the model formulation, changing the weight of conformity from a global variable to a local variable would improve realism in the model, representing differing learning styles in human communities. Experimenting with the timescale of learning in both the training and testing phase can give insight into the manner in which these timescales affect the state of the resource. There are also plans to run simulations while periodically adding agents that harvest at a predefined level, reintroducing the binary concept of cooperators and defectors, to observe how the agents respond to this type of perturbation. Adding 'naive' agents, represented by untrained neural networks into a system of trained agents is also a possibility for future simulations and has already been implemented. Finally, repeating all simulations as well as the proposed extensions in a much larger system using

discrete actions informed by the action distributions of the DDPG model can offer valuable insight to the effect of community size and greater complexity in spatial structure on the persistence of a common-pool resource.

References

[1] Thomas Dietz, Elinor Ostrom, and Paul C Stern. The struggle to govern the commons. *Science*, 302(5652):1907–1912, 2003.

[2] Elinor Ostrom, Joanna Burger, Christopher B Field, Richard B Norgaard, and David Policansky. Revisiting the commons: local lessons, global challenges. *science*, 284(5412):278–282, 1999.

[3] Andrew A Rosenberg. Managing to the margins: the overexploitation of fisheries. *Frontiers in Ecology and the Environment*, 1(2):102–106, 2003.

[4] Mark Appiah, Dominic Blay, Lawrence Damnyag, Francis K Dwomoh, Ari Pappinen, and Olavi Luukkanen. Dependence on forest resources and tropical deforestation in ghana. *Environment, Development and Sustainability*, 11(3):471–487, 2009.

[5] Elinor Ostrom. The challenge of common-pool resources. *Environment: Science and Policy for Sustainable Development*, 50(4):8–21, 2008.

[6] Jouni Paavola. Climate change: the ultimate 'tragedy of the commons'? *Property in land and other resources*, pages 417–434, 2011.

[7] Raja Rajendra Timilsina, Koji Kotani, and Yoshio Kamijo. Sustainability of common pool resources. *PloS one*, 12(2):e0170981, 2017.

[8] Chris Brozyna, Todd Guilfoos, and Stephen Atlas. Slow and deliberate cooperation in the commons. *Nature Sustainability*, 1(4):184, 2018.

[9] Hannelore Brandt, Christoph Hauert, and Karl Sigmund. Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1519):1099–1104, 2003.

[10] Joshua S Weitz, Ceyhun Eksin, Keith Paarporn, Sam P Brown, and William C Ratcliff. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences*, 113(47):E7518–E7525, 2016.

[11] Hendrik Santoso Sugiarto, John Stephen Lansing, Ning Ning Chung, CH Lai, Siew Ann Cheong, and Lock Yue Chew. Social cooperation and disharmony in communities mediated through common pool resource exploitation. *Physical review letters*, 118(20):208301, 2017.

[12] György Szabó and Christoph Hauert. Phase transitions and volunteering in spatial public goods games. *Physical review letters*, 89(11):118101, 2002.

[13] Friedrich Burkhard von der Osten, Michael Kirley, and Tim Miller. Sustainability is possible despite greed-exploring the nexus between profitability and sustainability in common pool resource systems. *Scientific reports*, 7(1):2307, 2017.

[14] Edward W Tekwa, Andrew Gonzalez, and Michel Loreau. Spatial evolutionary dynamics produce a negative cooperation–population size relationship. *Theoretical population biology*, 125:94–101, 2019.

[15] Benjamin Kerr, Claudia Neuhauser, Brendan JM Bohannan, and Antony M Dean. Local migration promotes competitive restraint in a host–pathogen'tragedy of the commons'. *Nature*, 442(7098):75–78, 2006.

[16] Hardin Garrett. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[17] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[18] Matto Mildenberger. The tragedy of the tragedy of the commons. *Sci Am Blogs. Accessed*, 12, 2019.

[19] Southern Poverty Law Center. Garrett hardin. https://www.splcenter.org/fighting-hate/extremist-files/individual/garrett-hardin. Accessed: 2021-05-19.

[20] Henry F Lyle and Eric A Smith. The reputational and social network benefits of prosociality in an andean community. *Proceedings of the National Academy of Sciences*, 111(13):4820–4825, 2014.

[21] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3):137–158, 2000.

[22] J Terrence McCabe. Turkana pastoralism: A case against the tragedy of the commons. *Human ecology*, 18(1):81–103, 1990.

[23] Bryce Morsky and Erol Akcay. Evolution of social norms and correlated equilibria (vol 116, pg 8834, 2019). *Proceedings of the National Academy of Sciences*, 117(14):8213–8213, 2020.

[24] Rachata Muneepeerakul and John M Anderies. The emergence and resilience of self-organized governance in coupled infrastructure systems. *Proceedings of the National Academy of Sciences*, 117(9):4617–4622, 2020.

[25] Roger B Myerson. *Game theory*. Harvard university press, 2013.

[26] Zhigang Cao and Xiaoguang Yang. Symmetric games revisited. *Mathematical Social Sciences*, 95:9–18, 2018.

