Reaching Consensus through Theory of Mind in Social Networks with Locally Distributed Interactions

Daphne Barretto (daphnegb@alumni.princeton.edu)
Department of Computer Science, Princeton University, Princeton, NJ, United States

Raja Marjieh (raja.marjieh@princeton.edu)
Department of Psychology, Princeton University, Princeton, NJ, United States

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, United States

Abstract

How people reach consensus in social networks with locally distributed interactions is relevant to understanding collective group decision-making and problem-solving. However, while the importance of theory of mind in consensus problems has been hypothesized, little work has been done to test it systematically. We present both computational modeling and behavioral experiments designed to test the impact of theory of mind on individual choices within such consensus networks. We test 2,108 computational models informed by theoretical work on a graph-coloring consensus task to compare models using theory of mind to other behavioral parameters. We then use behavioral responses from 107 participants in a similar task to evaluate support for theory of mind in consensus formation. We find that the computational model that best accounts for prior behavioral data uses theory of mind, and our behavioral results likewise support use of theory of mind over other potential decision-making models.

Keywords: Psychology; Decision Making, Group Behavior; Theory of Mind; Computational Modeling

Introduction

How people reach consensus is relevant to understanding collective decision-making, problem-solving, and memory (Kearns, 2012; Kearns & Tan, 2008; Kearns, Judd, Tan, & Wortman, 2009; Bullo, 2020; Balietti, Getoor, Goldstein, & Watts, 2021; Centola & Baronchelli, 2015; Coman, Momennejad, Drach, & Geana, 2016). Consensus problems involve individuals making choices based on local interactions with only a subset of other individuals within a social network, until global agreement on a decision is reached. Despite the limitation of local interactions, groups are still remarkably good at reaching consensus without global interactions.

Theoretical work has suggested many potential factors that influence performance in consensus tasks, including network structure (Enemark, McCubbins, Paturi, & Weller, 2011; Jackson, 2005), memory (Duong, Wellman, Singh, & Kearns, 2012), heterogeneous behaviors (Judd, Kearns, & Vorobeychik, 2010), and theory of mind (Kearns & Tan, 2008; Kearns, 2012). However, empirical studies have typically been limited to modeling success rates with relatively simple heuristics (e.g., myopically choosing the option that minimizes conflict; Judd et al., 2010), rather than designing experiments to test behavioral factors individually.

In this paper, we use both computational modeling and behavioral experiments to explore the factors that influence reaching consensus in social networks with locally distributed interactions. We test 2,108 computational models informed by theoretical work on a graph-coloring consensus task that emulate various decision strategies, including the use of theory of mind. We then use a novel behavioral experiment to evaluate support for theory of mind in consensus formation.

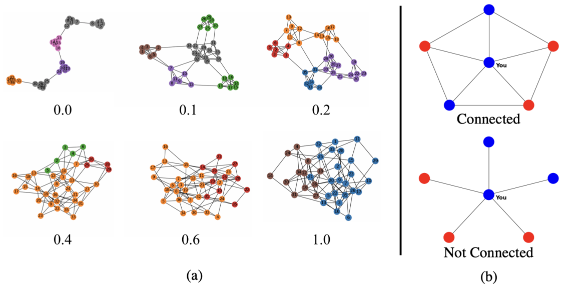

Figure 1: Example graph-coloring consensus social networks for (a) computational modeling experiments with different rewiring probabilities, q, and (b) human behavioral experiments with either `Connected` or `Not Connected` neighbors.

We find that the computational model that best accounts for prior behavioral data uses theory of mind, and that other tested factors, including memory, decision noise, stubbornness, and sub-optimal decisions, were unable to reach similar performance alone. Our behavioral results likewise support use of theory of mind over other potential decision-making models, particularly a majority model that considers one's own observations but not the observations of others.[1]

Background

Consensus Problems

Prior empirical work on consensus problems encompasses many settings including graph coloring (Judd et al., 2010; Kearns, Suri, & Montfort, 2006), biased voting (Kearns et al., 2009; Kearns & Tan, 2008), trade (Judd & Kearns, 2008), and network formation (Kearns, Judd, & Vorobeychik, 2012). Studies suggest that people perform well collectively (Kearns, 2012) and that their performance is shaped by social network connectivity (Enemark et al., 2011; Jackson, 2005), individual differences in decision-making strate-

[1] Code and data are available upon request.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

gies, personalities, and particularly "stubbornness" (Judd et al., 2010), bargaining (Chakraborty, Judd, Kearns, & Tan, 2010; Chakraborty, Kearns, & Khanna, 2009), conflict and fairness (Judd, Kearns, & Vorobeychik, 2011), and theory of mind (Kearns & Tan, 2008; Kearns, 2012).

