Origins of Bias in Perceived Social Norms from Social Network Formation

Jennifer E. Dannals

WORK IN PROGRESS

DO NOT CIRCULATE

Perceptions of social norms affect individual action but are often biased. The extent of such bias may depend on the manner in which people with different attributes come to be connected in social networks. Here, we study 2160 people in 120 experimentally-generated networks while they play a public goods game and observe others' behavior. After each decision, people report their perception of the average community behavior (the social norm), and we measure the actual behavior, locally and globally in the network. We find that when networks are structured with high similarity (propensity to connect with others who contribute similar amounts to the public good) and "main-effect" attraction (propensity to connect with others who contribute a lot to the public good) central individuals form significantly more biased perceptions of the social norm of contribution. Norms may be misperceived in communities, especially by individuals in particular locations and for particular attributes.

Research Transparency Statement:

Conflicts of interest: All authors declare no conflicts of interest. Funding: This research was supported by University Funding. Artificial intelligence: ChatGPT was used in streamlining and commenting R code. Ethics: This research received approval from a local ethics board (ID: 2000033732).

Preregistration: The hypotheses and methods and analysis plan were preregistered prior to data collection. Materials: All study materials are publicly available Data: All primary data are publicly available Analysis scripts: All analysis scripts are publicly available. All of the above can be found here: https://researchbox.org/4315&PEER_REVIEW_passcode=TALYUJ.

People have a fundamental drive to understand the opinions and behaviors of their peers (Bicchieri, 2006; Tankard & Paluck, 2016). Descriptive social norms, or what most peers do and think, are a powerful influence on behavior across situations and societies (Gelfand et al., 2011). However in many cases, perceptions of these norms are biased, a pattern which has been demonstrated across a diverse variety of behaviors and beliefs, including an over-perception of binge drinking on a college campus (Prentice & Miller, 1993), an under-perception of prevalence of concern for climate change in American adults (Sparkman, Geiger & Weber, 2022), and an under-perception of acceptability of female participation in the labor force in Saudia Arabia men (Bursztyn, González, & Yanagizawa-Drott, 2020). Biased perceptions of social norms are a contributing factor to a wide variety of consequential societal problems (Burtszyn & Yang, 2022; Miller & Prentice, 2016). Understanding how and why these perceptual biases emerge is critical for developing more accurate models of norm perception and for anticipating when, where, and for whom norm misperceptions are most likely to arise.

The perceptions of those in prominent positions in a community are of special importance. Central individuals, defined as individuals with many or well-positioned social ties (Bonacich, 1987), are often influential in their community by virtue of their position offering them more visibility and influence (Kim et al., 2015). Interventions targeting larger proportions of central people have greater ability to change community behavior because central individuals function as reference points for their community (Airoldi & Christakis, 2024; Paluck, Shepherd, & Aronow, 2013). Thus, understanding when central individuals hold biased perceptions of social norms is critical for understanding how social norms come to be misperceived in society.

An individual's perception of a social norm is influenced by two primary factors: the social information they encounter from others and their own behavior (Dannals & Li, 2024).

Prior research suggests that central individuals might be positioned to perceive norms more accurately because of their greater access to social information (Banerjee et al., 2013; Christakis & Fowler, 2010; Paluck et al, 2013) or, alternatively, may be biased if they anchor on their own personal characteristics or behavior which is often non-representative of their community (Aral & Walker, 2012; Feiler & Kleinbaum, 2015; Marks & Miller, 1987; Valente, Unger & Johnson, 2006). We consider and account for each of these influences and propose a new, inter-related pathway. Specifically, we argue that the greater amount of social information available to central individuals may itself be non-representative under predictable circumstances, because central individuals' unusual personal characteristics may attract a biased set of network connections.

In an experiment with 2160 individuals and 120 different networks, we vary the underlying forces of network formation to experimentally demonstrate how central people come to hold biased perceptions in their community under predictable circumstances as outlined in our theory. Our theory focuses on attributes of the norms themselves, and specifically the degree to which the behavior underlying the norm features two widely-studied phenomena: similarity attraction, also called homophily (Fu et al, 2012; McPherson, Smith-Lovin & Cook, 2001), or a propensity to form ties with others who do the behavior a similar amount to oneself, and main-effect attraction[1], or a propensity to form ties with others who do a behavior to a great extent (Goldenberg et al., 2023; Feiler & Kleinbaum, 2015; Feld, 1991). All social norms can be considered to vary with regards to the degree that the relevant behavior exhibits these forces. Some social norms, like drinking alcohol, might feature strong same-effect attraction; people

[1] We use the term main-effect attraction to describe an aggregate category of effects from prior research. These effects include acrophily, or the attraction to extreme political views within one's party (Goldenberg et al., 2023); popularity effects, or the increased likelihood of forming ties with those high in trait extraversion (Feiler & Kleinbaum, 2015); and social status effects, where individuals with social capital are more attractive ties for others in the network (Lin, 1999). This also encompasses the related friendship paradox, wherein a person friends have more friends on average than the person themselves, due to over-representation of those with a high number of ties (Airoldi & Christakis, 2024; Christakis & Fowler, 2010; Feld, 1991; Jackson 2019).

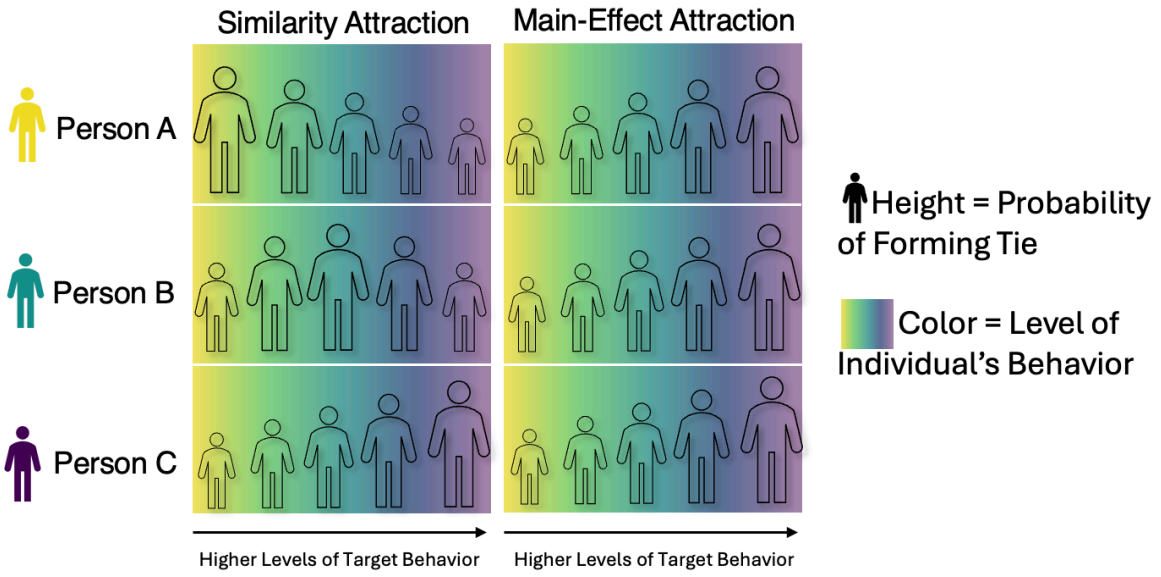
tend to be connected to others who drink a similar amount. Other social norms, like installing solar panels, may feature weaker same-effect attraction; people tend to be as likely to be connected to similar as dissimilar others. Similarly, all social norms vary in the degree to which they feature main-effect attraction. Some social norms, like number of hours worked, might feature stronger main-effect attraction; employees in an organization tend to be connected more often with others who work a lot compared to a little. Other social norms, like saving for retirement, might feature weaker main-effect attraction; people tend to be as frequently connected with others who do this behavior a lot as others who do this a little.

Each force can be either psychological—such that people hold a preference for others who strongly exhibit a behavior or are similar to them, sometimes called choice homophily—or structural—such that the environment makes similar or extreme individuals more salient or available than others, sometimes called induced homophily (Kossinets & Watts, 2009; e.g. when a music school recruits talented musicians who then choose as friends others they find nearby at the school). While not all behaviors are likely to feature a high degree of similarity and main-effect attraction, by considering where a given behavior falls along each spectrum, research can predict when the associated social norm is vulnerable to biased perceptions by central actors.

When a behavior features both similarity and main-effect attraction, we propose that social networks are likely to evolve such that those in central positions have ties to others whose behavior is not representative of the typical behavior in the community, leading to biased perceptions of social norms. The core intuition, as outlined in Figure 1, is that similarity and main-effect attraction act as compounding biasing forces for those in central positions but counteracting forces for those in more peripheral positions, leading the connections of central individuals to have biased social information from connections when these attributes influence

the network. If central individuals have biased social information from their network connections, then we propose their perceptions of the social norm will be biased as well, above and beyond other biasing forces.

Figure 1. Similarity and main-effect attraction shape both centrality and social information for individuals high, moderate, and low on a behavior.



Note: Here we compare how similarity and main-effect attraction of a given behavior shape three individuals' likelihood of forming connections with other individuals in their community as a function of that behavior. If we consider each row laterally, we can see the likely pattern of social information for each individual. Consider Person A. They are low on the target behavior and thus similarity and main-effect attraction balance each other leading to a generally representative sample of all colors in their probable connections. In contrast, Consider Person C. Similarity and main-effect attraction reinforce each other, leading to a likely over-representation of other purple ties. If we consider each column of color, we can see the likely patterns of centrality in the resulting network by summing the heights of each color individual in the gradient. While similarity attraction gives each individual similar probabilities of forming ties with others (assuming a uniform distribution of behavior), main-effect attraction preferences those high in the behavior (purple), leading those higher in the behavior to be more likely to be central in the resulting network.

Our paper makes three contributions to theory. First, we contribute to an ongoing line of work considering when individuals form well-calibrated perceptions of social norms (Savani, et al, 2022; Madan et al, 2025). Where prior work considers personal characteristics such as metacognition and stress reactivity, we complement this by considering aspects of the social

situation, namely when individuals, as a function of network position and social norm features, are likely to have representative social information. Second, we build upon prior work such as the Social Sampling Model (Galesic, Olsson & Rieskamp, 2018) to further theory on which attributes and behaviors are most susceptible to biased perceptions. Where the SSM focuses on distribution shape and level of homophily, we consider levels of both homophily and main-effect attraction and aim to predict differential bias as a function of centrality rather than average levels of bias across a population. Finally, we contribute to research aiming to causally disentangle the reflection problem in social influence (Manski, 1993) by running a network experiment which allows us to calibrate the effect of one's own behavior via social projection, relative to social influence, while still considering the influence of similarity and main-effect attraction in shaping the social information available for such influence.

Methods

We run a large-scale network experiment in order to causally identify the effects of similarity and main-effect attraction in social networks on individuals' perceptions of social norms as a function of the perceiver's centrality. All participants play a five-round networked public goods game. Public goods games are often used in prior research on social norms to model the structure of consequential norms in society (Fehr & Fischbacher, 2004; Ostrom, 2000; Villeval, 2020). Unlike network experiments which seed treatments in existing social networks (e.g. Paluck et al., 2016), we use temporally static networks generated for the purpose of the experiment to allow us to disentangle the causal effects of individual's own behavior from the effects of the information such individuals have access to in the network. Past research has used randomly generated networks with similar goals (Hasan & Koning, 2019; Suri & Watts, 2011), however such research generally fixes network structure and then randomly assigns participants

to that structure without further sorting. While this illuminates the causal effect of network position alone, it does not address the causal role of social information from one's network connections as a function of this position because the network connections in such experiments are random samples absent the influence of similarity or main-effect attraction.

