Impacts of Self-Interested Mediators on Cooperation in Networks

Modelling recommender systems with evolutionary game theory and comparing reward functions w.r.t. their effects on societies of agents

Francisco Lopes Pereira de Carvalho

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Manuel Lopes
Prof. Francisco Santos

Examination Committee

Chairperson: Prof. Name of the Chairperson
Supervisor: Prof. Manuel Lopes
Members of the Committee: Prof. Name of First Committee Member
Dr. Name of Second Committee Member
Eng. Name of Third Committee Member

Month 20XX

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Acknowledgments

I would like to thank my parents for their friendship, encouragement and caring over all these years, for always being there for me through thick and thin and without whom this project would not be possible. I would also like to thank my grandparents, aunts, uncles and cousins for their understanding and support throughout all these years.

I would also like to acknowledge my dissertation supervisors Prof. Some Name and Prof. Some Other Name for their insight, support and sharing of knowledge that has made this Thesis possible.

Last but not least, to all my friends and colleagues that helped me grow as a person and were always there for me during the good and bad times in my life. Thank you.

To each and every one of you – Thank you.

write for real

Abstract

Where we model recommendation systems as mediators in a setting of evolutionary game theory in networks. We then attempt to use reinforcement learning with graph neural networks to train mediators that make recommendations in order to optimize pro-social and self-interested metrics. Then we make different mediators compete and analyze their impacts on cooperation and network topology.

not done yet

Keywords

recommender systems; network science; evolutionary game theory; reinforcement learning

Resumo

(fazer depois)

Palavras Chave

sistemas de recomendação; redes complexas; teoria dos jogos evolucionária; aprendizagem por reforço

Contents

1 Introduction 1

1.1 Context . . . 3

1.2 Related Work . . . 5

1.3 Contributions . . . 7

2 Modeling 9

2.1 Model requirements . . . 11

Agents with partial information and mediators with complete information . . . 11

Agent-mediator interaction - recommendations . . . 12

Agent-agent interaction . . . 12

Utility functions for mediators, agents, and society . . . 12

2.2 Model definition . . . 12

2.2.1 Base model . . . 12

Games . . . 13

Evolution . . . 13

Rewiring . . . 13

Timescale . . . 13

Network . . . 14

2.2.2 How it fits model requirements . . . 14

2.2.3 Simulation algorithm . . . 14

2.2.4 Metrics . . . 14

Cooperation and total payoff . . . 15

Heterogeneity and CDD . . . 15

Rewire rate . . . 15

2.2.5 Fixed Rewiring Strategies . . . 16

RANDOM . . . 16

NO MED . . . 16

GOOD . . . 16
BAD . . . 16
FAIR . . . 16
2.3 Baseline results . . . 16
2.3.1 No rewiring . . . 17
2.3.2 Rewiring with no mediator . . . 17
2.3.3 Heuristic mediators . . . 17
3 Competition 23
3.1 Model of Competition . . . 25
3.2 Baseline Results . . . 27
4 Training 31
4.1 Environment: Mediating EGT . . . 33
4.2 Training Architecture . . . 33
4.3 Utility Functions . . . 34
4.4 Training . . . 34
4.5 Results . . . 34
5 Discussion 35
5.1 Single mediators . . . 37
5.2 Competition . . . 37
5.3 Training . . . 37
6 Conclusion 39
6.1 Limitations and Future Work . . . 41
Bibliography 43

List of Figures

1.1 Recommender systems as multi-stakeholder environments. 4

1.2 Taxonomy of ISA-human interactions. [Burr et al., 2018a] 6

2.1 UML of the relevant entities in our model. 11

2.2 Evolving the neighborhood. [Santos et al., 2006b] 13

2.3 Evolution of cooperation in a uniform random network. 17

2.4 Co-Evolution for Different Dilemmas and Time Scales. Each row: W=0.5,1,2,3,4, each column: final fraction of cooperators, final heterogeneity, final k$_{max}, rewire_n$. 18

2.5 Co-Evolution of cooperation for Different Dilemmas (t,s axes) and Time Scales (rows), for Different Mediators (columns). 19

2.6 Co-evolution of cooperation and structure in the prisoner's dilemma (t=2,s=-1) as a function of W for different mediators. 20

2.7 Heterogeneity for Different Dilemmas (t,s axes) and Time Scales (rows), for Different Mediators (columns). 21

3.1 Average final fraction of mediator populations after competition between NO MED and exclusive fixed mediators. For $W1 \in \{0.5, 1, 2, 3, \inf\}$ and $W2 \in \{0.01, 0.03, 0.1, 0.5, \inf\}$. For higher values of W2, we see a relatively low adoption rate and uniform distribution of mediators besides NO MED. Initialized at 90% NO MED and 10% split between the rest. 27

3.2 Final frequencies of mediators as functions of W2. 28

3.3 Metrics as a function of W2 for different values of W1. 29

4.1 UML of the relevant entities in our model, with added information from the current implementation.. 33

NO MED has a different value range. Fix? Use other pallete?

List of Tables

List of Algorithms

2.1 Simulation algorithm 15

3.1 Simulation algorithm for mediator competition 26

Listings

Acronyms

1

Introduction

Contents

1.1 Context 3

1.2 Related Work 5

1.3 Contributions 7

Recommender systems (RS) are software systems that assist users in interacting with large spaces of items, usually by presenting them with smaller personalized sets based on information such as past user behavior, user attributes, and features of the underlying items. User experience on social media, content platforms, and online stores is largely determined by RS.

Most recommender systems optimize metrics that are easy to measure and improve, like number of clicks, time spent, or number of daily active users. They are selected to do this by powerful optimization processes involving thousands of engineers and a significant fraction of global computing power. As hinted at by Goodhart's law [Goodhart, 1981], using simple metrics that are imperfectly aligned with the best-interest of users has resulted in a host of hard-to-measure side-effects like addiction, reduced cognitive capacity, and political radicalisation.

citations needed

RS serve a purpose for each individual user, but most importantly they act as mediators of human interaction and consumption online. Because of this, we want RS to act in the best interest of humanity at a system level and expand our ability to coordinate at a every scale in a way that causes as many positive externalities as possible. To approach this problem, we compare the differently aligned recommender systems with respect to the effects they have on the evolution of cooperation and social graph topology in society.

To that end, we build a toy model and use reinforcement learning with different reward functions to train agents for the task of recommendation. Afterwards we examine competition between RS: observing which mediators would dominate or be extinguished in a competition setting, and whether the presence of competition would be a net gain to society when compared with monopolistic mediators.

1.1 Context

Classically, the task of recommendation has been framed as providing ”relevant” items to users. In practice, RS are multi-stakeholder environments as illustrated in Figure 1.1, where multiple parties derive different utilities from recommendations - users, content providers, and system operators, for example. [Milano et al., 2020]

From these, RS implicitly favor their operators, [Burr et al., 2018b] who naturally have the ability to tune and replace such recommenders. Explicitly, RS in social media platforms often make their content recommendations in ways that maximize metrics like ad-clicks or user engagement, used as proxies for relevance and user satisfaction. “Conflating retention and satisfaction has allowed developers to mediate the tension between users (whom they wanted to help) and business people who wanted to capture them.” [Seaver, 2019]

The information people are exposed to radically biases their beliefs, preferences and habits with both short-term and long-term effects. [Vendrov and Nixon, 2019] For example, proxy metrics used

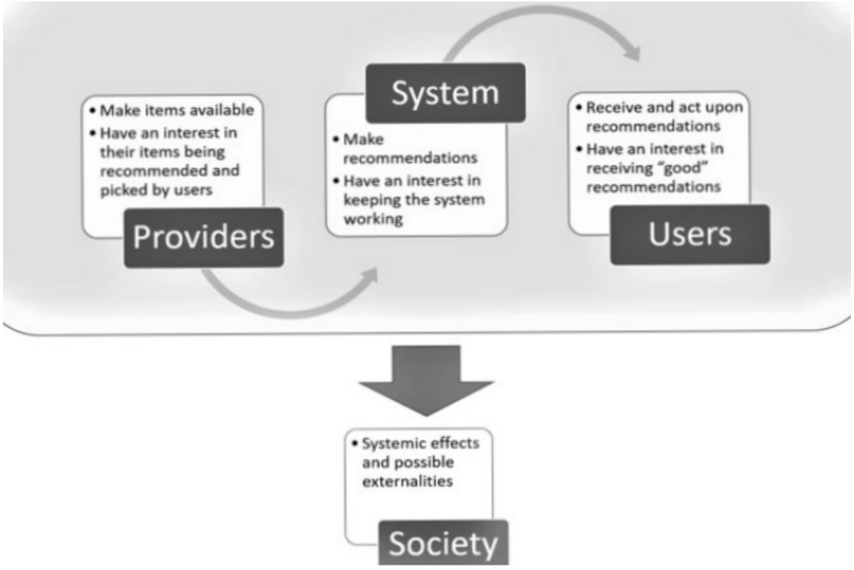

System
• Make recommendations
• Have an interest in keeping the system working
• Make items available
• Have an interest in their items being recommended and picked by users
Providers
• Receive and act upon recommendations
• Have an interest in receiving "good" recommendations
Users
• Systemic effects and possible externalities
Society

Figure 1.1: Recommender systems as multi-stakeholder environments. [Milano et al., 2020].

in social media, are only weakly correlated with what users care about, and users often regret time spent in social media. [Andreassen, 2015] Beyond taking a pre-existing preference profile and tailoring recommendations to it, RS contribute to the construction of user identity dynamically. [Floridi, 2011] Preference drift from social interactions has been validated empirically. [Zafari et al., 2018] This begs the question: Will RS be able to pursue their goals by strategically shaping our experiences to manage the evolution of our preferences?

Goodhart's Law states that "as soon as a measure becomes a target, it ceases to be a good measure" [Goodhart, 1981]. The misalignment between proxy metrics cause problems as side-effects that happen to increase the metric imposed by RS operators, often to the detriment of users or negative externalities in society.