Theory of Mind

Theory of mind refers to humans' ability to form meaningful inferences about the unobserved mental states of others (Gopnik & Meltzoff, 1997; Jara-Ettinger, 2019). This capacity plays a key role in a variety of cognitive skills such as decision-making (Lucas et al., 2014) and planning (Ho, Saxe, & Cushman, 2022). As such, empirical research into theory of mind has been the center of considerable work in the cognitive sciences (Fawcett & Markson, 2010; Lucas et al., 2014; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016) and machine learning (Rabinowitz et al., 2018) communities. Computational models of theory of mind have used frameworks such as inverse decision theory (Lucas et al., 2014; Jara-Ettinger et al., 2016) and inverse reinforcement learning (Jara-Ettinger, 2019; Ng, Russell, et al., 2000), which are based on Bayesian inference. In these models, Bayes' rule is used to update a probability distribution over the mental states of others given information about their behavior.

Computational Models for Consensus

We focus on modeling the graph-coloring consensus task, a common consensus problem. In this task, a network of agents attempt to reach consensus by all agents selecting a single color while only observing the colors previously selected by their neighbors. This task is known to be computationally tractable and has high human success rates, but understanding the factors influencing human decisions in the task has been limited to evaluating a myopic heuristic model that selects the color with the least conflict from its local neighbors (Judd et al., 2010; Kearns, 2012). While models incorporating signaling and preference inference components have been suggested (Kearns, 2012), a systematic evaluation of candidate models that incorporate theory of mind has not yet been conducted. Understanding a computationally simple task such as the graph-coloring consensus task could help us understand consensus decision-making more broadly.

Testing the impact of theory of mind on consensus requires using models that go beyond simple myopic heuristics, which do not precisely capture human behavior. To do this, we develop models that factor in the anticipated decisions of others, incorporating Bayesian inference and evaluating their predictions against human behavior, along with other baseline models that capture different decision-making strategies.

Computational Modeling Methods

We created a simulation framework to mimic previous behavioral experiments on the graph-coloring consensus problem (Judd et al., 2010; Kearns, 2012). In this task, a social network of agents attempt to reach consensus by all selecting the same color from a set of valid colors. However, the social network is limited to local interactions such that each agent can only observe the selected colors of its neighbors, which limits information on the global state of the network. Color selection iterates until consensus is reached or the task is timed out. Using this framework, we ran 531,480 trials to test 2,108 model settings with different behavioral parameters, analyzing their objective performance and their correlation with data from previous behavioral experiments.

Simulation Framework

The simulation framework is written in Python using `networkx` (Hagberg, Swart, & S Chult, 2008). The framework runs a trial of the graph-coloring consensus task by taking model behavioral parameters (see Model Design and Implementation) and trial setup parameters as inputs. The trial setup parameters specify the number of cliques in a network, the number of nodes within each clique, the rewiring probability, the set of colors to choose from, and the maximum number of iterations before the experiment ends (Figure 1a).

Specifically, a social network graph is created with `cliques` number of cliques, each with `nodes_per_clique` number of nodes. Each node represents a single agent, an instantiation of a model representing a person in the network. Each clique is connected to two other cliques, each by a single edge, in a chain of cliques–except for the two end cliques, which are each connected to only one other clique. Each edge in the graph is rewired with rewiring probability `q`, such that one of its nodes is changed to a random node that would create a new edge not currently in the graph. The process is repeated until a connected graph is created (Figure 1a).

All agents in the graph are initialized to a random color from a set of `colors` number of valid colors for the trial. The trial begins and continues to iterate until consensus is successfully reached or `max_iterations` number of iterations are complete. During each iteration, in an arbitrary but static order, every agent conducts a color selection process, determined by the individual agent and its model. At the end of each iteration, a Bayesian update occurs for all agents.