In contrast, we use a novel design in which network structure is not fixed, but rather generated probabilistically from a fixed algorithm. This allows for exogenous variation in network position while also having networks form with experimentally manipulated levels of similarity and main-effect attraction on the same given behavior, participant propensity to contribute to the public good. This approach is similar to Centola (2011), an experiment in which the structure of the network was fixed but the degree of similarity attraction was experimentally varied by shifting participant location in the network. We build on this research by simultaneously varying both similarity and main-effect attraction and by allowing network structure to vary minimally across networks while maintaining general structural similarity across conditions. This experiment allows us to causally identify the effects of similarity and main-effect attraction on perceived social norms as a function of centrality. In all networks, the underlying behavior featuring the experimentally set level of similarity or main-effect attraction is kept constant in order to avoid potential confounds of comparing perceptions of norms for different behaviors across conditions.

All materials, data, code and registrations are available here: https://researchbox.org/4315&PEER_REVIEW_passcode=TALYUJ.

Participants

We recruited 2160 participants, 46.57% Women, 52.64% Men, M_{Age} = 38.64, 70.19% White, 13.66% Black, 7.13% Asian, 5.65% Mixed Race, from Prolific Academic to take part in

the study. Due to the simultaneous nature of the experiment, data collection took place in batches with 4-7 networks collected during each batch on a given day. In each batch, participants were randomly assigned to one of four experimental conditions in groups of eighteen, based on pilot testing for feasibility, with each group then forming its own network, creating 120 unique networks, 30 each per experimental condition across the full data collection in a counterbalanced order at the network-level according to arrival time in the experiment. We found no differences in the order in which networks across experimental conditions filled with players within a given day of data collection, $ps > 0.738$. To ensure that networks promptly filled with participants, extra participants were recruited for each batch of data collection and a researcher monitored Prolific's messaging platform in order to respond to participant queries when waiting for the experiment to begin to ensure participant retention throughout the experiment.

Procedure

Each social network was formed according to a network generation algorithm designed to create networks with high or low levels of similarity and main-effect attraction for the same target behavior in a 2x2 design, while not significantly affecting general features of the network typology across conditions. In all conditions the target behavior used for determining similarity and main-effect attraction was a participant's indicated willingness to contribute to the public good prior to beginning the study. The algorithm calculated dyadic similarity and main-effect scores for all possible dyads and then used these scores, and the assigned experimental condition, to form network ties. For example, in the high similarity and high main-effect condition, the algorithm would increase the probability of individuals forming ties to those with similar intended levels of contribution and also would increase the probability of individuals forming ties with those with a high level of intended contribution.

There are two features to note about this algorithm. First, because our theory focuses on features of the behavior and not features of the network structure, we sought to keep network structure similar across all conditions. To accomplish this, we calculated dyadic similarity and main-effect scores for both the target behavior, contributing to the public good, and then assigned a "control" value that was randomly determined to all individuals to calculate similar, but uncorrelated, "control" dyadic similarity and main-effect scores. When in the high similarity or main-effect attraction conditions, the target behavior score was used to determine ties; when low, the control behavior was used instead. This allows for similar networks to be generated across conditions, but for individuals to emerge in different network positions within these structures as a function of experimental condition and the individual's indicated behavior level.

Second, within this algorithm the formation of network ties is partially stochastic. This means that two ego-alter pairs with identical individual behaviors have the same probability of forming a network tie, but in some cases this tie may manifest, while in others it does not due to random chance. This allows us to causally disentangle the effect of an individual's own behavior from the effect of his or her social information from network ties, while still allowing for those ties to be generated in part because of similarity or main-effect attraction to that individual's behavior. Because tie generation was stochastic, in the event that the algorithm resulted in an individual having no ties at all (n = 9 out of 2160 participants), a tie was randomly assigned for the individual to allow the experiment to continue. For the full algorithm and further description, see Supplemental Materials.

Prior to being randomly assigned to a condition, participants completed a series of four comprehension checks about the structure and payoffs of a network public goods game. All participants were given two tries to answer the comprehension checks correctly in order to

advance in accordance with Prolific's requirements (78.29% of participants completed all four comprehension checks correctly and were able to advance; 4.98% needed a second try on average across all questions). After completing the checks successfully, participants indicated how many tokens they would choose to donate in such a game which then became the target attribute for generating the experimentally determined level of similarity-attraction and main-effect-attraction in the networks. One can therefore consider the high similarity and high main-effect condition as a world in which people who contribute to the public good are of greater salience than those who do not, and one in which people tend to form ties with those who are similar to them in contribution level. In the low similarity and low main-effect condition, individuals were assigned another random value and formed ties based on their similarity and main-effect on the unrelated behavior rather than on contributions to the public good. Note that we do not suggest that contributions to the public good necessarily is a high similarity and high main-effect behavior—individuals might prefer to all connect to high contributors rather than similar level contributors (e.g. Rand, Arbesman, & Christakis, 2011)—however, in our study we are able to simulate worlds in which the behavior takes on these properties in order to examine the impact on perceived norms.

After the network was initialized, participants advanced to a screen in which they learned their randomly assigned animal moniker for the experiment as well as the monikers of all the other participants with whom they had a network tie. Animal monikers were used to help participants remember the other participants in the experiment. (Assigned animal moniker did not impact the results, see Supplemental Table 3 for regressions with assigned-moniker controls.)

Participants then played five rounds of a network public goods game (Suri and Watts, 2011). A network public goods game is similar to a standard public goods game with one notable

deviation. In a network public goods game, the community good is defined at the level of an individual's network ties rather than at the level of the entire group. Thus, an individual can benefit from and offer benefit to only those other individuals with whom they are connected in the network. Participant payoffs followed the following structure:

$$Payoff_i = 100 - c_i + \frac{2}{n}\sum_{j=1}^{n} c_j$$

where c_i is the contribution chosen by participant i and n is the number ties for participant plus one, to account for one's own inclusion in the group. All decisions were incentive compatible, and participants were paid a bonus on the basis of their accumulated tokens at the end of the experiment in order to ensure that decisions were of consequence to all participants.

Participants then decided how much they wished to contribute to the group out of their endowment of 100 tokens, in increments of 25 tokens. Participants had one minute to make a decision before the experiment would auto-advance in order to avoid delays given the simultaneous nature of the study. (In the event of auto-advance, the last indicated contribution would be carried forward, however in all cases in the data collected, participants entered a new contribution in time.) After all participants made a selection, participants viewed a results screen in which the contribution amounts for all of one's ties were listed, as well as the total amount earned as a result of these contributions. We then solicited participant perceptions of the descriptive norm, again displayed for a maximum of one minute, after which the experiment auto-advanced to the next round. Some participants failed to enter a guess on some rounds resulting in 4.63% of participant-round observations recorded as missing. Missing data did not vary significantly by condition, $ps > 0.475$. After the fifth round, all participants completed an exit survey with exploratory measures and were paid.

Measures

Starting Propensity to Donate. After completing comprehension checks, participants were asked, "Round 1 has not yet begun. Before it starts, please indicate using the buttons below how many tokens you wish to contribute to the collective pot. You will then enter the game, learn about your fellow players and begin Round 1." They were given options from 0 to 100 tokens in increments of 25.

Network Centrality. We operationalize network centrality as eigenvector centrality. Following Bonacich (1972), we define eigenvector centrality as:

$$C_i = \frac{1}{\lambda}\sum x_{ij}C_j$$

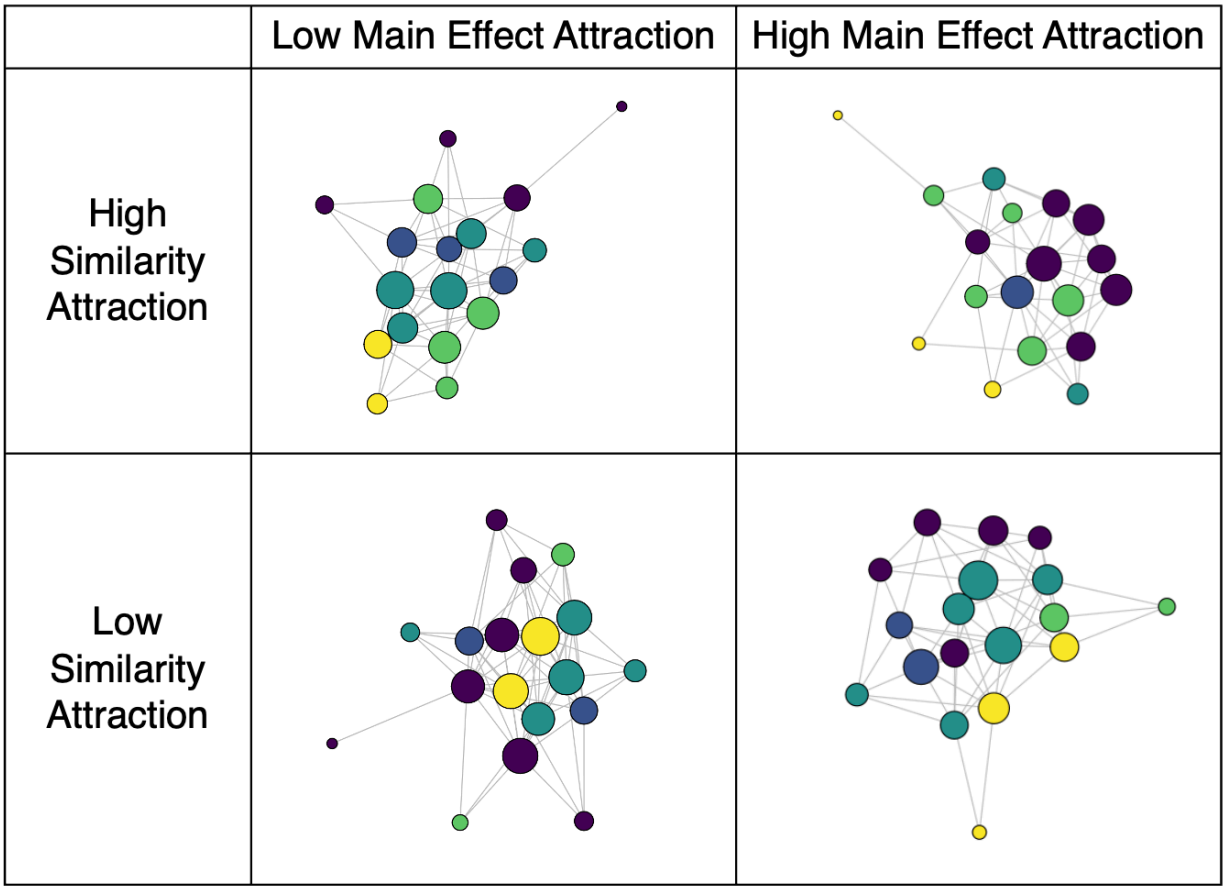
where C_i is the eigenvector centrality of an individual i; λ is a constant; x_{ij} indicates the presence (1) or absence (0) of a tie between person i and each other person j; and C_j is the eigenvector centrality of person j.

Own Contribution. In each of five rounds participants were given a choice of how much to donate to the collective pot from 0 to 100 tokens in increments of 25.

