Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-025-09396-3

ORIGINAL PAPER

Characterizing the roles of preference homophily and network structure on outcomes of consensus games

Pratyush Arya[1] · **Nisheeth Srivastava**[1]

Accepted: 3 January 2025

Abstract
This paper presents results from *in silico* experiments trying to uncover the mechanisms by which people both succeed and fail to reach consensus in networked games, for network structures produced by a variety of generative mechanisms. We find that the primary cause for failure in such games is preferential selection of information sources. Agents forced to sample information from randomly selected fixed neighborhoods eventually converge to a consensus, while agents free to form their own neighborhoods and forming them on the basis of homophily frequently end up creating balkanized cliques. Small-world structure attenuates the drive towards consensus in fixed networks, but not in self-selecting networks. Preferentially attached networks show the highest convergence to consensus, thereby showing resilience to balkanization even in self-selecting networks. We investigate the reasons for such behavior by altering graph properties of generated networks. We conclude with a brief discussion of the implications of our findings for representing behavior in socio-cultural modeling.

Keywords Social preference · Preference learning · Agent-based modeling · Clique formation · Balkanization · Filter bubbles · Polarization · Network hubs · Opinion dynamics.

1 Introduction

The rapidly advancing digital era of the 21st century reveals profound socio-cognitive divides, driven by polarization, filter bubbles, and the formation of cliques. Polarization reflects the increasing divergence of societal and political viewpoints, fragmenting ideological landscapes into opposing extremes (DiMaggio et al. 1996). In the realm of social media, this phenomenon is amplified by algorithmic

✉ Pratyush Arya
aryapratyush@gmail.com

[1] Indian Institute of Technology, Kanpur 208016, India

Published online: 23 January 2025

personalization, transforming it into a potent force that drives societies towards divisiveness (Zuiderveen Borgesius et al. 2016). In the same vein, filter bubbles, a term birthed by Pariser (2011), encapsulate the unsettling reality of intellectual isolation, which now pervades the World Wide Web. Algorithmically generated digital echo chambers present users with content that aligns with their preexisting preferences, reinforcing self-confirming information loops. Clique formation materializes as individuals, sharing common attributes or beliefs, clustering together in cyberspace, resulting in islands of homogeneity (McPherson et al. 2001). Homophily, an age-old sociological phenomenon, has experienced an exponential surge due to the reduction of friction in communication in cyberspace, exerting profound influences on society, politics, and cognitive processes (Sunstein 2017).

Previous efforts to comprehend and address these phenomena have predominantly adopted social and cultural perspectives, examining how societal structures, media environments, and cultural contexts shape their manifestations (DiMaggio et al. 1996; McPherson et al. 2001). However, there has been a noticeable lack of focus on how these phenomena impact and are influenced by individual cognitive and information processing mechanisms. This gap in the literature signals an uncharted frontier in our understanding of polarization, filter bubbles, and clique formation. The intricate interaction between external stimuli and internal cognitive processes is at the heart of how individuals navigate their social and informational environments. As such, understanding these phenomena from an information processing standpoint is crucial to understand why and when polarization is likely to result in networks of individuals.

In the context of this paper, we operationalize networks of individuals as graphs produced by three different mechanisms, two of which make sociological assumptions: Erdos–Renyi (ER), Barabasi–Albert (BA), and Watts–Strogatz (WS). ER graphs, characterized by random connections between nodes, offer a baseline mathematical graph model, with no socio-cultural appurtenances, for studying network dynamics. On the other hand, BA graphs, generated via a preferential attachment mechanism, exhibit power-law degree distributions, meaning the probability of encountering highly connected nodes is relatively higher. This aligns with the structure of social media and other digital networks, where influential individuals gather larger followings. BA graphs thus have a clear sociological connotation and provide insights into the dynamics of online communities wherein low social friction easily permits large inequalities in the degree distribution of connectivity between individuals. WS graphs, with their small-world architecture, strike a balance between local clustering and global connectivity, reflecting networks in the real-world wherein higher social friction reduces the range of degree distributions accessible to individuals. The small-world property further reflects the interconnected nature of real-world social networks, wherein individuals can establish connections with others through short paths, akin to the "six degrees of separation" concept, without requiring to make a large number of direct connections personally. We examine how social preferences vary over the course of networked consensus games in all three categories of graphs in this paper.

2 Empirical background for this work

The motivation, and the empirical background, for this work comes primarily from Michael Kearns' paper, "Behavioral Experiments on a Network Formation Game". Kearns et al. (2012) The paper talks about a series of behavioral experiments where 36 human participants had to solve a competitive coordination task (of biased voting) for monetary compensation. Communication, in these games, happens only via the game GUI, and only with individuals in one's assigned social neighborhood. It has been found that in such cases, where the social neighborhoods are explicitly fixed, and participants are then asked to achieve a collective goal, human participants tend to perform well - subjects are able to extract almost 90% of the value that is available to them in principle. This has led researchers to conclude that humans are quite good at solving a variety of challenging tasks from only local interactions in an underlying network (Kearns et al. 2009).

However, when Kearns made a slight change to the game, human performance deteriorated. The slight change entailed participants having to build the network during the experiment, via individual players purchasing links whose cost is subtracted from their eventual task payoff. A striking finding is that the players performed very poorly compared to behavioral experiments in which network structures were imposed exogenously. Despite clearly understanding the biased voting task, and being permitted to collectively build a network structure facilitating its solution, participants instead built very difficult networks for the task. This finding is in contrast to intuition, case studies and theories suggesting that humans will often organically build communication networks optimized for the tasks they are charged with, even if it means overriding more hierarchical and institutional structures (Burns and Stalker 1994; Nonaka and Nishiguchi 2009).

These results suggest that humans are able to achieve a collective goal if a network structure is imposed on them, and they are restricted to communicating within the fixed neighborhood itself; however, when they are free to choose people to communicate with, instead of selecting people that will maximize the chances of global coordination, human participants end up building sub-optimal networks and fail to coordinate effectively.

3 Social preference formation

Central to our model is the assumption that the inference of social preferences occurs through the same information processing mechanisms as the inference of individual preferences. Building upon this assumption, our account relies on two specific information-processing assumptions.

Firstly, we embrace the principle of inductive inference, which posits that individuals make decisions by inferring what to do based on their past choices involving similar options. In our model, agents exhibit this inductive reasoning

by updating their color preferences based on previous interactions and outcomes, thereby gradually adjust their preferences over time, resulting in the emergence of distinct color clusters.

Secondly, our model incorporates the concepts of memory growth and memory decay. Inspired by the workings of human memory, we assume that agents' memories of past interactions can both strengthen and fade. Memory growth reflects the reinforcement of memory traces associated with interactions that led to similar color preferences, promoting the formation of social ties with like-minded individuals. On the other hand, memory decay represents the natural process of forgetting, allowing agents to adapt and respond to changing social dynamics. These memory dynamics contribute to the evolution of the network structure and the emergence of distinct color clusters in the dynamic network case.

