Snakes in the Plane:

Mitigating the Effect of Bad Actors on a Network Coordination Task

Matt Nicholson

Advised by: Kyosuke Tanaka and Noshir Contractor

Abstract:

Coordination is vital to the human experience in everything from a group of friends picking a lunch spot to billion dollar construction projects. Similarly, the human experience is networked, as people have only local connections to those around them. In certain cases, agents in the network can act in opposition to the objectives of all those around them. Whether out of self-interest, or malintent, the effect can be modelled similarly. In this work, we present an agent-based model of networked coordination in the presence of adversaries, and aim to shed light on the underlying mechanism of the formation (or lack thereof) of consensus. We perform 100 runs of the model for each experimental (__ total) the model and find that _____.

Introduction:

Coordination and cooperation are important in almost every aspect for human life, and this is seen time and time again. Examples come from a wide range of domains, from economics, with Hardin's Tragedy of the Commons, and Tucker's Prisoners Dilemma. From Social Psychology, where the effects of framing a scenario differently can lead to different results. This even includes a task as mundane as driving on the highway. At any given time, a driver needs to coordinate (pick a lane) and cooperate (let others into their lane). Coordination and cooperation are vital to the human experience.

However, humans to not operate in a vacuum. We are connected in various social networks, and this can influence our behavior. Conversely, our behavior can also influence our social networks. Thus, this problem of coordination and cooperation depends heavily on the social environment where individuals find themselves.

In some cases, incentives to cooperate or coordinate are not uniform. Consider the discussion of zero-sum games, where one player's payoff is the opposite of another's. Clearly, global coordination and cooperation is much more difficult to achieve when compared to cases with identical payoffs (though I note that this is still difficult! – see literature on Social Dilemma problems). In other cases, the discrepancies between incentives could come from a different understanding of the world, or simply a lack of awareness of the truth. In a 2014 study on the spread of misinformation on Twitter, Starbird found that corrections lag behind rumors in both time and volume. This implies that even acting out of goodwill, behavior can act counter to the overall goals of a system.

Perhaps the most salient example of coordination in daily life is that of civic discourse, where the general public seeks to align on what they believe is best. However, much of this process occurs in online communities, where there exist a multitude of different voices with a variety of different intents. (SOMETHING ABOUT ECHO CHAMBERS and TROLLS).

This leads to the question of what is a bad actor. I operationalize the term "bad actor" more closely with the second player in a zero sum game, and define a bad actor as an agent whose incentive is contrary to that of the first. In the context of my research question, a bad actor is an agent that hopes to prevent coordination.

While there exists a good deal of prior work in the space of coordination on networks, it is not clear what the mechanisms for the emergence of consensus or lack thereof on networks in the presence of adversaries.

Prior work:

Related work in this field of study can be classified by the degree to which competition plays a role in the completion of the task. Many works have been focused in setting a social dilemma setting, wherein individual actors will benefit from choosing a selfish action, unless others do as well, in which case everyone is worse off. The most famous among this class of task is the prisoner's dilemma and it is the domain of many studies.

Gallo and Yan study the effect of reputation information and social information on cooperation. They found that people cooperation much more when they have access to the reputation information, but also the network information correlates to the level of cooperation. Both social information and global information affect the distribution of cooperative behavior in the group, one cooperative group and one not cooperative group. Yet, they note here is a gap in the theory because theory suggests that global reputational knowledge is not necessary, but experimental result differs.

Hajaj et al.'s work is among the only studies to consider the presence of agents with a malicious intent in a social learning context. To investigate this, the researchers used a color coordination task where the objective was to reach global consensus on a particular color, even in the presence of adversarial nodes who attempt to prevent consensus. In their experiments, they find adversarial nodes to be quite effective in accomplishing their aim. While this paper addresses a good deal of topics in the area of focus, there are key opportunities for further study. A dynamically forming network might alleviate the structural impact of adversarial nodes. Additionally, the presence of different policy rule governing gameplay (social norms, policing, sanctions etc.) might have an effect on the coordination.