Attempts to align RS with users have historically addressed problems reactively and one at a time (fake news, fairness, diversity, addiction, polarisation) and only after they were already widespread, often causing new harder problems. Even loftier high-level goals have been pursued, like user well-being, [Khwaja et al., 2019] or self actualization. [**?**] Higher-level approaches to RS alignment have been proposed, namely, the incorporation of well-being metrics, participatory objective design, interactive value learning, and optimizing for informed and deliberative preferences. [Stray et al., 2021].

citations needed

We propose competition between RS as a further higher-level approach. Even after techniques for building more aligned recommenders are developed, there's no guarantee they would be implemented by big platforms. Although it can be argued that it would be in their interest to work towards the ultimate benefit of users and society, [Hohnhold et al., 2015] that might not be the case in the presence of strategic dynamics between stakeholders in these systems [Kurland, 2019]. Especially in our current setting of effective monopolies - protected by large-scale network effects - on what should effectively be public goods. (e.g. search, social networking, instant messaging) We expect that decoupling recommenders from applications and allowing users to choose from a marketplace of algorithms will lead to the widespread adoption of ever more aligned RS as the field advances.

1.2 Related Work

To better understand the effects that differently aligned mediators have on their mediated populations, we must produce a toy model of society to be mediated. Users are modelled as greedy agents; while RS can be modelled as the very types of algorithms that see deployment in the real world. [Calero Valdez and Ziefle, 2018]

Strategic dynamics in content production and consumption may lead to the failure of classical principles of RS in maximizing social welfare. The need to avoid such a failure by revisiting those principles with game theory and multiple stakeholders in mind has inspired a whole research agenda. [Kurland, 2019] In the same game theoretic paradigm, competing recommenders have been studied [Izsak et al., 2014], and even the cost that strategic mediators in competition would impose on the population of agents they mediate [Babaioff et al., 2015], although not in the context of recommendation systems.

The dynamics of our toy model are informed by a taxonomy of interactions between humans and Intelligent Software Agents (ISA) [Burr et al., 2018a]. Both ISA and user may be assumed to pursue approximate maximisation of expected utility. (See Figure 1.2) ISA can impose control over (or persuade) a system of users through trading (pursuing its goals while increasing utility for the user) and nudging (exploiting user heuristics and biases to steer their behavior), with second-order effects over the user's beliefs and value function. (e.g. behavioral addiction)

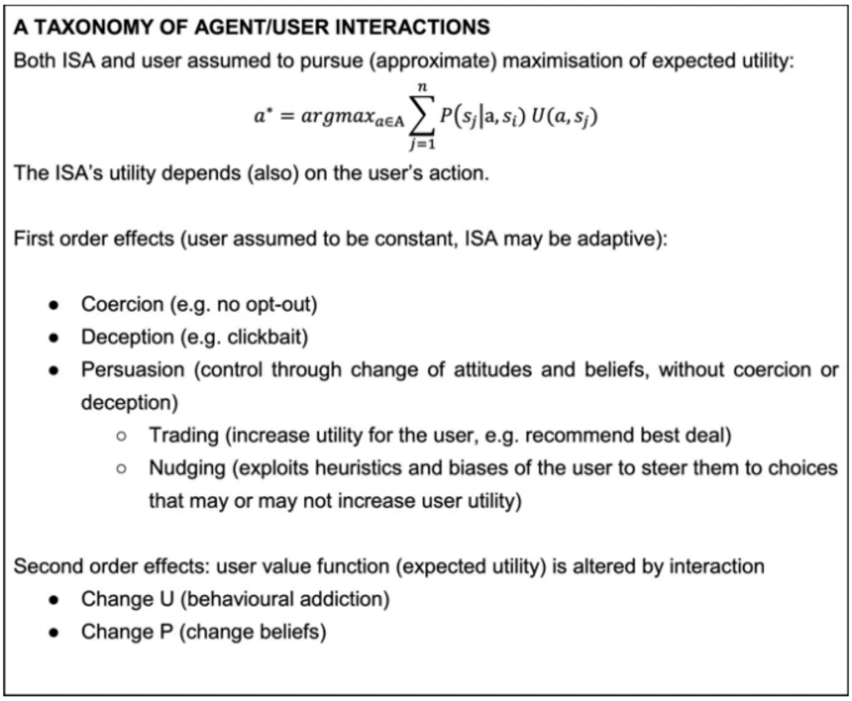
should clarify in what way?

In addition to concepts of utility for RS and utility for users and society, an adequate toy model must include mechanics for both user-user interaction and recommendations. (information flow from mediator to user) In peer-to-peer settings like Twitter - where users both write and read tweets - it makes sense to simplify content producers and consumers (see Figure 1.1) into just "users" and that's what we do.

Given this, a setup where networked agents play evolutionary games with rewiring of social ties [Santos et al., 2006b] fits our requirements as a minimal base model. Users interact in 2-by-2 matrix games, and recommendations are represented by potential new neighbors to rewire to. (Chapter 2) Recommendation mechanisms in spatial public goods games have been modeled before, although the recommendations were made by other agents instead of a central mediator. [Yang et al., 2013]

The powerful processes through which RS are optimized may already explicitly involve reinforcement learning, [Chen et al., 2020] or at least be optimized implicitly through the selection that engineers make when designing and choosing which models to deploy. Deep RL has been applied to in large discrete

citations needed



A TAXONOMY OF AGENT/USER INTERACTIONS

Both ISA and user assumed to pursue (approximate) maximisation of expected utility:

$$a^* = argmax_{a\in A} \sum_{j=1}^{n} P(s_j|a, s_i)\, U(a, s_j)$$

The ISA's utility depends (also) on the user's action.

First order effects (user assumed to be constant, ISA may be adaptive):

- Coercion (e.g. no opt-out)
- Deception (e.g. clickbait)
- Persuasion (control through change of attitudes and beliefs, without coercion or deception)
    - Trading (increase utility for the user, e.g. recommend best deal)
    - Nudging (exploits heuristics and biases of the user to steer them to choices that may or may not increase user utility)

Second order effects: user value function (expected utility) is altered by interaction

- Change U (behavioural addiction)
- Change P (change beliefs)

Figure 1.2: Taxonomy of ISA-human interactions. [Burr et al., 2018a]

action spaces [Dulac-Arnold et al., 2016] and specifically to train recommenders, from using RL to consider the long-term effects of recommendations [Liu et al., 2019] and optimizing them for long-term user engagement [Ie et al., 2019].

To the best of our ability, we couldn't find previous work using RL to solve a task of mediation in evolutionary game theory. The closest application we've been able to find was RL being used to learn policies of partner selection for individual agents in the iterated prisoner's dilemma. [Anastassacos et al., 2019]

Graph neural networks (GNN) preserve useful symmetries characteristic of graphs, making them ideal to handle graph data. GNNs have been used in conjunction with deep reinforcement learning to solve graph optimization problems. [Darvariu et al., 2020] We base our training architecture on recent work on the control of dynamical processes in graphs through node-level interventions, [Meirom et al., 2021] and train it with Proximal Policy Optimization (PPO) [Schulman et al., 2017] to obtain rewiring strategies that optimize different metrics in our model. We use PPO instead of an action-value approach like Q-learning because PPO doesn't require us to assign an approximate value to each possible action, as the action space is prohibitively large. (number of nodes). Most other policy-gradient algorithms use entropy to define a trust region. Calculating entropy is expensve, making PPO more efficient as it is not based on an explicit evaluation of entropy.

Comparing differently aligned mediators means comparing policies trained on different reward functions. Several ad-hoc methods have been used to compare reward funictions in work around reward learning. The earliest work on inverse reinforcement learning (IRL) evaluated rollouts of a policy trained on the learned rewards, [Ng and Russell, 2000] while recent work compared reward functions by making scatter plots of returns. [Ibarz et al., 2018]

1.3 Contributions

We compare the effects of differently aligned mediators, as well as simple fixed rewiring policies, on a toy model of society. After that we study the impact of competition between mediators on cooperation and other metrics. We are primarily concerned with the evolution of cooperation and network topology in our model under different combinations of mediators. Our hypotheses are that misaligned reward functions produce RS that lead to worse outcomes than their aligned counterparts, and that competition would drive people away from misaligned RS, lending weight to the idea that people should be able to modularly switch which RS they use in the real world.

speak of the concrete results instead of the hypotheses once you have them

Introduce this idea somewhere else first?

Our toy model of society is based on a previous model where networked agents play evolutionary games and rewire social ties (Santos, 2006), extended by introducing arbitrary rewiring policies to represent mediators. We then define new fixed rewiring policies and study their effects on the evolution of cooperation and network topology.

Highlight this is a contribution on its own??

We also define the task of making recommendations in our toy model as a Markov decision process: maximizing some reward function U while mediating a population of agents playing evolutionary games by recommending new neighbors to them. Then we apply an actor-critic RL algorithm to train policies in our task. Further, we extend our model to allow a notion of competition between rewiring policies and again study the evolution of cooperation and network topology under competition.

This thesis is is organized as follows: In Chapter 2, we use the literature to derive a set of requirements that a toy model of society should fulfill to be used in studying user-RS misalignment; we then present a model which fulfils those criteria. In Chapter 4 we define the task of recommending in our toy environment as an MDP and then we train agents with selfish and pro-social reward functions using deep reinforcement learning and graph neural networks. Each of these chapters contains an analysis of the evolution of cooperation and network properties of our model: Chapter 2 under fixed rewiring strategies (as baselines), and then Chapter 4 under the strategies obtained from training. Chapter 3 introduces a notion of competition between mediators to see if that would lead to better outcomes for society. Finally, we discuss our results in Chapter 5 and conclude in Chapter 6.

2

Modeling

Contents

2.1 Model requirements 11

2.2 Model definition 12

2.3 Baseline results 16

In this section, we present a framework to study the impacts of different RS mediators on individuals and society. We describe our modelling process, first deriving modelling requirements from the context present in Section 1.1, then extending an existing framework [Santos et al., 2006b] of evolutionary game theory in networks (where agents can rewire social ties) by allowing arbitrary rewiring policies to be used instead of the single one defined in the original work. Finally, we define a few fixed heuristic rewiring policies and test them on our framework to obtain baselines.