By integrating inductive inference, memory growth, and memory decay into our model, we aim to provide a more comprehensive understanding of how cognitive processes shape social behavior. While our model is a simplified representation of complex human decision-making, it offers insights into the mechanisms underlying social preferences and network dynamics.

3.1 Preference inference per iteration

There is now substantial evidence to believe that inductive inference underpins the construction of several (if not all) mental attributes (Tenenbaum et al. 2011). This Bayesian approach to cognition was recently applied to the problem of preference learning (Srivastava and Schrater 2012). Following their notation, an agent's preference for an option is identical to the probability that it is desirable, p(r|x), and can be calculated by summing out across evidence of desirability observed in multiple contexts,

$$p(r|x) = \frac{\sum_{c \in C} p(r|x,c)p(x|c)p(c)}{\sum_{c \in C} p(x|c)p(c)} \tag{1}$$

Here C is the set of all contexts offering x as a possible choice. The desirability probability p(r|x, c) simply considers the frequency with which the agent had previously preferred option x in context c, the option probability p(x|c) expresses the frequency with which the option x is observed in context c, and the context probability p(c) expresses the base rate of context c in the agent's environment.

3.2 Memory decay and memory growth through iterations

In the context of the model, memory decay and memory growth are parameters that control how the memory matrix evolves over time. The role of these parameters comes in particularly in the case of dynamic network.

Memory decay signifies the gradual decrease in the strength of an agent's memory of past interactions. It models the natural forgetting process in human memory.

A higher memory decay rate means that memories of past interactions fade more quickly, while a lower decay rate means that memories persist for a longer time.

$$new\_memory = memory \times (1 - memory\_decay) \tag{2}$$

Memory growth, on the other hand represents the strengthening of an agent's memory of past interactions that have led to similar color preferences. It captures the idea that repeated experiences of similarity reinforce memory traces. A higher memory growth rate means that agents are more likely to remember and interact with agents who have similar color preferences, while a lower growth rate means that memory is less influenced by past interactions.

$$\begin{aligned} new\_memory[i,j] = memory[i,j] + \\ (similar\_preferences[i,j] \times memory\_growth) \end{aligned} \tag{3}$$

where:

- new_memory[i, j] is the updated memory value for agent i's memory of agent j,
- memory[i, j] is the previous memory value for agent i's memory of agent j,
- similar_preferences[i, j] is a measure of the similarity between agent i's and agent j's color preferences,
- memory_growth is a parameter controlling the rate at which memory is reinforced.

We introduce an exponential decay factor to the memory distances, which represents the influence of memory decay (Ramamurthy et al. 2006). The memory weights are then calculated as the product of the exponential decay factor and the corresponding memory values between agents. This way, we emphasize stronger memories while accounting for the decay process.

The use of the exponential decay factor ensures that closer memory distances and stronger memory values lead to higher memory weights, indicating a higher probability of selecting an agent as a neighbor. The normalization step ensures that the memory weights sum up to 1, providing a valid probability distribution for neighbor selection. In doing so, the neighborhood selection process takes into account both memory growth and memory decay, resulting in the formation of connections based on the strength and recency of agents' memories.

Let's go through the structure and flow of the simulations to understand exactly where these mechanisms are invoked, and how.

4 Simulation details

We run our simulations in two conditions - fixed and dynamic. For both conditions, agent initialization, preference assignment, and neighborhood assignment are the same. The important difference that comes into picture in case of dynamic conditions is that the agents are free to sample and select their own neighbors. This process is driven by homophily.

The simulations begin with initialization of ER, BA, and WS graphs having 36 nodes each. Each of these nodes represents an agent, and at the outset, is randomly assigned preference for one out of four colors. The agent in our simulations has three major attributes - preference, neighborhood, and memory. Preference is initialized as a vector of 0 s, with 1 at the preferable index. Neighborhood is represented as an adjacency matrix of agents. Likewise, memory is also initialized as an adjacency matrix, to indicate the memory strength being the highest between two agents that have a link from one to the other. The memory matrix evolves over iterations using two parameters - memory decay and memory growth - both of which have been discussed earlier. These calculations give us the updated memory matrix for each agent, which is then used for finding new neighbors for the next iteration (in the case of dynamic condition). The overall goal of the simulations is to record consensus, and see how the trends of consensus or convergence vary among different graph types. So, the task is for all agents to converge to one color, in both fixed and dynamic conditions. Check Fig. 1 for a quick walkthrough of the simulations.

In the fixed network conditions, the neighborhood of an agent is fixed from the beginning - it is the neighborhood that was assigned during network initialization. As a result of this, the agent is forced to interact only with their immediate neighbors. The agents update their preferences based on the preferences of their immediate neighbors, based on the preference inference mechanism described above. This goes on for a fixed number of iterations (which is kept at 50 for most simulations).

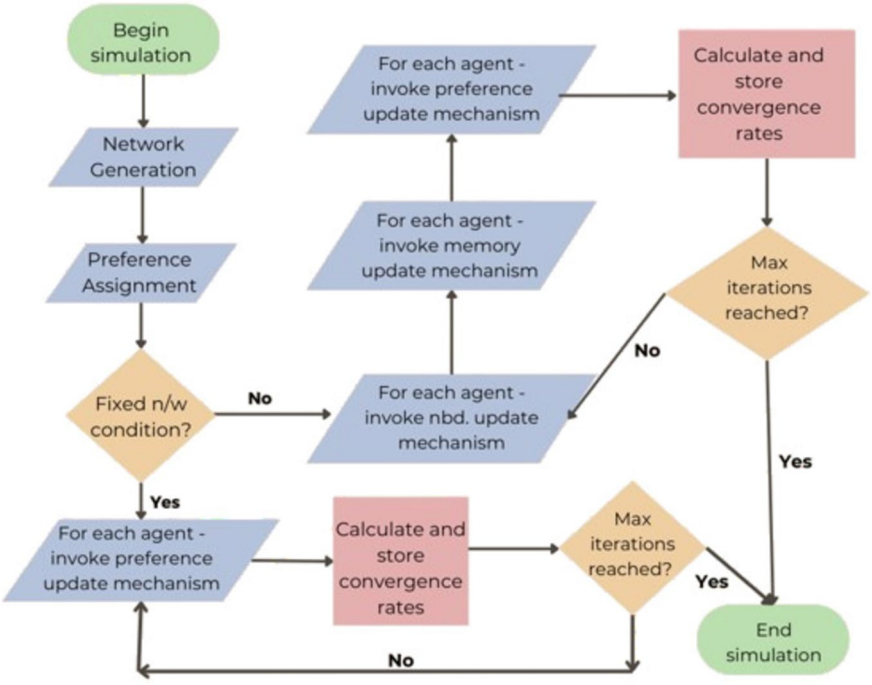

Fig. 1 Flowchart of the simulation

The convergence is calculated and recorded across iterations and across simulations to get a holistic idea of the trend.

