Adversarial Coordination on Social Networks

Chen Hajaj
Industrial Engineering & Management, Ariel University
Ariel, Israel
chenha@ariel.ac.il

Sixie Yu
Computer Science & Engineering, Washington University
St. Louis, MO
sixie.yu@@wustl.edu

Zlatko Joveski
Computer Science, Vanderbilt University
Nashville, TN
zlatko.joveski@vanderbilt.edu

Yifan Guo
Capital One
Arlington, VA
yifan.guo@vanderbilt.edu

Yevgeniy Vorobeychik
Computer Science & Engineering, Washington University
St. Louis, MO
yvorobeychik@wustl.edu

ABSTRACT

Extensive literature exists studying decentralized coordination and consensus, with considerable attention devoted to ensuring robustness to faults and attacks. However, most of the latter literature assumes that non-malicious agents follow simple stylized rules. In reality, decentralized protocols often involve humans, and understanding how people coordinate in adversarial settings is an open problem. We initiate a study of this problem, starting with a human subjects investigation of human coordination on networks in the presence of adversarial agents, and subsequently using the resulting data to bootstrap the development of a credible agent-based model of adversarial decentralized coordination. In human subjects experiments, we observe that while adversarial nodes can successfully prevent consensus, the ability to communicate can significantly improve robustness, with the impact particularly significant in scale-free networks. On the other hand, and contrary to typical stylized models of behavior, we show that the existence of trusted nodes has limited utility. Next, we use the data collected in human subject experiments to develop a data-driven agent-based model of adversarial coordination. We show that this model successfully reproduces observed behavior in experiments, is robust to small errors in individual agent models, and illustrate its utility by using it to explore the impact of optimizing network location of trusted and adversarial nodes.

KEYWORDS

Decentralized coordination; social networks; robust consensus

ACM Reference Format:
Chen Hajaj, Sixie Yu, Zlatko Joveski, Yifan Guo, and Yevgeniy Vorobeychik. 2019. Adversarial Coordination on Social Networks. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Coordination is one of the fundamental problems faced by teams, organizations, and societies. Such coordination problems are often decentralized and involve limited local information and interaction, with locality naturally captured by a network structure. A prominent example for the special case of consensus is blockchain, which enables verifiable decentralized transactions [30].

Considerable prior research has been devoted to understanding and modeling human behavior in networked coordination settings, such as networked consensus [19–21, 40], coloring [19, 28], bargaining [7], and social dilemma games [15, 26], among others. However, decentralized coordination problems often take place in adversarial predicaments. For example, organizations attempting to coordinate on a strategy may also compete with other organizations (legal and illegal), and coordination in combat mission planning and execution inherently faces adversarial entities in the form of enemy combatants. Moreover, adversaries often attempt to exert their influence covertly, such as by bribing insiders, taking control of network nodes through cyber attacks, and spreading malicious influence tacitly through social networks, for example, by means of fake news [4]. Consequently, an important consideration in decentralized coordination is resilience to adversarial tampering with the process. While much prior research has been devoted to the study of robust coordination protocols, these rely on simple stylized models of individual behavior [2, 6, 24, 25]. However, many settings feature humans in the loop who play an important role in reaching consensus. Surprisingly, the question of human behavior in adversarial coordination settings has received little prior attention.

We investigate the problem of decentralized consensus on networks in the presence of adversarial nodes, first using human subject experiments with 556 participants, and subsequently through the *data-driven agent-based modeling (DDABM)* methodology [44]. Our experiments focus on two design factors: allowing neighboring nodes to communicate, and embedding a small set of trusted nodes in the network. While communication has been a major subject of inquiry in prior research [8, 10, 11, 29], recent research suggests that communicating solely among network neighbors has limited value in facilitating consensus [40]. On the other hand, much prior research, using stylized models of individual behavior, has argued that the presence of trusted nodes can significantly facilitate decentralized coordination [1, 2, 39]. Our results run counter to both of these observations. First, we demonstrate that communication helps a great deal, especially as we increase the number of adversarial nodes. Second, we show that the presence of trusted nodes does not, in the aggregate, help, reinforcing the need to develop better models of individual and collective behavior in such settings.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

Next, we develop a data-driven agent-based model of adversarial decentralized consensus on networks, following the DDABM methodology [44]. In DDABM, individual agent models are derived from data, and are then instantiated in an agent-based framework via features that capture behavioral interdependencies among network neighbors. For us, these serve three purposes. First, they provide further insight into individual behavior. For example, we observe that adversarial nodes clearly engage in deliberate attempts to manipulate outcomes. Second, the resulting agent-based model effectively captures our experimental observations *at the macro level*, and is quite robust to small errors in the individual agent models. Third, we demonstrate the usefulness of the derived computational platform as a means for further simulation-based investigation of the adversarial consensus problem by studying the impact of optimized network location of both trusted and adversarial nodes. We find that optimizing location is particularly beneficial for adversarial nodes, even when the placement of trusted players is similarly optimized *before* we choose where to place adversaries (i.e., in a Stackelberg fashion). Consequently, and counter to prior observations with stylized behavioral models, trusted nodes appear to have only a limited value in facilitating decentralized human consensus in adversarial settings.

Related Work. Our study of networked coordination follows our study follows the recent increasing interest in the adversarial side of artificial intelligence [18, 37, 38, 41], and a number of prior efforts that investigate a variety of decentralized coordination problems on networks using human subjects methodology [7, 19–21, 23, 28, 40]. The impact of communication on human coordination and cooperation has extensive, parallel literature, using both human subjects [31, 35, 36] and theoretical methods [10, 11, 13, 14, 29]. However, in most of this literature, communication is grafted on as a distinct pre-play stage; moreover, much of this literature studies simple, two-player games. A recent exception, is the work of Vorobeychik et al. [40], combining both threads, but investigating only non-adversarial settings. Regarding human behavior, Coviello et al. [9] took a more algorithmic approach to look at the matching behavior of a human in social networks. While using the same experimental design as ours, the authors focus on the case where players have to divide into pairs, when the structure of the network is unknown, with a collective goal of maximizing the number of teams. Still, similar to our work, the authors use the experimental data to produce an algorithmic model and analyze its properties by simulations.