[27] Holger Ebel and Stefan Bornholdt. Coevolutionary games on networks. *Physical Review E*, 66(5):056118, 2002.

[28] Víctor M Eguíluz, Martín G Zimmermann, Camilo J Cela-Conde, and Maxi San Miguel. Cooperation and the emergence of role differentiation in the dynamics of social networks. *American journal of sociology*, 110(4):977–1008, 2005.

[29] Carlos Gracia-Lázaro, Alfredo Ferrer, Gonzalo Ruiz, Alfonso Tarancón, José A Cuesta, Angel Sánchez, and Yamir Moreno. Heterogeneous networks do not promote cooperation when humans play a prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 109(32):12922–12926, 2012.

[30] Martin A Nowak and Robert M May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826, 1992.

[31] Martin A Nowak, Sebastian Bonhoeffer, and Robert M May. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences*, 91(11):4877–4881, 1994.

[32] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652, 2017.

[33] Rajiv Sethi and Eswaran Somanathan. The evolution of social norms in common property resource use. *The American Economic Review*, pages 766–788, 1996.

[34] Naoki Masuda and Kazuyuki Aihara. Spatial prisoner's dilemma optimally played in small-world networks. *Physics Letters A*, 313(1-2):55–61, 2003.

[35] Francisco C Santos and Jorge M Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences*, 108(26):10421–10425, 2011.

[36] Jorge Marco and Renan Goetz. Tragedy of the Commons and Evolutionary Games in Social Networks: The Economics of Social Punishment. ETA: Economic Theory and Applications 259486, Fondazione Eni Enrico Mattei (FEEM), July 2017.

[37] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006.

[38] Raymond Chiong and Michael Kirley. Imitation vs evolution: Analysing the effects of strategy update mechanisms in n-player social dilemmas. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE, 2010.

[39] Daniele Vilone, José J Ramasco, Angel Sánchez, and Maxi San Miguel. Social imitation versus strategic choice, or consensus versus cooperation, in the networked prisoner's dilemma. *Physical Review E*, 90(2):022810, 2014.

[40] Joao A Moreira, Jorge M Pacheco, and Francisco C Santos. Evolution of collective action in adaptive social structures. *Scientific reports*, 3(1):1–6, 2013.

[41] Peter J Deadman. Modelling individual behaviour and group performance in an intelligent agent-based simulation of the tragedy of the commons. *Journal of Environmental Management*, 56(3):159–172, 1999.

[42] György Szabó and Gabor Fath. Evolutionary games on graphs. *Physics reports*, 446(4-6):97–216, 2007.

[43] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[44] Gerard Van der Laan and Xander Tieman. Evolutionary game theory and the modeling of economic behavior. *De Economist*, 146(1):59–89, 1998.

[45] Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10810–10817, 2014.

[46] Karl Sigmund. Introduction to evolutionary game theory. *Evolutionary game dynamics*, 69:1–26, 2011.

[47] György Szabó and Kristóf Hódsági. The role of mixed strategies in spatial evolutionary games. *Physica A: Statistical Mechanics and its Applications*, 462:198–206, 2016.

[48] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[49] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24(1):3–25, 1980.

[50] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.

[51] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[52] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440, 1998.

[53] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[54] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. *Vorabversion eines Lehrbuchs*, 5, 2016.

[55] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[56] Timothy Killingback and Michael Doebeli. Self-organized criticality in spatial evolutionary game theory. *Journal of theoretical biology*, 191(3):335–340, 1998.

[57] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality. *Physical review A*, 38(1):364, 1988.

[58] Dimitrije Marković and Claudius Gros. Power laws and self-organized criticality in theory and nature. *Physics Reports*, 536(2):41–74, 2014.

[59] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[60] Yuxi Li. Deep reinforcement learning. *arXiv preprint arXiv:1810.06339*, 2018.

[61] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[62] Simon Haykin. *Neural networks and learning machines, 3/E*. Pearson Education India, 2010.

[63] David Kriesel. A brief introduction on neural networks. 2007.

[64] Nils J Nilsson. Introduction to machine learning. an early draft of a proposed textbook. 1996.

[65] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[66] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[67] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

[68] Alain Degenne and Michel Forsé. *Introducing social networks*. Sage, 1999.

[69] Lars Carlsson and Annica Sandström. Network governance of the commons. *International Journal of the Commons*, 2(1):33–54, 2008.

[70] Arun Agrawal, Daniel G Brown, Gautam Rao, Rick Riolo, Derek T Robinson, and Michael Bommarito II. Interactions between organizations and networks in common-pool resource governance. *Environmental Science & Policy*, 25:138–146, 2013.

[71] Yong Min, Yuchen Du, and Cheng Jin. The effect of link rewiring on a coevolutionary common pool resource game. *Physica A: Statistical Mechanics and its Applications*, 512:935–944, 2018.

[72] Carlos Gracia-Lázaro, José A Cuesta, Angel Sánchez, and Yamir Moreno. Human behavior in prisoner's dilemma experiments suppresses network reciprocity. *Scientific reports*, 2:325, 2012.