Model Design and Implementation

At the end of each iteration of a trial, each agent conducts a Bayesian update based on the observed colors of their neighbors by integrating past observations within a given memory horizon, possibly subject to decision noise. Intuitively, the model estimates the probabilities of choosing each color by each neighbor, and then integrating those probabilities to decide which color is most likely to achieve consensus.

Formally, the model uses the observed colors of each neighbor within the last `memory` iterations to calculate the probability that each of their neighbors will choose a given color by counting the number of times that color was chosen per neighbor. To avoid zero probabilities, we implement one pseudo-observation count for every color. The model then creates its own color selection distribution of what it believes is most likely to reach consensus based on the `distribution` `update type`, which is either the `sum` or `product` of each of its neighbors' probabilities for selecting a given color. The model then conducts its color selection based on the `decision-making type`, which is either `deterministic`, selecting the largest sum or product value from the colors, or `probability matching`, probabilistically selecting a color from a normalized version of the distribution. Overall, these parameters correspond to 724 base model settings.[2]

We designed an additional set of 724 'theory-of-mind' models based on the base models. The set is identical to the base models, except that these models have the `theory of mindedness` parameter, which mimics considering the decision-making process of one's neighbors. If the `theory of mindedness` parameter is present, instead of actually using its own calculated color distribution as described above, a model is given access to its neighbors' decision distributions which it then uses to select the color that is most likely to reach consensus. In this way, each agent is selecting its color based on the perceived decision-making calculations of its neighbors and not only its own decision-making calculations from observations.

Finally, as another baseline that incorporates individual variations in agents' behavior, we included (based on pilot simulations to eliminate poorly performing models) a set of 660 'heterogeneous' models with `memory` parameters ranging from 0 to 10, a `distribution update type` of `product`, either `decision-making type`, and one of the following heterogeneous parameters: `decision noise`, `stubbornness`, and `sub-optimal decisions for collaboration`. The `decision noise` parameter is the probability that an agent will select a random valid color from a uniform distribution during any given iteration, ignoring all observations and memory. The `stubbornness` parameter is the probability that an agent will select the same color that it selected most recently during any given iteration, ignoring any observations and memory. The `sub-optimal decisions for collaboration` parameter is the probability that the agent will ignore the colors with the highest value in their own color distribution; `deterministic` agents select a color with the second highest probability for being selected by all neighbors, and `probability matching` agents randomly draw a color from the distribution of remaining normalized probabilities. The heterogenous parameters varied from 0.1 to 1.0 in increments of 0.1. When an agent is instantiated from a heterogeneous model, a value is drawn from a uniform distribution between 0.0 and the selected heterogeneous parameter value, such that each agent in a trial have different individual values of those parameters.

Trial Design and Implementation

We ran 531,480 graph-coloring consensus trials using our three sets of computational models. We placed each of our 2,108 models in trial setups identical to the prior behavioral experiments: `cliques` = `nodes_per_clique` = 6, `q` = {0.0, 0.1, 0.2, 0.4, 0.6, 1.0}, and `colors` = 9. Likewise, we set `max_iterations` to 180, similar to the behavioral experiment max time to consensus of 180 seconds. We used identical models and uniform initial prior color distributions for every agent in our trials, with the heterogeneous models adding differing individual behaviors between agents in the same trial. We ran 100 trials per each of the six `q` values for each of our base models with the `distribution update type` of `sum`, and 30 trials per each of the six `q` values for all other models.

Evaluating the Models

For each model with its set of behavioral parameters, we calculate the success rate and the average time to consensus in iterations for each value of `q`. (An unsuccessful trial's time to consensus is considered to be `max_iterations`.) As a comparison, we estimated the average time to consensus in seconds over `q` for the prior behavioral experiments of the graph-coloring consensus tasks and its heuristic model based on results presented in Judd et al. (2010). We compare success rates and average time to consensus across `q` for our models to analyze how the different parameters affect model performance, both in terms of objective performance (i.e., the highest success rates and lowest average time to consensus) and correlation with the results of Judd et al. (2010) (i.e., the correlation between a model's average time to consensus to the reported human average time to consensus). We also evaluated the heuristic model based on results presented in Judd et al. (2010), finding a correlation of 0.65 with the human data.