Robust coordination has been analyzed by several efforts, but theoretically and in simulations, using highly stylized behavior models [24, 25, 43]. Specifically, [24, 25] focus on design of a consensus protocol that is resilient to worst-case security breaches assuming the compromised nodes have full knowledge of the network and the intentions of the other nodes. In this work, we provide a behavioral analysis using extensive human subject experiments using a well-known crowdsourcing platform. Furthermore, we relax the assumption of full knowledge and the knowledge about the intentions of different nodes in the network. Several prior efforts study the importance of trusted nodes in such settings [1, 2, 39]. Our results suggest that stylized models used in these efforts may be limited in evaluating the efficacy of trusted nodes.

Finally, data-driven or empirical agent-based modeling has been proposed as a means of performing simulations that reliably reflect actual behavior data [42, 44]. Our simulation-based analysis follows in the spirit of these efforts.

2 EXPERIMENTAL METHODOLOGY

2.1 General Setup

We designed a human subject experiment to study *adversarial coordination on social networks*. Specifically, the experiment builds on networked consensus games [19, 22], in which a collection of players (human subjects) act as nodes on an exogenously specified graph, choosing between two colors: RED and GREEN. These games proceed for 60 seconds, with individuals able to make changes to their color choice in essentially real time. Each player has an egocentric view of the game illustrated in Figure 1, where their node is displayed at the center, and their network neighbors are shown surrounding the "Me" node, along with their color choices, as well network connections among them. Any node is displayed as white prior to actively choosing a color. The display screen also shows time remaining in the game. Each player receives a base payment for each game played ($0.15), as well as a bonus of $0.20 if a global consensus on either color is reached. The game ends as soon as consensus is reached.

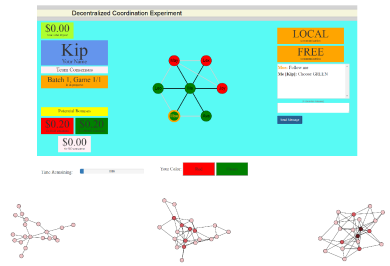


Figure 1: Top: an example graphical interface from the point of view of an experimental subject, who is represented by a node in the network. Bottom: example instances of networks, where darker colors indicate higher node degrees.

The game description so far replicates features from all prior experiments in networked consensus. A new feature, introduced by [40], allows network neighbors to communicate through an instant message-style interface, shown on the right in Figure 1. To facilitate such communication, when allowed, each player is assigned a 3-letter name at the beginning of each game, and this name serves as their unique identifier in communicating with others. Specifically, when a player sends a message through this interface, all their immediate network neighbors receive the message (this mode of communication was termed *local* communication by [40]).

We made one change to this general setup, which turns out to be quite consequential. In all prior experiments, the interface featured a *progress bar*, which shows how close the overall state is to global consensus (measured by the number of nodes disagreeing with majority color). In our setting, however, such a progress bar communicates too much information, particularly when adversaries

are present, and we consequently removed it (particularly since it doesn't have a clear motivation and was just a design artifact of prior experiments). As we observe below, removing the progress bar increases the importance of communication, relative to findings reported by Vorobeychik et al. [40].

2.2 Design of Adversarial Consensus Games

Starting with the basic experimental framework described above, we augment the experimental platform with several features in order to study how adversarial nodes impact the ability of the rest (i.e., the non-adversarial sub-network) to reach global consensus. For this purpose, we divide players into two teams: a *consensus* team and a *no-consensus* team (in our parlance, these are *adversaries*). The goal of the *consensus* team is to reach global consensus *among members of this team only*, captured by the bonus payment structure described above. The goal of the *no-consensus* team is to prevent consensus among members of the *consensus* team, which we incentivize by paying a $0.40 bonus to members of this team if and only if consensus fails. At the beginning of the game, each player is assigned to one of these teams, and this assignment is indicated in their view of the game (see left part of Figure 1).

We fixed the number of *consensus* players in each game to 20, to control the baseline difficulty of the task (the underlying consensus problem on networks becomes more difficult as the network size grows, other things being equal). In addition, we introduced in each game a *no-consensus* players, where $a \in \{0, 2, 5\}$. The value of a was not disclosed to the players at the beginning of a game; although an omniscient observer can infer it from the size of the network (which is $20 + a$), no player could, in fact, do this, since players could only observe their direct neighbors, and we limited the maximum degree to 15 to facilitate effective visualization.

A crucial part of our design was the invisibility of adversaries (no-consensus nodes) to others, including other adversaries, and vice versa. On the other hand, it is often possible to have a small number of known *reliable* or *trusted* nodes on the network, for example, nodes which are particularly difficult to compromise due to a high amount of investment in their security, and conventional wisdom is that such nodes can greatly facilitate consensus [2]. To allow for this, we vary the number of *visible* members of the *consensus* team (henceforth, *visible nodes*), $v \in \{0, 1, 2, 5\}$.[1] However, these nodes are visible only to their immediate network neighbors, highlighted by an orange circle around the corresponding nodes, as in Figure 1 for the player with an assigned name "Moe".

2.3 Network Topologies

For each game, we exogenously specify a network topology, stochastically generated from one of three random graph models: two variations of Erdos-Renyi (ER) graphs [12], and a Barabasi-Albert (BA, also known as preferential attachment) model [5]. The two variations of the ER model differ in network density: one we term ER-dense, and the other ER-sparse. The 20-node version of the ER-dense model has average degree 5.1, while the ER-sparse networks have an average degree of 2.6. BA networks have an average degree of 5.1 (same as ER-dense). Average degrees increase slightly when we add adversarial nodes. Figure 1 shows example networks for each of the three network generative models.

[1]Recall that there are always 20 nodes in the consensus team. Thus, when 2 members are visible, there are 18 regular nodes in this team.