From now, we will refer to RS as ”mediators” and to users as ”agents”, to preserve the generality of our framework as it can apply in other domains where agents with partial information interact while being mediated by an agent with complete information, as is likely to happen with AI service ecosystems. [Drexler, 2019]

should I keep this sentence?

2.1 Model requirements

In Section 1.1, we discussed a few dynamics relevant to our problem: RS as multi-stakeholder environments where parties derive different utilities from recommendations; how recommendations can determine interactions between users that then shape preferences and values; and how the interests of RS operators and society have been misaligned in the past.

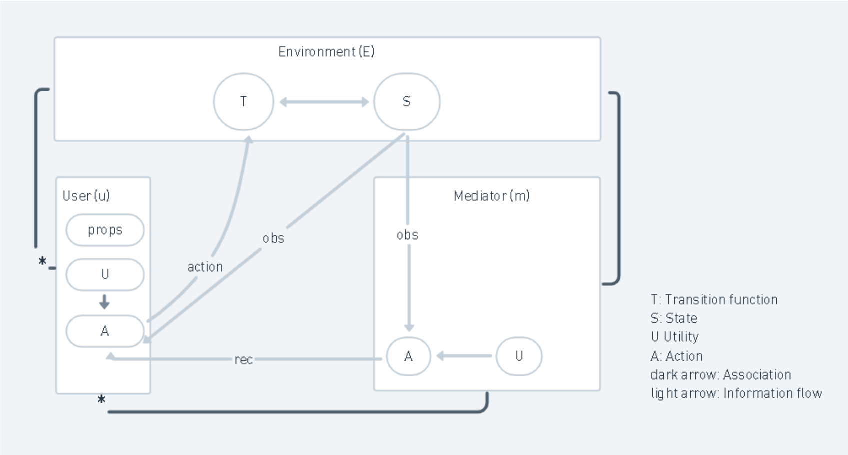


Figure 2.1: UML of the relevant entities in our model.

From our assessment of these phenomena, we have compiled a minimal list of requirements that a model intended to study RS alignment should fulfill:

Agents with partial information and mediators with complete information The need for recommender systems emerges from user inability to observe all existing items. Therefore, users should be modelled as only having information about a subset of their environment. Since the task of RS is to parse large spaces of items and present personalized subsets to users, it is a reasonable approximation to model mediators as if they have global information, even if in practice that's not strictly true.

Agent-mediator interaction - recommendations RS mediate the relationship between users and their non-local environment by providing information about it. Usually in the form of a recommendation of content or another user.

Agent-agent interaction The dynamical system of society. People exist and pursue their goals in the world. On social media platforms, experience consists of interacting with other users either directly or by consuming and producing content. The RS largely mediates this experience. Of course the offline world represents a source of externalities away from the reach of RS but an increasing part of people's lives and the economy is conducted online under mediation.

Utility functions for mediators, agents, and society RS are selected by their operators to improve some metric, either implicitly by humans or explicitly by reinforcement learning algorithms. In the case of many social media platforms this metric is a proxy for the ultimate benefit of whoever is in charge of the RS - usually related to click-through rate on advertisements or time spent by users on the platform. Defining utility for even a single individual is challenging under the full complexity of the real world. There has been work on optimizing for self-actualization or self-reporting of well-being, but ultimately, realistic notions of individual utility must include regular feedback and redefinition. A notion of utility for society is even more challenging to achieve. People often have incompatible goals and values that must be traded-off when making decisions with complete information.

citation needed

citation needed - utility aggregation

2.2 Model definition

We present a model of society consisting of networked agents playing 2-by-2 social dilemmas, evolving their strategies and rewiring social ties.

2.2.1 Base model

There are two types of individuals - cooperators and defectors - who engage in social dilemmas of cooperation - specifically 2-player symmetric games - where players can either cooperate or defect when interacting. Individuals only interact with their neighbors on the network. By comparing fitness, individuals can change their strategies to those of their neighbors. Individuals can also rewire their social ties if unsatisfied with their neighbors.

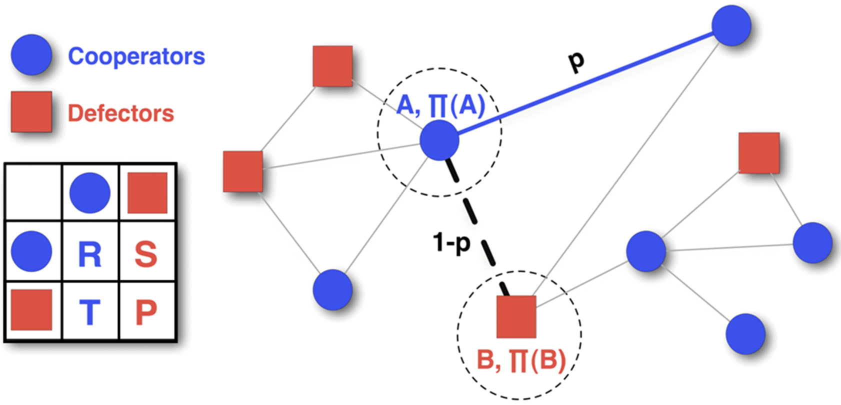


Figure 2.2: Evolving the neighborhood. [Santos et al., 2006b]

Games Agents interact by playing social dilemmas: symmetric, 2-player, 2x2 matrix games. (as seen in Figure 2.2) We normalize the difference between mutual cooperation (R) and mutual defection (P) to 1, making R = 1 and P = 0, respectively. As a consequence, we investigate all dilemmas in a 2-D parameter space where the payoff T (temptation to cheat) satisfies 0 $\leq$ T $\leq$ 2 and the payoff S (disadvantage of being cheated) satisfies $-1 \leq$ S $\leq$ 1.

Evolution The strategy of a node x evolves through imitation of a neighbor y. A node updates its strategy according to a Fermi update probability based on the difference between the fitness of each player. [Traulsen et al., 2006] (eq. (2.1)) Fitness corresponds to the cumulative payoffs of a node, resulting from the sum of payoffs from playing each of one's neighbors.

$$p = \frac{1}{1 + e^{-\beta(f_B - f_A)}} \tag{2.1}$$

Rewiring Given an edge between A and B, we say A is satisfied with the link if B is a cooperator, being dissatisfied otherwise. If A is satisfied, they will keep the link. If dissatisfied, A will compete with B to rewire the link. (Figure 2.2) The action taken is contingent on the fitness $\Pi(A)$ and $\Pi(B)$ of A and B respectively. A redirects the link to a new neighbor given by its rewiring strategy with probability p given by eq. (2.1). With probability $1 - p$, A either stays linked to B - if A is a cooperator - or B rewires its link with A to one of A's neighbors.

Timescale Strategy evolution and structural evolution can occur at different timescales, T_a and T_e respectively. The ratio $W = T_e/T_a$, leads to different outcomes for cooperation. In realistic situations, the two time scales should be of comparable magnitude. W serves as a measure of agents' inertia to react to their conditions: large values of W reflect populations where individuals - on average - react promptly to adverse ties, whereas smaller values reflect some inertia for rewiring social ties. (compared with strategy change)

Network The network is always initialized as a uniform random graph and its topology is allowed to evolve.

2.2.2 How it fits model requirements

This EGT framework fits our requirements as a model of society and interaction with mediators.

Agents only have information about their neighborhoods through cumulative game payoffs and strategies of neighbors when deciding whether to rewire. Mediators can easily make use of complete information by using the whole graph to make rewiring recommendations.

Interactions between agents are social dilemmas and players derive utility from their interactions. Agents are both providers and consumers of content, as occurs in peer-to-peer social networks. (e.g. Twitter) (fig. 1.1) Mediators interact with agents by providing recommendations of new neighbors for agents to rewire to. Although in realistic settings items of content are usually modelled, we abstract over these by representing exposure to another user's content as an edge between them in the graph.

The natural utility function for agents is their payoff obtained from games, whereas a social utility function can aggregate individual payoffs. Full cooperation in the population maximizes aggregate payoff. Furthermore, a society of cooperation could also be considered inherently desirable over a society where defection is common. Utility for mediators can be defined as any function of the environment, as we will see. (cooperators, rewires, total agent payoff)

2.2.3 Simulation algorithm

The pseudo code for the process we use in our simulations can be found in Algorithm 3.1. A Graph G is defined as a tuple (V, E) of its vertices and edges. $fermi(A, B, beta)$ is the function that calculates eq. (2.1) given A, B, and temperature term $beta$. $cumulativePayoff(x)$ returns the sum of payoffs a node x gets after playing a game with each of its neighbors. W is the ratio between the timescales of structural evolution and strategy evolution: $W = T_e/T_a$. $Strat$ is a vector of strategies (taking values in C, D) and $rewireStrat$ is a vector of rewiring strategies, each of these has a length $\#V$. A rewiring strategy is a function from a pair of nodes $(agent, neighbor)$ to a recommended node z for rewiring. $doRewire(G, x, y, z)$ deletes the edge (x, y) from G and adds edge (x, z).

2.2.4 Metrics

We introduce the metrics we are interested in studying as we vary rewiring policies.