[73] Tanya L Chartrand and Jessica L Lakin. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308, 2013.

[74] Hanwei Zhu and Michael Kirley. Deep multi-agent reinforcement learning in a common-pool resource system. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 142–149. IEEE, 2019.

[75] Attila Szolnoki and Matjaž Perc. Antisocial pool rewarding does not deter public cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 282(1816):20151975, 2015.

[76] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.

[77] Juan Camilo Cárdenas. Human behavior and the use of experiments to understand the agricultural, resource, and environmental challenges of the xxi century. *Agricultural Economics*, 47(S1):61–71, 2016.

[78] Munyaradzi J Mutenje, Gerald F Ortmann, and Stuart RD Ferrer. Management of non-timber forestry products extraction: Local institutions, ecological knowledge and market structure in south-eastern zimbabwe. *Ecological Economics*, 70(3):454–461, 2011.

[79] Maja Schlüter, Alessandro Tavoni, and Simon Levin. Robustness of norm-driven cooperation in the commons. *Proceedings of the Royal Society B: Biological Sciences*, 283(1822):20152431, 2016.

[80] Marten Scheffer, Frances Westley, and William Brock. Slow response of societies to new problems: causes and costs. *Ecosystems*, 6(5):493–502, 2003.

[81] Paul R Ehrlich and Simon A Levin. The evolution of norms. *PLoS Biol*, 3(6):e194, 2005.

[82] Elizabeth Dougall. Revelations of an ecological perspective: Issues, inertia, and the public opinion environment of organizational populations. *Public Relations Review*, 31(4):534–543, 2005.

[83] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Physical review E*, 66(3):035103, 2002.

[84] Amartya Sen, Master Amartya Sen, Sen Amartya, James E Foster, James E Foster, et al. *On economic inequality*. Oxford University Press, 1997.

[85] Emilio F Moran and Elinor Ostrom. *Seeing the forest and the trees: human-environment interactions in forest ecosystems*. Mit Press, 2005.

[86] Per Olsson, Lance Gunderson, Steve Carpenter, Paul Ryan, Louis Lebel, Carl Folke, and Crawford S Holling. Shooting the rapids: navigating transitions to adaptive governance of social-ecological systems. *Ecology and society*, 11(1), 2006.

[87] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

[88] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[89] Thomas M Bury, Chris T Bauch, and Madhur Anand. Charting pathways to climate change mitigation in a coupled socio-climate model. *PLoS computational biology*, 15(6), 2019.

[90] Robert Boyd, Peter J Richerson, and Joseph Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10918–10925, 2011.

[91] Andrew R Tilman, Joshua B Plotkin, and Erol Akçay. Evolutionary games with environmental feedbacks. *Nature communications*, 11(1):1–11, 2020.

[92] Christoph Hauert, Camille Saade, and Alex McAvoy. Asymmetric evolutionary games with environmental feedback. *Journal of theoretical biology*, 462:347–360, 2019.

[93] Xin Wang, Zhiming Zheng, and Feng Fu. Steering eco-evolutionary game dynamics with manifold control. *Proceedings of the Royal Society A*, 476(2233):20190643, 2020.

[94] David A Kim, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H Fowler, and Nicholas A Christakis. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153, 2015.

[95] Feng Fu, Nicholas A Christakis, and James H Fowler. Dueling biological and social contagions. *Scientific reports*, 7:43634, 2017.

[96] Alessio Cardillo and Naoki Masuda. Critical mass effect in evolutionary games triggered by zealots. *Physical Review Research*, 2(2):023305, 2020.

[97] Charles W Cobb and Paul H Douglas. A theory of production. *The American Economic Review*, 18(1):139–165, 1928.

[98] Paul H Douglas. The cobb-douglas production function once again: its history, its testing, and some new empirical values. *Journal of political economy*, 84(5):903–915, 1976.

[99] Abraham Charnes, WW Cooper, and AP Schinnar. A theorem on homogeneous functions and extended cobb-douglas forms. *Proceedings of the National Academy of Sciences*, 73(10):3747–3748, 1976.

[100] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social networks*, pages 179–197. Elsevier, 1977.

[101] Bruce Kogut and Gordon Walker. The small world of germany and the durability of national networks. *American sociological review*, pages 317–335, 2001.

[102] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[103] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the national academy of sciences*, 99(suppl 1):2566–2572, 2002.

[104] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[105] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster,

Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[106] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[107] Steven J Lade, Susa Niiranen, Jonas Hentati-Sundberg, Thorsten Blenckner, Wiebren J Boonstra, Kirill Orach, Martin F Quaas, Henrik Österblom, and Maja Schlüter. An empirical model of the baltic sea reveals the importance of social dynamics for ecological regime shifts. *Proceedings of the National Academy of Sciences*, 112(35):11120–11125, 2015.

[108] Graeme S Cumming, David HM Cumming, and Charles L Redman. Scale mismatches in social-ecological systems: causes, consequences, and solutions. *Ecology and society*, 11(1), 2006.