In the dynamic network condition, on the other hand, first the neighborhood update mechanism is invoked for each agent, keeping in context the memory till the previous iteration. The neighborhood update happens based on the distance between agents (which is used as a measure of similarity between agents). The idea here is to facilitate connection and link building between agents with similar preferences. Then, based on the new neighborhood, the memory matrix is updated. This is done using equations 2 and 3. The similar preferences parameter that is used is a measure of how similar two agents are. The similarity of preferences plays a role in increasing the memory strength between two agents. This increase happens by a factor of memory growth, which can be played around with to increase the effect of homophily on memory update. This updated memory then becomes the basis for neighborhood update for the next iteration, and this goes on for a fixed number of iterations.

Memory strength is bounded between 0 and 1 to prevent numerical instability, and also consistent with known limitations on the upper bound of memory strength for human memory. Memory growth and decay parameters were selected for various network sizes by trial and error, with values of between 0.01 and 0.1 for the memory decay parameter and between 0.1 to 0.5 for the memory growth parameter producing results consistent with those presented in this paper. All simulation and analysis code relevant to the reproduction of these results are available at this OSF repository to enable researchers to follow our parameter choices and reproduce our results exactly.

4.1 Distance metric

In this model, memory-based neighbor selection is guided by the Euclidean distance between agents' preferences. The Euclidean distance is computed using the formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

where d_{ij} represents the distance between agents i and j, and p_{ik} and p_{jk} are the preference vectors of the agents. These distances are used to rank potential neighbors, with agents selecting their closest neighbors based on these computed values. This ensures that agents with more similar preferences are prioritized in the neighborhood formation process, aligning with the principle of homophily.

Unlike what might be expected in other models, the current implementation does not apply an exponential decay function to these distances when selecting neighbors. Instead, raw Euclidean distances are directly used for sorting and identifying the closest agents. However, an exponential decay transformation could be incorporated in future iterations to further emphasize the impact of proximity in preference space.

For memory updates, the model uses cosine similarity between agents' preferences, reflecting the psychological basis of similarity in cognitive alignment. The cosine similarity is computed as:

$$\text{similarity}(i,j) = \frac{\sum_{k=1}^{n} p_{ik} p_{jk}}{\sqrt{\sum_{k=1}^{n} p_{ik}^2} \sqrt{\sum_{k=1}^{n} p_{jk}^2}}$$

This similarity measure allows for the reinforcement of memory between agents with highly aligned preferences. Together, the use of Euclidean distances for neighborhood selection and cosine similarity for memory updates ensures that the model dynamically adapts to both spatial proximity and cognitive alignment, providing a robust framework for the evolution of social preferences.

4.2 Agent link changes in dynamic settings

In dynamic networks, agents adapt their links based on the evolving psychological salience of their interactions and the principle of homophily, which emphasizes preference similarity. Each agent is limited to a maximum number of neighbors (e.g., 4 in a network of 36 nodes, or 10%). This constraint ensures the network remains sparse and interpretable while reflecting real-world limitations on social connections. Link formation is driven by a memory-based probabilistic mechanism where memory weights for potential neighbors are calculated using an exponential decay function that incorporates both the psychological similarity (preference distance) between agents and the strength of their past interactions. These weights are normalized to ensure they form a valid probability distribution, with agents more likely to connect with those who share similar preferences and have stronger interaction histories.

When agents reach the maximum number of allowable neighbors, they periodically reassess their network connections. In such cases, agents evaluate their current neighbors against potential new ones based on updated memory weights and preference similarities. If a new neighbor with higher salience or similarity is identified, the agent replaces one of their existing connections with the new one. This dynamic mechanism ensures that agents' networks evolve in response to changes in their preferences and interactions, reflecting the adaptive nature of social ties. Memory growth strengthens connections with similar neighbors, while memory decay allows the fading of less relevant connections over time.

This dynamic link adaptation mechanism, combined with the homophily-driven selection of neighbors, promotes the formation of clusters based on psychological salience. It should be noted that there is no explicit cost for link formation for the agents in our network – which one would expect in real-world settings, in the form of cognitive costs if nothing else. Considering this is a simplistic model, we have not added incentives or punishments for link building, but doing that surely constitutes a future direction for this project.

5 Demonstrations and results

In a typical consensus game, members of a group are permitted to preferentially assign themselves one of a small set of colors, but the entire group is rewarded if it eventually converges to one color. Kearns et al. (2012) finds that people are very good at maximizing the group's welfare across a variety of network structures and incentives, so long as the set of their neighbors is held constant: human subjects achieved approximately 90% of the theoretically maximum payout attainable by a perfectly coordinated group.

To assess the behavior of our social preference learning agents, we simulated an environment containing 36 agents, each randomly endowed with one of four color preferences. In other words, for a given agent i, the initial $p_i(r|x) = 1$ for one x, and $= 0$ for the three other xs (colors). The agents could interact with any of the other agents in a sequence. The possible agents with which the initiator i interacts with are, from his perspective, the context; thus, interaction partners (responders) are considered c and the interaction is selected by sampling the available neighbors. For simulations using fixed networks, each agent's neighborhood was specified and it could not be changed during the course of the iterations. During an interaction, the responder indicates to the initiator his preferred color ($\arg\max[p(r|x, c)]$), and the responder received no information. At each time step, the initiator updates their own color preferences by marginalizing across the preferences expressed by their neighbors using the preference inference computation mentioned earlier.

We simulate neighborhoods randomly using all three types of graphs - ER, BA, and WS - 1000 times, and report results using the average convergence (the greatest number of nodes converging to a particular color divided by the total number of nodes in the graph at any point in time) obtained for 50 iterations of the consensus game played on each graph for all three categories of graphs mentioned. Even in the absence of an explicitly specified reward for group consensus, our simulation results show that individual agents use the preferences of their neighbors to change their personal preferences, until consensus is reached. See Fig. 2 to get an idea of how the network evolved during the fixed neighborhood condition. On the left you see the initial state of a WS-generated network in the fixed condition. The network starts off with different agents being randomly assigned preference for one out of four colors. As you can see on the right of Fig. 2, the final network state shows all agents reaching consensus to one preference. Figure 3 shows the convergence trend for all three network types in the fixed neighborhood condition. As can be observed, all network types converge to one opinion well within the maximum number of iterations. Consistent with the existing literature (Tang et al. 2013), we find that the color with the greatest representation in the initial condition of each graph wins most frequently (this result simply verifies that under a fixed network structure, our model appropriately propagates beliefs). Out of the 1000 simulations we ran, we find that regardless of the starting network type, the proportion of simulations where initially dominant color ended up winning were high across graph types. We find that 622,

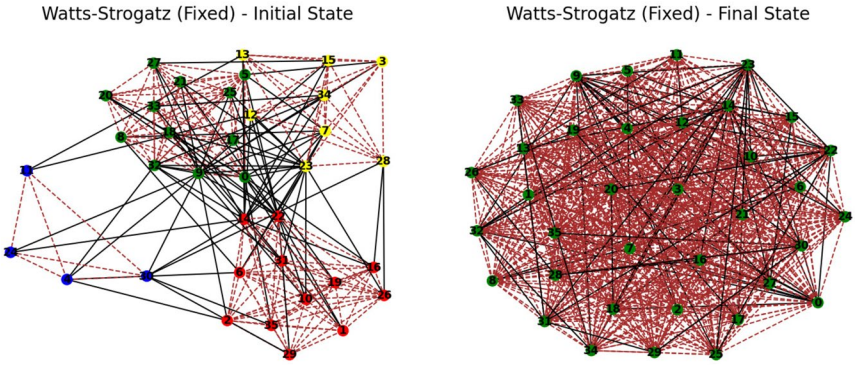

Fig. 2 On the left is the initial network structure for the fixed condition in one of the trials using the WS graph. On the right is the final network after agents iteratively went through a series of preference updates with a fixed neighborhood. We see the network converges to one preference well within the maximum number of iterations for the simulation