Bias in Perceived Descriptive Social Norms. After each round's decisions, participants learned how other participants to whom they were connected behaved. They were then asked, "Before proceeding to the next round, please enter your best guess of the average amount of tokens contributed in the previous round across all 18 players currently playing the game." Participants were given an open-ended text box with a button next to it labeled, "Submit Guess." Any responses that were not numerical or were below 0 or above 100 triggered an error message prompting participants to update their guess. To calculate bias in this perceived norm, we calculate the true average donation amount by network for each round and subtract this value from participants' entered guesses Positive values therefore indicate overestimation; negative values indicate underestimation.

Additional Measures. All participants completed an exit survey in which they made one final guess of the descriptive social norm and rated their desire to continue playing with each of their network connections in the future. We report results from these exploratory measures in Supplemental Materials. Participant demographic variables were collected from the Prolific Academic Platform directly.

Figure 2: Example Social Networks by Experimental Condition and Starting Participant Attributes



Note: Node colors indicate starting propensity to contribute to the public good prior to network formation, with yellow representing intentions to contribute the minimum number of tokens, and purple representing intentions to contribute the maximum number of tokens, while node diameter represents eigenvector centrality. Though network structure did not significantly vary across conditions, location of players with these attributes shifts to generate higher or lower levels of similarity and main-effect attraction. Note that in the top quadrant those with larger diameter nodes (high centrality) are also more likely to be those with high levels of the behavior (purple) as compared to the other conditions. Note also that in the top row those with similar color nodes are more likely to share ties.

Results

We first checked that our randomization was successful and that our conditions were balanced on observable attributes. A regression with standard errors clustered by network revealed no significant main effects or interactions for starting levels of intended contributions to the public good by assigned condition, *ps* > 0.406.

We further checked to determine that networks generated in each condition did not differ significantly in key structural features. We found no significant main effects or interactions predicting differences for average network degree, *ps* > 0.658, clustering, *ps* > 0.110, modularity, *ps* > 0.235, or diameter *ps* >0.492. See Figure 2 for sample networks and for a visual depiction of all networks by condition and further discussion of the verification of their differences in levels of similarity and main-effect attraction, see Supplemental Materials. We then turned to our pre-registered analyses.

Do central individuals have more biased perceptions of the social norm when the norm features similarity and main-effect attraction?

Our primary pre-registered analysis to test for this bias was a linear regression using the interaction of similarity attraction, main-effect attraction and eigenvector centrality (and all subordinate two-way interactions) to predict bias in descriptive social norm perception in the first-round decisions in the experiment with standard errors clustered at the network level. We chose to focus on only the first round of decisions because later rounds might introduce potential bias given the potential interdependencies between an individual's own behavior and the observed behavior of others in the previous round. In this regression we control for a participant's propensity to donate as measured before the interactive portion of the experiment began. This allows us to causally isolate the effect of network position, similarity attraction and

main-effect attraction. As predicted, we find a significant three-way interaction, $B = 1.08$, $SE = 0.40$, $t(1960) = 2.71$, $p = 0.007$. We then probed this interaction to better understand its form. For individuals with centrality one standard deviation below the mean, there was no significant increase in bias as a result of being part of a network featuring similarity and main-effect attraction, $B = -0.51$, $SE = 0.59$, $t(1960) = -0.87$, $p = 0.387$. However, for individuals with centrality one standard deviation above the mean, the interaction of similarity and main-effect attraction in network formation caused significantly greater bias in perceived norms, $B = 1.66$, $SE = 0.60$, $t(1960) = 2.75$, $p = 0.006$. Furthermore, when both similarity and main-effect attraction were high, higher levels of centrality lead participants to become more biased in their perceptions, significantly overestimating the average contributions in their network, $B = 2.72$, $SE = 0.77$, $t(1960) = 3.53$, $p < 0.001$. However, when both similarity and main-effect attraction are low, centrality is unrelated to bias in perceived norms, $B = 0.33$, $SE = 0.73$, $t(1960) = 0.45$, $p = 0.652$, see Figure 3 for overall pattern.

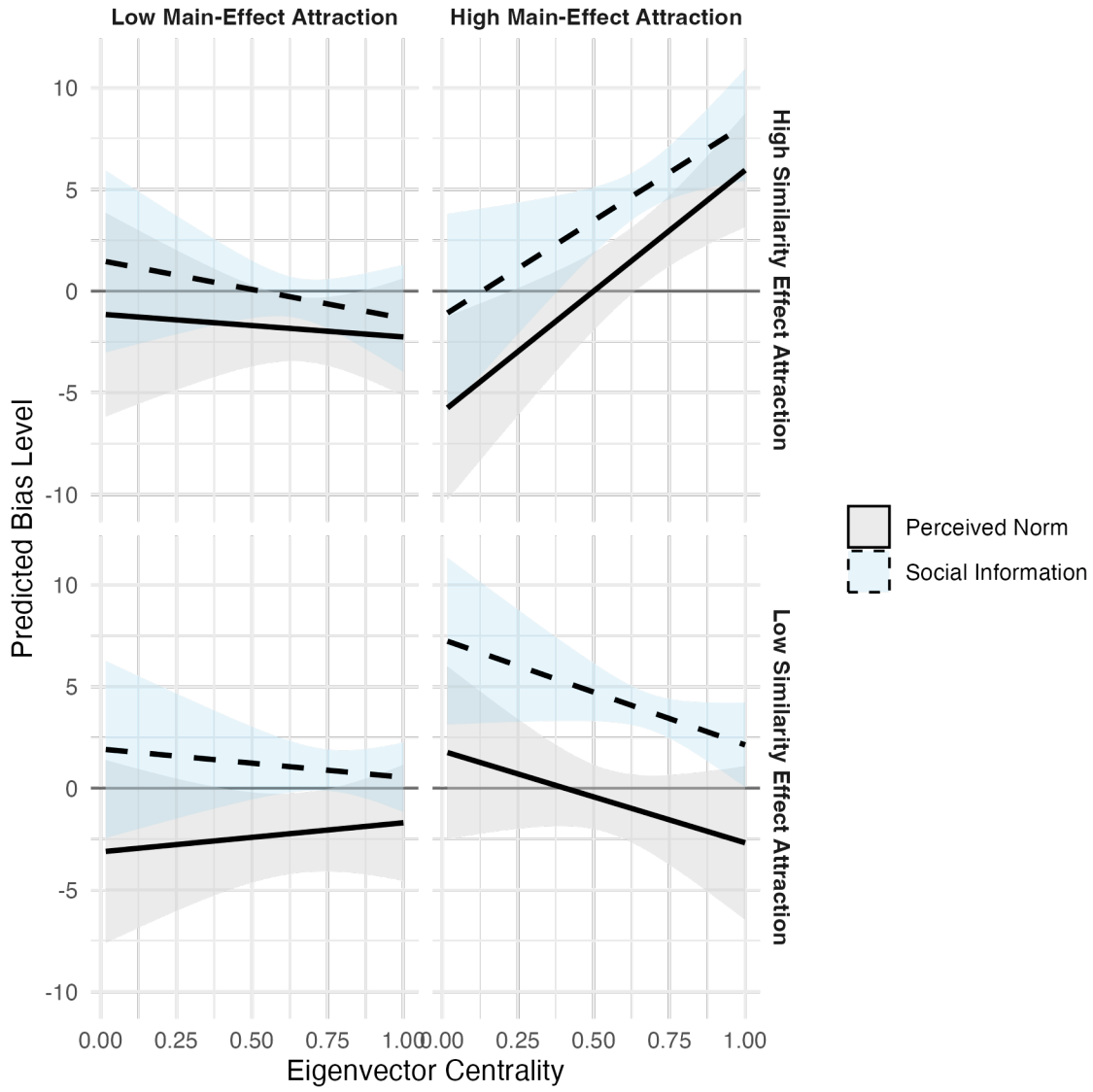
We also pre-registered a secondary analysis examining the biased norm perception longitudinally across all five rounds[2] using the same three-way interaction and controlling for starting propensity to donate and round of decision-making with standard errors three-way clustered by participant, network and round using the inclusion-exclusion procedure. In this procedure standard errors are built by summing the variance contributions calculated when one clusters on each of the three dimensions separately, subtracting the three possible two-way clustered variances to remove double-counting, and then adding back the variance clustered simultaneously on all three dimensions to yield a heteroskedasticity-robust covariance matrix

[2] All participants knew that the experiment would have five rounds and thus contributions in round 5 were substantially lower than contributions in the previous rounds. However, this does not affect our primary analyses which relate to participant perceptions of the norm rather than their contributions or strategic decisions. See Supplemental Materials for contributions per round, per condition.

that remains consistent even when errors are correlated within any combination of the three clustering dimensions (Cameron, Gelbach & Miller, 2011). This analysis resulted in a non-positive definite error matrix, likely due to the inclusion of round of decision making as both a cluster and a covariate, requiring us to modify the pre-registered analysis. In the interest of robustness, we therefore run two alternative regressions. In the first, we use fixed effects for the decision round and cluster standard errors by network and individual and find a similar pattern to the first-round decisions, though weaker, across all five rounds of the experiment, $B = 0.72$, $SE = 0.33$, $t(10287) = 2.16$, $p = 0.031$. In the second, we remove round of decision as a covariate and three-way cluster standard errors by network, participant and round. When doing so the estimates are of similar size, but now of marginal significance, $B = 0.72$, $SE = 0.39$, $t(10291) = 1.86$, $p = 0.063$, see Figure 4 and 5 for overall pattern.

To test the robustness of these findings to alternative specifications we repeat both of these analyses (1) using degree (a simple count of the number of one's network ties) instead of eigenvector centrality, (2) controlling not just for pre-experiment propensity to donate but also a participant's contribution in the relevant round, (3) controlling for all available participant characteristics and demographic features (including age, race, gender, employment status, education status, language fluencies, time taken and prior approvals on the Prolific platform) and (3) controlling for the assigned animal moniker given to each participant. In all four alternative specifications significance levels were consistent or stronger than the pre-registered specifications (see Supplemental Materials for full results).

Figure 3. Bias in social information and perceived norms as a function of centrality and experimental condition.