In related work, Shirado and Christakis study a networked color coordination game human participate jointly with autonomous agents. The game is as follows: each subject is allowed to choose one of several different colors with the goal of choosing a different color that each of its neighbors. This is another example of a task where local maxima present a challenge. The authors recruited participants online and assigned them to one of 11 different treatment groups, with varying placement of the autonomous agents on the network and variable degrees of randomness in bots' decision making process. Each participant only only sees their own color, as well as the color of their neighbors. They found that in many cases, some degree of randomness (roughly 10% of actions being random) was helpful in reducing time to complete the task because it helped to avoid being caught in local maxima. Zero noise bots encouraged less randomness in their human counterparts and higher noise (30%) served to destabilize the network. This work provides one example of a context in which acting outside of the direct best interest of the task some of the time counterintuitively provides benefit. Additional work, can be done in exploring the network topology, or whether the networks are dynamic or static.

ALSO TALK ABOUT THE REDDIT STUDY ON HATE SPEECH HERE

Research questions:

To study the problem of coordination on a network, we make use of a variation on the networked version of the color matching game (bricker1995multiplayer). The problem is as follows: n agents exist on a network, each with an internal type which is one of the cooperator or adversarial. K of these agents are adversarial and aim to prevent the remaining n-k nodes from reaching consensus by displaying the same color. Each agent only can see their own color and then colors of their immediate neighbors. In each time step, the actions available to the agents are selecting a color to display, and in certain conditions, this could also be changing connections. The game ends when all of the cooperators are displaying the same color for five consecutive time steps and are all a part of the same connected component, or after 500 time steps.

In order to understand the mechanisms for the emergent behavior of reaching consensus, we first aim to understand the effect of adversary on the ability for the global network to reach consensus.

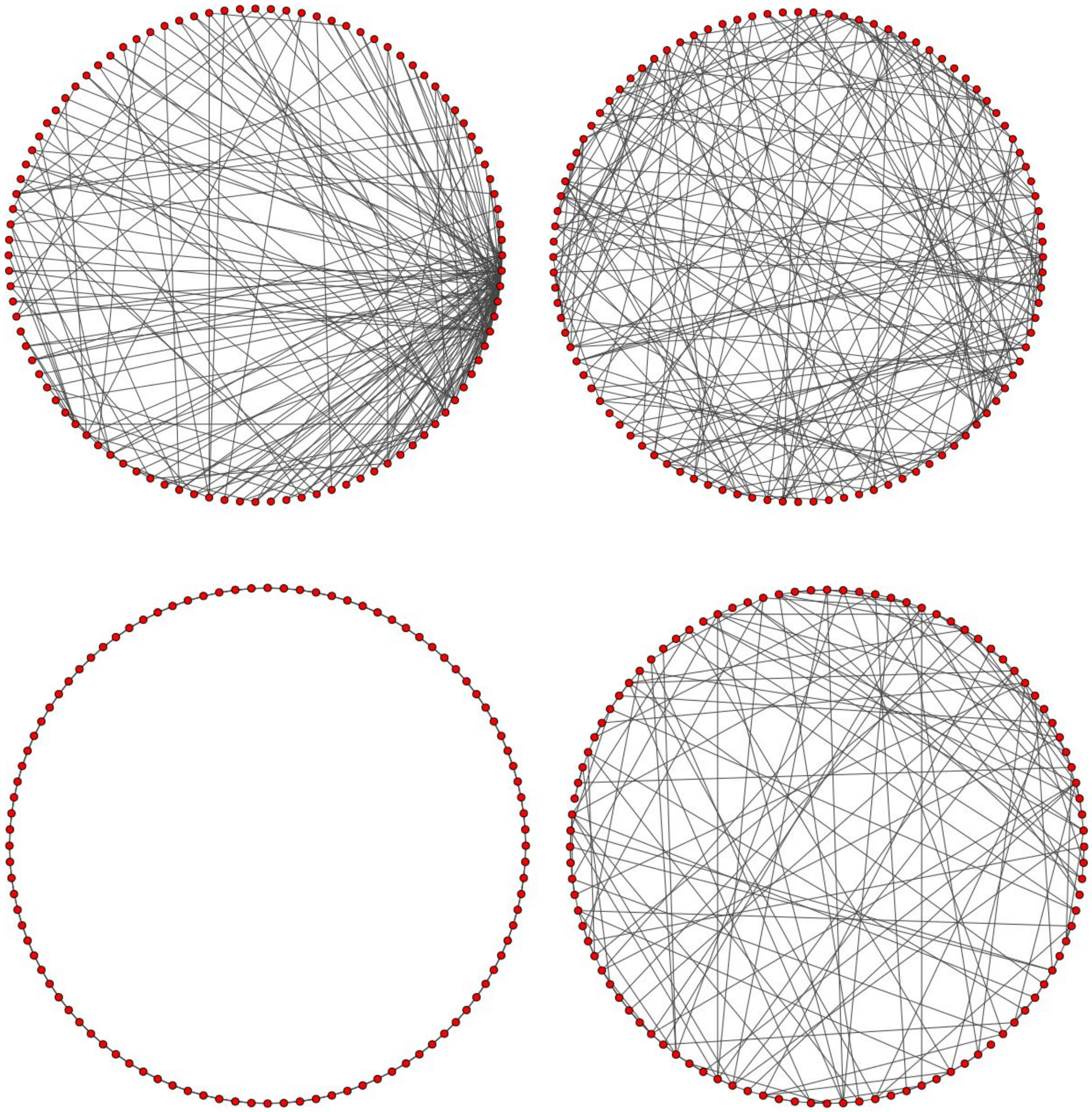
Thus, we have the first broad research question:

RQ 1: What is the effect on the proportion of runs and speed of which of global consensus is reached in the presence of bad actors on a network coordination task?

We break down this into a number of subquestions:

RQ 1.1: What is the effect of network structure on the proportion of runs and speed of which of global consensus is reached in the presence of bad actors on a network coordination task?

We consider four distinct classes of network: Barabasi (barabasi1999emergence), Erodos-Renyi(erdos1959random), Lattice and Watts-Strogratz (watts1998collective). We generate (See Appendix A: Network Generation Code) an example of each network. See Table 1 for more details on the networks.



Barabasi1 (top left), Erdos-Renyi1(top right), Lattice (bottom left), and Watts-Strogatz1

Table: Initial Network Statistics.

Network	Barabasi1	Erdos_Renyi1	Lattice1	Watts_Strogatz1
Number of edges	197	200	200	200
Average Degree (std.)	3.94 (5.21)	4.00 (1.78)	4.00 (0.00)	4.00 (1.45)
Average Path Length	2.71	3.43	12.88	3.50

RQ 1.2: What is the effect of number of adversaries on the proportion of runs and speed of which of global consensus is reached in the presence of bad actors on a network coordination task?

We consider two different conditions: 10 and 50 adversaries. The low end comes from the ___ literature, where 10% (of the 100 total nodes) is typically considered ____, and range to complete parity between what we deem cooperators and adversaries.

RQ 1.3: What is the placement of the adversaries on the proportion of runs and speed of which of global consensus is reached?

We consider two different conditions, adversaries placed at the k vertices of the graph with the highest degree centrality. Ties are broken randomly. We refer to as low and high, respectively. This rises directly from what is referred to in sociology and political science literature as the core-periphery theory (hojman2008core, friedmann1967general).

Similarly, of interest is the examination of how different individual actions lead to emergent behavior on the network as a whole. This leads to the second main objective of this investigation.

RQ 2: What is the effectiveness of different mitigation strategies on local and global performance?

RQ 2.1: What is the impact of network dynamism on both local and global consensus?

We consider five different conditions where cooperator nodes are in control of the changing network structure. First of these is the static condition, which is to say no network dynamism. The next condition considered is randomly selecting another node in the graph from the uniform distribution and forming an undirected edge. The final class of connection strategy is reputation, where nodes use the past behavior of nodes with which they have been previously connected. This is very much inspired from information cascades and the field of game theory in general, where agents with imperfect information aim to uncover some private information of other player based on their past actions (bikhchandani1992theory). In this case, the cooperating agents aim to identify whether a given neighbor is another cooperator node with different information about what the choice of color should be, or whether the node is an adversary. We mechanize this with a single parameter called tolerance, the proportion of previous color mismatches a given agent will tolerate before deciding to no longer connect.

RQ 2.2: What is the effect of cooperator color change strategy on both local and global consensus?

The strategies considered are all variant on a majority vote. The first case is naive majority vote, which is to say simply change color based on the most popular color of your neighbors. The following two strategies are similar, but involve a weighting. We call the first case social-majority-vote, where the colors of nodes with a high degree are more considered more than those of nodes with a small degree. This final is a majority vote based on reputation, where nodes that have past history of matching colors are weighted more heavily.

RQ 2.3: What is the effect of adversary color change strategy on both local and global consensus?