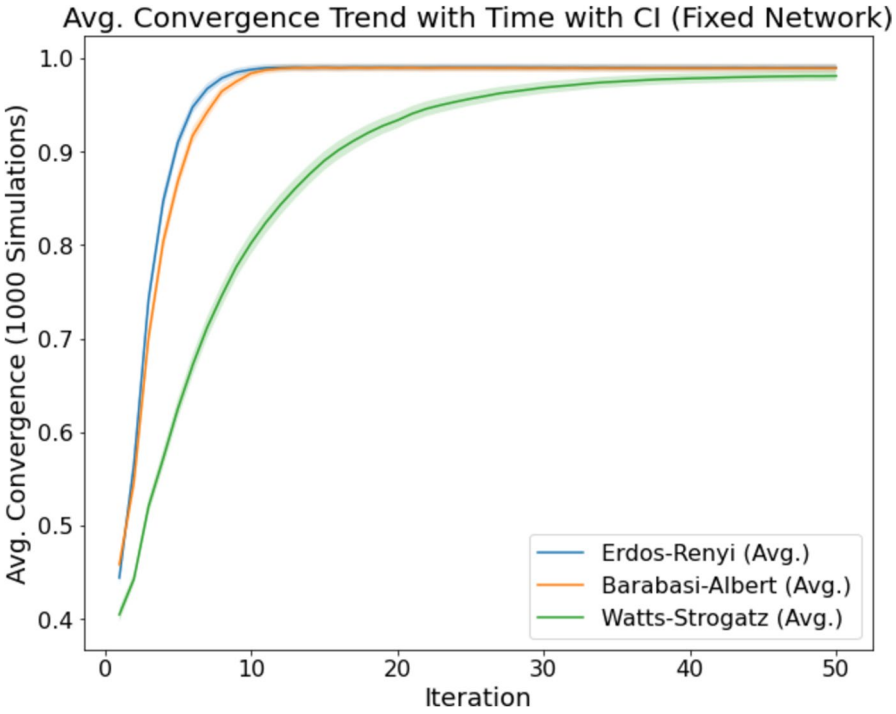

Fig. 3 The plot shows convergence over time for all the fixed network simulations for all three types of graphs. Shaded area represents 95% CI after 1000 simulations. We see convergence for all three graph types

690, and 712 simulations resulted in color with the greatest initial representation winning for WS, ER, and BA networks respectively. It should also be noted that the simulations where initially dominant color did not win were the simulations where intiailly, two colors had almost equal representation - so we see the other color winning in those runs. We also find that the rate of convergence to consensus is directly proportional to the degree of nodes on average in all three types of graphs.

But what happens when agents are free to choose their neighbors? When (Kearns et al. 2009) relaxed the fixed network structure, such that subjects could select which of their neighbors they wished to receive information about, they found that coordination suffered massively, with efficiency dropping to about 40%. It turns out that while humans are extremely good at adapting their preferences to existing network structures, something about the process of social link formation causes this facility of coordination to break down.

We find similar results from our simulation experiment across a broad range of parameter values for memory growth and memory decay. Since network connections were now permitted to be dynamic, agents updated their neighborhoods using encounter information throughout the simulation. At each model iteration, the propensity for interacting with other agents changed, and so did their current preference, using the computation for p(r|x) as above. See Fig. 4 - agents start out with a fixed WS network, and are then allowed to sample from other agents to update their neighborhood and connections based on preference homophily. As a result of this, the final network state (on the right of Fig. 4) turns out to be balkanized. Similar results stand for other two networks - BA and ER - too in the dynamic condition. In the fixed condition, on the other hand, the initial networks converge to one color, as the convergence trend from Fig. 3 shows, which is mostly the majority color of the initial network.

When updating preferences in fixed network conditions, agents performed the computation as suggested by Equation 1, and that was enough to get them to global

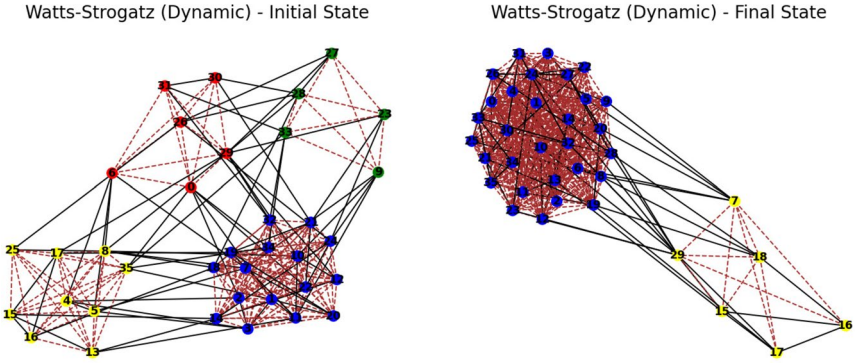

Fig. 4 On the left is the initial network structure for the dynamic condition in one of the trials using the WS graph. On the right is the final network after agents iteratively went through a series of neighborhood updates and preference updates. We see the network stabilize into a balkanized state, instead of converging on one preference, well within the maximum number of iterations for the simulation

convergence - even in absence of any specified rewards. However, in the case of dynamic networks, when agents were free to choose their neighbors in every iteration, agents retain memories of past interactions, enabling them to recall and potentially favor agents with whom they have had shared color preferences in the past. This memory retention allows for the persistence of social ties and the potential formation of clusters based on shared preferences. This contributes to the reinforcement of existing social ties, potentially leading to the emergence of cohesive clusters of agents with similar color preferences. This is the case for ER, BA, as well as WS graphs. However, there is a curious differentiation that can be observed when we look at the convergence asymptote value for the three types of graphs across all simulations and all iterations - see Fig. 5 above.

We see that Barabasi-Albert networks show convergence to a higher asymptotic value compared to Watts-Strogatz as well as Erdos-Renyi networks. Considering the structural differences in how the three graphs are generated, we find an interesting explanation for this difference. What makes the BA graph different from the other two is its degree distribution, which follows a power law - thereby increasing the probability of finding nodes that are thickly connected with many neighbors, compared to ER graphs, where the degree distribution is binomially (approximately normally) distributed. Likewise, with WS, we have a small world structure, yielding a close to uniform degree distribution.