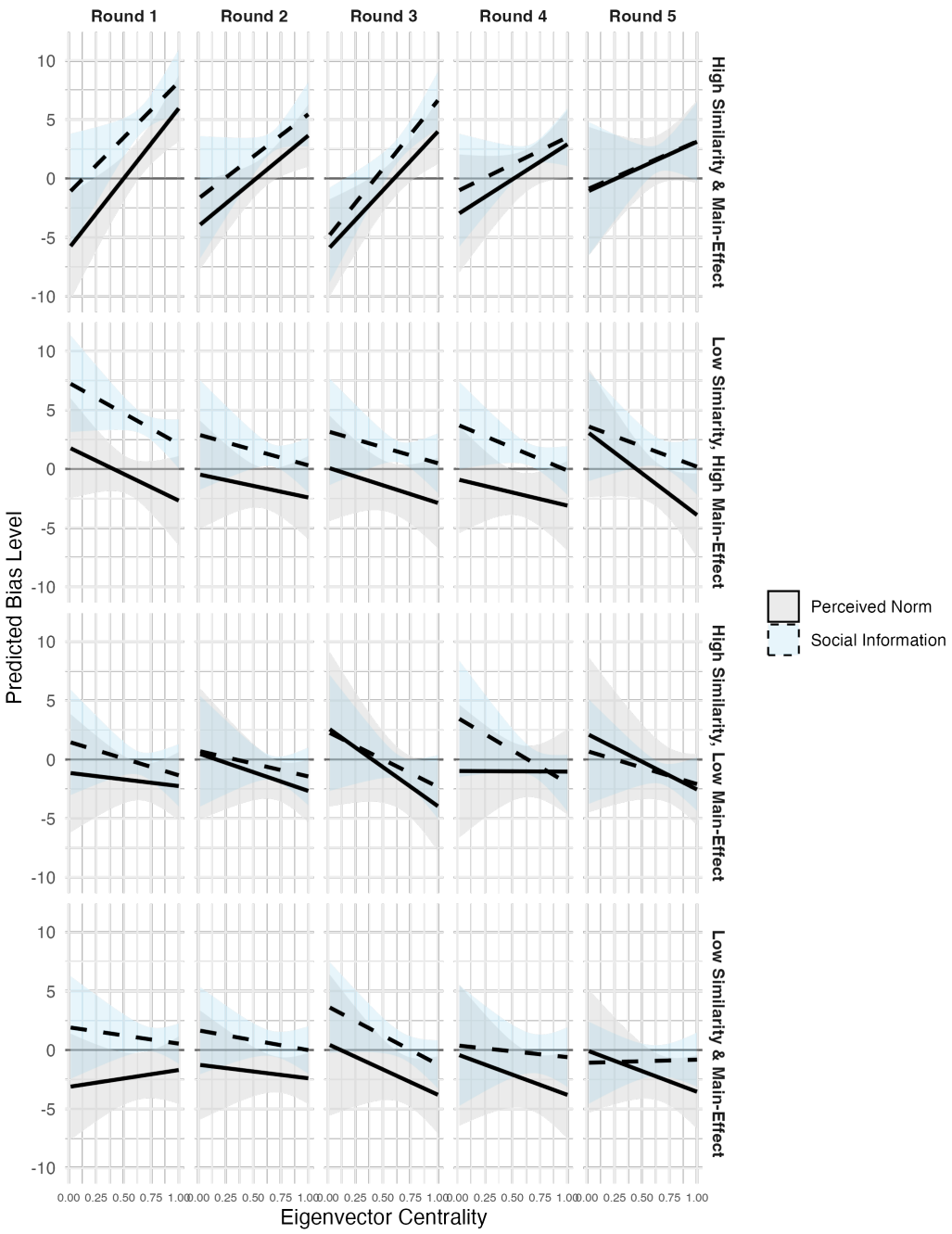
Note: Here, we show how, in networks with high similarity and main-effect attraction, central individuals are more likely to have access to a biased sample of behavior from their network ties (dotted line) as compared to both more peripheral individuals and central individuals in differently formed networks. This information in turn was associated with their perception of the social norm. Shaded regions reflect confidence intervals with standard errors clustered at the network level.

Why are central individuals more biased in their perceptions in these circumstances?

To better understand why central individuals hold biased views of social norms in networks with high similarity and main-effect attraction, we ran exploratory analyses examining the social information available to those in central positions across the different experimental conditions. To do this we calculate a new variable per participant per round which represents the average contribution by a participant's direct connections and then subtract the average network contribution across all eighteen players from this variable in order to represent the degree of bias in a participant's social information. We can then examine how the resulting variable varies as function of centrality and experimental condition, with the same controls and clustered standard errors as in previous analyses. As predicted, central individuals in networks with high similarity and main-effect attraction had significantly more biased social information from their network connections in Round 1, $B = 0.92$, $SE = 0.38$, $t(2151) = 2.42$, $p = 0.015$, and this pattern continued, though it weakened, over the course of all five rounds, $B = 0.73$, $SE = 0.36$, $t(10790) = 2.05$, $p = 0.040$, see Figure 4 and 5 for overall pattern.

We then examined whether bias in social information mediated bias in participants' perceptions of the social norm. Using the "mediation" package in R, we find that for both starting bias in perceived norms, *ACME* = 0.50, *95 CI*: [0.14, 0.87], $p = 0.010$, and bias after five rounds of interactions, *ACME* = 0.48, *95 CI*: [0.10, 0.87], $p = 0.020$, the social information from connections mediated the three-way interaction of centrality, main-effect attraction and similarity attraction on perceived descriptive norm.

Figure 4. Bias in social information and perceived norms as a function of centrality and experimental condition over five rounds of interaction.



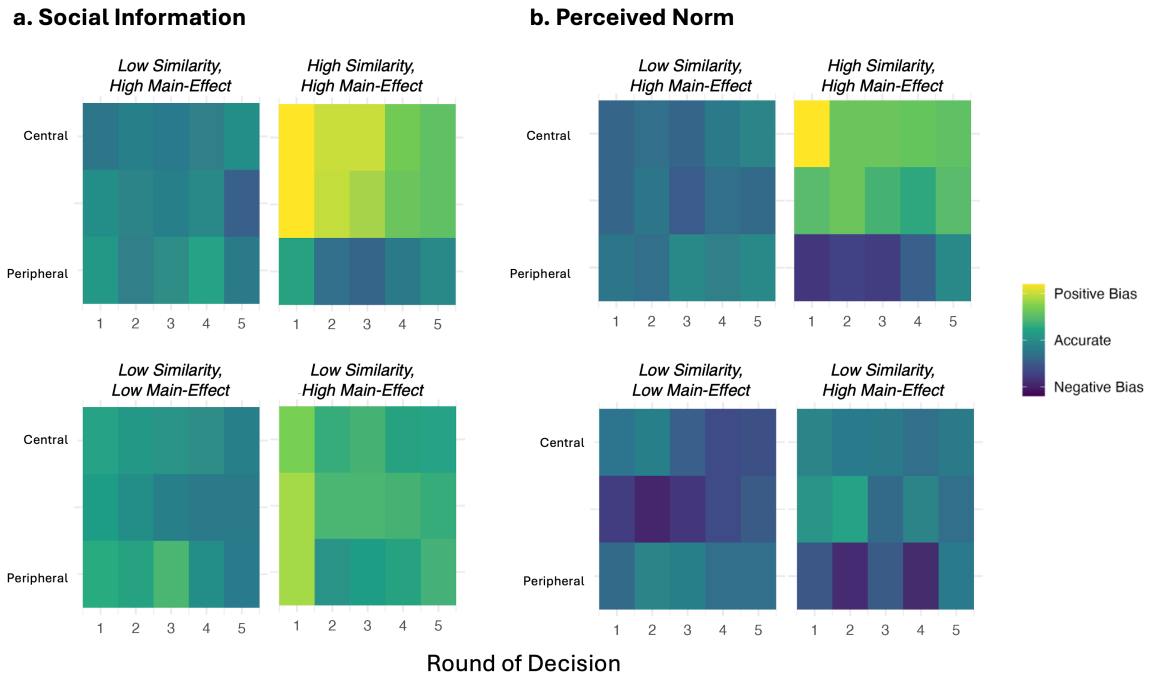
Note: Here, we display participant perceived social norms (solid line) and the social information they had access to from network ties (dotted line) across all five rounds of decisions. In the top row, one can see the positive slopping line persists, but weakens, across all five rounds reflecting the persistence of bias in perceived norms by central individuals as a function of network. In addition, across conditions, the dotted and solid lines become generally more aligned as time goes on, reflecting social learning by all participants.

How do biased perceptions affect behavior?

Having identified that networks with high similarity and main-effect attraction cause central individuals to hold biased perceptions of the associated social norm and that this due, in part, to central individuals in these networks having access to misrepresentative social information, we then sought to confirm that one's perception of the social norm was associated with one's behavior in later rounds of the experimental game, as prior research on social norms and behavior would predict.

To examine this question, we ran a series of four exploratory regressions predicting participants' contributions to the public good in rounds two through five using participants' perception of the norm from the previous round while controlling for participants' contributions in the previous round. In each regression we clustered standard errors at the participant and network level. Controlling for participant behavior in the first round, a one standard deviation increase in one's perception of the norm was associated with contributing about five more tokens on average in the next round, $B = 5.12$, $SE = 0.65$, $t(1966) = 7.87$, $p < 0.001$. The pattern continued through all five rounds of the experiment with similar or greater magnitude and significance, see Supplemental Materials for full table of results.

Figure 5. Social information travelling from the periphery to the center over time to influence bias in perceived social norms.



Note: Here, we display the same data as in Figure 4 but with an alternative representation. The three vertical rows represent participants grouped within network by rank-ordered centrality, with the bottom row representing the averaged values of the six participants per network with the lowest eigenvector centrality, the middle those with the middle six centrality scores, and the top those with the highest 6 centrality scores. The horizontal columns represent each round of decision within the experiment. In Panel A, we display the values for a participants' social information, i.e. what these participants see their connections doing. In Panel B, we display the players' perceptions of the descriptive norm. Two patterns are of note. First, bias is much more likely for central individuals when contributions to the public good exhibit high similarity and main-effect attraction in network formation. Second, over time, in the that condition more accurate information filters through the periphery and towards central individuals leading to less biased perceptions of the norm over time.

Discussion

Perceived social norms greatly influence behavior across a variety of contexts, but extant knowledge of when, for what behaviors, and for whom such perceptions are likely to be biased is strikingly limited. Our experiment creates four types of universes to compare, creating each thirty times across over two-thousand individuals. In one fourth of these universes, contributing to the public good features both high similarity and main-effect attraction, and under these circumstances we observe that central individuals have access to more biased social information

compared to their peers in other worlds or other network positions, and that this leads them to hold more biased perceptions of the social norm. Though connected to the most people in the community, they form incorrect impressions of that community's behavior, because their own connections are not a representative sample. Our methods allow us to isolate the role of this misrepresentative social information from any influence of an individual's own behavior, enabling us to identify the causal role of similarity and main-effect attraction in perceived social norms.

Over time, information from the periphery filters through the network allowing central individual's perceptions of their community to improve, albeit not entirely. Several features of this experiment provide insight into the ways perceptions of social norms might improve, and faster than in real-world communities. First, the networks here are small and dense. A walktrap algorithm (Brusco, Steinley, & Watts, 2024) reveals an average modularity across all networks of only 0.13, a relatively low score, indicating that the experimental networks do not contain strong subgroups. The task of learning the social norm is thus simpler because the community is both more observable and less siloed. In larger networks, when such subgroups exist and are combined with similarity and main-effect attractiveness, echo chambers within subgroups may allow biased information to persist for longer whereas here they are quashed quickly.

Second, in the experiment, everyone can observe connections' behavior with full accuracy and in a bidirectional fashion. In everyday life, social information from one's connections may be incomplete or false, either due to chance, intentional self-censoring or misrepresentation (Cowan, 2014). In addition, central actors may pay less attention to those on the periphery than vice versa, whereas here all ties are bidirectional. Both features are likely to skew social information in the direction of the current social norm as perceived by central

individuals, because individuals out of step with norms may avoid publicizing their behavior when possible. This would slow one's ability to form an accurate norm perception and cause early bias, particularly from those in central positions, to persist for longer. Collectively, this suggests that the pattern observed over five rounds is a conservative estimate of the trajectory of bias in perceptions over time. Future research might explore these dynamics more to better determine situations in which early bias might ameliorate or instead become exacerbated.

In this experiment we held the target behavior, contributions to the public good, constant while introducing experiment-generated similarity and main-effect attraction on that behavior. It is thus worth considering how the phenomena observed in this internally valid context might translate to more externally valid ones. Different behaviors are likely to naturally exhibit different levels of similarity and main-effect attraction in networks. Future research might benefit from categorizing common norms with regards to their degree of similarity and main-effect attraction. Whether structural or psychological, both forces might be measured via tie structure in existing networks, or via survey questions that ask about propensity to form ties with similar or extreme others. This would allow for better predictions of which social norms are vulnerable to pluralistic ignorance (Prentice & Miller, 1993).