[109] James M Acheson. *Capturing the commons: devising institutions to manage the Maine lobster industry*. Upne, 2003.

[110] Raul P Lejano and Helen Ingram. Place-based conservation: lessons from the turtle islands. *Environment: Science and Policy for Sustainable Development*, 49(9):18–27, 2007.

[111] Tim H Clutton-Brock and Geoffrey A Parker. Punishment in animal societies. *Nature*, 373(6511):209, 1995.

[112] Reuven Cohen and Shlomo Havlin. Scale-free networks are ultrasmall. *Physical review letters*, 90(5):058701, 2003.

[113] S Eubank. Synthetic data products for societal infrastructures and protopopulations: Data set 3.0. Technical report, Technical Report NDSSL-TR-07-003, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, 2008.

[114] Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415(6868):137, 2002.

[115] Ariel Fernández and Hugo Fort. Catastrophic phase transitions and early warnings in a spatial ecological model. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):P09014, 2009.

[116] C Pagnutti, M Anand, and M Azzouz. Lattice geometry, gap formation and scale invariance in forests. *Journal of theoretical biology*, 236(1):79–87, 2005.

[117] Mendeli H Vainstein and Jeferson J Arenzon. Disordered environments in spatial games. *Physical Review E*, 64(5):051905, 2001.

[118] Martin J Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004.

APPENDICES

Appendix A

A.1 Additional figures

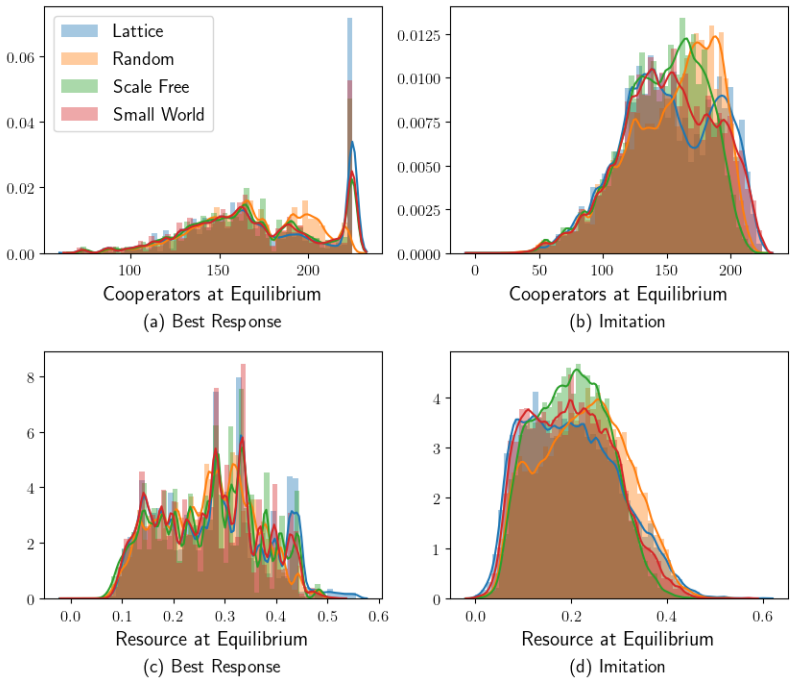


Figure A.1: The frequency distribution of cooperators (a, b) and resource (c, d) at equilibrium, normalized with a Gaussian kernel density estimate

Appendix B

B.1 Extrema of the payoff function

We can find global extrema for the collective payoff function below:

$$\frac{\partial \Pi}{\partial X} = \alpha\beta X^{\alpha-1} R^{1-\alpha} - c \tag{B.1}$$

$$\frac{\partial^2 \Pi}{\partial X^2} = \alpha(\alpha-1)\beta X^{\alpha-2} R^{1-\alpha} < 0 \tag{B.2}$$

$$X^* = \left(\frac{c}{\alpha\beta}\right)^{\frac{1}{\alpha-1}} R \tag{B.3}$$

$$\Pi^* = \left(\frac{c^\alpha}{\beta}\right)^{\frac{1}{\alpha-1}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right) R \tag{B.4}$$

From this collective payoff maximum, we can solve for an upper bound on an agents individual payoff:

$$\pi_i^* = \frac{x_i}{X^*} \Pi^* \tag{B.5}$$

$$\pi_i^* = x_i c(\alpha^{-1} - 1) \tag{B.6}$$

Since π_i^* is a linear function of x_i, the highest possible value of $x_i = x_{\max}$ will provide the maximum value for an agent's payoff.