Computational Modeling Results

Figure 2 compares the average time to consensus for humans reported in Judd et al. (2010) with our best performing models, as measured by their Pearson correlation (*r*) with the human data. The models that are shown were best performers in each of the following categories: the two best performing theory of mind models with either `decision-making type` across `distribution update type` and `memory`, the two best performing base models with either `decision-making`

		Success Rate over q					
Model	*r*	**0.0**	**0.1**	**0.2**	**0.4**	**0.6**	**1.0**
1	0.96	0.03	0.77	1.00	1.00	1.00	1.00
2	0.70	0.00	0.00	0.03	0.77	0.90	0.87
3	0.78	0.00	0.07	0.62	0.99	0.93	0.83
4	0.70	0.00	0.00	0.04	0.65	0.93	0.96
5	0.75	0.00	0.27	0.60	0.60	0.37	0.30
6	0.84	0.00	0.20	0.73	0.97	0.77	0.77
7	0.81	0.00	0.00	0.03	0.20	0.50	0.30

Table 1: Correlations (*r*) and success rates for best performing models, as labeled in the legend of Figure 2. We see that our best performing model, 1, has the highest success rate across all values of `q`, especially over `q` lower values. This correlates with the behavioral data from Judd et al. (2010).

[2] The base model with no memory and which uses the `product` and `deterministic` parameters is functionally equivalent to the prior work heuristic model (Judd et al., 2010; Kearns, 2012).

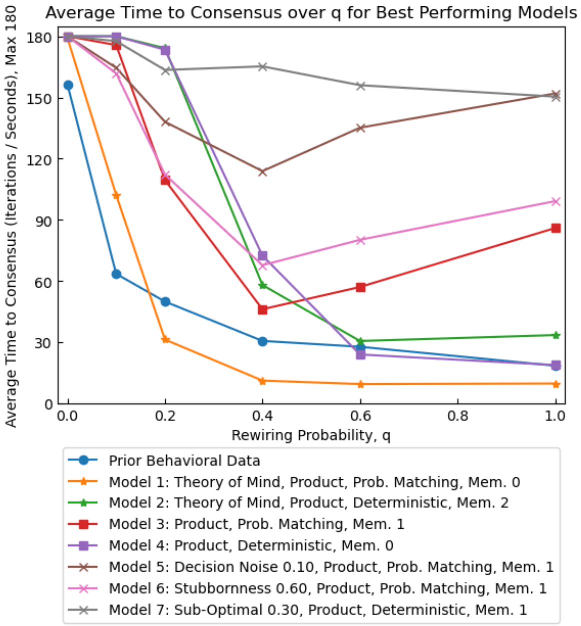

Figure 2: Comparisons of the average time to consensus over q for the prior behavioral data from Judd et al. (2010) and the best performing models in several categories. Our best performing model is a theory of mind model with the product, probability matching, and memory= 0 parameters, which has a correlation of 0.96 with the behavioral data.

type across distribution update type and memory, and the three best performing heterogeneous models with either decision-making type across memory and their heterogeneous parameters. The correlations between the prior behavioral experimental results and the seven models and the model success rates over q are provided in Table 1.

In general, models performed better in networks with higher q values, which indicates higher connectedness. By analyzing the simulations, we observed that the primary obstacle to success in the task was subgroups of agents agreeing on a single color but being unable to reach consensus with other subgroups in the network, such as the examples shown in Figure 1a for q values of 0.0, 0.1, and 0.2. This explains why networks that are more interconnected can reach consensus quicker and with a higher success rate, with and without theory of mind, because it is less likely for color subgroups to form without global consensus, which is a common failure case for majority models. Conversely, in networks that are less interconnected, color subgroups are more prone to getting stuck as they receive less information about other parts of the network. Nonetheless, the behavioral data (Figure 2) indicates that people had consistently high success rates at this task across different q values, suggesting that they used some strategy to overcome this state.