2.4 Recruiting and Scheduling

We recruited subjects for the experiment using the Amazon Mechanical Turk (AMT) platform [27, 32], now in common use for economic experiments with human subjects [16, 17, 27, 33, 34]. Recruited subjects were directed to read detailed experiment instructions and consent to participate in the experiment (which was collected online). Once we had a large enough pool of consented subjects, we scheduled experiment sessions. For each experiment session, we recruited 30-35 subjects, to ensure that we have a sufficient number even when there are no-shows. Upon arrival, subjects were placed in a waiting room, and if there were more subjects than nodes in a graph, they were randomly rotated each game. Each session began with a series of 5 practice games, followed by 50-65 actual games in which we systematically varied 4 experimental variables:

(1) Number of adversaries (*no-consensus* players): $a \in \{0, 2, 5\}$.
(2) number of visible nodes (within the *consensus* team): $v \in \{0, 1, 2, 5\}$.
(3) network topology: ER-dense, ER-sparse, and BA.
(4) communication: allowed or not allowed.

The full study protocol was approved by Vanderbilt University IRB. We recruited a total of 556 participants who jointly played 1080 games.

3 EXPERIMENTAL RESULTS

We now analyze the results of the experiments. Throughout, we focus on consensus rate, or proportion of games reaching global consensus on a single color among the *consensus* players, as a measure of coordination success.[2]

3.1 The Impact of Adversarial Players on Consensus Rate

One would naturally expect that having adversarial players participate in the game would have a deleterious impact on consensus rate. This intuition is readily confirmed in Figure 2 (left), with all

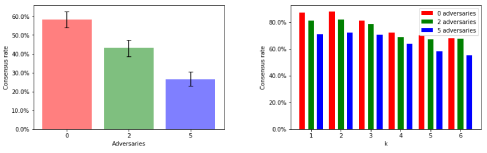


Figure 2: Impact of adversaries on the consensus rate. Left: overall consensus rate, as function of the number of adversaries. Right: For each network distance, proportion of pairs of nodes with this distance between them who agree on a color at the end of the game.

differences statistically significant ($p < 0.01$). However, this observation obscures a crucial distinction between two kinds of impact adversaries can have in our setting:

[2]We note that a t-test was used to extract the statistical significance of our results.

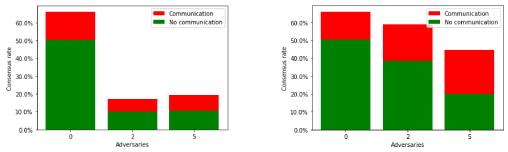
(1) *Structural impact*: the adversarial nodes change network structure—in the extreme case, disconnecting the network among the *consensus* team members, and
(2) *Behavioral impact*: behavior of adversarial nodes impacts the ability of the good nodes to reach consensus.

There is a clear structural impact: 16% of games with 2 adversaries, and 34% of games with 5 adversaries become disconnected if we were to remove adversarial nodes. In the cases in which adversarial nodes disconnect the graph [3], as depicted in Figure 2, consensus rate drops to 14-15%, roughly what one would expect by random chance (if we only have two connected components, and use the consensus rate of 58% which obtains with no adversaries for each component, the expected consensus rate is 17%). Of course, it is worth remembering that the network is not, in fact, disconnected, and adversarial nodes need to deliberately prevent the information about network state from spreading through them. Indeed, not only do adversaries do so, the resulting consensus rates are slightly below expected, suggesting that adversarial behavior itself has an additional deleterious impact on the ability of nodes to coordinate.

To isolate the behavioral impact, in Figure 2 (right) we plot the proportion of times a pair which is k network hops apart agrees on a color at the end of the game, as a function of network distance k (we only include k with at least 100 instances), where network distance is defined as the number of nodes between a pair. Here, we can still see a systematic decrease in coordination success, as a function of the number of adversaries, no matter how far apart nodes are. For example, even network neighbors (i.e., $k = 1$) are finding it increasingly more difficult to agree on a color, on average, as we increase the number of adversaries.

3.2 Communication Improves Resilience

Next, we consider the impact that allowing players to communicate with their network neighbors has on their ability to coordinate successfully. Figure 3 shows that communication makes a clear impact

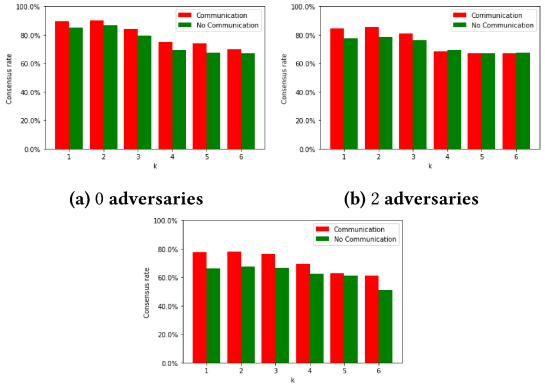


(a) Disconnected networks **(b) Connected networks**

Figure 3: The impact of communication on consensus rate.

(pooling broken and unbroken networks, all results are significant with $p < 0.01$). In the aggregate, the value of communication increases with the number of adversaries: when no adversaries are present, communication increases consensus rate by 23.5%, with 2 adversaries improvement rises to 35.1%, and with 5 adversaries games that feature communication are 54.5% more likely to reach consensus than those that don't. Moreover, Figure 3 breaks these results into two plots: one when networks are disconnected if we were to remove adversarial nodes (3a), and one for the remaining connected networks (3b). One would have expected that with disconnected networks consensus occurs largely by chance, and consequently, communication should have no impact. We can observe that this is not so: even when networks are disconnected by adversaries, communication increases consensus rate, nearly doubling it when there are 5 adversaries. To understand this result, observe that with no communication, consensus rates in disconnected networks are well below what it should be by *random chance*, whereas communication raises them to approximate parity with random chance. In other words, in this setting communication successfully parries the *behavioral* impact of adversaries.