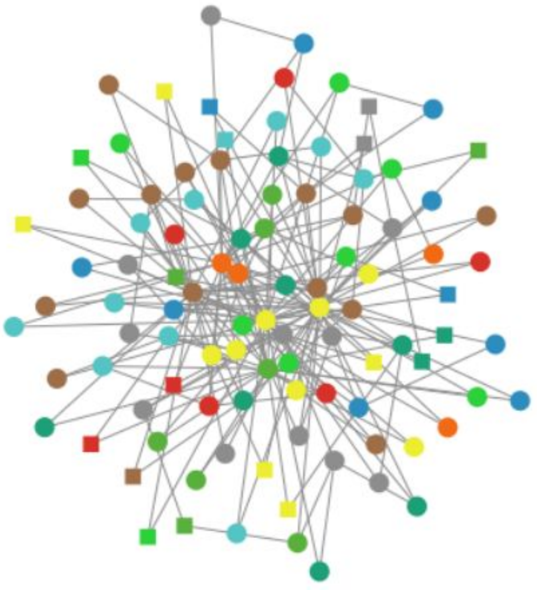
We consider two separate classes of strategy, with a total of three strategies. The first of these is the disengaged strategy, where adversaries simply do nothing after receiving their initial random assignment of color. The second of these is what we call blend in, where adversaries select a random color with either probability 1 or 0.5, and try to act as a cooperator with the remaining probability.

Model:

The model itself is written in Netlogo (wilensky1999netlogo), a program language designed specifically for agent-based modelling. First, an initial network structure is created from a locally saved adjacency matrix, with each node being marked as a cooperator. Certain nodes are then changed to be adversarial according to the particular experimental conditions.

When the game begins, each node decides whether to and how to change their color and ties. (create a table for each, with citations of where they came from). The game ends when all of the cooperator nodes are a part of the same connected component and are also displaying the same color for five consecutive time steps. This constitutes a consensus according to (wikipedias editing guidelines and the Article IX of the Agreement Establishing the World Trade Organization). [Aside: I am currently looking into differences between what I am calling transient and permanent consensus, with the former being the a consensus where the effect is lost in the following time step. I have not completed that analysis yet, so it is not going to be included in this version] At each time step in the model, the proportion of cooperator node that are displaying the same color is reported. Additionally, to a allow for further analysis of the structural properties of the network, as well as the performance of the individual nodes over time, the full network is reported twice every 250 time steps, and at the end of each run (either after converging, or 500 time steps.)

The actual runs of the model are performed on QUEST (cite this), Northwestern University's High Performance Computing Cluster. See Appendix for more information.



An example view of a network. There circular node represent cooperator node, while the squares represent adversaries. Lines between pairs of shapes represent undirected edges between nodes. A float chart of the process will go here, but I have not yet made it.

A summary of all the experimental conditions and parameters of the model can be found in Table _.

Table _.

Experimental Conditions	
Networks	Barabasi, Erdos-Renyi, Lattice, Watts-Strogatz
Number of Adversaries	10 or 50
Placement of Adversaries	Low or High
Cooperator Connection Strategy	Fixed, Random, Reputation(Low), Reputation(High)
Cooperator Color Change Strategy	Naive Majority Vote, Social Majority Vote, Reputation Majority Vote
Adversary Color Change Strategy	Disengaged, Blend(Low), Blend(High)

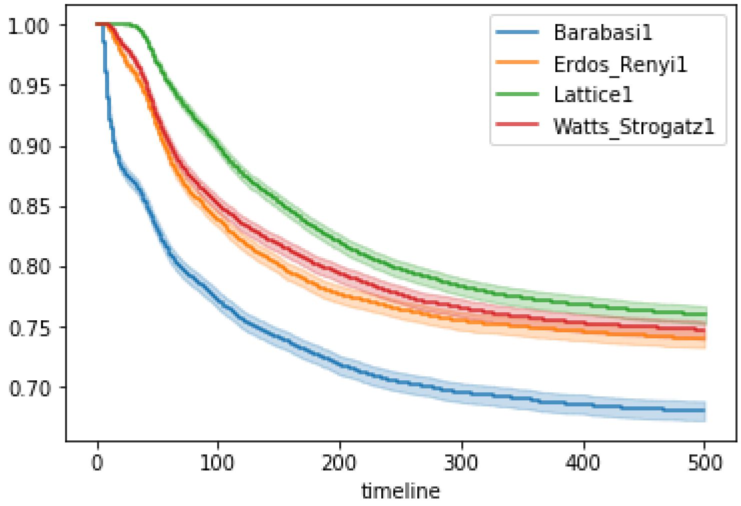
Results and Analysis:

I take inspiration from (shirado2017locally) for the methodological approach to this anaysis. [Aside: My models finished running <18 hours ago, so these next couple sections are still very much a work in progress and a lot of the interpretations are missing right now].