Algorithm 2.1: Simulation algorithm

```
begin
    while $t < timeLimit$ do
        $x \longleftarrow randomSample(V)$
        $y \longleftarrow randomNeighbor(x)$
        $P_x, P_y \longleftarrow cumulativePayoff(x), cumulativePayoff(y)$
        $p \longleftarrow fermi(P_x - P_y, beta)$
        if $random(0,1) < (1+W)^{-1}$ then
            if $random(0,1) < p$ then
                $Strat_x \longleftarrow Strat_y$
        else
            if $Strat_y == D \, and \, Strat_x == C$ then
                if $random(0,1) < p$ then
                    $z \longleftarrow rewireStrat_x(x,y)$
                    $doRewire(G,x,y,z)$
            if $Strat_y == D \, and \, Strat_x == D$ then
                if $random(0,1) < p$ then
                    $z \longleftarrow rewireStrat_x(x,y)$
                    $doRewire(G,x,y,z)$
                else
                    $z \longleftarrow rewireStrat_y(y,x)$
                    $doRewire(G,y,x,z)$
```

Cooperation and total payoff Our chosen utility functions for society. Summing payoffs is the trivial way of aggregating individual utility and, beyond also maximizing total payoff, cooperation is intrinsically desirable in human systems.

write something better / more grounded

Heterogeneity and CDD The topology of networks has many interesting implications. Heterogeneous graphs have been shown to improve the survival of cooperation. [Santos et al., 2006a] Thus, we compute the heterogeneity of the graph (eq. (2.2)) and the cumulative degree distribution $D(k) = N^{-1}\sum_{i=k}^{N-1} N_i.k_{max}$ the maximum value for node degree in the graph also provides a simple measure of heterogeneity.

$$h = N^{-1} \sum_{k} k^2 N_k - z^2 \tag{2.2}$$

Rewire rate We're interested in user-engagement because it's a metric used to optimize current RS, effectively tracking use of the RS itself. It's natural to make a parallel with the number of rewire attempts to obtain a metric for the self-interest of mediators. Effectively, this number encodes the fraction between the number of "rewire attempts" (structural updates where the neighbor of the node in focus is a defector, thus triggering a rewire) and the total number of opportunities. When varying the timescale term W, in order to track the ”rewire rate” caused by a mediator, we must consider that W will modulate the number of opportunities and so we obtain the $\#opportunities = T/(1-(W+1)^{-1})$ where T is the total time the simulations run for. Finally we get $rewireRate = \#attempts/\#opportunities$.

2.2.5 Fixed Rewiring Strategies

We define the following fixed rewiring strategies to study this framework in extreme rewiring environments, as well as to serve as baselines in our analysis of trained mediators.

RANDOM Recommends a random node from the whole graph. The baseline uniform distribution case.

NO MED The rewiring policy defined in [Santos et al., 2006b]. Given an edge (A, B), node A rewires to a random neighbor of node B. The intuition behind this reasoning is that simple agents, being rational individuals with partial information, are more likely to interact with nearby agents. [Kossinets and Watts, 2006] Moreover, selecting a neighbour of an inconvenient partner is also a good choice, since this partner also tries to establish links with cooperators, making it more likely that the rewiring results in a tie to a cooperator.

GOOD Always recommends a random cooperator. This is meant to represent the behavior of a mediator perfectly capable of identifying cooperators and defectors. This mediator is aligned with each individual user who requests a recommendation, without considering that it is harming cooperators by rewiring defectors to them.

BAD Like GOOD, but always recommends a random defector. This is meant to represent the behavior of a poorly aligned or exploitative mediator.

FAIR Recommends a random cooperator to cooperators and a random defector to defectors. Implements some notion of fairness by only benefiting cooperators and punishing defectors. Before our experiments, this policy was expected to lead to the fastest convergence.

2.3 Baseline results

Our simulations have produced an unexpected ordering of rewiring policies eq. (2.3) in terms of the speed of convergence in parameter space (T,S). All of our experiments in chapter 2 are initialized on a uniform random network with number of nodes $N = 500$, mean degree $z = 30$, and imitation temperature $\beta = 0.005$.

2.3.1 No rewiring

Our first baseline, a contour plot of the final fraction of cooperators (fig. 2.3) in the absence of rewiring (W=0) is plotted as a function of two game-parameters: S, the disadvantage of a cooperator being defected (when $S < 0$), and T, the temptation to defect on a cooperator (when $T > 1$). Absent any of these threats ($S \geq 0$ and $T \leq 1$; upper-left quadrant) cooperators trivially dominate. The lower-left quadrant ($S < 0$ and $T \leq 1$) corresponds to the Stag-Hunt dilemma, by definition. The lower triangle in the upper-right quadrant ($S \geq 0$, $T > 1$ and $(T + S) < 2$) corresponds to the Snowdrift game, also by definition. The lower-right quadrant ($S < 0$ and $T > 1$) corresponds to the Prisoner's Dilemma domain (PD). (Left) [Santos et al., 2006a]

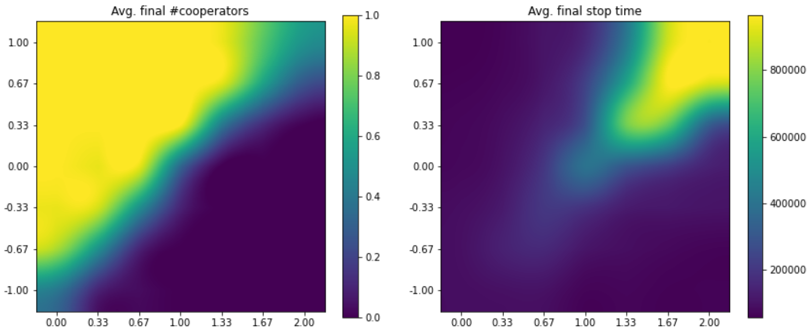


Figure 2.3: Evolution of cooperation in a uniform random network.

2.3.2 Rewiring with no mediator

Introducing rewiring ($W > 0$) according to NO_MED to the experiment in section 2.3.1, we recover the result from [Santos et al., 2006b]. fig. 2.4 shows the fraction of successful evolutionary runs ending in 100% cooperation for different values of the time-scale ratio W. Above a critical value ($W_{critical} \approx 4.0$) cooperators efficiently wipe out defectors.

add W labels

2.3.3 Heuristic mediators

In this section we present the results we obtained from running simulations where nodes rewire social ties according to fixed heuristic rewiring strategies. Primarily, we observe an ordering by convergence speed.

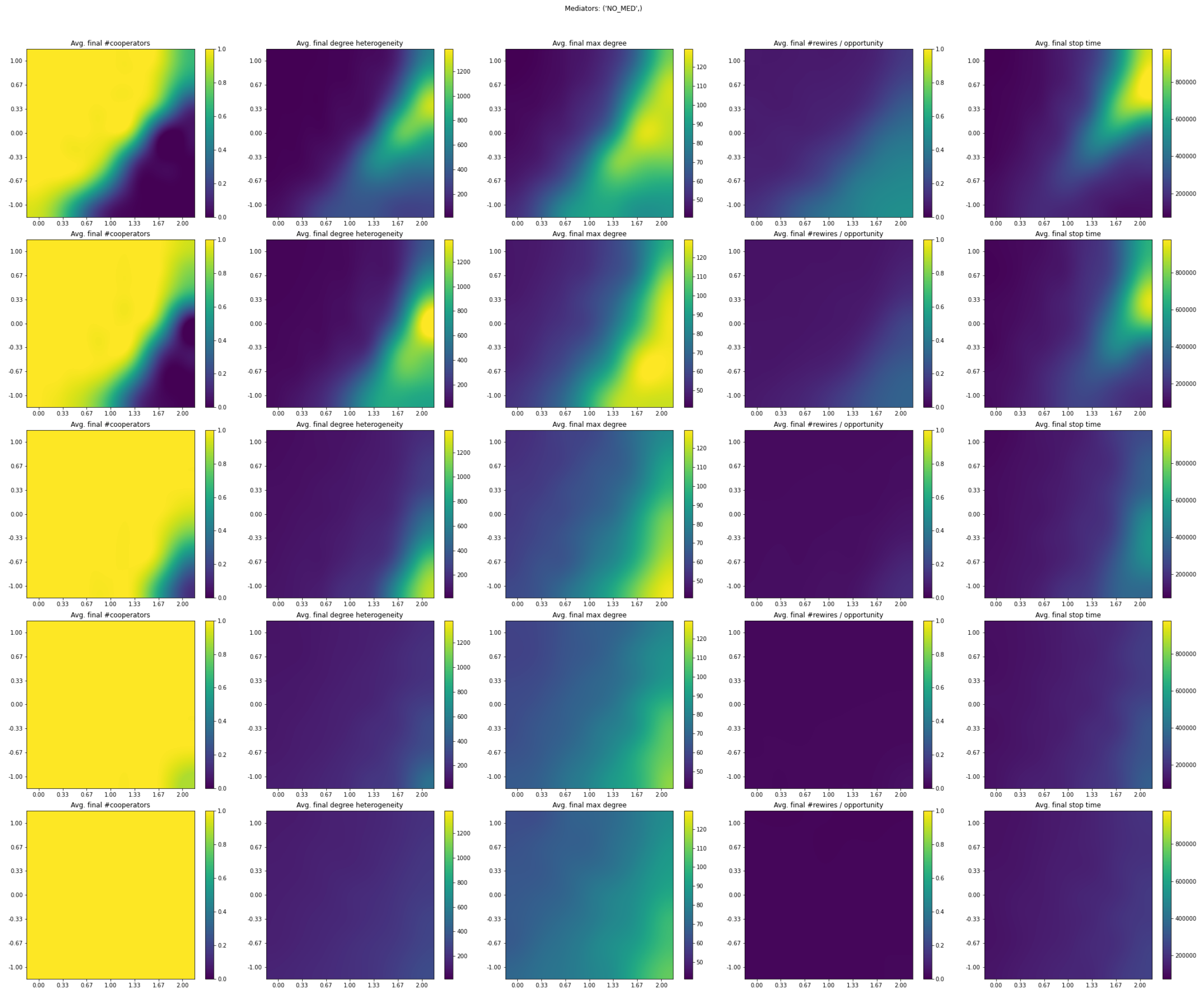


Figure 2.4: Co-Evolution for Different Dilemmas and Time Scales. Each row: W=0.5,1,2,3,4, each column: final fraction of cooperators, final heterogeneity, final k$_{max}, rewire_n$.

$$BAD < NOMED < FAIR < RANDOM < GOOD \tag{2.3}$$

We plot the evolution of cooperation for each of our types of mediator. (fig. 2.5) As expected, BAD produces little change in outcomes. We expected FAIR mediators to incentivize cooperation even more than RANDOM or even GOOD, as these didn't punish defection, but this was not the case. Our sense is that FAIR might isolate cooperators from defectors, leading to defectors having no cooperators to imitate. RANDOM doing better than NO MED also shows that escaping one's neighborhood by rewiring to a random place in the whole graph can lead to faster convergence in this model, although it obviously has drawbacks in the real world.

expand

We also plot our metrics solely for the Prisoner's Dilemma (T=2,S=-1) as a more granular function of W, to better compare the convergence speed under different mediators.