As for the heterogeneous models (i.e., those deploying the decision noise, stubbornness, and sub-optimal decisions for collaboration parameters), we found that those helped avoid deterministically remaining in color subgroups for some lower q values (i.e., 0.1, 0.2, 0.4), but often resulted in longer average time to consensus in higher q values. The best performing model was a theory-of-mind model with the product, probability matching, and memory $= 0$ parameters. It had success rates of [0.03, 0.77, 1.00, 1.00, 1.00, 1.00] over q, and a correlation of 0.96 with the behavioral data. We found that it outperformed the other models most at q values of 0.1 and 0.2, where the second-best success rates were only 0.27 and 0.73, respectively. Moreover, unlike the other behavioral parameters that performed well at lower q values, the best model maintained high success rates and low average time to consensus through the higher q values too. As shown in Figure 2, this generally aligns with the human data in Judd et al. (2010), which similarly exhibits a substantial drop in average time to consensus from q of 0 to 0.1. However, even the best model still had a very low success rate (0.03) at q = 0, potentially due to the theory of mindedness behavior not being weighted heavily enough for the limited amount of interconnectedness, so that individuals generally still decided on the same color as their own subgroup; different subgroups still generally selected their own previous colors and did not select the same color as all other subgroups within the iteration limit, suggesting that there is still room for improvement with this behavioral parameter. Nonetheless, by demonstrating that our best performing model is substantially aligned with the behavioral data, our results support the idea that people use theory of mind in consensus problems, allowing them to overcome conflicts that challenge many heuristic models.

Behavioral Experiment Methods

Behavioral Paradigm

The results of our comprehensive testing of the different models suggest that incorporating theory of mind is essential for capturing human behavior on consensus tasks. To further test this idea, we conducted a behavioral experiment on the impact of theory of mind for individual choices within the graph-coloring consensus problem. To do this, we tested the color selection responses of 107 participants to 52 graph-coloring consensus scenarios, using one-shot decision stimuli that display a set of observations for one's own node and neighbors' nodes (Figure 1b). In each trial, the participant is associated with a central node that is connected to five neighboring nodes, and each node can be either red or blue. The goal of the participant is to choose the color for the hypothetical next iteration so as to achieve consensus. We considered two network structures: a Connected structure, meaning that each of the participant's neighbors is connected to two other neighbors, and a Not Connected structure, meaning that the participant is the only node each of its neighbors can observe (Figure 1b).

We ran simulated theory of mind and majority decision-making models against all 2×2^6 combinations of red and blue nodes for the two network structures to determine what perfectly rational versions of those models would respond to each stimulus. The simulated majority model selects a color that aligns with the majority of its observations, i.e., its previous color selection and those of its neighbors. The simulated theory of mind model considers the computations of its neighbors and approximates their next decisions based on the majority model, and then chooses the color that aligns with the majority of those expected decisions.

From all potential stimuli, we selected the 20 `Connected` and 32 `Not Connected` networks where following a rational majority model and a rational theory of mind model would result in deterministically different responses. Note that we allow for potential redundancy in stimuli due to rotational and color symmetry to test if these differences result in substantial variations in results.

Analysis

Based on the simulated rational majority and theory of mind models, we assign a theory of mindedness value for each participant response to each stimulus, such that the value is 1 if the response aligns with the simulated rational theory of mind model and 0 if not (i.e., the response aligns with the simulated majority model). We calculate mean participant-level theory of mindedness values by averaging the individual values for each participant across all of their responses to stimuli. Likewise, we calculate mean stimulus-level theory of mindedness values by averaging the individual values for each stimulus across all of the 107 participants. Finally, we performed a bootstrapping analysis with 100 iterations over the participant and stimuli theory of mindedness values, and used these in our analysis.

Participants

We recruited $N = 107$ online participants through Amazon Mechanical Turk (AMT). To ensure data quality, participants were required to reside in the United States and to have successfully completed at least 5,000 AMT tasks. Moreover, participants were required to pass an Ishihara color blindness test (Clark, 1924). All participants provided informed consent prior to participation in accordance with an approved Princeton University Institutional Review Board (IRB) protocol (#10859), and they provided up to 30 judgments each to randomly-selected one-shot decision scenarios (out of the 52 stimuli). Participants additionally had the option of describing their strategies in free-text form at the end of the study.

Behavioral Experiment Results

Our data yielded a participant-level mean theory of mindedness value of 0.66, with a 95% confidence interval of [0.60, 0.72]. Figure 3 shows the distribution of those values. Notably, 29/107 participants had a mean participant theory of mindedness value > 0.95, and 55/107 participants had a value > 0.70, while only 18/107 participants had a value < 0.30.