(A summary of everything with some descriptive statistics will go here)

Here is where I do Kaplan-Meier survival analysis, looking at having reached consensus in time step t as dying.

RQ 1.1: What is the effect of network structure on the proportion of runs and speed of which of global consensus is reached in the presence of bad actors on a network coordination task?

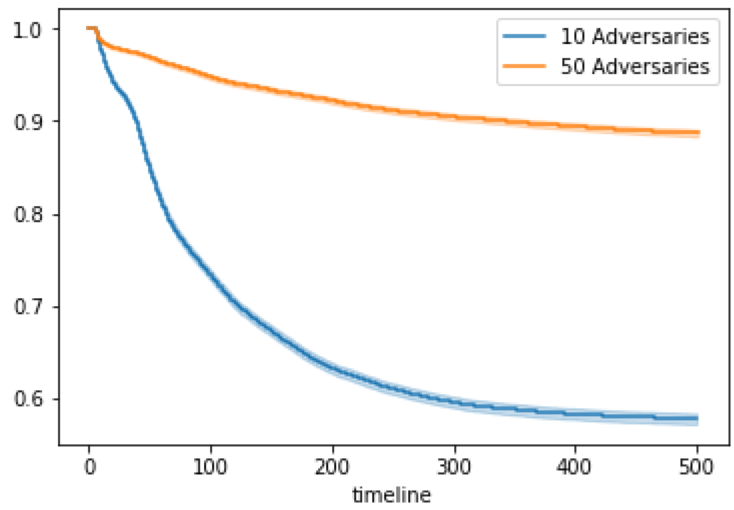


Test Statistics for the Logrank Test for the Differences between Initial Network Structure

	Barabasi	Erdos_Renyi	Lattice	Watts_Strogatz
Barabasi		158.39 **	308.94 **	200.45 **
Erdos_Renyi			29.02 **	2.73 *
Lattice				13.73 **
Watts_Strogatz				

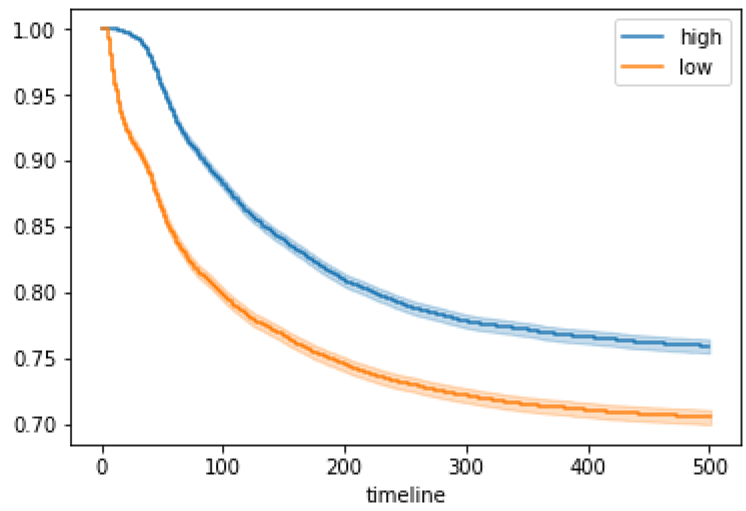
A single asterisk (*) denotes significance at the 0.1 level
Two asterisks (**) denote significance at the 0.01 level

RQ 1.2: What is the effect of number of adversaries on the proportion of runs and speed of which of global consensus is reached in the presence of bad actors on a network coordination task?