$$\pi_{\max} = x_{\max} c(\alpha^{-1} - 1) \tag{B.7}$$

Additionally, as H is strictly increasing in R, the minimum collective payoff must occur when $R = R_{\min} = 0.01K$. Since there is only one local optima of Π in X, the minimum collective payoff must also either occur at $X_{\max}$ or $X_{\min} = 0$. As $\Pi(X, R_{\min}) < 0$ for $X > R_{\min}\left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}}$, which is always true for the parameters in this system, then $\Pi_{\min}$ must occur at $(X_{\max}, R_{\min})$. Therefore, the minimum individual payoff occurs at:

$$\pi_{\min} = \min_{x_i} \frac{x_i}{X_{\max}} \Pi_{\min} \tag{B.8}$$

$$\pi_{\min} = \frac{x_{\max}}{X_{\max}} \Pi_{\min}, \qquad \text{since } \Pi_{\min}/X_{\max} < 0 \tag{B.9}$$

$$\pi_{\min} = x_{\max}\left(\beta X_{\max}^{\alpha-1} R_{\min}^{1-\alpha} - c\right) \tag{B.10}$$

B.2 Parameters in the reward function

To make sure that the wealth goal saturates at the maximum attainable payoff for a given round, we can choose values for k and s such that $\xi_i(\pi_{\max}) = 0.99$ and $\xi_i(\pi_{\min}) = 0.01$:

$$k = \frac{\ln(\gamma)}{\pi_{\max} - \pi_{\min}} \tag{B.11}$$

$$\gamma = \frac{\frac{1}{0.01} - 1}{\frac{1}{0.99} - 1} \tag{B.12}$$

$$s = \frac{\ln\left(\frac{1}{0.99} - 1\right)}{k} \tag{B.13}$$

The group conforming goal, λ_i is a scaled normal distribution with $\mu = \hat{x_i}$ and

$$\sigma = \begin{cases} \frac{x_{\max} - \hat{x_i}}{3} & \hat{x_i} < \frac{x_{\max}}{2} \\ \frac{\hat{x_i}}{3} & \hat{x_i} \geq \frac{x_{\max}}{2} \end{cases} \tag{B.14}$$

This ensures that λ_i is maximized at $\hat{x_i}$ and reaches a value close to 0 at the boundary furthest from $\hat{x_i}$.

B.3 Change of variables for the profit reward

As payoff π_i is a function of effort x and the resource R, we can simplify the reward function, r_i by reducing the input variables to be only in terms of effort. Since $\pi_i = x_i\left(\frac{H}{X} - c\right)$ and $0 \leq x_i \leq x_{\max}$, we have three cases for $\pi_{\min}$ and $\pi_{\max}$.
If $\frac{H}{X} - c > 0$, then $\pi_{\max} = x_{\max}\left(\frac{H}{X} - c\right)$ and $\pi_{\min} = 0$.
If $\frac{H}{X} - c < 0$, then $\pi_{\max} = 0$ and $\pi_{\min} = x_{\max}\left(\frac{H}{X} - c\right)$.
If $\frac{H}{X} - c = 0$, then $\pi_{\max} = \pi_{\min} = 0$:

$$\xi_i(x_i) = \begin{cases} \frac{1}{\exp\left(-\frac{\ln(\gamma)}{x_{\max}\left(\frac{H}{X}-c\right)}\left(x_i\left(\frac{H}{X}-c\right)-\left(x_{\max}\left(\frac{H}{X}-c\right)-\frac{\ln\left(\frac{1}{0.99}-1\right)x_{\max}\left(\frac{H}{X}-c\right)}{\ln(\gamma)}\right)\right)\right)+1} & \frac{H}{X} - c > 0 \\ \frac{1}{\exp\left(\frac{\ln(\gamma)}{x_{\max}\left(\frac{H}{X}-c\right)}\left(x_i\left(\frac{H}{X}-c\right)+\frac{\ln\left(\frac{1}{0.99}-1\right)x_{\max}\left(\frac{H}{X}-c\right)}{\ln(\gamma)}\right)\right)+1} & \frac{H}{X} - c < 0 \\ 0 & \frac{H}{X} - c = 0 \end{cases} \tag{B.15}$$

$$\xi_i(x_i) = \begin{cases} \frac{1}{\exp\left(-\ln(\gamma)\left(\frac{x_i}{x_{\max}}-\left(1+\frac{\ln\left(\frac{1}{0.99}-1\right)}{\ln(\gamma)}\right)\right)\right)+1} & \frac{H}{X} - c > 0 \\ \frac{1}{\exp\left(\ln(\gamma)\left(\frac{x_i}{x_{\max}}+\frac{\ln\left(\frac{1}{0.99}-1\right)}{\ln(\gamma)}\right)\right)+1} & \frac{H}{X} - c < 0 \\ 0 & \frac{H}{X} - c = 0 \end{cases} \tag{B.16}$$

An equivalent condition for $\frac{H}{X} - c > 0$ is $X < \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R$. Similarly for $\frac{H}{X} - c < 0$, $X > \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R$ and $\frac{H}{X} - c = 0$, $X = \left(\frac{c}{\beta}\right)^{\frac{1}{\alpha-1}} R$.

Since this is in terms of the resource rather than cumulative harvest, we will use these conditions for the reward function instead.