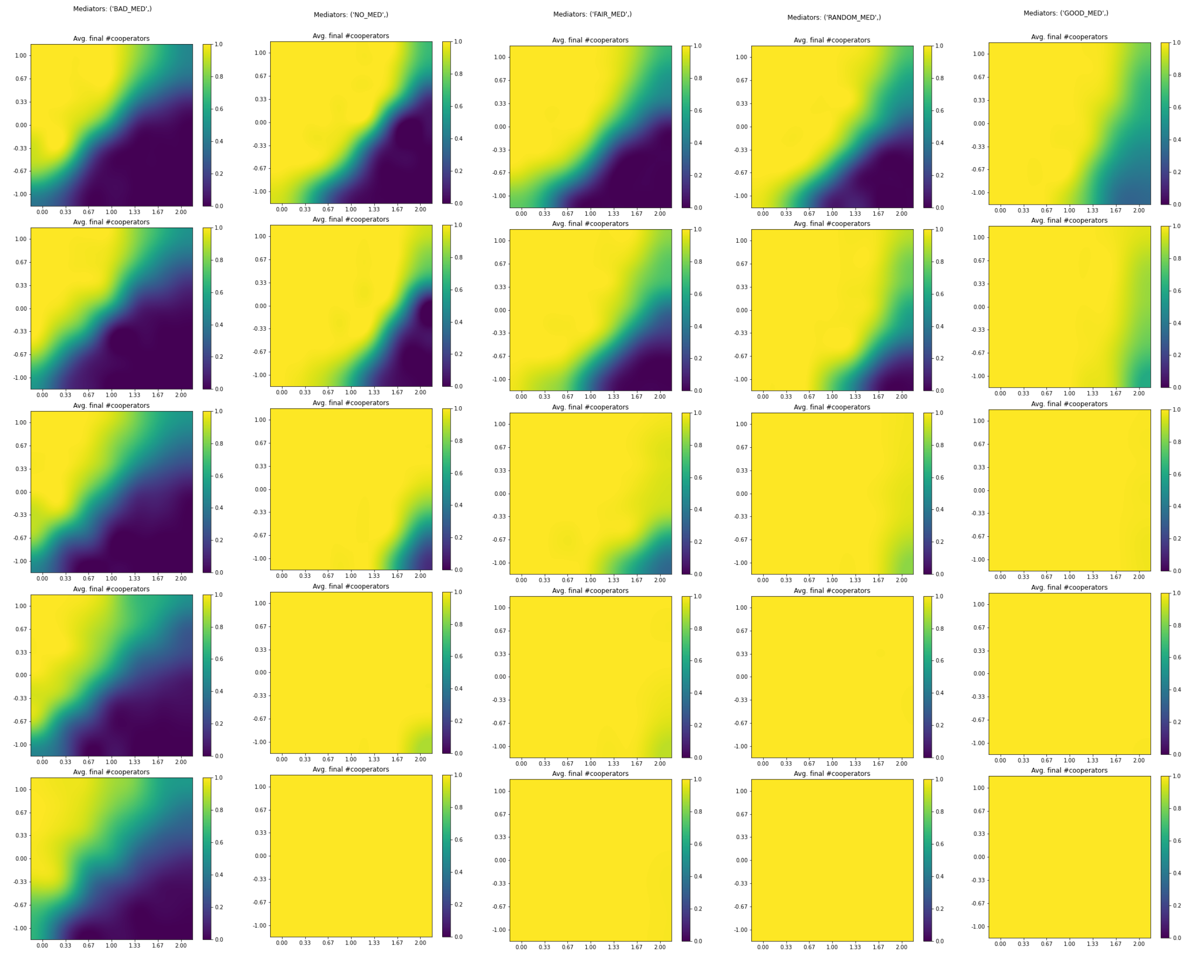


Figure 2.5: Co-Evolution of cooperation for Different Dilemmas (t,s axes) and Time Scales (rows), for Different Mediators (columns).

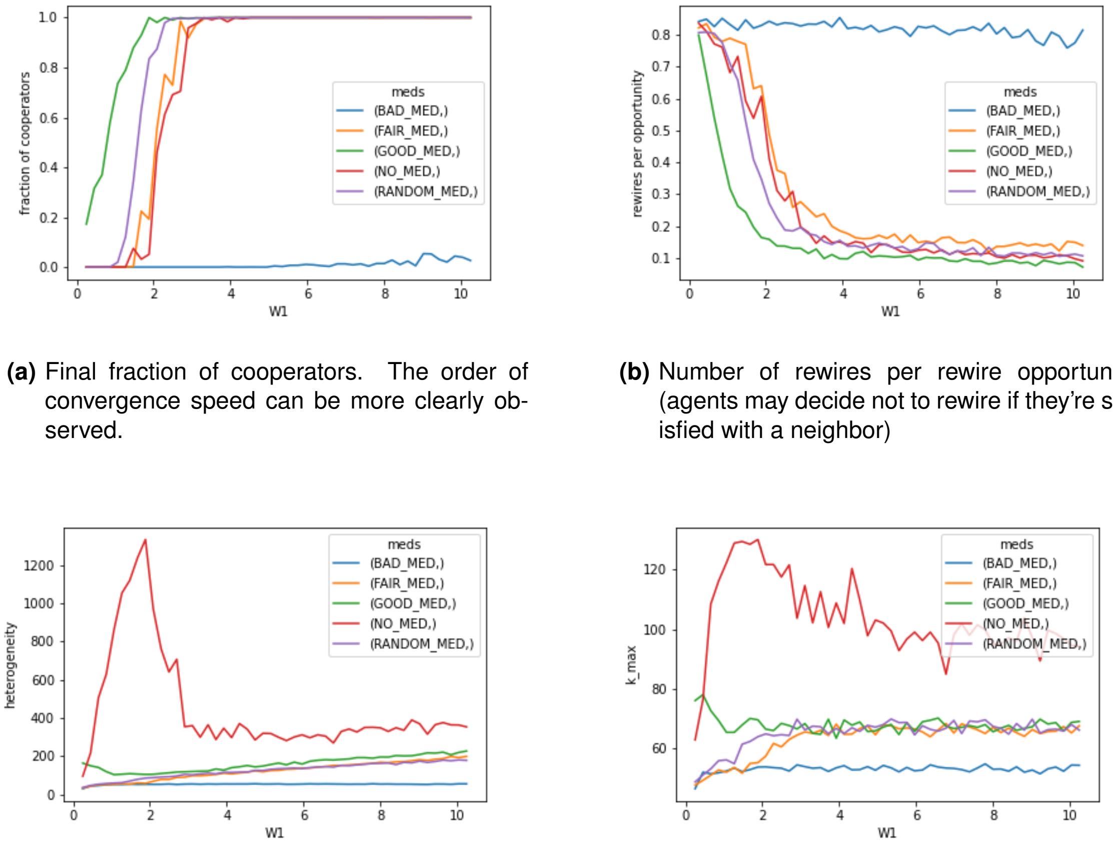
In terms of rewiring, BAD leads to the highest rate, while GOOD produces the lowest, with the remaining mediators being comparable. One of the reasons is that agents will only want recommendations if the tie they're rewiring is a Defector, so a population that converges to D will always want rewires, while a population that converges to C will cease to want new neighbors.

NO MED produces by far the most heterogeneous graphs, presumably due to its local rewiring dynamics. (fig. 2.7) Among the fixed mediators, GOOD produces the highest heterogeneity around the prisoner's dilemma (T=2, S=-1) and especially at low and high Ws.

include rewire contour plots

worth running local-first mediators to test hypothesis?

NO MED has a different value range. Fix? Use other pallete?



(a) Final fraction of cooperators. The order of convergence speed can be more clearly observed.

(b) Number of rewires per rewire opportunity. (agents may decide not to rewire if they're satisfied with a neighbor)

(c) Heterogeneity. NO MED clearly leads to more heterogeneous networks. We expect this to be due to its local focus, rather than global recommendations.

(d) Max degree, another metric of heterogeneity. We can see BAD leads to the absolute lowest heterogeneity.

Figure 2.6: Co-evolution of cooperation and structure in the prisoner's dilemma (t=2,s=-1) as a function of W for different mediators.