Test Statistics for the Logrank Test
test_statistic p
6718.46 <0.005

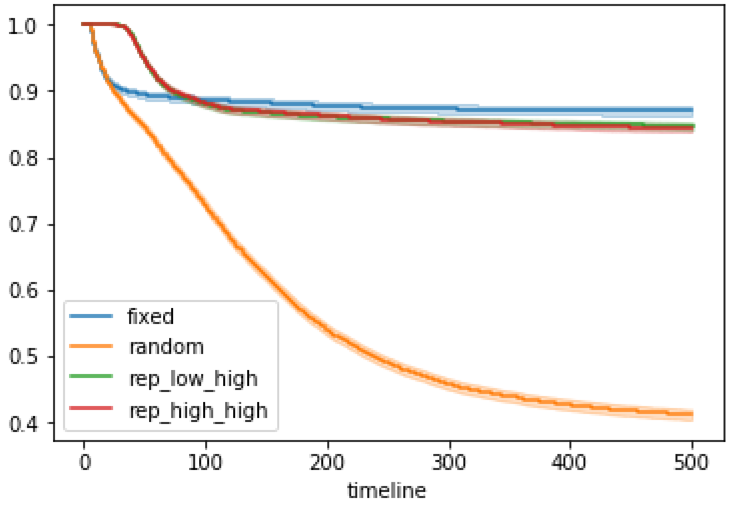
RQ 1.3: What is the placement of the adversaries on the proportion of runs and speed of which of global consensus is reached?



Test Statistics for the Logrank Test
test_statistic p
291.34 <0.005

RQ 2: What is the effectiveness of different mitigation strategies on local and global performance?

RQ 2.1: What is the impact of network dynamism on both local and global consensus?

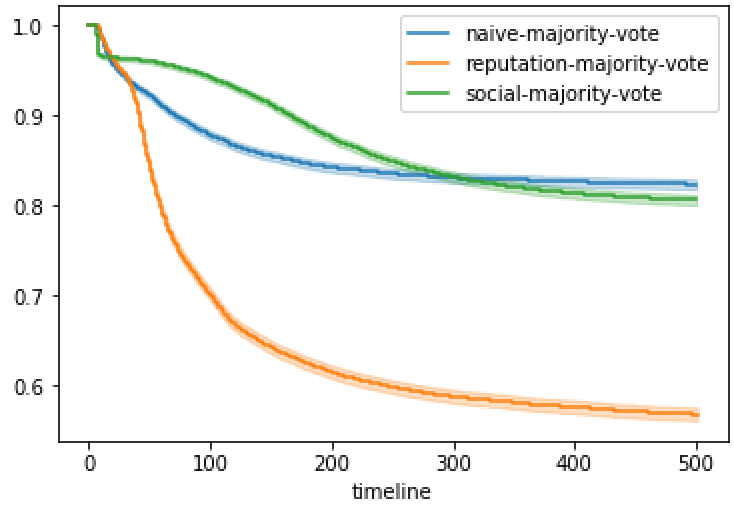


Test Statistics for the Logrank Test

	fixed	random	rep(low)	rep(high)
fixed		4505.11 **	13.08 **	17.02 **
random			5586.65 **	5541.22 **
rep(low)				0.30
rep(high)				

Two asterisks (**) denote significance at the 0.01 level

RQ 2.2: What is the effect of cooperator color change strategy on both local and global consensus?

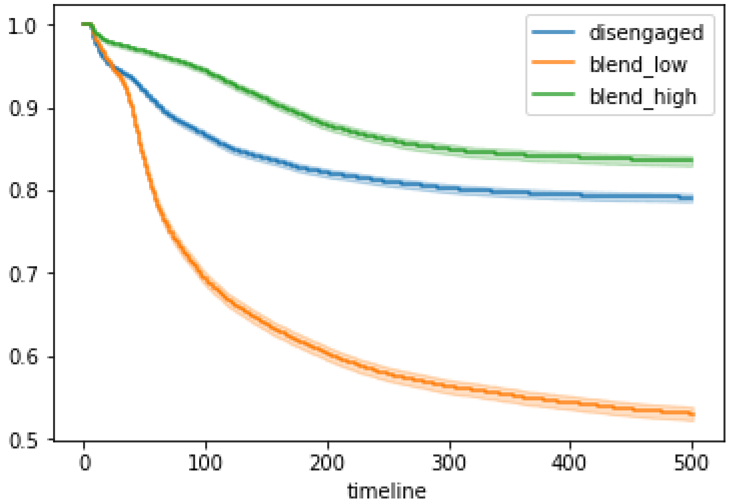


Test Statistics for the Logrank Test for the Differences between Color Change Strategies

	Naive	Reputation	Social
Naive		139.75 **	286.99 **
Reptation			240.13 **
Social			

A single asterisk (*) denotes significance at the 0.1 level
Two asterisks (**) denote significance at the 0.01 level

RQ 2.3: What is the effect of adversary color change strategy on both local and global consensus?