The observed individual-level variation was also reflected in the reported participant strategies. In particular, we found descriptions that aligned with both a majority decision-making model which corresponds to lower mean theory of mindedness values (e.g., “I tried to go with the color that already had the majority, most of the time.”, “I just went with the color that has more [representation].”) and descriptions indicating some form of theory of mind that aligned with higher mean theory of mindedness values (e.g., “I looked at each node and figured out what they saw and then figured the probability each would choose.”, “In general, I looked at what the other nodes were connected to and guessed on whether they would change their colors based on what they saw.”, “I tried to think that everyone [cooperated and picked] what they saw more of.”). Finally, we also observed one participant strategy that explicitly described a stubborn decision-making strategy: “I always chose the color of me.” All reported strategies appeared to align with one of these three models.

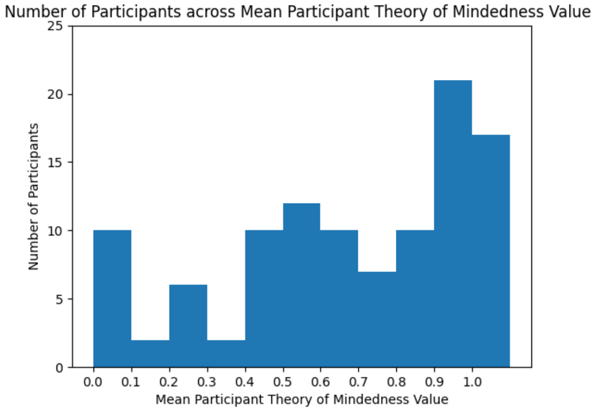

Figure 3: Distribution of the mean participant theory of mindedness values across participants. The overall bootstrapped mean theory of mindedness value across all participants is 0.66, supporting the use of a theory of mind model over a majority model across participants.

In addition to differences across participants, we also found variation in the mean stimulus theory of mindedness value. Figure 4 shows the distribution of stimuli and their mean stimulus theory of mindedness values. Notably, overall 45/52 stimuli have values > 0.50, with a large majority having values between 0.60 and 0.80.

Discussion

Our work provides clear support for the role of theory of mind for reaching consensus in social networks with locally distributed interactions. Through computational modeling, we find that across an exhaustive search of 2,108 model settings informed by prior theoretical work, our best performing model deployed theory of mind and improved over prior heuristic models. Moreover, this model outperformed an array of additional baselines that incorporated features such as different probability updates, memory, and agent heterogeneity, as these were unable to achieve similarly high success

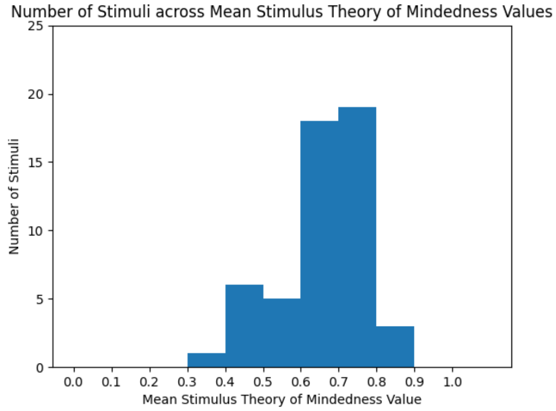

Figure 4: Distribution of the mean stimulus theory of mindedness values across stimuli. 45/52 stimuli have values over 0.50, indicating the use of a theory of mind model over a majority model across different stimuli.

rates and low average time to consensus for lower values of the rewiring parameter q while maintaining performance at higher values. While these features may still affect consensus in social networks with locally distributed interactions, our results highlight the importance of considering theory of mind in explaining how humans reach consensus.

Through our behavioral experiments, we further found support for a theory of mind model over a majority model by carefully selecting one-shot decision scenarios that lead to diverging predictions for each of these models. We also found individual differences in the participants' deployed strategies which were also reflected in the text descriptions provided by those participants.

Limitations and Future Work

While we tested a large set of computational models informed by theoretical work, there is still room for testing more complex models, in particular ones that can cope with highly clustered networks like the $q = 0$ case, or even adaptive models that depend on the topology of the network. Additionally, our models were primarily analyzed in comparison to prior behavioral data, which had a limited sample size of three trials per q value, and against our own behavioral experiments which were limited to instantaneous decision-making problems, rather than tracking an entire graph-coloring consensus problem. Our one-shot two-color study also does not differentiate well between a potential stubbornness model and a rational theory of mind model (though the reported verbal strategies of participants suggest that the former was not a prevalent strategy as it was only mentioned by one participant). Future work could address these limitations in tandem, focusing on gathering more behavioral data in varied structures for the graph-coloring consensus task, computationally searching for sets of models that align with this behavioral data, and iterating between these two steps until converging on a more complete explanation of the role of theory of mind in decision-making within social networks with locally distributed interactions. This work can also be expanded to other consensus problems outside of the graph-coloring family to evaluate whether it generalizes.