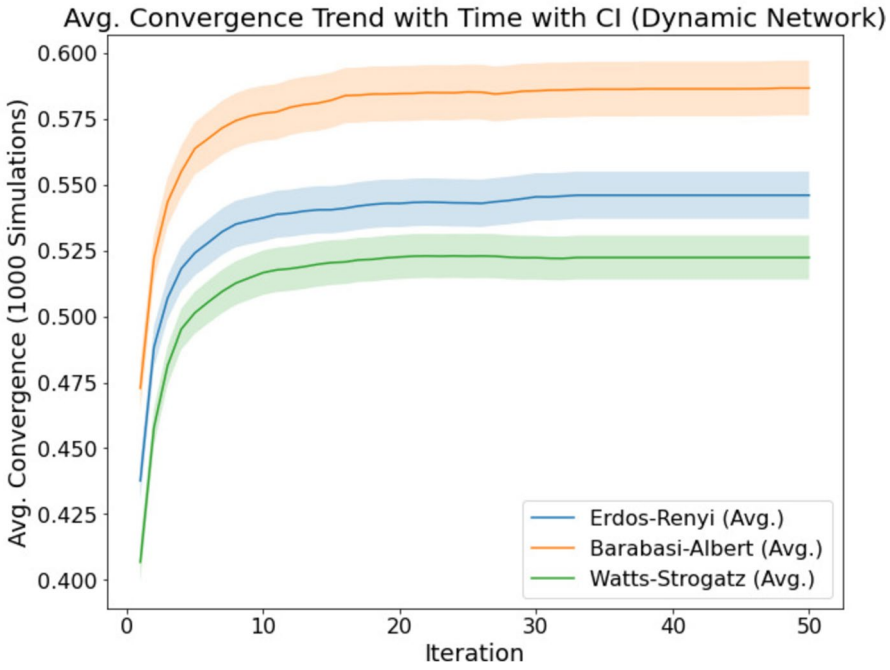

Fig. 5 The plot shows convergence over time for all the dynamic network simulations for all three types of graphs. Shaded area represents 95% CI after 1000 simulations. We see all three network types fail to reach consensus, with the BA-generated networks showing more resilience to balkanization, or a higher degree of convergence, than the other two

For the consensus game, all that matters is the local neighborhood - so, if a node is thickly connected, there is a high chance that it is connected to nodes that have varying colors. If such a node switches over, it's going to have a lot of impact on the rest of the graph. Since we are more likely to see this sort of highly trusted or highly influential node in a BA network than in ER or WS graphs, we see a higher convergence asymptote for BA than ER or WS graphs.

Thus, we find that the same algorithm, when allowed to work with a fixed network structure, performs information coordination efficiently, whereas when allowed freedom to preferentially create local network neighborhoods, agents behave in locally optimal ways that reduce global coordination. We believe these findings explain to a considerable extent the mysterious gap in coordination performance in Kearns' networked game experiments: Agents, and likely humans, assure themselves that they have equilibrated to the consensus preference through sampling the preference of their neighbors. When forced to consider all neighbors, they must necessarily engage with all the information present in their neighborhood; when free to choose, they end up restricting communication with neighbors who share their preference.

These findings, and the crucial difference in network behavior between BA vs. the rest of the graphs, particularly WS, in the dynamic neighborhood condition, led us to hypothesize the role of presence of hubs in the network in facilitating the overall consensus. The presence of hubs, and as a result of it, a power-law degree distribution in BA networks as opposed to WS networks could, perhaps, have something to do with this difference? We tested out this hypothesis by structurally modifying BA- and WS-generated networks to bring about the desired structural changes, i.e., introduce hubs to WS, making the degree distribution more power-law-like and remove hubs from BA, making the distribution more flat, so to say.

6 Role of hubs in overall network consensus

To understand the distinct behaviors of agents in various network structures, particularly in the BA and WS models, we developed a new hypothesis. This hypothesis posits that the presence of hubs, which are highly-connected nodes and a defining feature of the BA model, may significantly influence these behavioral patterns. Drawing on insights from social (network) psychology, we suggest that hubs could have a substantial impact on the dissemination of opinions and preferences within the network.

To rigorously test this hypothesis, we embarked on a series of computational experiments, aiming to modify the structural properties of networks generated by both the BA and Watts-Strogatz (WS) models. The primary objective was to alter the BA networks to make them more homogenous, effectively 'flattening' the network by reducing the prominence of hubs. Concurrently, we aimed to adjust the WS networks to introduce a higher degree of skewness, akin to a power-law distribution, which is typically observed in BA networks. These modifications were intended to examine how changes in network topology affect the dynamics of opinion formation and consensus-reaching processes.

6.1 Implementing network modifications

The modifications to the network structures involved algorithmic alterations to the network generation processes of both BA and WS models. For the BA model, this entailed tweaking the preferential attachment mechanism to limit the growth of hubs. In contrast, for the WS model, we introduced mechanisms to encourage the development of certain nodes into more hub-like entities, thereby inducing a skewness in the network's connectivity distribution.

Here's how we went about implementing these modifications to WS and BA generated graphs. For BA graphs, we started by finding the highest-degree node. Then, we removed one of its edges. Then, to maintain the flatness of the degree distribution, we find the lowest-degree agent and connect it to a random agent, based on a coin toss. Based on this implementation, we were able to get a flatter degree distribution for graphs generated using BA generation mechanism. Figure 6 shows how the degree distributions changed for the BA graph before and after implementing these structural modifications.

Likewise, for the WS networks, the goal was to make the degree distribution more skewed, close to BA's original degree distribution. To induce such a power-law-like behavior to our WS graphs, we first went about creating hubs in the WS generated graphs. We picked the top few nodes having high-degrees and sought to make them hubs. This was done by adding extra edges to these selected high-degree nodes. We also probabilistically thinned out the medium-degree nodes to make the distribution more BA-like and push more nodes towards having fewer degree. See Fig. 7 for a graphical representation of how the degree distributions changed before and after the modifications.

If our hypothesis about the role of hubs in driving network consensus has to stand, the earlier set of simulations should be run on these modified networks, too. If we find the new set of simulation results to be reflecting the difference in network behavior such that the modified WS networks facilitate more convergence than the modified BA networks, we can safely say that structural properties of the network do indeed play a significant role in driving a network of agents to consensus.

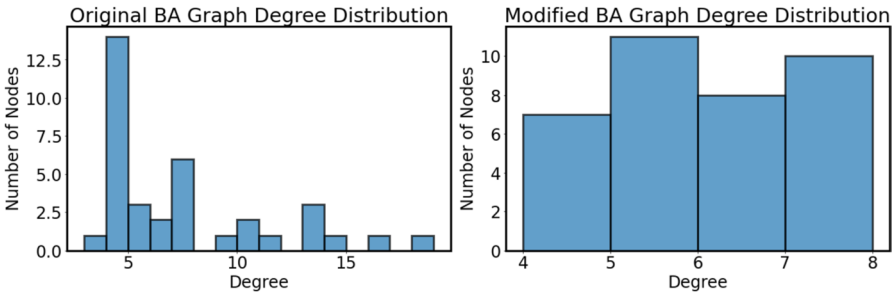

Fig. 6 The graph on the left shows the degree distribution for a BA generated network with 36 nodes. The graph on the right shows the degree distribution of the structurally modified BA generated graphs