It is noteworthy that communication helps even when there are no adversaries, in contrast with prior results [40]. The key distinction in our setting is the absence of the progress bar: now that this source of global information is missing, communication becomes considerably more informative.



(a) 0 adversaries **(b) 2 adversaries**

(c) 5 adversaries

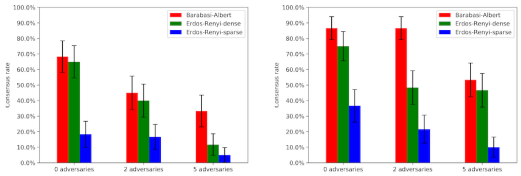
Figure 4: The impact of communication on pairs of nodes agreeing in color choice, by node distance.

Figure 4 unpacks the analysis of the impact of communication further by isolating, again, the behavioral impact of the adversaries, and the result is generally consistent, with communication increasing the likelihood of a given pair of nodes agrees on a color at the end of the game, particularly when they are relatively close to each other in the network.

3.3 The Impact of Network Structure

Next, we consider what impact the network structure has on the ability of players to reach consensus with and without adversaries aiming to sabotage coordination. Figure 5 shows the results, broken up by network (BA, ER-dense, and ER-sparse), number of adversaries, and whether or not communication was allowed. Perhaps the most dramatic impact that communication has is on BA networks: when communication is enabled, 2 adversaries are unable to significantly impact consensus rate, in contrast with games with no communication, where consensus rates of BA networks drop by over 30%. This suggests that with few adversarial nodes, the ability to communicate endows scale-free networks with resilience *even in the face of behavioral manipulation* by adversaries (which we observe to have a significant overall effect otherwise). This finding complements the already well-known resilience of BA networks to random node removal [3].

[3] A graph is disconnected if it is composed of more than a single connected component after removing the adversarial nodes.



(a) No communication. (b) Communication allowed.

Figure 5: The effect of adversary players and network type on the consensus rate.

3.4 The Value of "Trusted" Nodes

Lastly, we look at the value of "trusted" or visible nodes, that is, nodes whose intention of achieving coordination is visible. Prior

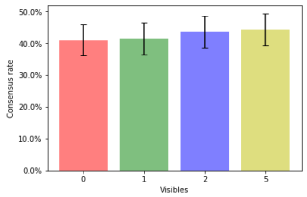


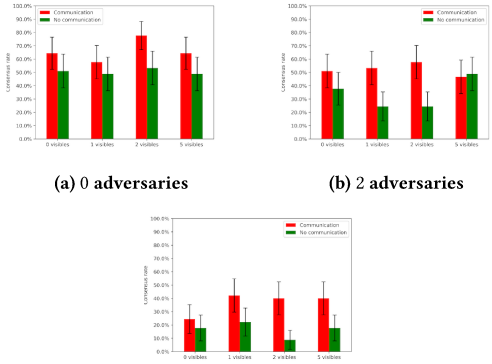
Figure 6: The effect of visible players on the consensus rate.

research using stylized models of node behavior demonstrated that the presence of trusted nodes in a network can significantly improve resilience to attacks [1, 2, 39]. It is thus natural to hypothesize that nodes which are visibly on the *consensus* team (we can view these as trusted nodes, in the sense that they are known not to be adversarial) would significantly facilitate consensus. Remarkably, Figure 6 shows that this is not the case: as we increase the number of visible nodes, the impact on consensus rates is almost undetectable. The reason for the difference is that typical models assume that trusted nodes cannot be attacked. In our case, trusted nodes (as any other node) have no information about who the adversaries are, and, consequently, can also be influenced by the attackers, albeit indirectly.

To understand the impact of visible (trusted) nodes in greater depth, we unpack the results in Figure 7 by the number of visible nodes, the number of adversaries, and whether or not communication is allowed. With 0 or 2 adversaries, it is difficult to see any systematic improvement in performance as we increase the number of trusted nodes. However, with 5 adversaries and communication, having visible nodes constitutes a clear improvement over having none ($p < 0.05$). Thus, merely having trusted nodes is of dubious value, but allowing players (as well as the trusted node) to communicate can improve resilience when there are many adversarial nodes.

4 DATA-DRIVEN AGENT-BASED MODELING AND ANALYSIS

Our observations of collective behavior in adversarial consensus games provides a starting point for the next step: the development of a data-driven agent-based model (DDABM) of this scenario. The



(a) 0 adversaries (b) 2 adversaries

(c) 5 adversaries

Figure 7: The effect of visible and adversarial players given the type of communication on the consensus rate.

DDABM methodology builds agent-based models from data ground up: first, data of individual *human* behavior is used to learn computational models of this behavior, and second, such models are tied together in an agent-based simulation through variables that take as input observed behavior by other agents (in our case, network neighbors and visible nodes) [44]. Crucially, model validation must be performed at both the individual and aggregate levels.

4.1 Modeling and Analysis of Individual Behavior

We start by using the data generated in our experiments to develop computational individual agent models that will give rise to a credible agent-based simulation model with more predictive power than the conventional stylized models. An additional benefit of these models is that they will provide qualitative insight into human behavior in adversarial networked consensus. While we found communication as an important factor in our analysis of the experiments, it is not clear how to model it in simulation, and we therefore focus on the setting with no communication and defer the issue of modeling communication to future work.

Given that the players in our game only choose between two colors, the modeling task before us may seem simple at first glance. This simplicity, however, is quite misleading. In particular, there are several complications in modeling human behavior in our settings. The first is the fact that individuals may have three distinct roles:

(1) Adversarial node: a member of the *no-consensus* team, whose goal is to prevent consensus among the "good" nodes (i.e., nodes on the *consensus* team),
(2) Visible ("trusted") node: a member of the *consensus* team who is visibly a member of this team (that is, all neighbors can see that this node is on the *consensus* team), and
(3) Regular node: all other members of the *consensus* team.