Conclusion

In this work, we conducted both extensive computational modeling and human behavioral experiments to test the impact of theory of mind on individual choices within consensus problems on social networks, using the graph-coloring consensus problem as a case study. Our results underscore the importance of theory of mind in distributed decision making problems, highlighting that effective collective behavior requires effective models of individuals, and we hope that our work will inspire other researchers to explore the interaction between individual and collective intelligence.

Acknowledgements

This work was supported by grants from the Office of Naval Research (ONR grant number N00014-22-1-2813) and the NOMIS Foundation.

References

Balietti, S., Getoor, L., Goldstein, D. G., & Watts, D. J. (2021). Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, *118*(52), e2112552118.

Bullo, F. (2020). *Lectures on network systems* (Vol. 1) (No. 3). Kindle Direct Publishing Seattle, DC, USA.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994.

Chakraborty, T., Judd, S., Kearns, M., & Tan, J. (2010). A behavioral study of bargaining in social networks. *Proceedings of the 11th ACM Conference on Electronic Commerce*, 243–252.

Chakraborty, T., Kearns, M., & Khanna, S. (2009). Network bargaining: algorithms and structural results. *Proceedings of the 10th ACM Conference on Electronic Commerce*, 159–168.

Clark, J. (1924). The Ishihara test for color blindness. *American Journal of Physiological Optics*, *5*, 269–276.

Coman, A., Momennejad, I., Drach, R. D., & Geana, A. (2016). Mnemonic convergence in social networks: The emergent properties of cognition at a collective level. *Proceedings of the National Academy of Sciences*, *113*(29), 8171–8176.

Duong, Q., Wellman, M. P., Singh, S., & Kearns, M. (2012). Learning and predicting dynamic networked behavior with graphical multiagent models. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, *1*, 441–448.

Enemark, D. P., McCubbins, M. D., Paturi, R., & Weller, N. (2011). Does more connectivity help groups to solve social problems. *Proceedings of the 12th ACM Conference on Electronic Commerce*, 21–26.

Fawcett, C. A., & Markson, L. (2010). Children reason about shared preferences. *Developmental Psychology*, *46*(2), 299.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. MIT Press.

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring Network Structure, Dynamics, and Function using NetworkX* (Tech. Rep.). Los Alamos National Lab (LANL), Los Alamos, NM (United States).

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Jackson, M. O. (2005). A survey of network formation models: stability and efficiency. *Group Formation in Economics: Networks, Clubs, and Coalitions*, *664*, 11–49.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Judd, S., & Kearns, M. (2008). Behavioral experiments in networked trade. *Proceedings of the 9th ACM Conference on Electronic Commerce*, 150–159.

Judd, S., Kearns, M., & Vorobeychik, Y. (2010). Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences*, *107*(34), 14978–14982.

Judd, S., Kearns, M., & Vorobeychik, Y. (2011). Behavioral conflict and fairness in social networks. *International Workshop on Internet and Network Economics*, 242–253.

Kearns, M. (2012, Oct). Experiments in social computation. *Communications of the ACM*, *55*(10), 56–67.

Kearns, M., Judd, S., Tan, J., & Wortman, J. (2009). Behavioral experiments on biased voting in networks. *Proceedings of the National Academy of Sciences*, *106*(5), 1347-1352.

Kearns, M., Judd, S., & Vorobeychik, Y. (2012). Behavioral experiments on a network formation game. *Proceedings of the 13th ACM Conference on Electronic Commerce*, 690–704.

Kearns, M., Suri, S., & Montfort, N. (2006). An experimental study of the coloring problem on human subject networks. *Science*, *313*(5788), 824–827.

Kearns, M., & Tan, J. (2008). Biased voting and the democratic primary problem. *Internet and Network Economics*, 639–652.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., … Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *Public Library of Science One*, *9*(3), e92160.

Ng, A. Y., Russell, S., et al. (2000). Algorithms for inverse reinforcement learning. *International Conference on Machine Learning*, *1*, 2.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *International Conference on Machine Learning*, 4218–4227.