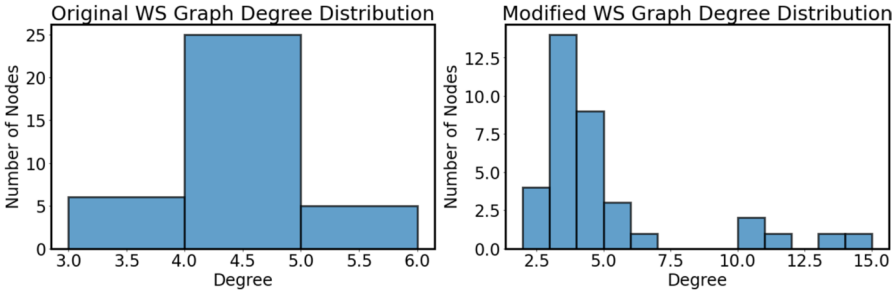

Fig. 7 The graph on the left shows the degree distribution for a WS generated network with 36 nodes. The graph on the right shows the degree distribution of the structurally modified WS generated networks

6.2 Running simulations on modified networks

Once the graph were restructured as per our requirements, using the modification mechanisms described above - such that the BA networks now showed BA-like properties, and vice versa - we ran the same two-conditions simulation on the new networks. As earlier, in the fixed condition, agents had a fixed assigned neighborhood from the start, with no flexibility to change it. In the dynamic condition, the neighborhood selection was based on homophily and on memory mechanisms. Again, the goal of the game was overall convergence of the network.

Our hypothesis predicts that keeping everything else the same, if we just run the modified networks through our simulations, things should remain essentially unchanged for the fixed network condition, while there should be some more convergence for WS networks, and less for BA networks. Essentially, the behavior of modified-BA should mimic the behavior of original WS in the previous simulations, and likewise for the modified-WS networks, since the structural properties have been changed so. The results of simulations, averaged out over 1000 runs, for the fixed condition are shown in Fig. 8. The trend shows the progression of how converges evolves in the network as a function of iterations. As the prediction would go, all three network types converge. But interestingly enough, this time, the WS network converges before the BA network. The WS network also seems to converge way quicker than the original, un-modified-WS network, as shown in Fig. 2.

In the condition where agents are free to choose their neighbors per iterations, i.e., the dynamic condition, we find that the effect of hubs and lack of them shows up in the overall convergence trend. As Fig. 9 would show, The modified-WS network converges quicker than the BA network this time, with both having modified structural properties. The results remained consistent across a wide range of number of simulations over which the results were averaged. Further, the modified WS network seem to be performing almost equivalent to the actual BA networks that were used in the previous set of simulation runs. Likewise, the modified BA networks seem to be faring similar to the original WS networks. In the previous simulations, the convergence trend graph for the BA networks asymptotes to an average convergence of close to 58%, while for WS it does at close to 51%. In the simulations with modified

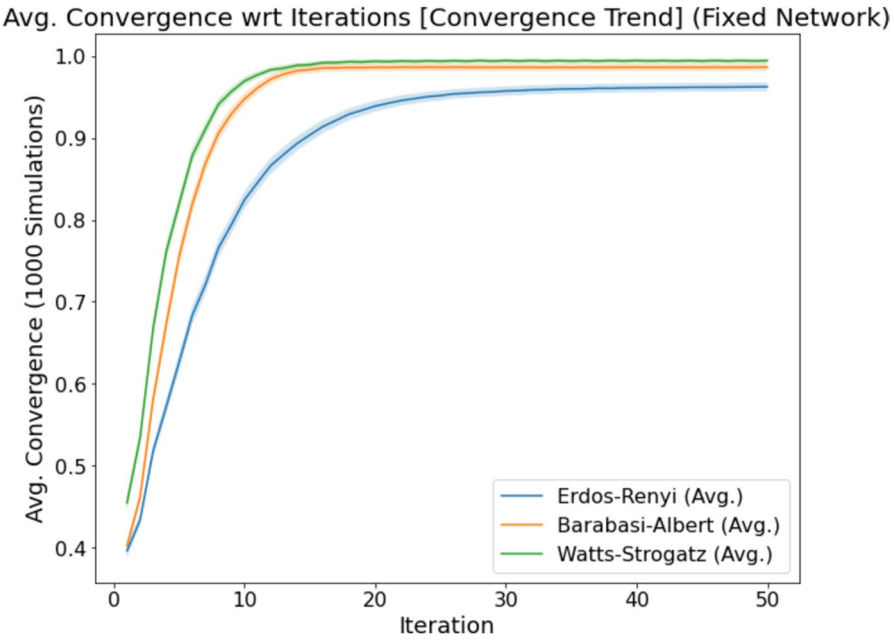

Fig. 8 BA and WS here refer to structurally modified BA and WS networks. The plot shows convergence over time for all the fixed condition simulations for all three types of graphs, including modified BA and WS. Shaded area represents 95% CI after 1000 simulations. We see convergence for all three graph types

network structures, the modified-BA graph outperforms and the convergence graph asymptotes at close to 56%, as compared to close to 65% for the modified-WS graph.

The convergence trends plotted above give a picture of how the convergence evolves across iterations for each network type. In simple terms, the graph represents the trend of how the network state changes from the first iteration to the last iteration, i.e., how the convergence of the network changes iteratively, averaged over all the simulation runs. So, the convergence trend plot for a particular network type asymptoting at, say, 0.60, means that for that network type, on an average across all simulations, close to 60% of the agents ended up converging to one color.

Apart from this, we also plotted average convergence across simulations on a line graph, to give a better visualization of the rise and fall of convergences across network types, with and without structural changes. The interpretation of values on this plot is the same as that of the convergence trend plot, just that the latter gives an overall, iteration-by-iteration, picture of how convergence evolved, while the former just plots the average proportion of agents that ended up converging to one color averaged across all simulations. Figure 10 and 11 give a visualization of the said line graph showing the final value at which each network type asymptotes, averaged over 1000 simulation runs. We see the iteration-wise trend reflect in the average convergence over simulations picture, too - and reasonably so. In earlier simulations, the average convergence value for BA networks was consistently more than that for WS networks, both in the dynamic as well as the fixed neighborhood conditions. This