Test Statistics for the Logrank Test

	Disengaged	Blend(low)	Blend(high)
Disengaged		2947.77 **	141.78 **
Blend(low)			3338.43 **
Blend(high)			

Two asterisks (**) denote significance at the 0.01 level

Separately, do a bunch of t tests on the local vs. global agreement measure within a condition to see if some are better or worse for polarization. (include a table)

Quantitatively note differences in final network structure. Look at differences conditional on reaching consensus and not.

Discussion:

Obviously need to more analysis before I can fully flesh this section out, but this is my proposed structure.

Implications for people with people trying to cooperate:

- Overall, coordination without communication is hard, but it's not impossible to do

Implications for adversaries:

- Blocking consensus is not particularly difficult and it is relatively easy to disguise the adversarial behavior as a good faith attempt

Design implications for networked systems:

- Network dynamism is a feasible method of creating a robust distributed system

Conclusion:

How the presence of adversaries do or don't make a big difference and when. Come up with a take for what it means

Future work:

Clearly, this work is limited and there are numerous opportunities for further exploration of the topics discussed. The easiest extension of the model is the inclusion of further strategies. The following are in no way exhaustive, but do present intriguing opportunities to add to the existing body of work:

- Some ties in the network cannot change like (harrell2018strength)
- Ban nodes altogether from being a member of a certain connected component, similar to what is described in the case of certain hate-speech filled communities on reddit, as in (chandrasekharan2017you)
- Limit the frequency that nodes can change color, similar to how communities like reddit fundamentally operate to prevent spam
- Change the objective of the model to converging on particular color, rather than any color. In this case, the color is a stand in for some objective truth (friedkin2017truth)
- Adding the ability for agents to play mixed strategies, or create heterogeneous populations of both adversaries and cooperators.

A similarly useful endeavour would be to run an empirical and qualitative study with the same set up. This would provide an opportunity to calibrate the agent-based model, and would provide an opportunity to examine the decision individuals make in a real world context.