B.4 Additional figures

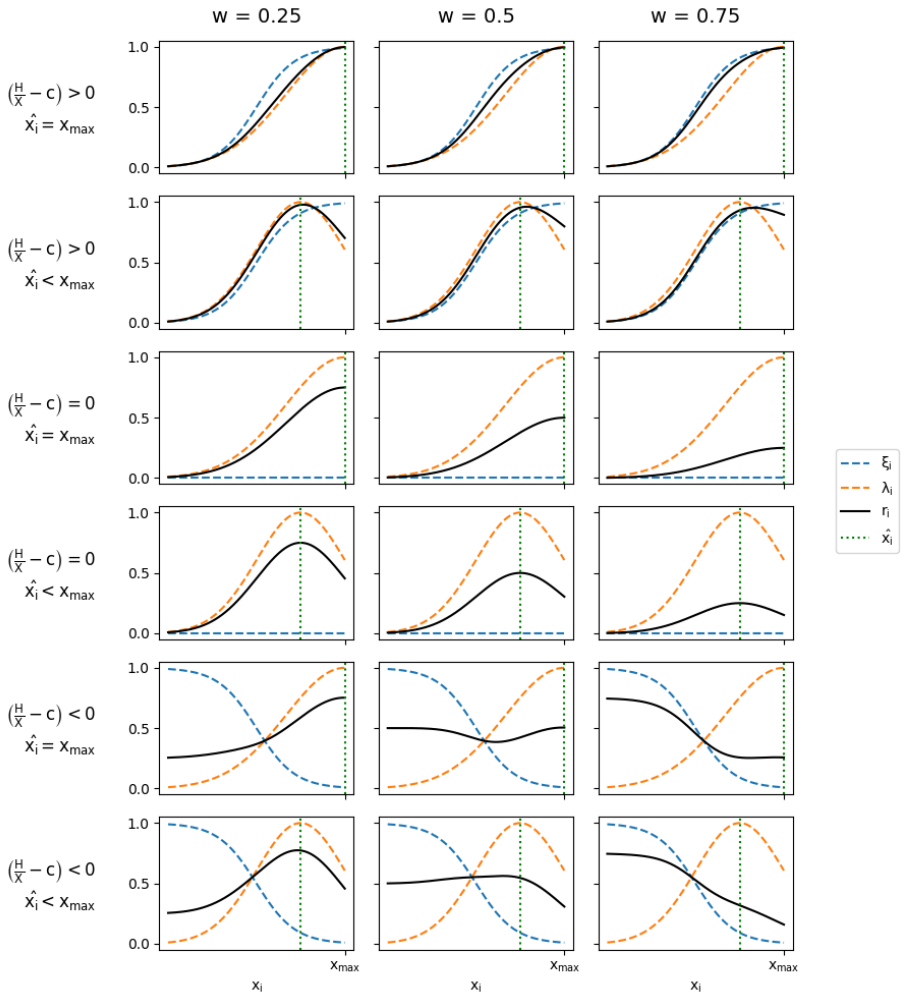


Figure B.1: Reward function (r_i), profit goal (ξ_i) and group conforming goal (λ_i) over all cases of X, R, and $\hat{x_i}$ and three different weights (w).