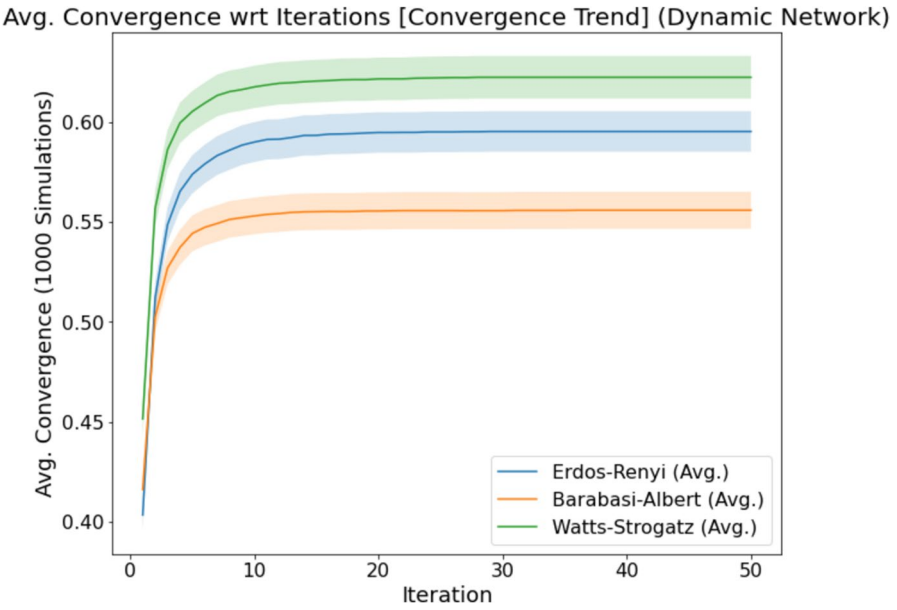

Fig. 9 BA and WS here refer to structurally modified BA and WS networks. The plot shows convergence over time for all the dynamic condition simulations for all three types of graphs, including modified BA and WS. Shaded area represents 95% CI after 1000 simulations. We see the modified-WS graphs outperform the modified-BA graph in terms of overall convergence across simulations

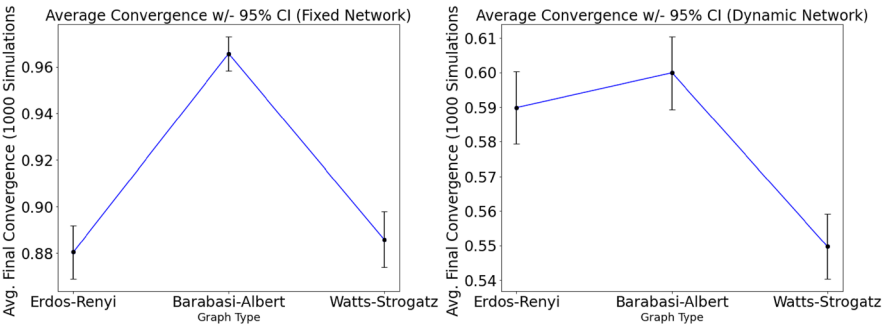

Fig. 10 BA and WS here refer to the original, unmodified networks generated using respective mechanisms. The plot shows the average convergence after 1000 simulations for three graph types - ER, original BA, and original WS. The error bars represent 95% CI. As we can see, with structural properties intact, BA networks outperform WS networks in consensus games

trend is depicted clearly in Fig. 10 which plots the average convergence after 1000 simulations for all three graph types, including original BA and WS networks.

After structural modifications, now WS networks had a few nodes with a high-degree, essentially hubs, and a greater number of nodes with a low-degree. This introduced hubs in the networks created by the WS generative mechanism. The modified BA networks, on the other hand, lost their power-law degree advantage,

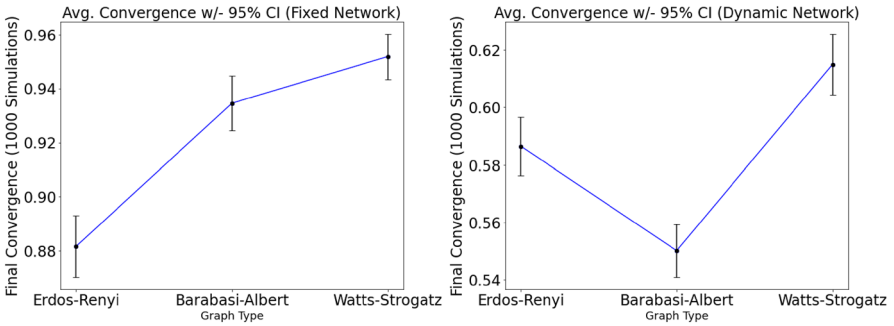

Fig. 11 BA and WS here refer to structurally modified BA and WS networks. The plot shows the average convergence after 1000 simulations for three graph types - ER, modified BA, and modified WS. The error bars represent 95% CI. As we can see, now with structural modifications, WS networks are outperforming BA networks in consensus games

and had a more flatter distribution. This introduced a removal of hubs from the networks created using the BA generative mechanism.

As Fig. 11 depicts, so does Fig. 9 above, these structural changes brought about hypothesized behavioral changes in consensus games. We see that the average convergence for original BA networks is more or less comparable to the modified WS networks, and vice versa. A quick side-by-side view of Fig. 10 and Fig. 11 will give you a clearer sense of this claim. Without structural modifications, BA-generated networks converged to an average value of 60 These results, to a considerable degree, support the hypothesis about the influence of structural properties of the network on global network convergence, generally, and the importance of presence of hubs in the network in driving overall consensus, particularly.

The next interesting question, or area of exploration, from here could be to understand hub behavior and hub dynamics. It is evident that hubs are somehow facilitating more convergence, and therefore their presence is making the network more resilient to network balkanization. Unearthing the mechanisms behind this facilitation could be interesting, and easily doable with slight modifications to our model. In terms of how hubs have been defined in the literature, the hubs that are playing a role in our simulations - both in the case of original BA and modified WS networks - can be seen as information mavens, having a large in-degree, which enables them to aggregate information. If many agents in the network are close to such information mavens, asymptotic learning is said to happen. Since the network structures for now don't differentiate between inbound and outbound connections, having only bidirectional connection as of now, it is difficult to say whether these hubs could be acting as social connectors, too, which have a large out-degree, and that enables them to communicate information to a large number of agents. Social connectors are presumably useful for asymptotic learning if they are close to mavens, so that information distribution can happen easily. Acemoglu and Ozdaglar (2011) All of these claims can be tested with some modifications to our proposed model of social learning and preference formation.

7 Discussion

In this paper, we used a memory-based model of social preference learning to reproduce both the success and failure of agents to attain consensus in a networked game, based on whether agents were permitted to select their social neighborhood. We showed that networks of agents forced to play with neighborhoods assigned to them nearly always converged to a consensus color in the game, although this process was slower for Watts-Strogatz small-world neighborhoods. We also showed that networks of agents permitted to create their own neighborhoods failed to converge to a consensus, with Barabasi-Albert style preferentially attached networks reaching more majority consensus than alternative types.

We also saw one possible reason - appealing to the network structure - of such a difference in behavior between BA and WS networks in the dynamic network condition. We tested the hypothesis that the presence of hubs could be a factor in pushing BA to convergence but not WS. Our modified-BA and modified-Ws networks had degree distribution similar to actual WS and actual BA network, respectively. We then ran the same simulations on modified networks. We found that the effect of hubs shows up in two important ways. One, the WS network, in fixed condition, now converges significantly faster than it did before the structural modifications. Two, in the dynamic condition, the behavior of modified BA and WS was essentially a mirror image of their behavior without structural modifications. These results could potentially point towards the importance of hubs in networks in order to drive global convergence, or at least facilitate higher convergence levels across the network.