This paper offers an experimental examination of when central individuals in networks will misperceive descriptive social norms by manipulating network forces in order to causally isolate the impact of biased social information on social norm perception. In 1936, Muzafer Sherif described the process of learning social reality by observing the behavior of one's peers and updating. We suggest one way of understanding a critical step in this process, specifically what behavior one can observe as a function of one's network position and attribute qualities,

and the consequences of this for the emergence of bias in perceived social norms and its amelioration.

References

Airoldi, E. M., & Christakis, N. A. (2024). Induction of social contagion for diverse outcomes in structured experiments in isolated villages. *Science*, *384*(6695). https://doi.org/10.1126/science.adi5147

Aral, S., & Walker, D. (2012). Identifying Influential and Susceptible Members of Social Networks. *Science*, *337*(6092), 337–341. https://doi.org/10.1126/science.1215842

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, *341*(6144), 1236498.

Bicchieri, C. (2006). The grammar of society: The nature and dynamics of social norms (1st ed.).

Bonacich, P. (1972). Technique for analyzing overlapping memberships. *Sociological Methodology*, *4*, 176. https://doi.org/10.2307/270732

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170–1182. https://doi.org/10.1086/228631

Brusco, M., Steinley, D., & Watts, A. L. (2024). Clustering methods: To optimize or to not optimize? *Psychological Methods*. https://doi.org/10.1037/met0000688

Bursztyn, L., & Yang, D. Y. (2022). Misperceptions about others. *Annual Review of Economics*, *14*(1), 425–452. https://doi.org/10.1146/annurev-economics-051520-023322

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, *110*(10), 2997–3029. https://doi.org/10.1257/aer.20180975

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. https://doi.org/10.1198/jbes.2010.07136

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, *334*(6060), 1269–1272. https://doi.org/10.1126/science.1207055

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLoS ONE*, *5*(9). https://doi.org/10.1371/journal.pone.0012948

Dannals, J. E., & Li, Y. (2024). A theoretical framework for social norm perception. *Research in Organizational Behavior*, *44*, 100211. https://doi.org/10.1016/j.riob.2024.100211

Fehr, E., & Fischbacher, U. (2004). Social Norms and Human Cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Feiler, D. C., & Kleinbaum, A. M. (2015). Popularity, similarity, and the network extraversion bias. *Psychological Science*, *26*(5), 593–603. https://doi.org/10.1177/0956797615569580

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, *96*(6), 1464–1477. https://doi.org/10.1086/229693

Fu, F., Nowak, M. A., Christakis, N. A., & Fowler, J. H. (2012). The evolution of homophily. *Scientific Reports*, *2*(1). https://doi.org/10.1038/srep00845

Galesic, M., Olsson, H., & Rieskamp, J. (2018). A sampling model of social judgment. *Psychological Review*, *125*(3), 363–390. https://doi.org/10.1037/rev0000096

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M. S., Fischlmayr, I. C., . . . Yamaguchi, S. (2011). Differences between tight and loose cultures: a 33-Nation study. *Science*, *332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Goldenberg, A., Abruzzo, J. M., Huang, Z., Schöne, J., Bailey, D., Willer, R., Halperin, E., & Gross, J. J. (2023). Homophily and acrophily as drivers of political segregation. *Nature Human Behaviour*, *7*(2), 219–230. https://doi.org/10.1038/s41562-022-01474-9

Hasan, S., & Koning, R. (2019). Prior ties and the limits of Peer Effects on Startup Team Performance. *Strategic Management Journal*, *40*(9), 1394–1416. https://doi.org/10.1002/smj.3032

Jackson, M. O. (2019). The friendship paradox and systematic biases in perceptions and social norms. *Journal of Political Economy*, *127*(2), 777–818. https://doi.org/10.1086/701031

Kim, S., Lee, J., & Yoon, D. (2015). Norms in social media: The application of theory of reasoned action and personal norms in predicting interactions with Facebook page like ads. *Communication Research Reports*, *32*(4), 322–331. https://doi.org/10.1080/08824096.2015.1089851

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology*, *115*(2), 405–450. https://doi.org/10.1086/599247

Lin, N. (1999). Social networks and status attainment. *Annual Review of Sociology*, *25*(1), 467–487. https://doi.org/10.1146/annurev.soc.25.1.467

Madan, S., Savani, K., Mehta, P. H., Phua, D. Y., Hong, Y. Y., & Morris, M. W. (2025). Supplemental material for stress reactivity and sociocultural learning: More stress-reactive individuals are quicker at learning sociocultural norms from experiential feedback. (2025). *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000487.supp

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531. https://doi.org/10.2307/2298123

Marks, G., & Miller, N. (1987). Ten Years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72–90. https://doi.org/10.1037//0033-2909.102.1.72

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, *67*(1), 339–361. https://doi.org/10.1146/annurev-psych-010814-015013

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158. https://doi.org/10.1257/jep.14.3.137

Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology*, *103*(6), 899–915. https://doi.org/10.1037/a0030015

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A Social Network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571. https://doi.org/10.1073/pnas.1514483113

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243–256. https://doi.org/10.1037//0022-3514.64.2.243

Savani, K., Morris, M. W., Fincher, K., Lu, J. G., & Kaufman, S. B. (2022). Experiential learning of cultural norms: The role of implicit and explicit aptitudes. *Journal of Personality and Social Psychology*, *123*(2), 272–291. https://doi.org/10.1037/pspa0000290

Sparkman, G., Geiger, N., & Weber, E. U. (2022). Americans experience a false social reality by underestimating popular climate policy support by nearly half. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-32412-y

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0016836

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for Social Change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Valente, T. W., Unger, J. B., & Johnson, C. A. (2005). Do popular students smoke? the association between popularity and smoking among middle school students. *Journal of Adolescent Health*, *37*(4), 323–329. https://doi.org/10.1016/j.jadohealth.2004.10.016

Valente, T. W., Unger, J. B., Ritt-Olson, A., Cen, S. Y., & Johnson, A. (2006). The interaction of curriculum type and implementation method on 1-year smoking outcomes in a school-based prevention program. *Health Education Research*, *21*(3), 315–324. https://doi.org/10.1093/her/cyl002

Villeval, M. C. (2020). Performance feedback and peer effects: A Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3543371

Supplemental Materials for

"Origins of Bias in Perceived Social Norms from Social Network Formation"

Table of Contents

Network Assignment Algorithm

Players in each game were assigned network ties as a function of the following algorithm,

$$p(tie_{ij}) = \left(a\left(\frac{p_i+p_j}{2}\right) + (1-a)\left(\frac{q_i+q_j}{2}\right)\right) * \left(h(1 - |p_i - p_j|) + (1-h)(1 - |q_i - q_j|)\right)$$

where p_i is player *i's* indicated preference to contribute to the public good prior to network assignment and q_i is a randomly drawn value that is identically and independently distributed to *p*.

The first factor captures the role of main-effect attraction in tie formation, while the second factor captures the role of similarity attraction in tie formation. According to this formula, all *ij* parings will then probabilistically take on a value of 0, did not connect, or 1, did connect. The variables *a* and *h* capture the experimentally assigned value of main-effect and similarity attraction respectively. High main-effect attraction conditions take on a value of $a = 0.8$, low take on a value of $a = 0$; high same-effect attraction conditions take on a value of $h = 0.8$, low take on a value of $h = 0$.

To illustrate how this might operate in day-to-day life, imagine that in order for a connection to form, player *i* must first meet player *j* and then must also like player *j*. The pair might meet due to similar interests, in which case the similarity attraction term could be thought of as capturing the probability of meeting: the more similar their interests the more likely they are to meet. After they meet, what determines whether *i* likes *j*? This would be captured by the main-effect attraction term, which is to say, that if *a* is high, agents like people with similar but stronger interests. In this scenario, one might interpret the equation above as, *P(tie) = P(like)*P(meet)*.

Alternatively, imagine that *i* meets *j* because *j* strongly exhibits a behavior that makes *j* very visible or popular, such as extraversion. Here, the main-effect attraction term more accurately describes the probability of *i* meeting *j* given that *j's* high value increases meeting chances. Once *i* and *j* meet, however, *i* may like *j* more if they are more similarly extraverted, such that liking is determined by similarity attraction. In this scenario, *P(tie) = P(meet)*P(like)*. In both cases the probability of connection is mathematically equivalent, but the interpretation of the two network forces underlying such connections is different. We explain each term in greater depth below.

Probability of main-effect attraction in ties

For each individual *i*, the probability that *i* will form a tie with individual *j* due to main-effect attraction is determined by the first factor in Equation 1 above:

$$a\left(\frac{p_i + p_j}{2}\right) + (1 - a)\left(\frac{q_i + q_j}{2}\right)$$

There are three things to note about this expression. First, a key component is *i* and *j*'s average level of *p*, $\left(\frac{p_i+p_j}{2}\right)$, and likewise for the average level of *q*. The decision to average *i* and *j*'s average level of *p* may seem counterintuitive at first glance given that main-effect attraction is defined as an attraction to extreme values. This decision means that two players, *j* and *k,* with values of *p* of 0 and 1, the most extreme difference possible, would have an equal probability of ties due to main-effect attraction as players, *l* and *m,* with values of *p* of 0.49 and 0.51. The goal of the main-effect attraction term in the model is to capture the degree to which certain behaviors or attributes when held in greater amounts, lead one to be more readily encountered or more readily liked, thus leading to more tie formation. In order to consider this on a dyadic level, one must not only consider the main-effect attraction of *i* towards *j* but also the main-effect attraction

of *j* towards *i*. To illustrate this, consider the two previous dyads *jk* and *lm*. When main-effect attraction of attribute *p* is high, player *k* is a very valuable or very visible connection within the network, however *k* is not more valuable or visible to player *j* than to any other (similarly low on *p*) agent because main-effect attraction is a directional preference rather than a comparative preference (Goldenberg et al., 2023). Player *j* on the other hand, is a connection of lower value and visibility, who should therefore be less likely to connect to *k* than a connection with a higher value of *p* when main-effect attraction is high. In networks in which nodes have infinite capacity to link with other nodes (e.g., the internet), *j*'s preference to connect with *i* could forge a tie even if *i* is indifferent (Barabási & Albert 1999); but in social networks, where individual have finite capacity for forming ties (Roberts, Wilson, Fedurek & Dunbar, 2008), a low value of p_j makes a tie less likely. On balance, this leaves the probability of a tie *jk* at only a moderate level: *k* is likely to make ties with many agents, *j* is likely to make ties with fewer agents. Given this, one can see how *jk* may have a similar likelihood of main-effect attraction-based ties as the moderate-*p* dyad *lm*.

Second, note that a is the weight placed on the first term in the equation and $(1 - a)$ is the weight placed on the second term. The intuition is that *a* determines the degree to which probability of meeting is dictated by *i* and *j*'s average level of *p* (the target characteristic, contribution to the public good) as opposed to *q* (other behaviors). When *a* is high, then the probability of *i* connecting *j* is primarily determined by the extent to which those two individuals are high in behavior *p*. In the opposite extreme, when *a* is low, behavior *p* is less relevant for the probability of meeting.

Probability of similarity attraction-based ties

For each individual i, the probability that i will form a tie with another individual j due to similarity attraction is drawn from the second factor in Equation 1 above:

$$h * \left(1 - \left|p_i - p_j\right|\right) + (1 - h) * \left(1 - \left|q_i - q_j\right|\right)$$

There are two things to note about this expression. First, a key component is the average similarity between persons i and j on behavior p captured by the term $(1 - \left|p_i - p_j\right|)$. The greater the distance between p_i and p_j, the less similar they are, and the smaller this term will be. second, parameter h is the weight placed on the first term in the equation and $(1 - h)$ is the weight placed on the second term. Analogous to role of a in the previous equation, h determines the degree to which probability of connection is dictated by i and j's similarity in p (the focal behavior) as opposed to similarity in q (other behaviors). When h is high, similarity attraction acts on p to affect connection likelihood; when is low, then behavior p is less relevant for the probability of connection (and similarity attraction occurs on other attributes and behaviors in the social environment).