It is intuitive that adversarial nodes behave differently from others. For example, adversarial nodes change color more often than others: 2.9 times per game, in comparison with visible consensus team players, who make only 2.1 changes in a game, and non-visible nodes, who change their color only twice a game, on average. Below, we observe that visible nodes also behave differently from regular

nodes. The second challenge is that nodes in *any* of these roles may behave differently depending on whether they see visible nodes among their neighbors. The third is the fundamental challenge of how we should model real-time color choices by the players.

We address the third challenge by discretize time into 1 second intervals, so that there are (up to) 60 decision points in any game (as a game lasts 60 seconds).

To address the first two challenges, we created distinct behavioral models for the three roles, and distinct models for the situations when they have a visible node as a neighbor, and when they don't (thus, 6 individual agent models altogether).

Each of these cases raises an additional complication: agents make two kinds of decisions during the span of a game: first, as they start as "white" (non-commited), they must choose an initial color, and subsequently, they choose whether to switch their color. Consequently, we split the decision model into two parts: 1) choosing the initial color, and 2) switch the color. The rationale is that the initial decision is a deliberate choice of a particular color, and includes both the timing of changing from the initial default "white" color to either red or green, as well as the particular choice between these two. In contrast, once a color is chosen, players exhibit a considerable amount of inertia: they change color less frequently than once every 20 seconds on average. Thus, modeling the decision to switch (or, effectively, the *timing* of a color switch) naturally captures such inertia, and also cleanly captures the inherent symmetry of their decision at this point, since players do not have a preference for one color over the other beyond reaching consensus.

Finally, the initial decision was itself split into two models: the first modeling the timing of the initial color choice, and the second modeling which color is actually chosen. Consequently, altogether we learned 18 different behavior models, or 3 models for each of the 6 roles and neighborhood assignments. Next, we describe these 3 models (which are qualitatively the same for each of the role x neighborhood predicaments): *timing of initial color choice*, *choosing the initial color*, and *timing of color change*. We briefly note that all models below are highly effective: either they exhibit high accuracy ($90-95\%$), or large likelihood improvement over a frequency-based baseline (50%-100% improvement).

Timing of Initial Color Choice. Our first set of models predicts the timing of the initial choice of color, or, more precisely, the probability that the initial color is chosen in a discrete time unit. For these models, the features are: D_{inv}, the absolute difference between the fraction of a player's non-visible neighbors that picked *red* and the fraction that picked *green*; D_{vis}, the absolute difference between the fraction of a player's visible neighbors that picked *red* and the fraction of those who picked *green* (if the player has visible neighbors); N_{vis}, the number of a player's neighbors that are visible, and N_{inv}, the number of a player's neighbors whose are non-visible (note that $N_{vis} + N_{inv}$ is the total number of neighbors the player has). The decision model is represented by a logistic regression with these features, the parameters (coefficients) of which we learned from experimental data. We added l_1 (sparse) regularization to control for overfitting, with regularization parameter tuned using cross-validation. In all models, VN is a boolean feature indicating if a node has a visible neighbor. All feature were normalized.

Table 1: Color-picking model, P(pick a color).

Type	VN	Intercept	D_{inv}	D_{vis}	N_{inv}	N_{vis}
Reg	No	−1.952	1.29			
	Yes	−2.21	0.548	0.933	0.002	0.016
Vis	No	−2.045	1.742		0.04	
	Yes	−1.734	0.579	0.84	-0.061	0.048
Adv	No	−2.284	1.25		0.011	
	Yes	−2.744	0.802	0.662	0.025	0.155

The learned model coefficients for both the model with and without visible neighbors are given in Table 1. The results offer several interesting insights. First, we can see that disagreement among neighbors stimulates a player to make an initial color choice earlier. This is somewhat surprising, as we may expect players to wait until their neighbors had come to a near-consensus before making an initial move. Second, disagreement among visible nodes has a more significant, positive impact on the likelihood of choosing a color at a particular time point. Third, the behavior of adversarial nodes is broadly consistent with the first observation, but not with the second: such players appear to be more stimulated by disagreement among non-visible than among visible (trusted) neighbors.

Choosing the Initial Color. Conditional on deciding to choose the initial color in a particular discrete time unit (per our previous models), the next decision we model is which of the two colors the player chooses. We again use l_1-regularized logistic regression, where we predict the probability that a player chooses "red" as their initial color (conditional on choosing *some* initial color). As before, we use cross-validation to tune the regularization coefficient. For these models, the features are: G^{inv}_{local}, the fraction of a player's non-visible neighbors choosing *green*; G^{vis}_{local}, the fraction of a player's visible neighbors choosing *green*; R^{inv}_{local}, the fraction of a player's non-visible neighbors choosing *red*; and R^{vis}_{local}, the fraction of a player's visible neighbors choosing *red*. Note that $G^{inv}_{local} + R^{inv}_{local}$ and $G^{vis}_{local} + R^{vis}_{local}$ are not necessarily 1, since some of the neighbors may not have yet chosen a color. As before, all of the features were normalized.

Table 2: Red picking model, P(red| pick a color).

Type	VN	Intercept	G^{inv}_{local}	G^{vis}_{local}	R^{inv}_{local}	R^{vis}_{local}
Reg	No	0	-4.863		5.032	
	Yes	-0.066	-2.855	-2.022	3.453	1.733
Vis	No	0.109	-4.411		4.202	
	Yes	0.188	-3.215	-1.599	2.395	1.996
Adv	No	-0.023	0.817		-0.649	
	Yes	-0.286	0.172	0.732	-0.204	

The coefficients of the learned models are presented in Table 2. The results closely follow expectations here: the more neighbors (visible and not) are choosing *red* as opposed to *green*, the more likely the *consensus* team player to choose *red* as the initial color. On the other hand, adversarial players tend to act in opposition to their neighbors, with *red* prevalence in their local neighborhood generally leading them to choose *green*.