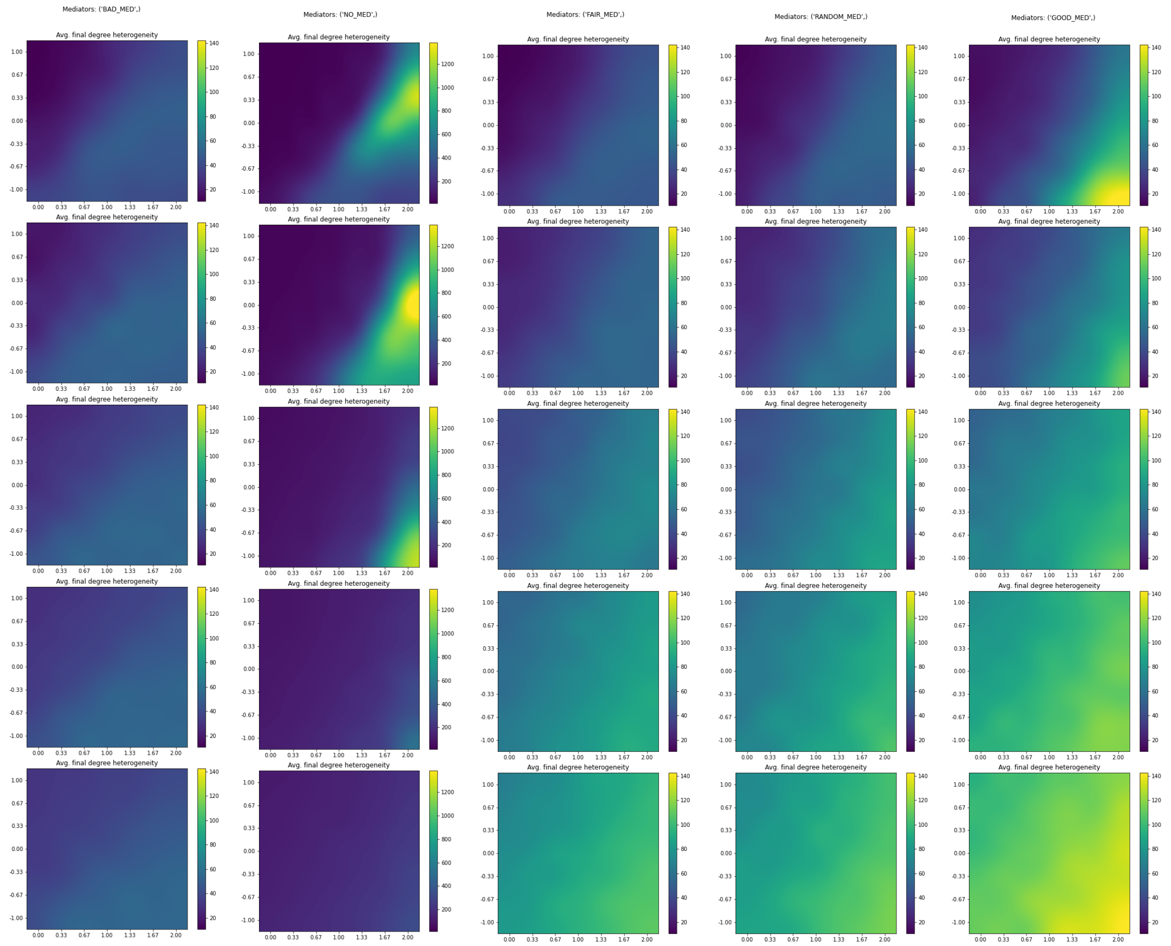


Figure 2.7: Heterogeneity for Different Dilemmas (t,s axes) and Time Scales (rows), for Different Mediators (columns).

3 Competition

Contents

3.1 Model of Competition 25

3.2 Baseline Results 27

In this section, we introduce the notion of competition between mediator strategies to our model and present baselines for the outcomes of competition between fixed rewiring strategies.

3.1 Model of Competition

We extend the model by allowing competition between mediators according to the pseudo-code in algorithm 3.1. This is achieved by attributing a rewiring strategy (mediator) to each node and allowing them to evolve in the same way as game strategies evolve. At each time-step, there is a chance that the update performed is a mediator update. In that case, an agent selects a random neighbor and imitates its rewiring strategy with probability p weighed on their fitness difference given by the Fermi update. (eq. (2.1)) Mediators are exclusive, meaning they will only recommend nodes from among their own users. This reflects what we see in the world, where systems (mostly) only have information about their own users.

Finally, we introduce a second timescale ratio $W2 = t_m/t_e + t_a$ to regulate the relative frequencies of mediator updates (t_m) and other two kinds of updates. (strategy t_a or structural t_e)

Algorithm 3.1: Simulation algorithm for mediator competition

```
begin
    while $t < timeLimit$ do
        $x \longleftarrow randomSample(V)$
        $y \longleftarrow randomNeighbor(x)$
        $P_x, P_y \longleftarrow cumulativePayoff(x), cumulativePayoff(y)$
        $p \longleftarrow fermi(P_x - P_y, beta)$
        if $random(0,1) < (1 + W2)^{-1}$ then
            if $random(0,1) < p$ then
                $rewireStrat_x \longleftarrow rewireStrat_y$
        else
            if $random(0,1) < (1 + W)^{-1}$ then
                if $random(0,1) < p$ then
                    $Strat_x \longleftarrow Strat_y$
            else
                if $Strat_y == D \, and \, Strat_x == C$ then
                    if $random(0,1) < p$ then
                        $z \longleftarrow rewireStrat_x(x, y)$
                        $doRewire(G, x, y, z)$
                if $Strat_y == D \, and \, Strat_x == D$ then
                    if $random(0,1) < p$ then
                        $z \longleftarrow rewireStrat_x(x, y)$
                        $doRewire(G, x, y, z)$
                    else
                        $z \longleftarrow rewireStrat_y(y, x)$
                        $doRewire(G, y, x, z)$
```

3.2 Baseline Results

Here we study the effects of competition between all our baseline strategies in a Prisoner's dilemma setting, over various timescale combinations W and W2. All competition runs are initialized with 1000 nodes, 90% of which use the NO MED rewiring strategy, while each node the remaining 10% has one of the others. Mediator updates use a different temperature parameter $beta_{med} = beta * 10 = 0.05$ because the impact of mediator updates was negligible using 0.005. Given the stochastic nature of our simulations, all presented results are averaged over 30 runs.

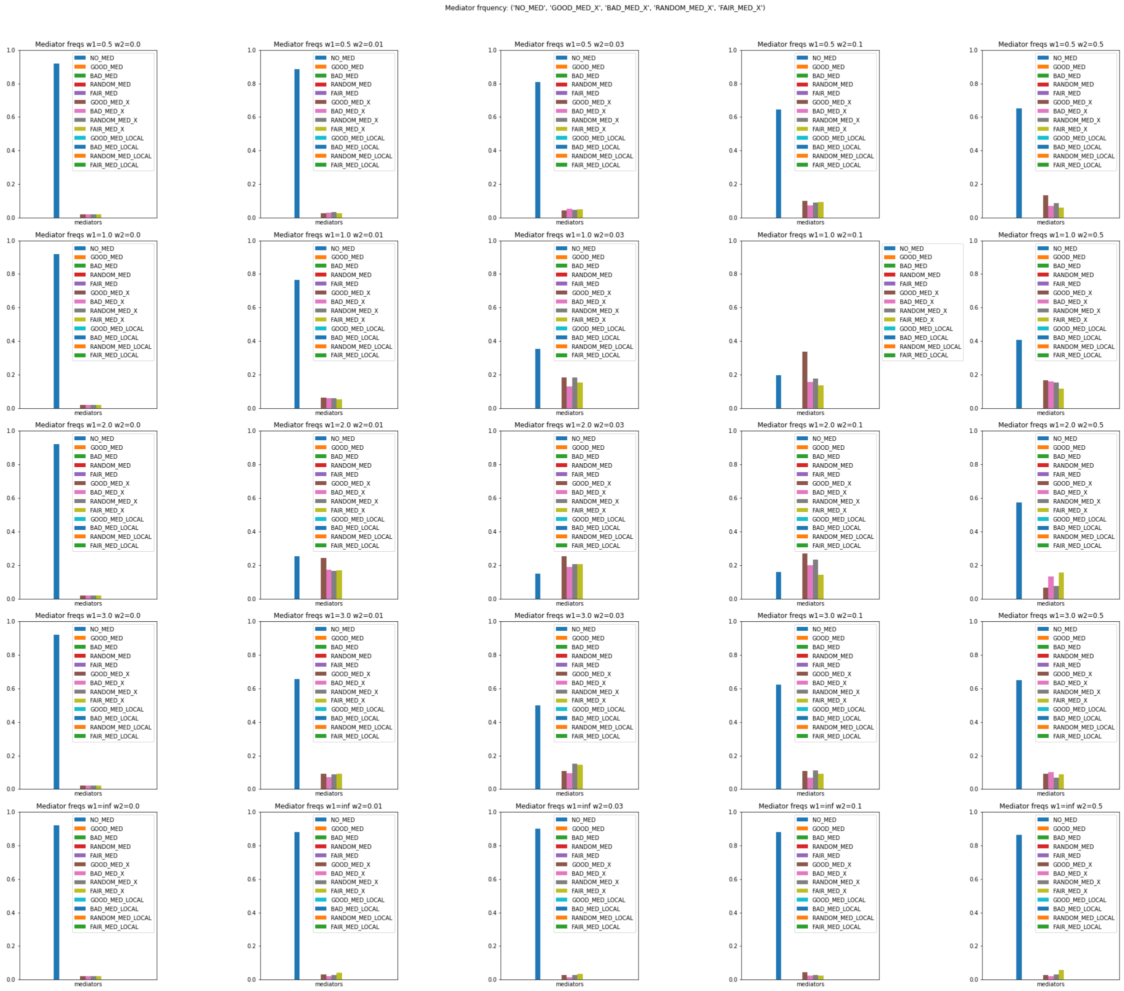
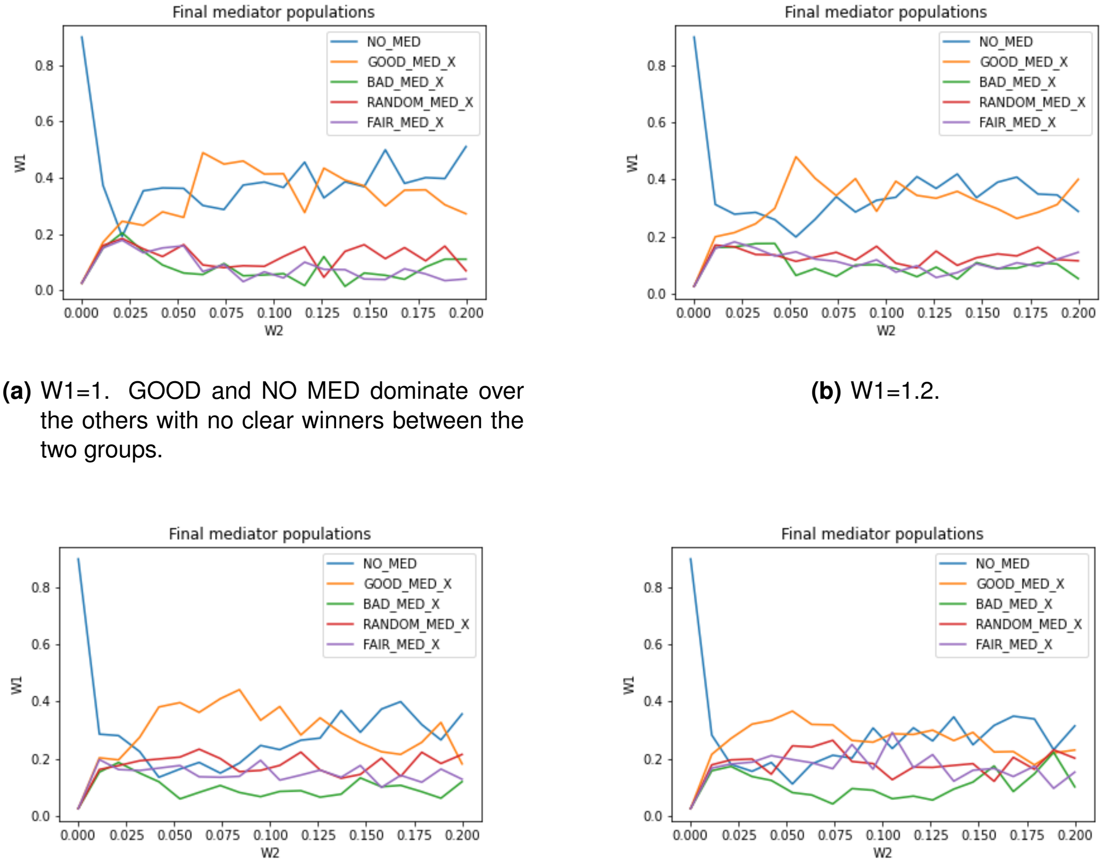


Figure 3.1: Average final fraction of mediator populations after competition between NO MED and exclusive fixed mediators. For $W1 \in \{0.5, 1, 2, 3, \inf\}$ and $W2 \in \{0.01, 0.03, 0.1, 0.5, \inf\}$. For higher values of W2, we see a relatively low adoption rate and uniform distribution of mediators besides NO MED. Initialized at 90% NO MED and 10% split between the rest.