Network Visualizations

For each network visualization below, node diameter is proportional to eigenvector centrality and node color represented a participant's indicated intended contribution to the public good. Yellow nodes represent zero token, bright green represent 25 tokens, teal represents 50 tokens, blue represents 75 tokens and purple represents 100 tokens. Though networks do not significantly differ on structure across conditions, in the visualizations the relative location of individuals shifts as a function of their starting value of the focal behavior, the intended contribution to the public good. Below each network is its unique "gameID" identifier which can be found in the dataset available online.

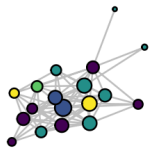
In order to verify that each network does in fact have differing degrees of main-effect and similarity attraction, we calculate two new variables, each representing the "achieved" level of attraction within the realized network.

For main-effect attraction we calculate the $\frac{p_i+p_j}{2}$ for every tie a given player has and then take the average of these values. This represents the average amount of main-effect attraction operating in a given player's set of network connections. We sum this for all players in the game and then divide but the total *possible* main-effect attraction in the game. To calculate the total possible main-effect attraction, we calculate $\frac{p_i+p_j}{2}$ for *all* possible pairs in the game, regardless of whether a tie exists between two players. The resulting value is thus the proportion of main-effect attraction which was realized or achieved by the network algorithm in a given network instantiation. Networks in the high main-effect conditions had significantly greater achieved main-effect attraction than those in the low main-effect conditions, $B = 0.02$, $SE = 0.002$, $t(117) = 9.88$, $p < 0.001$.

In the same vein, for similarity attraction we calculate the $1 - |p_i - p_j|$ for every tie a given player has and then take the average of these values. This represents the average amount of similarity attraction operating in a given player's set of network connections. We sum this for all players in the game and then divide but the total *possible* similarity attraction in the game. To calculate the total possible similarity attraction, we calculate $1 - |p_i - p_j|$ for *all* possible pairs in the game, regardless of whether a tie exists between two players. The resulting value is thus the proportion of similarity attraction which was realized or achieved by the network algorithm in a given network instantiation. Networks in the high similarity conditions had significantly greater achieved similarity attraction than those in the low similarity conditions, $B = 0.04$, $SE = 0.004$, $t(117) = 12.18$, $p < 0.001$.