References: [I am eventually going to write this up in Latex, so I apologize for pasting bibtex right now]

```
@article{mason2012collaborative,
  title={Collaborative learning in networks},
  author={Mason, Winter and Watts, Duncan J},
  journal={Proceedings of the National Academy of Sciences},
  volume={109},
  number={3},
  pages={764--769},
  year={2012},
  publisher={National Acad Sciences}
}

@article{gallo2015effects,
  title={The effects of reputational and social knowledge on cooperation},
  author={Gallo, Edoardo and Yan, Chang},
  journal={Proceedings of the National Academy of Sciences},
  volume={112},
  number={12},
  pages={3647--3652},
  year={2015},
  publisher={National Acad Sciences}
}

@article{hajaj2018adversarial,
  title={Adversarial Coordination on Social Networks},
  author={Hajaj, Chen and Yu, Sixie and Joveski, Zlatko and Vorobeychik, Yevgeniy},
  journal={arXiv preprint arXiv:1808.01173},
  year={2018}
}

@article{mao2017resilient,
  title={Resilient cooperators stabilize long-run cooperation in the finitely repeated Prisoner’s
Dilemma},
  author={Mao, Andrew and Dworkin, Lili and Suri, Siddharth and Watts, Duncan J},
  journal={Nature communications},
  volume={8},
  pages={13800},
  year={2017},
  publisher={Nature Publishing Group}
}

@article{shirado2017locally,
  title={Locally noisy autonomous agents improve global human coordination in network experiments},
  author={Shirado, Hirokazu and Christakis, Nicholas A},
  journal={Nature},
  volume={545},
  number={7654},
  pages={370},
  year={2017},
  publisher={Nature Publishing Group}
}

@article{kearns2006experimental,
  title={An experimental study of the coloring problem on human subject networks},
  author={Kearns, Michael and Suri, Siddharth and Montfort, Nick},
  journal={Science},
  volume={313},
  number={5788},
  pages={824--827},
  year={2006},
  publisher={American Association for the Advancement of Science}
}

@article{friedkin2017truth,
  title={How truth wins in opinion dynamics along issue sequences},
  author={Friedkin, Noah E and Bullo, Francesco},
  journal={Proceedings of the National Academy of Sciences},
  volume={114},
  number={43},
  pages={11380--11385},
  year={2017},
  publisher={National Acad Sciences}
}

@article{chandrasekharan2017you,
  title={You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech},
  author={Chandrasekharan, Eshwar and Pavalanathan, Umashanthi and Srinivasan, Anirudh and Glynn, Adam and Eisenstein, Jacob and Gilbert, Eric},
  journal={Proceedings of the ACM on Human-Computer Interaction},
  volume={1},
  number={CSCW},
  pages={31},
  year={2017},
  publisher={ACM}
}

@article{harrell2018strength,
  title={The strength of dynamic ties: The ability to alter some ties promotes cooperation in those
that cannot be altered},
  author={Harrell, Ashley and Melamed, David and Simpson, Brent},
  journal={Science advances},
  volume={4},
  number={12},
  pages={eaau9109},
  year={2018},
  publisher={American Association for the Advancement of Science}
}

@article{barbera2015tweeting,
  title={Tweeting from left to right: Is online political communication more than an echo
chamber?},
  author={Barber{\'a}, Pablo and Jost, John T and Nagler, Jonathan and Tucker, Joshua A and
Bonneau, Richard},
  journal={Psychological science},
  volume={26},
  number={10},
  pages={1531--1542},
  year={2015},
  publisher={Sage Publications Sage CA: Los Angeles, CA}
}
```

```
@article{harrell2018strength,
  title={The strength of dynamic ties: The ability to alter some ties promotes cooperation in those
that cannot be altered},
  author={Harrell, Ashley and Melamed, David and Simpson, Brent},
  journal={Science advances},
  volume={4},
  number={12},
  pages={eaau9109},
  year={2018},
  publisher={American Association for the Advancement of Science}
}
```

```
@article{friedkin2017truth,
  title={How truth wins in opinion dynamics along issue sequences},
  author={Friedkin, Noah E and Bullo, Francesco},
  journal={Proceedings of the National Academy of Sciences},
  volume={114},
  number={43},
  pages={11380--11385},
  year={2017},
  publisher={National Acad Sciences}
}

@article{hojman2008core,
  title={Core and periphery in networks},
  author={Hojman, Daniel A and Szeidl, Adam},
  journal={Journal of Economic Theory},
  volume={139},
  number={1},
  pages={295--309},
  year={2008},
  publisher={Elsevier}
}

@article{friedmann1967general,
  title={A general theory of polarized development},
  author={Friedmann, John},
  year={1967},
  publisher={ILPES}
}

@article{bricker1995multiplayer,
  title={Multiplayer Activities That Develop Mathematical Coordination.},
  author={Bricker, Lauren J and Tanimoto, Steven L and Rothenberg, Alex I and Hutama, Danny C and Wong, Tina H},
  year={1995},
  publisher={ERIC}
}

@misc{wilensky1999netlogo,
author = {Wilensky, U},
title = {Netlogo},
howpublished = {\url{ http://ccl.northwestern.edu/netlogo/}},
lab = {Center for Connected Learning and Computer-Based Modeling}
}
```

```
@article{barabasi1999emergence,
  title={Emergence of scaling in random networks},
  author={Barab{\'a}si, Albert-L{\'a}szl{\'o} and Albert, R{\'e}ka},
  journal={science},
  volume={286},
  number={5439},
  pages={509--512},
  year={1999},
  publisher={American Association for the Advancement of Science}
}

@article{erdos1959random,
  title={On random graphs, I},
  author={Erd{\"o}s, Paul and R{\'e}nyi, Alfr{\'e}d},
  journal={Publicationes Mathematicae (Debrecen)},
  volume={6},
  pages={290--297},
  year={1959}
}

@article{watts1998collective,
  title={Collective dynamics of ‘small-world’networks},
  author={Watts, Duncan J and Strogatz, Steven H},
  journal={nature},
  volume={393},
  number={6684},
  pages={440},
  year={1998},
  publisher={Nature Publishing Group}
}

@article{bikhchandani1992theory,
  title={A theory of fads, fashion, custom, and cultural change as informational cascades},
  author={Bikhchandani, Sushil and Hirshleifer, David and Welch, Ivo},
  journal={Journal of political Economy},
  volume={100},
  number={5},
  pages={992--1026},
  year={1992},
  publisher={The University of Chicago Press}
}
```

Appendices: These are all in various states of completion

Appendix A: Initial Network Formation

Appendix B: Mathematical Formalization of the Strategies

Appendix C: Information on QUEST and running the model

Appendix D: Full Data

Appendix E: Explanation and Math Formalization of the Statistical Tests