However, with regard to the adversarial players, we make a few noteworthy observations. First, note that adversaries are much

more influenced by visible nodes than non-visible neighbors (acting more strongly in opposition to these), whereas regular players tend to be less swayed by the behavior of visible neighbors as compared to others in their neighborhood. Presumably, the adversaries are deliberately trying to counter the presumed influence of the visible nodes, which they appear to over-estimate. Second, adversarial nodes act *relatively unaggressively*: the negative relationship between neighbor choices and their own initial color choice is relatively slight, in comparison with the magnitude of the positive relationships for the regular nodes (remember that features are normalized, so this comparison is meaningful). This observation that adversarial nodes are less aggressive in their activities aimed at thwarting consensus is surprising. We will return to it below, as we make a similar observation in the case of player decisions about when to change their previously chosen color.

Timing of Color Change. Our last set of models determine the timing of a color change by a player. More precisely, we again learn l_1-regularized logistic regression models which represent the probability that a player switches to the other color (either from *red* to *green*, or vice versa) at a given discrete time unit. For these models, the features are: O_l^i, the fraction of a player's non-visible neighbors choosing the opposite color from the one chosen by the player; O_l^v, the fraction of a player's visible neighbors choosing the opposite color from the one chosen by the player; C_l^i, the fraction of a player's non-visible neighbors choosing the same color as the player; C_l^v, the fraction of a player's visible neighbors choosing the same color as the player; N_v, the number of a player's neighbors who are visible; and N_i, the number of a player's neighbors that are non-visible players.

Table 3: Color-changing model.

Type	VN	Intercept	O_l^i	O_l^v	C_l^i	C_l^v	N_i	N_v
Reg	No	−3.98	2.65		−0.33		-0.01	
	Yes	-3.79	1.1	1.48	-0.87	0.09	0	-0.03
Vis	No	-4.11	2.7		-0.1		-0.01	
	Yes	-3.53	1.07	1.27	-0.33	-0.29	-0.06	0
Adv	No	-2.8	-1.13		1.19		0	
	Yes	-2.72	-0.6	-0.37	0.95	0.30	0	-0.2

The model coefficients are presented in Table 3. The broad results are again intuitive: as we would expect, when the local color choices oppose that of a player, a regular player tends to switch, whereas the adversary tends to stay with their current color choice. However, unlike their choice of the first color, here the adversaries less aggressively respond to visible node decisions as compared to those for their remaining neighbors.

Interestingly, as we had observed above, adversarial nodes appear to be somewhat less aggressive in acting *against* the neighborhood trends, as compared to *consensus* players in their decisions to switch to be better aligned with these. This is at first glance unexpected: why would adversaries hold back, rather than aggressively opposing an emerging consensus in their neighborhood? We conjecture that the explanation is that they are concerned also about being covert. If adversarial nodes act in a way that opposes neighborhood choices too aggressively, they run the risk of being discovered by their neighbors as such, at which point their behavioral influence would, presumably, be minimized. Consequently, adversarial nodes likely attempt to achieve their disruptive goals without being overly obvious to their non-adversarial neighbors.

4.2 Agent-Based Model

Given the computational models of human behavior described above, it is direct to construct an agent-based model (ABM): one simply instantiates each agent as a node on an exogenously specified network, with roles assigned randomly according to an exogenously specified model. In our case, we use the same random assignment model as in the human subjects experiments.

4.2.1 Model Validation. While statistical and face validity are essential steps in confirming that our individual behavior models are reasonable, we now add another dimension: validation in terms of *aggregate outcomes* of agent-based simulations. Specifically, We simulate identical environments as in our experiments using our constructed ABM, but now using artificial agents and in discrete time, for 60 iterations (since each time step in our models is equivalent to 1 second in the experiments). Finally, we compare both qualitative trends, and quantitative outcomes, to those reported in the experimental results section above (Section 3). Quantitatively, the agreement is reasonable, with the largest deviation between simulation outcomes and the experimental consensus rates are within 0.14. The qualitative agreement is even stronger, as we illustrate in

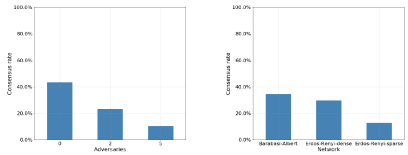


Figure 8: Coordination ratios as a function of single variable.

Figure 8, which shows predicted consensus rates (using simulations) as a function of the number of adversaries (left plot) and network topology (right plot). Comparing to corresponding results from the human subject experiments in Section 3, we can observe broad qualitative agreement. Note that the agreement between simulated and experimental results we achieve for games at this scale (at least 20 players, with considerable interdependencies in behavior) compares quite favorably with similar efforts for devising artificial agents to model coordination in prior literature [19].[4] The degree of consistency between simulations and experiments is particularly noteworthy in our case, if one considers that we had to construct 18 distinct behavior models to capture human behavior.

Despite strong agreement with experimental findings, it is still natural to wonder whether our models are robust to small changes in parameters. Such robustness is crucial if we are to trust the models to remain predictive as we significantly change the setup of the experiment, as we do below. We now show that our model is, indeed, robust to *worst-case* perturbations in the parameters of regular players (as these dominate the simulations).

[4] [42] is noteworthy as well. However, they consider a public goods game, and aim to predict average contribution. Predicting the probability of consensus using such data-driven agent-based simulations appears to be a more challenging problem.

Recall that for each non-adversarial player we have two models: the first when a player has at least one visible neighbor, and the second when they do not. Since we have two types of non-adversarial actors (visible and non-visible nodes), we optimize coefficients of the four associated models with the objective of maximizing consensus rate, with the constraint that the l_1 norm of the modification does not exceed an exogenously specified ϵ. We approximately solve this problem using *Coordinate Greedy (CG)* local search, which iteratively chooses a parameter to optimize, and attempts to find the best improvement of this parameter. To abide by the l_1 norm constraint, we subsequently project the result into the feasible space. Overall, we find that even for relatively large ϵ, the impact is surprisingly small: it appears that incremental changes in behavior of individuals has little impact on ability to successfully coordinate (the impact is generally $< 5\%$ even for ϵ as large as 0.2).