Low Main-Effect Attraction & Low Similarity Attraction

1 01JBVSR02489M0GMZXMWR0DWW5



10 01JD5BM67N90652YAQPCJA1JNH

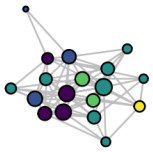
11 01JD7J0D4B20HQWP3RCPKETGAD

11 01JD7J1RT9ZTXDAF1QWVXV8FAF

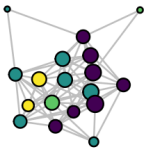
12 01JD800T3Y9KX6YQ97VADAKS82

13 01JKX94D1BB8W4YEXJ21X2E91H

13 01JKX94VD7M2Q0TT0M8BX0KS18



13 01JKXHZBTXYBSMAFX7SQA63C8G

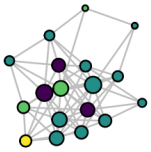


14 01JKZW9XZVQMDCQYXNAEZHVRQ9

15 01JM0EYSF87JKEGWYNMF0APKV1

2 01JBWCBKHC865BCP9JZ6PBBFW3

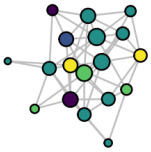
2 01JBWCC95F8J9HST0ZXDS7AKFJ



2 01JBYAHN2VY74Y2Q1755HC1WTP

2 01JBYT58HA0D0GMXC6QPYH6QPV

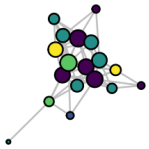
2 01JC0WJVPA06WKE0SPPRQPRVPF



3 01JC3X558ZZEDM6VBG2QRW9HFA

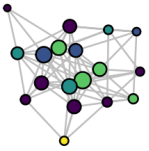
3 01JC605C6WV53QVM63EYAMP8BB

4 01JCDS9TMZNY3YXY0PNTCNN81D

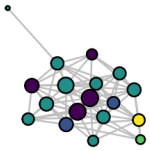


4 01JCE686JQANWP1VWJCZ7F3RED

4 01JCE6DG99DTRDD9YKCGT0Z6CE



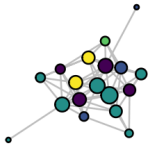
5 01JCGFVR0DCF7HECNQC4K61E7D



5 01JCGFW2D3QB4PXKAWE0P8ZZ7M

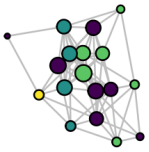
5 01JCK1AKNCD44KVGK1RNXVHGR7

5 01JCK1APA3Z2M7MGAKV2GD2E9R



6 01JCRR33EY80HGTGJN0182KYHN

7 01JCZTVQ4JV480E15M5JPF06V9



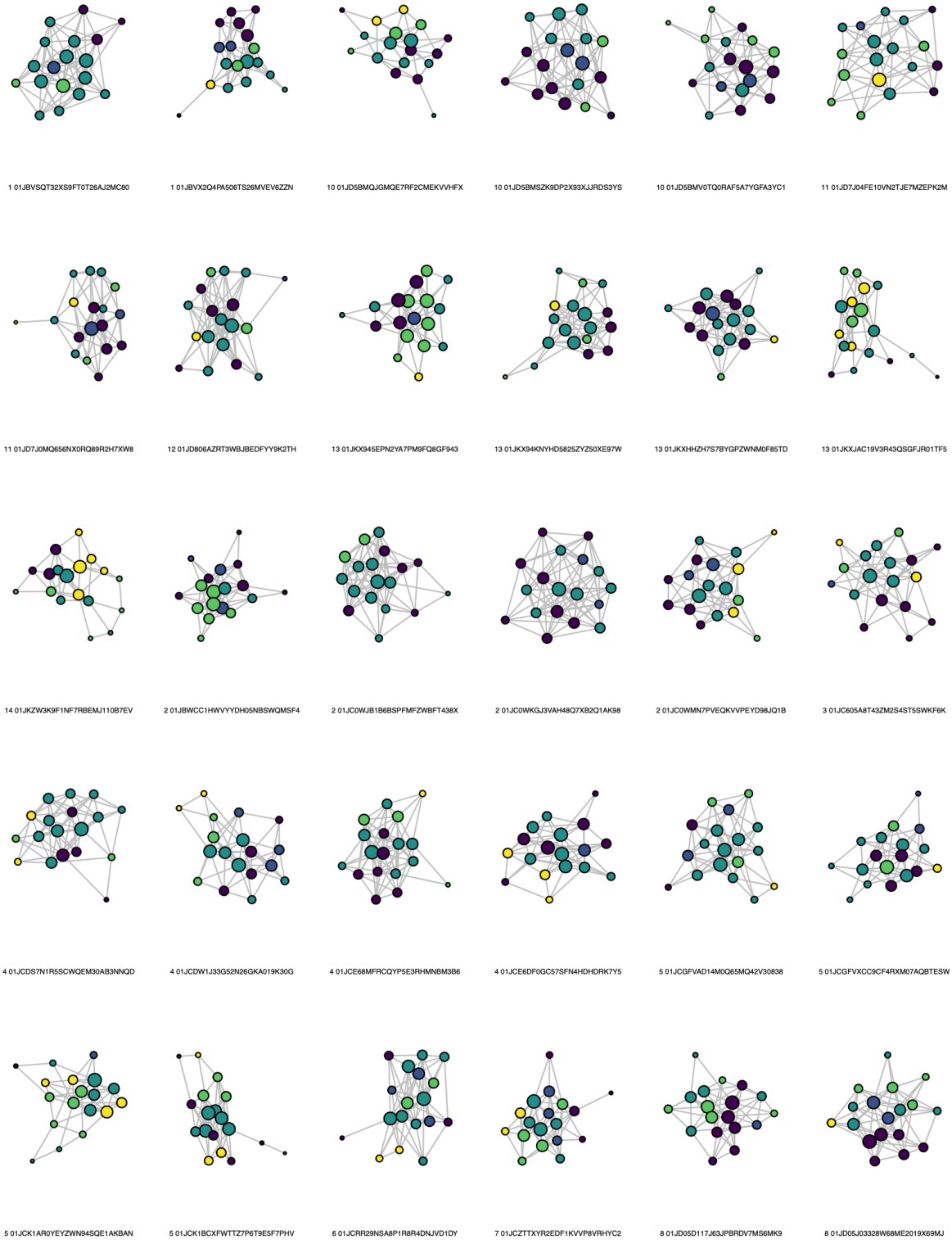
8 01JD05J9N6HK2Z8KTE7C5YBG0Y

8 01JD05JRWA0502ST45K39NE0TD

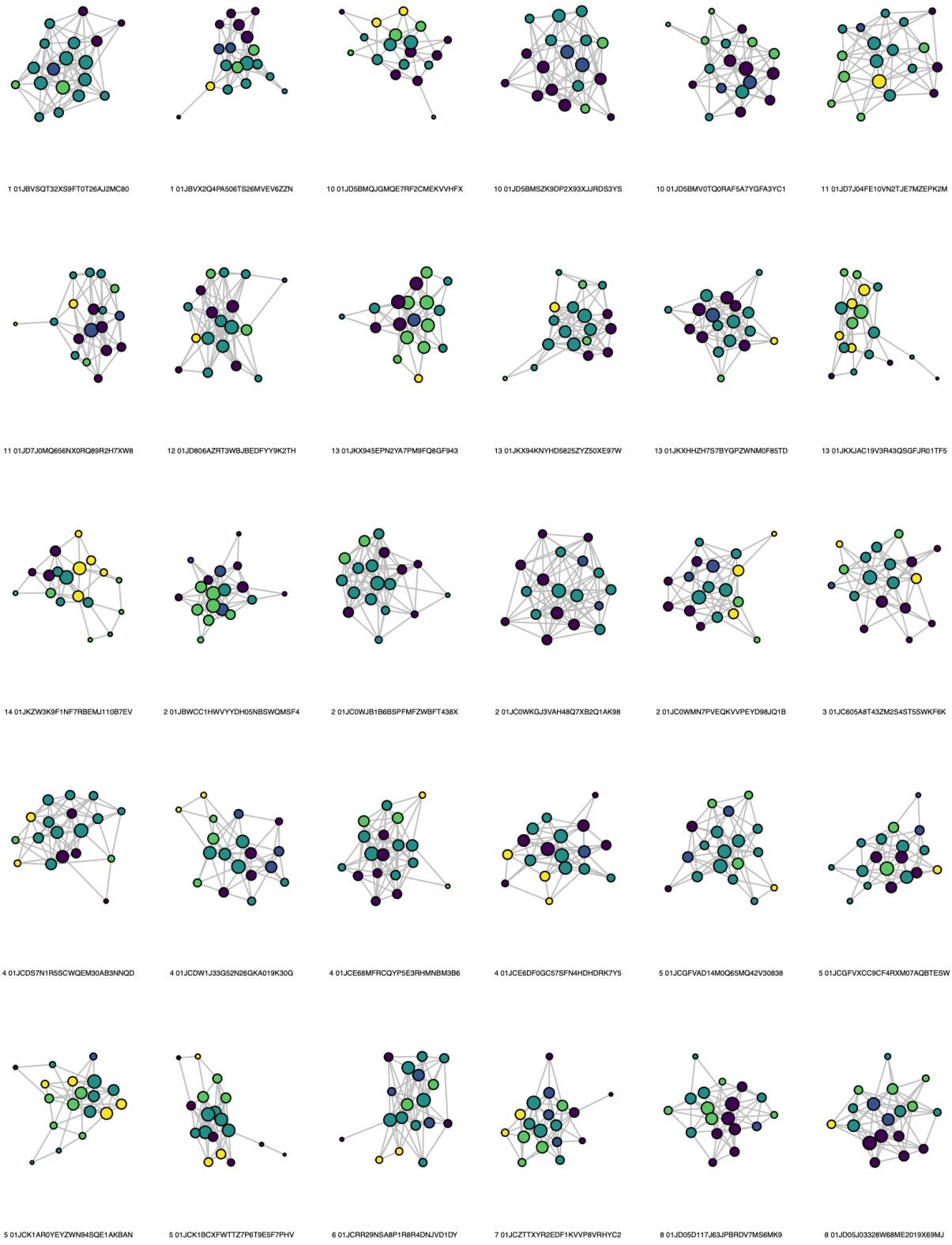
9 01JD2B566TQ4HQHT28BZF94HRV

9 01JD2B5AVB8AV605Q86JY1XVTJ

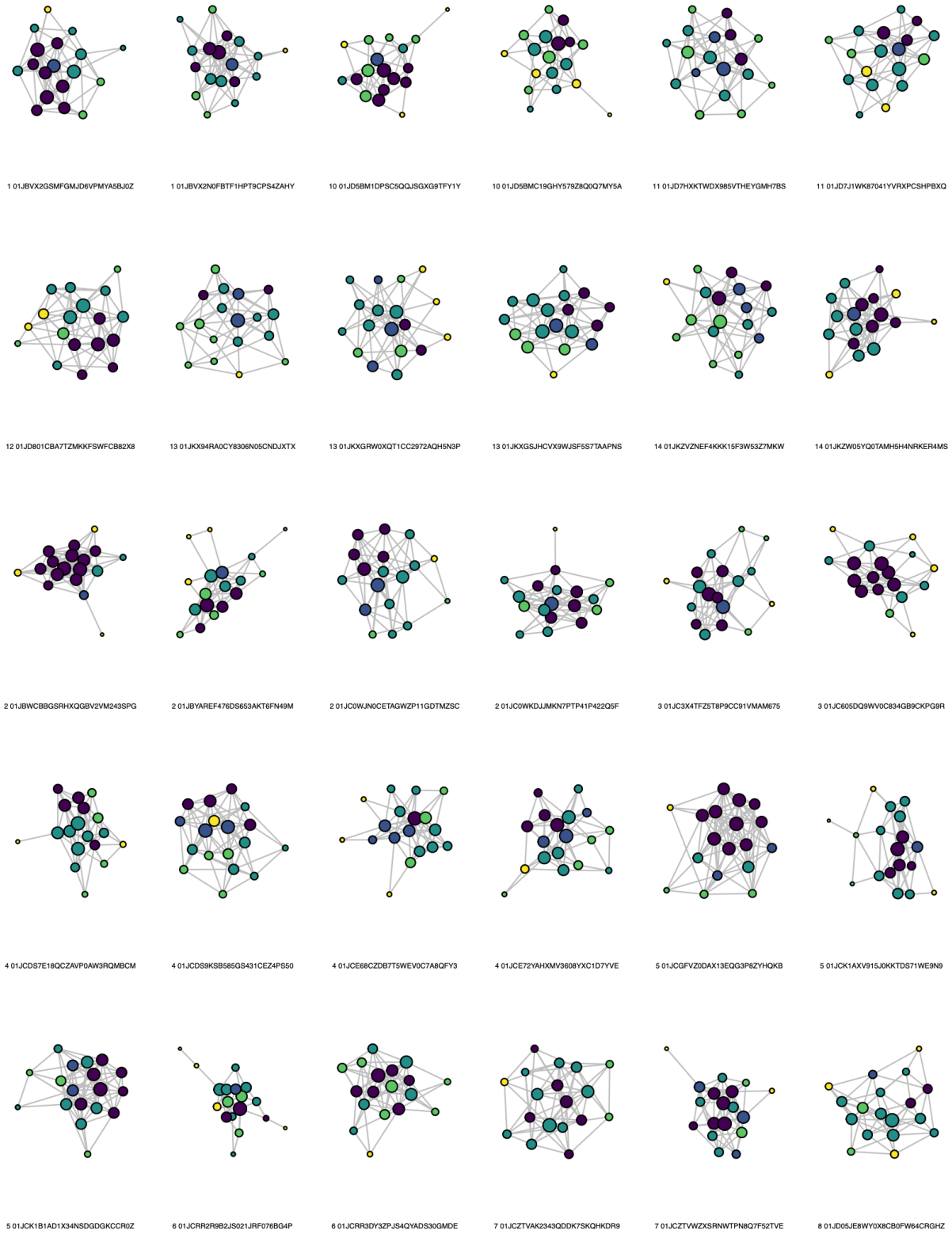
Low Main-Effect Attraction, High Similarity Attraction



High Main-Effect Attraction, Low Similarity Attraction

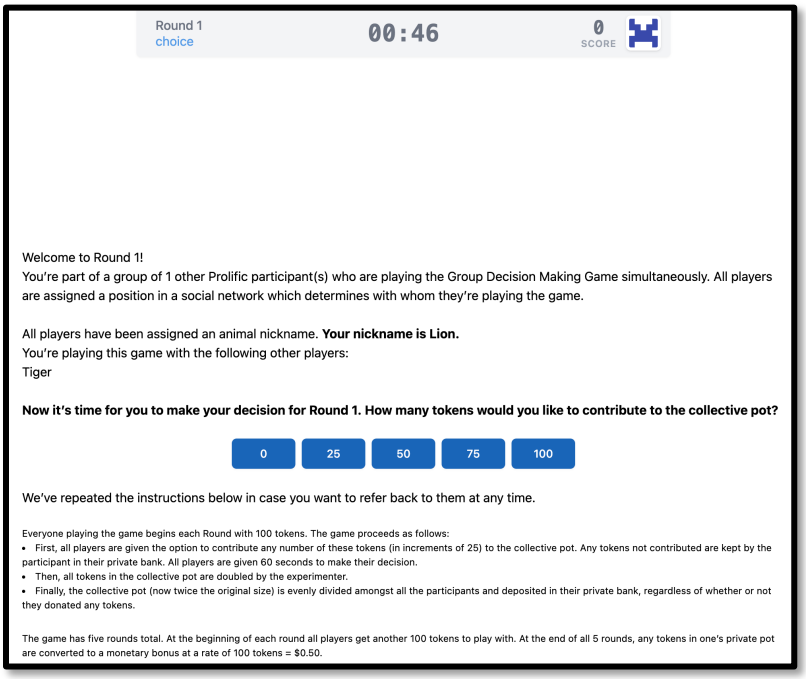


High Main-Effect Attract, High Similarity Attraction

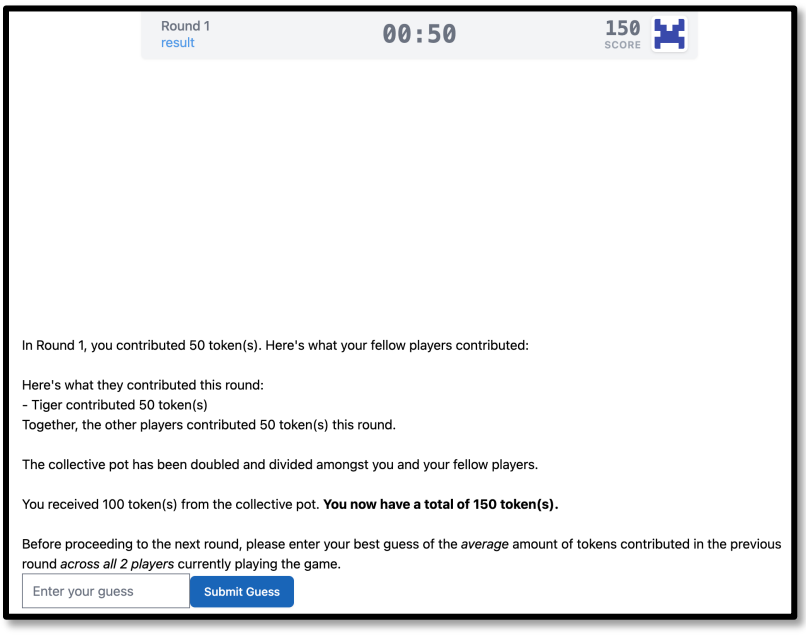


Screenshots of Game Play

a.

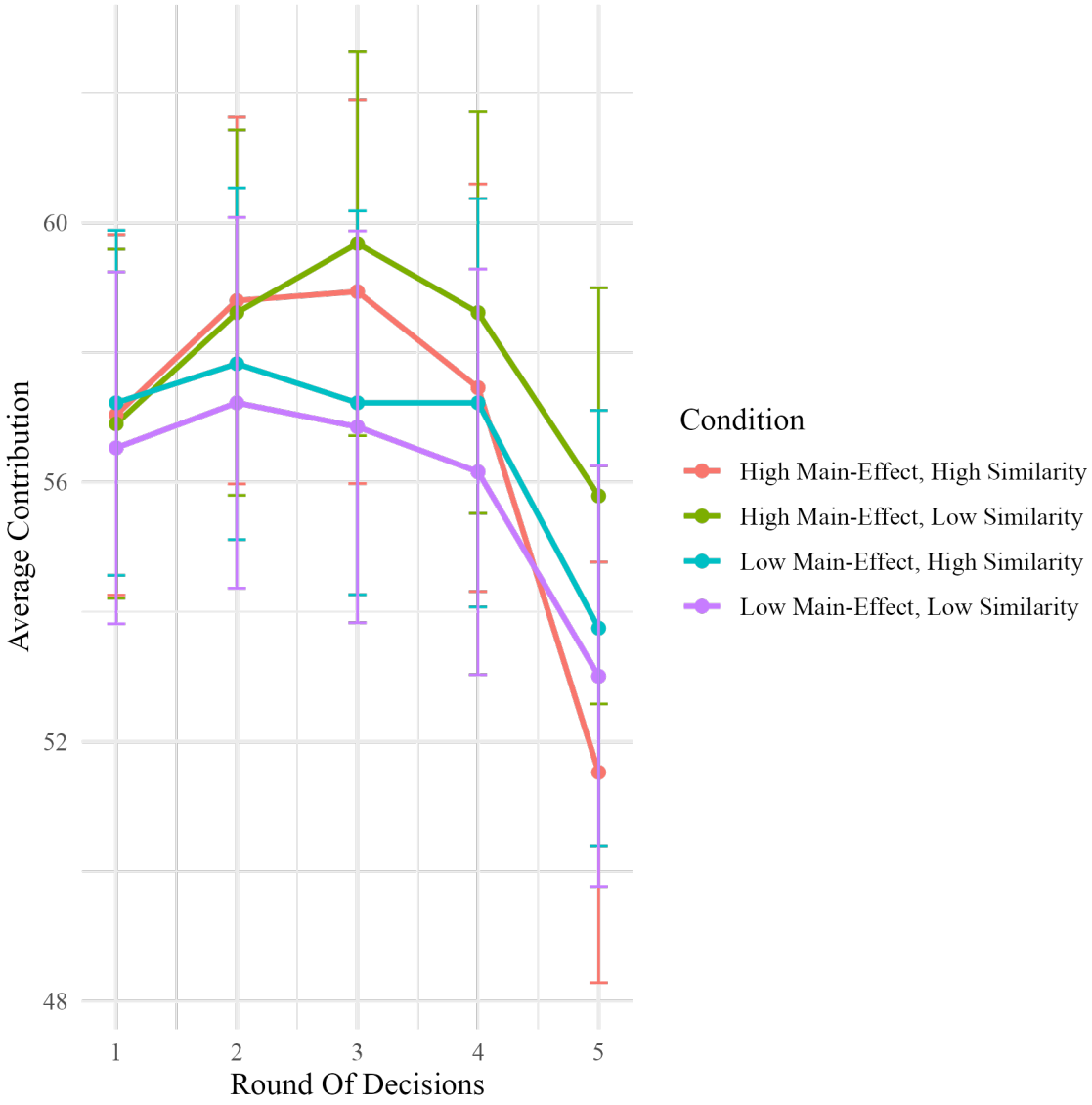


b.