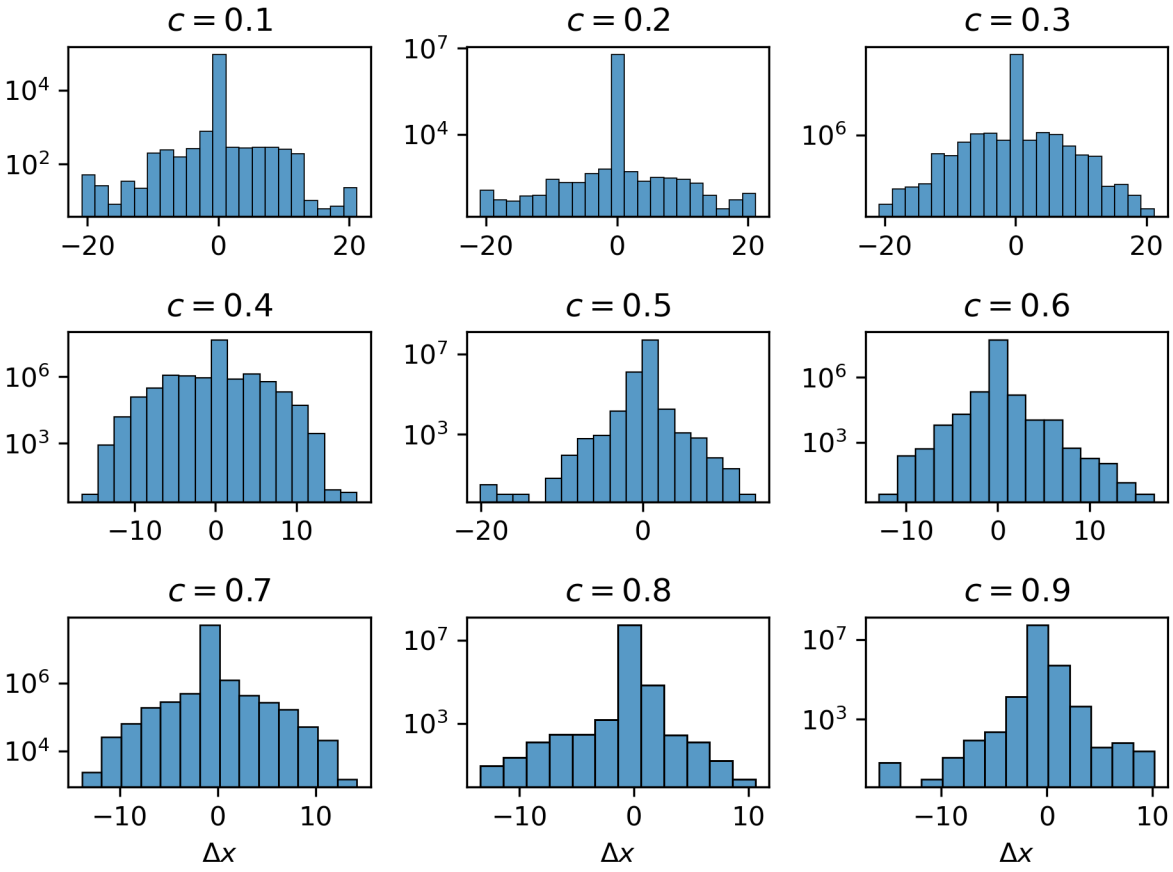


Figure B.2: The distribution of actions during testing for each cost value, c. As c increases, the range of actions chosen by agents decreases since in these systems, the conform goal has a higher influence on the social dynamics, incentivizing agents to stick with the status quo.

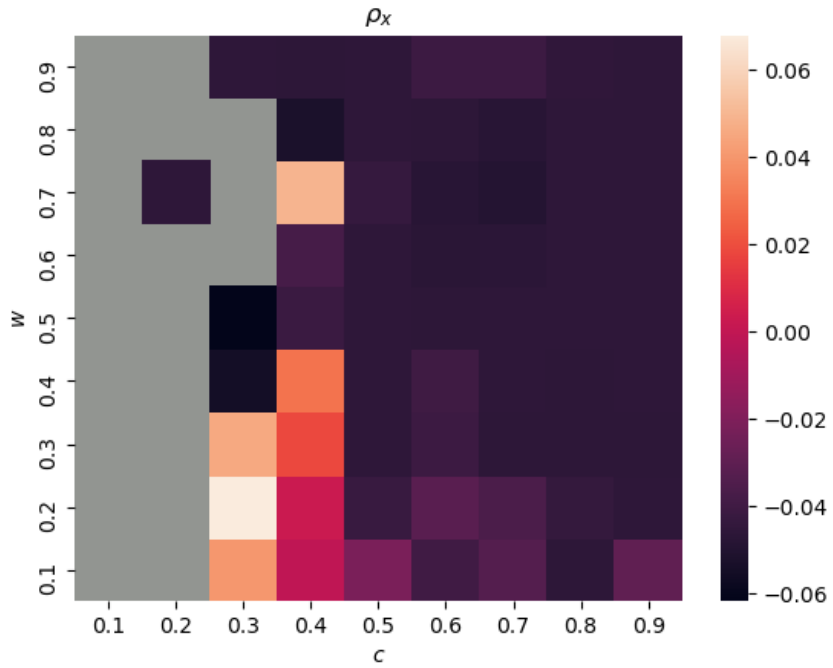
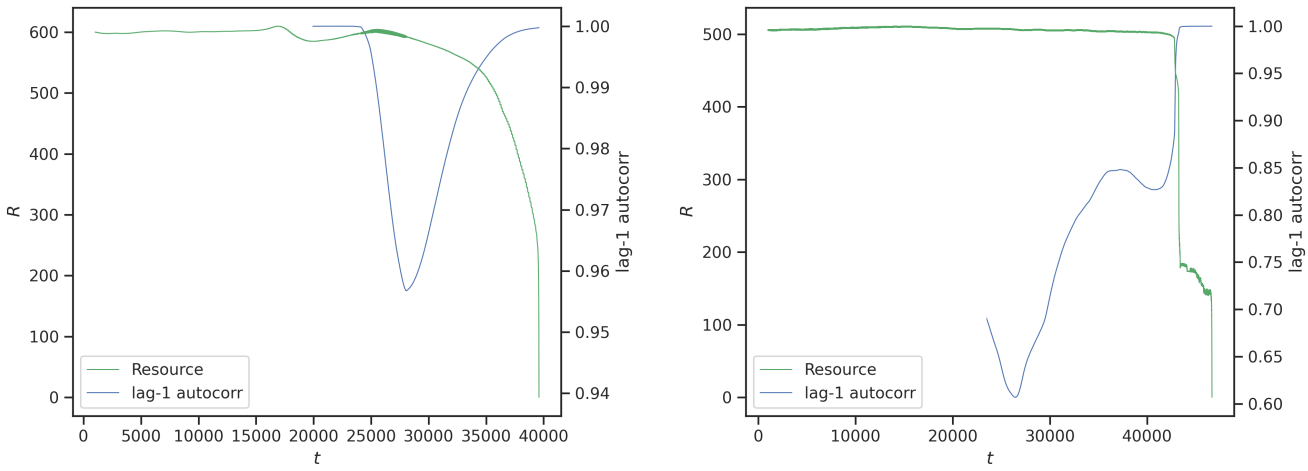