4.2.2 Optimizing Placement of Trusted and Adversarial Players. In our experiments, we randomly assigned trusted and adversarial players to nodes within the network. We now explore the alternative possibility where the assignment of these is more deliberate. To study the problem systematically, we consider the decision of where to place trusted (visible) and adversarial nodes as a Stackelberg game with two players, the coordinator (the Stackelberg leader) and the adversary (the follower). The coordinator first places the trusted nodes on the network, and, fixing this placement, the adversary places adversarial nodes. The goal of the coordinator is to maximize consensus rate, which the adversary aims to minimize. In order to avoid time-consuming simulations in the optimization phase for both the coordinator and the adversary, we use a proxy objective of choosing a set of nodes maximizing the number of *unique* neighbors; we call this *optimal* for either player. Since the game in our case is relatively small, we solve for optimality by exhaustive search. In addition, we create three baselines for comparison: first, when both players choose nodes randomly (as in our experiments), whereas in the second and third baseline, one player chooses nodes randomly, whereas the other optimizes.

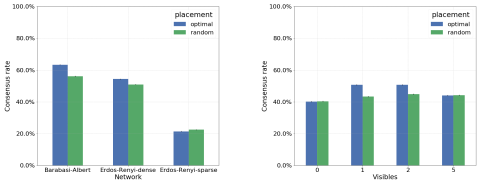


Figure 9: Consensus rate as a function of placement of visible nodes when no adversaries are present. Left: for different network topologies. Right: different number of visible nodes.

We first consider settings with no adversaries, and explore the impact of having an optimal placement of visible nodes, as compared with random placement. The results are presented in Figure 9, for different network topologies (left), and different numbers of visible nodes (right). The broad trend is that while optimal placement of visible nodes is typically helpful, the impact it has on consensus rate is quite muted, further bolstering our experimental observation that the value of having trusted nodes in this setting can be limited.

Figure 10 presents the results of considering the two placement strategies (random and optimal) for visible and adversarial nodes.

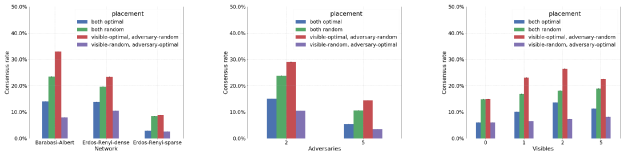


Figure 10: Consensus rate for different strategies of placing visible and adversarial nodes, as a function of: (Left) network topologies; (Right) the number of adversaries; and (Bottom) the number of visible nodes.

From this figure we can make several noteworthy observations. First, *adversarial players are highly effective with optimal placement*: consider blue and purple (first and last) bars in the plots, which correspond to adversaries placed optimally. In both cases, consensus rates are quite small, for all network topologies, and even with only 2 adversaries. This is especially surprising when we also consider the optimal placement of visible nodes, *which are placed before adversaries*, and can thereby ensure that networks remain connected even after adversarial nodes are added. While optimal placement of visible nodes clearly helps, the impact is smaller than we would have expected. Second, *optimally placing visible nodes helps*: consider the red bars (tallest in all plots), which correspond to the optimal placement of visible nodes, followed by random placement of adversaries. In this situation, we can observe a clear value of visible nodes, particularly for the scale-free (BA) topology. On the other hand, we can see that having 2 visible nodes is actually better than 5, which we conjecture is due to the increased potential for miscoordination among visible nodes themselves in the latter case.

5 CONCLUSION

We consider the problem of adversarial consensus on social networks both using human subjects and agent-based modeling methodologies. The overall goal of the subjects is to reach global consensus on a particular color, despite adversarial nodes who attempt to prevent consensus. We find that while the ability to communicate can significantly improve coordination success despite adversarial presence, embedding trusted nodes within the network is of limited value. We observe several strategies used by adversarial players to subvert coordination, such as choosing a color which opposes local majority. However, we also note that these malicious activities are used in a somewhat subdued manner, suggesting perhaps an attempt of adversarial players to remain covert. We use experimental data to construct and validate an agent-based model of adversarial consensus. Extensive simulations using an agent-based model created based on experimental data additionally show that the importance does increase when their network location is optimized, but this improvement is often small, particularly when adversarial nodes are also optimizing location, and even though adversaries do so *after* we choose where to place trusted nodes.

ACKNOWLEDGMENT

This research was partially supported by the National Science Foundation (IIS-1526860, IIS-1905558), Office of Naval Research (N00014-15-1-2621), and Army Research Office MURI (W911NF-18-1-0208).

REFERENCES

[1] Waseem Abbas, Aron Laszka, and Xenofon Koutsoukos. 2017. Improving Network Connectivity and Robustness Using Trusted Nodes with Application to Resilient Consensus. (2017), 1–7.

[2] Waseem Abbas, Yevgeniy Vorobeychik, and Xenofon Koutsoukos. 2014. Resilient consensus protocol in the presence of trusted nodes. In *International Symposium on Resilient Control Systems*. 1–7.

[3] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. 2000. Error and attack tolerance of complex networks. *Nature* 406 (2000), 378–482.

[4] Noga Alon, Michal Feldman, Omer Lev, and Moshe Tennenholtz. 2015. How Robust is the Wisdom of the Crowds?. In *IJCAI*. 2055–2061.

[5] Albert-Laszlo Barabasi and Reka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512.

[6] Gabriel Bracha and Sam Toueg. 1983. Resilient Consensus Protocols. In *ACM Symposium on Principles of Distributed Computing*. 12–26.

[7] Tanmoy Chakraborty, Stephen Judd, Michael Kearns, and Jinsong Tan. 2010. A Behavioral Study of Bargaining in Social Networks. In *Proceedings of the 11th ACM Conference on Electronic Commerce*. 243–252.

[8] Russell Cooper, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1992. Communication in coordination games. *Quarterly Journal of Economics* 107, 2 (1992), 739–771.