In this work, we have relied on a specific formulation of preference learning as the primary container of our modeling results (Srivastava and Schrater 2012). It is quite possible to expand the model to take additional aspects of preference formation, such as the role of alternative reward structures, into account. It is also possible to use alternative mathematical specifications for a preference learning model, such as Instance-Based Learning (Gonzalez et al. 2005), or Adaptive Control of Thought - Rational (ACT-R) (Anderson et al. 2004). However, models are ultimately just vehicles for psychological assumptions - and it is these assumptions that determine how accurately the model can predict phenomena in the real world. Thus, the choice of a specific model does not fundamentally change our conclusions, so long as the assumptions that guide our model are valid and are themselves representative of the phenomena we seek to understand.

Social networks have previously been analyzed using models rooted in sociology and mathematics, such as the Friedkin-Johnsen (FJ) model (Noah 1999) or the DeGroot model (Morris 1974), both of which mathematically model opinion change as arising from a consideration of the average of neighbors' opinions. Beyond these, other models have also tried to look at opinion dynamics using similar averaging mechanisms. Another example is the Hegselmann-Krause (HK) model (Hegselmann and Krause 2002) that introduces bounded confidence, where agents only average opinions within a certain distance of their own. Similarly, the Lehrer-Wagner model (Lehrer and Wagner 1981) incorporates rational

deliberation into the averaging process, providing insights into collective decision making. These models share a common underlying principle - description of opinion evolution as an iterative averaging process.

However, inspired by Kearns' experiments, we sought to model a scenario where opinion dynamics and network relationships co-evolve. To do so, we propose a model of social preference formation that involves active sensing of a limited number of neighbor opinions. This approach shifts epistemic focus from opinion dynamics already well-modeled by the FJ class of models to a more general view of the formation of social preferences, particularly emphasizing how cognitive mechanisms drive network evolution. Another key distinction between our model and opinion averaging models previously seen in the literature lies in the treatment of social influence. The FJ model, for instance, incorporates stubbornness, where agents retain a fixed weight for their own opinions while averaging neighbors' opinions based on social influence (Noah 1999). In our model, stubbornness is replaced by salience-weighted averaging, where the psychological relevance of interactions plays a pivotal role. This salience is determined by factors like the similarity of preferences, the strength of past interactions, and the decay of memory over time, drawing from principles of human cognition and memory research (John 1991), recognizing that individuals do not merely conform to social norms but are actively influenced by the perceived importance of specific social cues and experiences. By incorporating psychological constructs, our model bridges the gap between sociological and psychological approaches to network analysis. It thus offers a new perspective on how cognitive mechanisms such as inductive inference and memory modulation shape social preferences and influence clique formation. This focus on psychological salience and preference dynamics provides, as we see in our simulations, new insights about the role of social preferences in the emergence of echo chambers, as well as the role of high-degree nodes in the fate of consensus formation in self-selecting social networks.

A natural followup of this work could be to create networked game experiments to test our model's predictions about the interaction between network types and the drive to consensus. The results of such experiment could possibly inform the further development of computational models and our overall understanding of the role of structural properties in clique forming behavior in social networks. If the importance of hubs can be shown to exist with human participants, one could then possibly probe deeper into hub dynamics and see how hubs interact with other nodes, and how opinions spread across the network. This knowledge could then possibly be used to drive more convergence and think of ways of translating this work into actual human social networks. One big assumption in all of this, however, is that agents (humans or otherwise) are open to changing their opinions based on the opinions of their neighbors.

Our findings have theoretical as well as practical implications for enhancing group efficiency and cohesion, particularly in addressing the challenges posed by clique formation and balkanization. By understanding the mechanisms underlying network dynamics and their impact on group behavior, it may be possible to design social media platforms and online communities that foster a less balkanized environment. In particular, our results show that it is not necessary to impose fixed

networked structure to prevent balkanization. The presence of highly connected nodes in networks also protects communities from failures in consensus, so long as these nodes are open to changing their colors based on observing their local neighborhood's majority view. Interestingly, these results are consistent with recent empirical work showing that the effect of filter bubbles in large-scale social media may be overstated (Dahlgren 2021).

References

Acemoglu D, Ozdaglar A (2011) Opinion dynamics and learning in social networks. Dyn Games Appl 1(3):49

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111:1036

Burns T, Stalker GM (1994) The management of innovation. Oxford University Press, Oxford

Dahlgren PM (2021) A critical review of filter bubbles and a comparison with selective exposure. Nordicom Rev 42(1):15–33

DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? Am J Sociol 102(3):690–755

Friedkin NE, Johnsen E (1999) Social influence networks and opinion change, Advances in Group Processes

Gonzalez C, Lerch FJ, Lebiere C (2005) Instance-based learning in dynamic decision making. Cogn Sci. https://doi.org/10.1016/S0364-0213(03)00031-4

John R (1991) Anderson and Lael J. Schooler, Reflections of the environment in memory. Psychol Sci 2(6):396–408

Kearns M, Judd S, Tan J, Wortman J (2009) Behavioral experiments on biased voting in networks. Proc Natl Acad Sci U S A 106:1347–1352

Kearns M, Judd S, Vorobeychik Y (2012) Behavioral experiments on a network formation game, Proceedings of the ACM conference on electronic commerce

Lehrer K, Wagner C (1981) Rational consensus in science and society: a philosophical and mathematical study. D. Reidel Publishing Company, Berlin

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Socio 27:415–444

Morris H (1974) DeGroot, reaching a consensus. J Am Stat Assoc 69(345):118–121

Noah E (1999) Friedkin and Eugene C. Johnsen, Social Influence Networks and Opinion Change, Advances in Group Processes 16:1–29

Nonaka I, Nishiguchi T (2009) Fractal design: Self-organizing links in supply chain management, In: Knowledge Creation: A Source of Value, Ed. St. Martin's Press, p. 199-230

Pariser E (2011) The filter bubble: what the Internet Is hiding from you. Penguin Press, NY

Rainer H, Krause U (2002) Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation, Journal of Artificial Societies and Social Simulation 5

Ramamurthy U, D'Mello SK, Franklin S, (2006) Realizing forgetting in a modifiedSparce distributed memory system, Proceedings of the 28th Annual Conference of the Cognitive Science Society

Srivastava N, Schrater P (2012) Rational inference of relative preferences. Proceedings of Advances in Neural Information Processing Systems 26

Sunstein C (2017) #Republic: divided democracy in the age of social media. Princeton University Press, Princeton

Tang J, Wu S, and Sun J, (2013) Confluence: Conformity influence in large social networks, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 347-355

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: Statistics, structure, and abstraction. Science 331:1279–1285

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157):1124–1131

Zuiderveen Borgesius F J et al (2016) Should we worry about filter bubbles? Internet Policy Rev 5:1–6

Pratyush Arya is currently a doctoral student in the department of Cognitive Science at Indian Institute of Technology, Kanpur. His areas of interests are diverse, and include mechanisms of spontaneous thoughts and belief formation at an individual agent level and social preferences, opinion and knowledge sharing, and clique formation at a network of agents level.

Nisheeth Srivastava is an Associate Professor in Computer Science and Cognitive Science at the Indian Institute of Technology Kanpur. His primary research interests include the study of behavior and intelligence across temporal, spatial and agential scales using a combination of computational and experimental methods.