Figure B.3: The mean network assortativity of the agents' effort, $\bar{\rho_x}$. Note that the trends in assortativity were not significant as $\rho \in [-1, 1]$, however $\rho_x \in [-0.06, 0.06]$



(a) $w = 0.3$, $c = 0.9$

(b) $w = 0.8$, $c = 0.3$

Figure B.4: Trends in autocorrelation were observed and may prove useful as an early warning signal for resource depletion, however further analysis is required before conclusions are made.

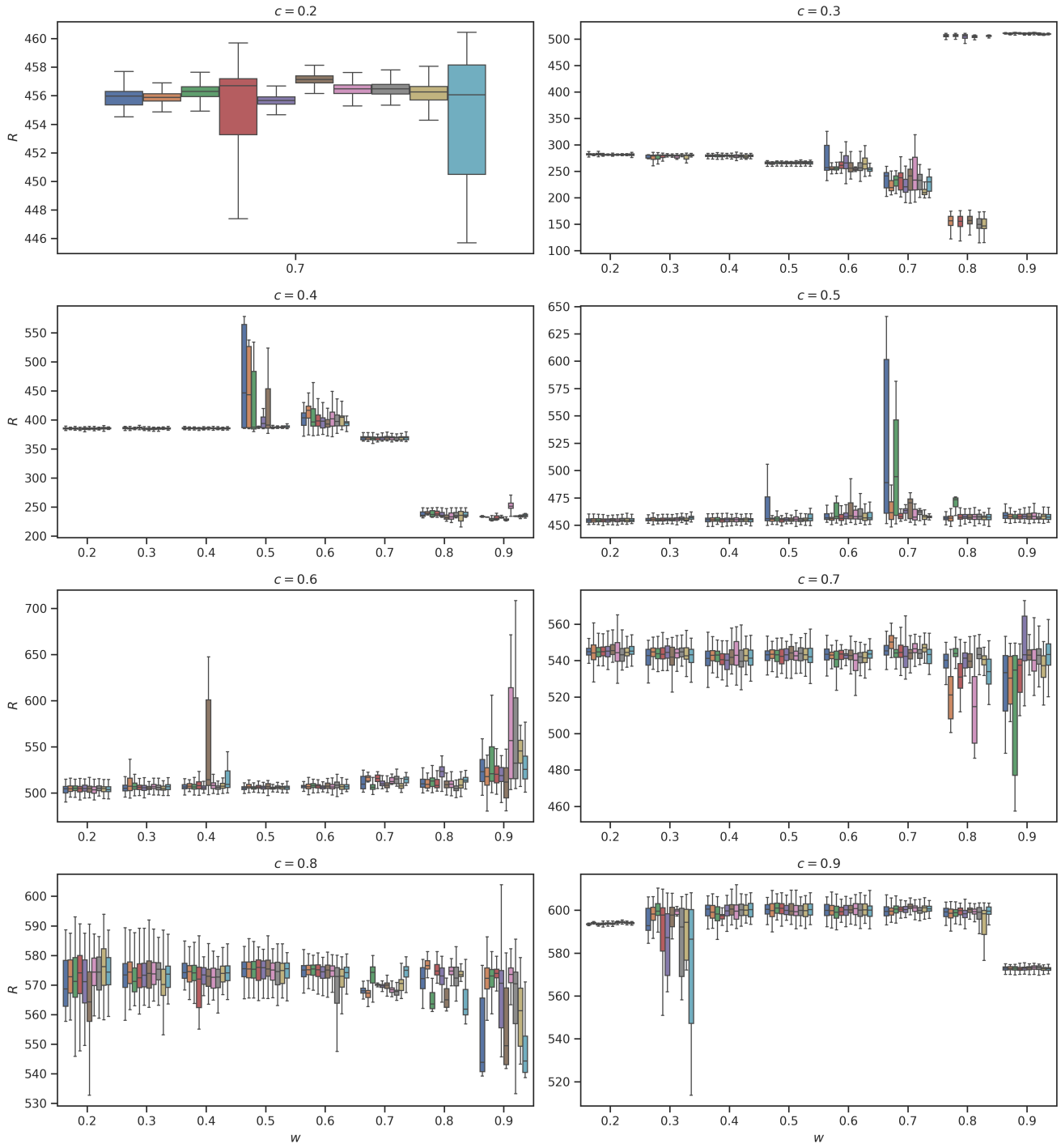


Figure B.5: This boxplot for the resource level, R, shows how variance differs with different weights of conforming, w. The significant difference in mean values also suggests the presence of alternative stable states at $c = 0.5$, $w = 0.8$ and $c = 0.9$, $w = 0.1$