[9] Lorenzo Coviello, Massimo Franceschetti, Mathew McCubbins, Ramamohan Paturi, and Andrea Vattani. 2012. Human matching behavior in social networks: an algorithmic perspective. *PloS one* 7, 8 (2012), e41900.

[10] Stefano Demichelis and Jorgen W. Weibull. 2008. Language, meaning, and games: A model of communication, coordination, and evolution. *American Economic Review* 98, 4 (2008), 1292–1311.

[11] Tore Ellingsen and Robert Ostling. 2010. When does communication improve coordination? *American Economic Review* 100 (2010), 1695–1724.

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

[13] Joseph Farrell. 1987. Cheap talk, coordination, and entry. *RAND Journal of Economics* 18, 1 (1987), 34–39.

[14] Joseph Farrell. 1988. Communication, coordination and Nash equilibrium. *Economic Letters* 27 (1988), 209–214.

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

[16] Chen Hajaj, Noam Hazon, and David Sarne. 2015. Improving comparison shopping agents' competence through selective price disclosure. *Electronic Commerce Research and Applications* 14, 6 (2015), 563–581.

[17] Chen Hajaj, Noam Hazon, and David Sarne. 2017. Enhancing comparison shopping agents through ordering and gradual information disclosure. *Autonomous Agents and Multi-Agent Systems* 31, 3 (2017), 696–714.

[18] Chen Hajaj and Yevgeniy Vorobeychik. 2018. Adversarial task assignment. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 3783–3789.

[19] Stephen Judd, Michael Kearns, and Yevgeniy Vorobeychik. 2010. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences* 107, 34 (2010), 14978–14982.

[20] Michael Kearns. 2012. Experiments in social computation. *Commun. ACM* 55, 10 (2012), 56–67.

[21] Michael Kearns, Stephen Judd, Jinsong Tan, and Jennifer Wortman. 2009. Behavioral experiments in biased voting in networks. *Proceedings of the National Academy of Sciences* 106, 5 (2009), 1347–1352.

[22] Michael Kearns, Stephen Judd, and Yevgeniy Vorobeychik. 2012. Behavioral experiments on a network formation game. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 690–704.

[23] Michael Kearns, Siddharth Suri, and Nick Montfort. 2006. An experimental study of the coloring problem on human subject networks. *Science* 313, 5788 (2006), 824–827.

[24] Heath J. LeBlanc and Xenofon D. Koutsoukos. 2012. Low Complexity Resilient Consensus in Networked Multi-agent Systems with Adversaries. In *Proceedings of the 15th ACM International Conference on Hybrid Systems: Computation and Control (HSCC '12)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/2185632.2185637

[25] Heath J. LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. 2013. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications* 31, 4 (2013), 766–781.

[26] Andreas Leibbrandt, Abhijit Ramalingam, Lauri Sääksvuori, and James M. Walker. 2015. Incomplete punishment networks in public goods games: experimental evidence. *Experimental Economics* 18, 1 (2015), 15–37.

[27] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[28] McCubbins Matthew, Paturi Ramamohan, and Weller Nicholas. 2009. Networked coordination: effect of network structure on human subjects' ability to solve coordination problem. *Am Polit Res* 37 (2009), 899–920.

[29] John H. Miller and Scott Moser. 2004. Communication and coordination. *Complexity* 9, 5 (2004), 31–40.

[30] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. 2016. *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press.

[31] Anne J Olmstead, Navin Viswanathan, Karen A Aicher, and Carol A Fowler. 2009. Sentence comprehension affects the dynamics of bimanual coordination: Implications for embodied cognition. *Quarterly Journal of Experimental Psychology* 62, 12 (2009), 2409–2417.

[32] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010).

[33] Noam Peled, Sarit Kraus, et al. 2015. A study of computational and human strategies in revelation games. *Autonomous Agents and Multi-Agent Systems* 29, 1 (2015), 73–97.

[34] Akshay R. Rao and Kent B. Monroe. 1989. The effect of price, brand name, and store name on buyers' perceptions of product quality: An integrative review. *Journal of marketing Research* (1989), 351–357.

[35] Peter J. Richerson and Robert Boyd. 2010. Why possibly language evolved. *Biolinguistics* 4, 2-3 (2010), 289–306.

[36] Szabolcs Szamado. 2011. Pre-hunt communication provides context for the evolution of early human languge. *Biological Theory* 5, 4 (2011), 366–382.

[37] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, and Yevgeniy Vorobeychik. 2017. A Framework for Validating Models of Evasion Attacks on Machine Learning, with Application to PDF Malware Detection. (2017). arXiv:cs.CR/1708.08327

[38] Liang Tong, Sixie Yu, Scott Alfeld, and Yevgeniy Vorobeychik. 2018. Adversarial Regression with Multiple Learners. In *International Conference on Machine Learning*. 4953–4961.

[39] James Usevitch and Dimitra Panagou. 2018. Resilient Leader-Follower Consensus to Arbitrary Reference Values. In *Annual American Control Conference*. 1292–1298.

[40] Yevgeniy Vorobeychik, Zlatko Joveski, and Sixie Yu. 2017. Does communication help people coordinate? *PloS One* 12, 2 (2017), 1–19.

[41] Yevgeniy Vorobeychik and Murat Kantarcioglu. 2018. *Adversarial Machine Learning*. Morgan & Claypool Publishers.

[42] Michael Wunder, Siddharth Suri, and Duncan J Watts. 2013. Empirical agent based models of cooperation in public goods games. In *Proceedings of the fourteenth ACM conference on Electronic commerce*. ACM, 891–908.

[43] Wente Zeng and Mo-Yuen Chow. 2014. Resilient distributed control in the presence of misbehaving agents in networked control systems. *IEEE transactions on cybernetics* 44, 11 (2014), 2038–2049.

[44] Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. 2016. Data-driven agent-based modeling, with application to rooftop solar adoption. *Journal of Autonomous Agents and Multiagent Systems* 30, 6 (2016), 1023–1049.