We observe (fig. 3.1) that for $W1 = \inf$ (no strategy updates, only rewires and mediator updates), the initial conditions remain practically the same, while for $W2 = 0$S the initial conditions remain the

remake plot with less labels

same due to there being no mediator updates. We find a critical region where NO MED does not have a majority around $W1 \in \{1, 2\}$ and $W2 \in \{0.03, 0.1\}$. We investigate this range more closely in fig. 3.2 and observe that GOOD and NO MED have a tendency to dominate over the others that is most pronounced at $W1 = 1$. As W1 increases, the gap between strategy populations becomes less evident, although BAD takes the bottom place more convincingly. For the same interval of $W2$, we study metrics of cooperation, heterogeneity, and rewires (fig. 3.3) and find that higher values of W1 lead to higher cooperation overall (as expected) but decrease heterogeneity and rewire number. Comparing these with the metrics we obtained for the simulations with single mediators, we observe much higher heterogeneity (around 500 rather than 200) and k_max (around 200 to 450 rather than 60 to 80) in the competition scenario.



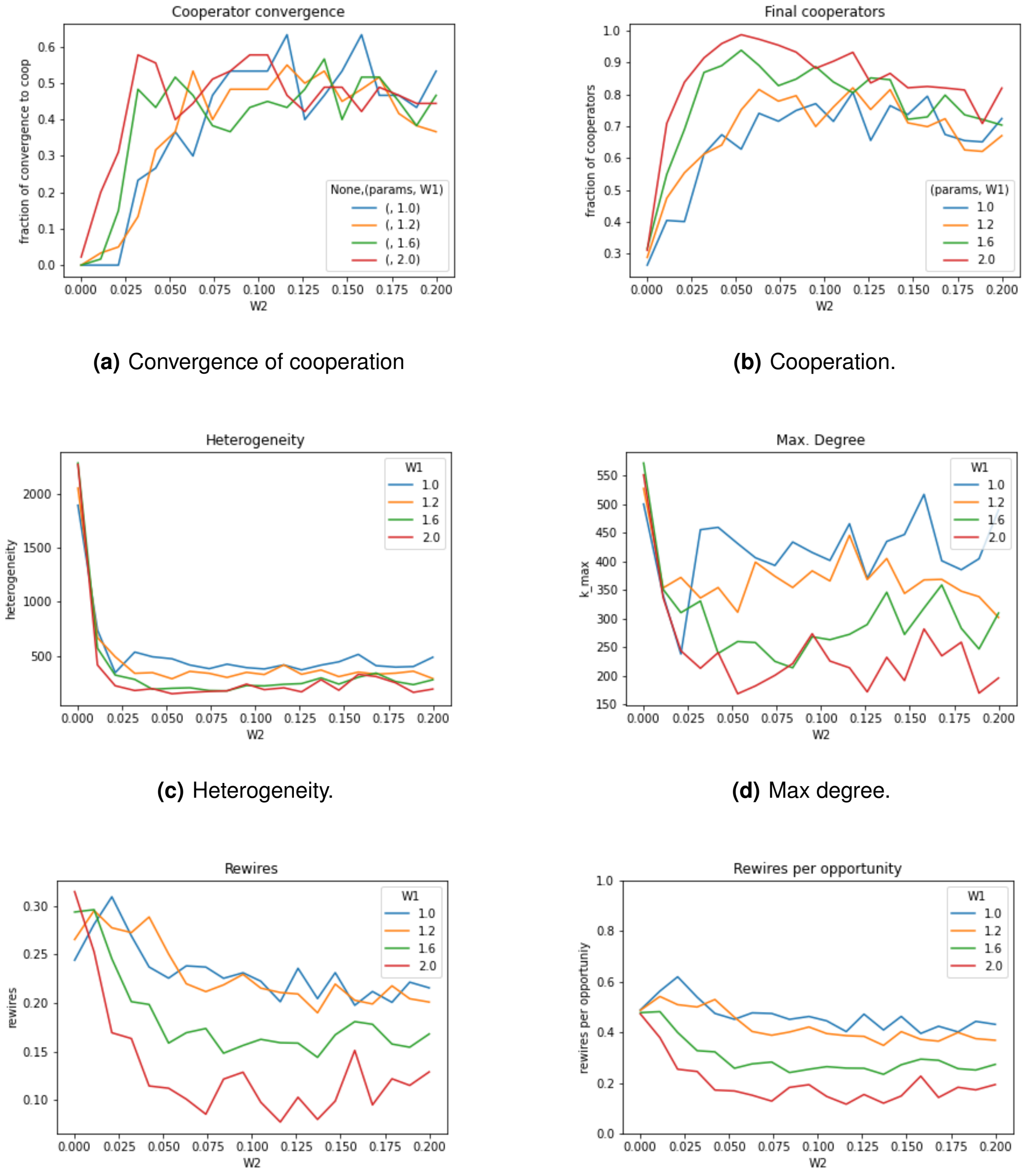
(a) W1=1. GOOD and NO MED dominate over the others with no clear winners between the two groups.

(b) W1=1.2.

(c) W1=1.6.

(d) W1=2. More difficult to discern than W1=1 but BAD is clearly on the bottom, while GOOD and NO MED seem to share the top but by a smaller margin.

Figure 3.2: Final frequencies of mediators as functions of W2.



(a) Convergence of cooperation

(b) Cooperation.

(c) Heterogeneity.

(d) Max degree.

(e) Rewires.

(f) Rewires per opportunity.

Figure 3.3: Metrics as a function of W2 for different values of W1.

4

Training

Contents

4.1 Environment: Mediating EGT 33

4.2 Training Architecture 33

4.3 Utility Functions 34

4.4 Training 34

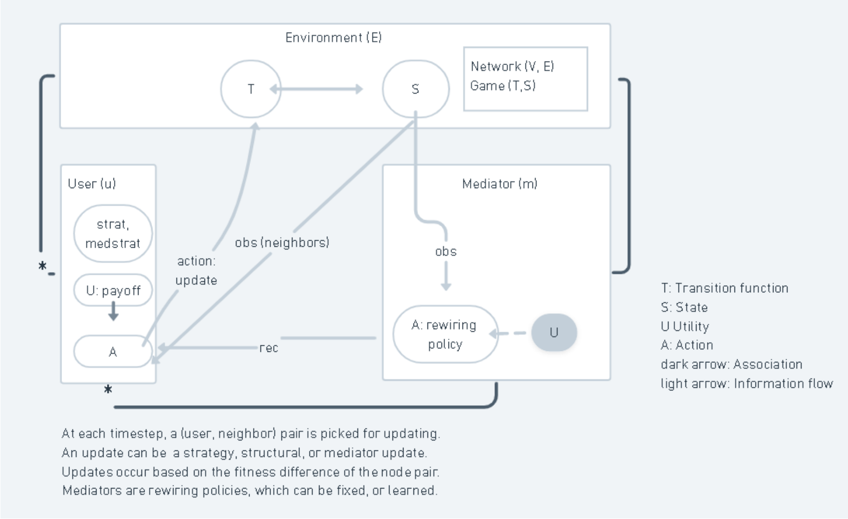
4.5 Results 34

In this section, we describe the task of recommendation as an MDP, the RL architecture that we used to train rewiring policies, detail the training process, and present the impacts of the trained mediators as applied to our toy model.

This section isn't done yet

4.1 Environment: Mediating EGT

We modelled the task of choosing new neighbors to rewire to (with the goal of maximizing some metric).



At each timestep, a (user, neighbor) pair is picked for updating.
An update can be a strategy, structural, or mediator update.
Updates occur based on the fitness difference of the node pair.
Mediators are rewiring policies, which can be fixed, or learned.

Figure 4.1: UML of the relevant entities in our model, with added information from the current implementation..

To train a rewiring policy using reinforcement learning, we must express the task of continually providing recommendations to the evolving system as a Markov decision problem. (MDP) (S, A, P, R, γ)

An MDP is one possible formalization of decision making processes. The decision maker, called agent, interacts with an environment. When in a state $s \in S$, the agent must take an action a out of the set $A(s)$ of valid ones, receiving a reward r governed by the reward function $R(s, a)$. Finally, the agent finds itself in a new state s' , depending on a transition model P that governs the joint probability distribution $P(s', a, s)$ of transitioning to state s' after taking action a in state s. This sequence of interactions gives rise to a trajectory. The agent's goal is to maximize the expected discounted sum of rewards it receives over all trajectories.

4.2 Training Architecture

We design a model that leverages graph neural networks (GNNs) and optimize it with proximal policy optimization (PPO), based on recent work done with the aim of controlling dynamical processes in graphs. [Meirom et al., 2021] The original work focuses on maximizing metrics in epidemic or influence maximization processes, whereas our underlying processes are evolutionary games.

4.3 Utility Functions

To compare aligned and misaligned mediators, we need reward functions. We model ”user-engagement” as the total number of rewire requests, as that corresponds to the number of times users are interacting with the mediator. In the real world, it would be very hard to find a reward function perfectly aligned with the best interest of society but in our model, maximizing the number of cooperators would do the most good.

Discuss these reward functions further in the appendix

4.4 Training

4.5 Results

5

Discussion

Contents

5.1 Single mediators 37

5.2 Competition 37

5.3 Training 37

Still a stub, will be significantly expanded

The goal of this investigation is to progress towards a world where RS mediate human systems in ways that help them flourish and achieve their goals, whatever those are. In that respect, we believe cooperation in 2-player social dilemmas is an acceptable abstraction for ”human flourishing” and one we can represent in a very simple model. Many hard problems that might be helped by mediation are problems of cooperation and coordination, so widespread cooperation is a reasonable metric to evaluate a mediator on.

A point we must address is the use of number of rewires as the utility function for self-interested mediators, which we chose in analogy to how real world RS maximize user retention. This needs to be mentioned because the state that maximizes number of rewires is that of convergence to all-Defectors. Rewires implying defectors wasn’t intended at the time of defining the model but it does mean that, in this model, user retention is in direct opposition to maximizing the number of cooperators. This is analogous to - in real life - the RS avoiding satisfying the user lest he stop searching. Of course, real people always have the option of not using the RS - which is not the case here - and they also have reasons to ask for recommendations that are not a defecting partner - e.g. new needs arising.

5.1 Single mediators

The heterogeneity of networks mediated by fixed heuristic mediators is always much lower than those mediated by NO MED. This is to be expected because mediators recommend random nodes from anywhere in the network.

explain what happens if we use local-first mediators

We observed that FAIR mediators that penalize Defectors actually produce slower convergence towards cooperation. We hypothesize that this is because defectors and cooperators become segregated, leaving defectors without ”good examples” to imitate strategies from. FAIR also handles higher S values than NO MED, but lower Ts, so under a FAIR mediator you can be slightly less afraid (of defection) but must also be less greedy.

test modularity of a network partition by strategy

verify this later

5.2 Competition

In competition, BAD is significantly out-competed, and the mediator that actually gets more rewires is BLANK. NO MED and GOOD dominate over other strategies in terms of presence in the population.

fill this in once you check

why NO MED and not FAIR or RANDOM?

5.3 Training

6

Conclusion

Contents

6.1 Limitations and Future Work 41

Still a stub, will be significantly expanded

We identified a problem in the world: misalignment between users and recommender systems, agents in networks and their mediators. We sought to compare how cooperation in society evolved under mediators with different goals. In order to make this comparison, we constructed a toy model to use as a substrate. We defined a few fixed heuristic mediators and analyzed their effects on the model to obtain baselines. Then we used RL to learn mediator policies to maximize different metrics and compare those. Finally, we set these mediators in competition to understand their adoption dynamics and the impacts of competition.

We identify an ordering between fixed heuristic mediators with respect to how fast a network mediated by them converges towards cooperation. We also conclude that there's a critical region in timescale space where competition between mediators leads to non-uniform results and which leads to mediators which correctly identify cooperators (GOOD) to dominate in tandem with a random local heuristic. (NO MED)

heterogeneity, rewires

include RL results when ready

6.1 Limitations and Future Work

stub, write more after full results

Our toy model is very simple, recommendation is optimal as long as one gets a Cooperator. A number of features could be added to agents to make personalized recommendation more challenging. We model user-user interaction and recommendation directly, rather than modelling individual items, which is more common in real systems.

This framework of comparison can be applied to any setting where agents with partial information interact on an environment and are able to request information from a mediator agent with global (or complete) information. It's useful for studying strategic dynamics between agents and mediator.

A direct extension of this work would be to replace 2-player games with public goods dilemmas, which are some of the most important problems we face. A more effortful one would be to do away with matrix games and explore these dynamics in sequential social dilemmas with multi-agent RL.

Bibliography

[Anastassacos et al., 2019] Anastassacos, N., Hailes, S., and Musolesi, M. (2019). Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. *arXiv:1902.03185 [cs]*.

[Andreassen, 2015] Andreassen, C. S. (2015). Online Social Network Site Addiction: A Comprehensive Review. *Current Addiction Reports*, 2(2):175–184.

[Babaioff et al., 2015] Babaioff, M., Feldman, M., and Tennenholtz, M. (2015). Mechanism Design with Strategic Mediators. *arXiv:1501.04457 [cs]*.

[Burr et al., 2018a] Burr, C., Cristianini, N., and Ladyman, J. (2018a). An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, 28(4):735–774.

[Burr et al., 2018b] Burr, C., Cristianini, N., and Ladyman, J. (2018b). An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, 28(4):735–774.

[Calero Valdez and Ziefle, 2018] Calero Valdez, A. and Ziefle, M. (2018). Human Factors in the Age of Algorithms. Understanding the Human-in-the-loop Using Agent-Based Modeling. In Meiselwitz, G., editor, *Social Computing and Social Media. Technologies and Analytics*, Lecture Notes in Computer Science, pages 357–371, Cham. Springer International Publishing.

[Chen et al., 2020] Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., and Chi, E. (2020). Top-K Off-Policy Correction for a REINFORCE Recommender System. *arXiv:1812.02353 [cs, stat]*. arXiv: 1812.02353.

[Darvariu et al., 2020] Darvariu, V.-A., Hailes, S., and Musolesi, M. (2020). Improving the Robustness of Graphs through Reinforcement Learning and Graph Neural Networks. *arXiv:2001.11279 [cs, stat]*. arXiv: 2001.11279.

[Drexler, 2019] Drexler, K. (2019). Reframing Superintelligence: Comprehensive AI Services as General Intelligence", Technical Report. Technical report, Future of Humanity Institute, University of Oxford.

[Dulac-Arnold et al., 2016] Dulac-Arnold, G., Evans, R., van Hasselt, H., Sunehag, P., Lillicrap, T., Hunt, J., Mann, T., Weber, T., Degris, T., and Coppin, B. (2016). Deep Reinforcement Learning in Large Discrete Action Spaces. *arXiv:1512.07679 [cs, stat]*. arXiv: 1512.07679.

[Floridi, 2011] Floridi, L. (2011). The Construction of Personal Identities Online. *Minds and Machines*, 21(4):477–479.

[Goodhart, 1981] Goodhart, C. (1981). *Problems of Monetary Management: The UK Experience*.

[Hohnhold et al., 2015] Hohnhold, H., O'Brien, D., and Tang, D. (2015). Focusing on the Long-term: It's Good for Users and Business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1849–1858, Sydney, NSW, Australia. Association for Computing Machinery.

[Ibarz et al., 2018] Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., and Amodei, D. (2018). Reward learning from human preferences and demonstrations in Atari. *arXiv:1811.06521 [cs, stat]*. arXiv: 1811.06521.

[Ie et al., 2019] Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., Cheng, H.-T., Lustman, M., Gatto, V., Covington, P., McFadden, J., Chandra, T., and Boutilier, C. (2019). Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology. *arXiv:1905.12767 [cs, stat]*.

[Izsak et al., 2014] Izsak, P., Raiber, F., Kurland, O., and Tennenholtz, M. (2014). The search duel: a response to a strong ranker. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, SIGIR '14, pages 919–922, Gold Coast, Queensland, Australia. Association for Computing Machinery.

[Khwaja et al., 2019] Khwaja, M., Ferrer, M., Iglesias, J. O., Faisal, A. A., and Matic, A. (2019). Aligning Daily Activities with Personality: Towards A Recommender System for Improving Wellbeing. *arXiv:1909.03847 [cs]*. arXiv: 1909.03847.

[Kossinets and Watts, 2006] Kossinets, G. and Watts, D. J. (2006). Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90. Publisher: American Association for the Advancement of Science Section: Report.

[Kurland, 2019] Kurland, Oren, M. T. (2019). Rethinking Search Engines and Recommendation Systems: A Game Theoretic Perspective.

[Liu et al., 2019] Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., Guo, H., and Zhang, Y. (2019). Deep Reinforcement Learning based Recommendation with Explicit User-Item Interactions Modeling. *arXiv:1810.12027 [cs]*. arXiv: 1810.12027.

[Meirom et al., 2021] Meirom, E. A., Maron, H., Mannor, S., and Chechik, G. (2021). Controlling Graph Dynamics with Reinforcement Learning and Graph Neural Networks. *arXiv:2010.05313 [cs]*. arXiv: 2010.05313.

[Milano et al., 2020] Milano, S., Taddeo, M., and Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & SOCIETY*.

[Ng and Russell, 2000] Ng, A. and Russell, S. (2000). Algorithms for Inverse Reinforcement Learning. *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*.

[Santos et al., 2006a] Santos, F., Pacheco, J., and Lenaerts, T. (2006a). Evolutionary Dynamics of Social Dilemmas in Structured Heterogeneous Populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3490–4.

[Santos et al., 2006b] Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006b). Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Computational Biology*, 2(10).

[Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv:1707.06347 [cs]*. arXiv: 1707.06347.

[Seaver, 2019] Seaver, N. (2019). Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 24(4):421–436.

[Stray et al., 2021] Stray, J., Vendrov, I., Nixon, J., Adler, S., and Hadfield-Menell, D. (2021). What are you optimizing for? Aligning Recommender Systems with Human Values. *arXiv:2107.10939 [cs]*. arXiv: 2107.10939.

[Traulsen et al., 2006] Traulsen, A., Nowak, M. A., and Pacheco, J. M. (2006). Stochastic Dynamics of Invasion and Fixation. *Physical Review E*, 74(1):011909.

[Vendrov and Nixon, 2019] Vendrov, I. and Nixon, J. (2019). Aligning Recommender Systems as Cause Area. Publication Title: EA Forum.

[Yang et al., 2013] Yang, Z., Li, Z., Wu, T., and Wang, L. (2013). Role of recommendation in spatial public goods games. *Physica A: Statistical Mechanics and its Applications*, 392(9):2038–2045.

[Zafari et al., 2018] Zafari, F., Moser, I., and Baarslag, T. (2018). Modelling and Analysis of Temporal Preference Drifts Using A Component-Based Factorised Latent Approach.