These screenshots were taken from a hypothetical game between only two players. In the real game the like "across all 2 players currently playing the game" would read "18 players." Note the timer at the top of the page counting down the one mine for submitting. In Pane a, the screenshot depicts a "decision" screen, with instructions repeated at the bottom of the screen. In Pane b, the screenshot depicts a "results" screen, in which participants would enter their guess of the descriptive social norm for the network to advance.

Contributions Per Round Per Condition



This figure depicts contributions to the public good across the course of the experiment on average across all players in a network. Average contributions do not significantly differ across condition, though all conditions show the same drop in contributions in round 5, a typical pattern in strategic games when the last round in known. Interesting, the drop-off is directionally greatest in the high-high condition, where over-estimation of generosity is also the highest and thus free-riding would have the highest expected value. However, as this is not a significant difference, we caution overinterpreting this difference.

Additional Exploratory Measures

At the end of the experiment participants were asked again, "Please enter your best guess of the average amount of tokens contributed in the previous round across all 18 players currently playing the game." This question is largely redundant with a participant's guess of the perceived social norm in Round 5 of the experiment, however it was asked again for the sake of redundancy. Some participants (44.10%) enter the same value as in Round 5, but others do not, perhaps having interpreted this question more globally to mean their estimate of the average contribution across the entire game.

In the interest of transparency, we thus report two additional analyses. In the first, we calculate bias in final perceived norm relative to the average behavior in the network in round 5 and regress this variable on the three-way interaction of eigenvector centrality, main-effect attraction, and similarity attraction, while controlling for participant starting contribution and clustering standard errors at the network level. We find a marginal three-way interaction B = 3.17, SE = 1.67, $t(2132) = 1.90$, $p = 0.057$ in line with the analyses presented in the paper. In the second analysis we calculate bias in final perceived norm relative to the average behavior across all 5 rounds of the experiment and then repeat the same regression specifications. We find a significant three-way interaction of similar magnitude, $B = 3.34$, $SE = 1.59$, $t(2132) = 2.10$, $p = 0.036$.

Participants also indicated their desire to repeat the experiment with each of their network ties on a scale from 1 to 7 where 1 indicated "I would NOT want to play with this player" and 7 indicated "I would DEFINITELY want to play with this player." We find two consistent effects with regards to rating other players. First, players who earned more points overall in the experiment rate all their network ties more highly. In a regression with standard errors clustered by network and participant, a one standard deviation increase in earnings is associated with about a quarter of a point increase, $B = 0.26$, $SE = 0.03$, $t(14655) = 7.83$, $p < 0.001$. Second, participants judged others who contributed below their estimate of the norm more harshly, $B = -0.22$, $SE = 0.023$, $t(14654) = -9.50$, $p < 0.001$. This pattern holds whether considering a participant's final perception of the social norm, or their perception in round 5 of the experiment and whether considering the alter's final contribution in round 5 or their average contribution across the entire experiment, $ps < 0.001$.

Behavioral Consequences

Predicting Contributions Across Rounds

	Dependent variable:			
	Round 2 Contribution	Round 3 Contribution	Round 4 Contribution	Round 5 Contribution
	(1)	(2)	(3)	(4)
Contribution in Round T-1 (z-scored)	18.415***	21.806***	22.831***	20.509***
	(0.665)	(0.611)	(0.682)	(0.798)
Perceived Norm in Round T-1 (z-scored)	5.119***	6.012***	5.405***	7.101***
	(0.665)	(0.698)	(0.739)	(0.747)
Constant	58.130***	58.302***	57.453***	53.555***
	(0.625)	(0.591)	(0.568)	(0.627)
Observations	1,969	2,079	2,090	2,082
R^2	0.380	0.484	0.469	0.387

Note: Standard errors clustered by network and participant. *p < 0.10, **p< 0.05, ***p<0.01

Robustness Checks

Supplemental Table 1

	Dependent variable:		
	Bias in Descriptive Social Norm Perception		
	(1)	(2)	(3)
Degree Centrality (z-scored)	-0.535		
	(0.424)		
Eigenvector Centrality (z-scored)		0.441	0.347
		(0.419)	(0.468)
Similarity Attraction (1 = high, -1 = low)	0.769*	0.761*	0.664
	(0.430)	(0.441)	(0.469)
Main-Effect Attraction (1 = high, -1 = low)	1.171***	1.164***	1.339***
	(0.430)	(0.440)	(0.455)
Pre-Experiment Propensity to Donate to Public Good	3.088***	2.599***	2.706***
	(0.458)	(0.779)	(0.486)
Degree * Similarity	0.957**		
	(0.396)		
Degree * Main-Effect	0.101		
	(0.396)		
Own Contribution to Public Good (z-scored)		0.078	
		(0.767)	
Time Taken on Experiment (z-scored)			1.715**
			(0.696)
Prior Approvals on Prolific (z-scored)			0.101
			(0.399)
Age (z-scored)			-0.335
			(0.487)
Language Fluency (0 = English, 1 = English and others)			0.953
			(0.932)
Sex (0 = Male, 1 = Female)			-0.422
			(0.866)
Sex (0 = Male, 1 = Prefer not to say)			-2.415
			(4.495)
Ethnicity (0 = White, 1 = Asian)			-2.482
			(1.767)
Ethnicity (0 = White, 1 = Black)			-0.118
			(1.723)
Ethnicity (0 = White, 1 = Mixed)			0.441
			(1.509)

Ethnicity (0 = White, 1 = Other)			-1.644
			(2.730)
Student Status (Not a student = 0, Student = 1)			-0.380
			(1.142)
Employment status (Full Time = 0, Due to start a new job within the next month = 1)			7.050**
			(3.003)
Employment status (Full Time = 0, Not in paid work (e.g. homemaker', 'retired or disabled) = 1)			-2.017
			(1.483)
Employment status (Full Time = 0, Other = 1)			0.547
			(1.989)
Employment status (Full Time = 0, Part Time = 1)			-0.060
			(1.170)
Employment status (Full Time = 0, Unemployed (and job seeking = 1)			-0.068
			(1.596)
Eigenvector * Similarity		0.792**	0.983**
		(0.397)	(0.441)
Eigenvector * Main-Effect		0.405	0.594
		(0.396)	(0.450)
Similarity * Main-Effect	0.602	0.576	0.103
	(0.430)	(0.441)	(0.454)
Degree * Similarity * Main-Effect	0.995**		
	(0.398)		
Eigenvector * Similarity * Main-Effect		1.081***	1.043**
		(0.400)	(0.450)
Constant	-0.846**	-0.870**	-0.103
	(0.430)	(0.441)	(0.868)
Observations	1,969	1,969	1,631
R^2	0.050	0.050	0.061
Adjusted R^2	0.046	0.045	0.047
Residual Std. Error	16.071 (df = 1960)	16.077 (df = 1959)	16.106 (df = 1606)
F Statistic	12.921*** (df = 8; 1960)	11.420*** (df = 9; 1959)	4.342*** (df = 24; 1606)

Note: Standard errors are clustered at the network level *p < 0.10, **p < 0.05, ***p<0.01

Supplemental Table 2

	Dependent variable:		
	Bias in Descriptive Social Norm Perception		
	(1)	(2)	(3)
Degree Centrality (z-scored)	-0.359***		
	(0.126)		
Eigenvector Centrality (z-scored)		0.009	-0.193
		(1.762)	(1.491)
Similarity Attraction (1 = high, -1 = low)	-0.955	-0.785	-0.918
	(1.019)	(1.150)	(1.062)
Main-Effect Attraction (1 = high, -1 = low)	0.202	-0.830	-1.394
	(1.077)	(1.266)	(1.148)
Pre-Experiment Propensity to Donate to Public Good	2.368***	1.051**	2.205***
	(0.353)	(0.454)	(0.384)
Round of Experiment (Round 1 = 0, Round 2 = 1)	-0.237		-0.471
	(0.353)		(0.407)
Round of Experiment (Round 1 = 0, Round 3 = 1)	-0.474		-0.697
	(0.485)		(0.524)
Round of Experiment (Round 1 = 0, Round 4 = 1)	-0.462		-0.666
	(0.493)		(0.532)
Round of Experiment (Round 1 = 0, Round 5 = 1)	0.124		-0.112
	(0.565)		(0.618)
Degree * Similarity	0.273**		
	(0.125)		
Degree * Main-Effect	0.078		
	(0.131)		
Round of Experiment (Continuous, 1-5)		0.045	
		(0.065)	
Own Contribution to Public Good (z-scored)		1.748***	
		(0.295)	
Time Taken on Experiment (z-scored)			1.412***
			(0.489)
Prior Approvals on Prolific (z-scored)			-0.393
			(0.341)
Age (z-scored)			-0.326
			(0.374)
Language Fluency (0 = English, 1 = English and others)			-0.470
			(0.847)
Sex (0 = Male, 1 = Female)			-0.045
			(0.640)

Sex (0 = Male, 1 = Prefer not to say)			-3.026
			(1.983)
Ethnicity (0 = White, 1 = Asian)			-1.136
			(1.233)
Ethnicity (0 = White, 1 = Black)			0.974
			(1.269)
Ethnicity (0 = White, 1 = Mixed)			1.693
			(1.516)
Ethnicity (0 = White, 1 = Other)			0.203
			(2.323)
Student Status (Not a student = 0, Student = 1)			-1.790*
			(1.029)
Employment status (Full Time = 0, Due to start a new job within the next month = 1)			5.125
			(3.405)
Employment status (Full Time = 0, Not in paid work (e.g. homemaker', 'retired or disabled) = 1)			-1.470
			(1.064)
Employment status (Full Time = 0, Other = 1)			0.817
			(1.714)
Employment status (Full Time = 0, Part Time = 1)			0.582
			(0.916)
Employment status (Full Time = 0, Unemployed (and job seeking = 1)			-0.360
			(1.154)
Eigenvector * Similarity		2.634*	3.295**
		(1.577)	(1.454)
Eigenvector * Main-Effect		2.448	3.867**
		(1.707)	(1.567)
Similarity * Main-Effect	-1.439	-1.485	-1.675
	(1.029)	(1.198)	(1.083)
Degree * Similarity * Main-Effect	0.277**		
	(0.126)		
Eigenvector * Similarity * Main-Effect		3.024*	3.050**
		(1.669)	(1.478)
Constant	1.627	-1.201	-0.156
	(0.993)	(1.048)	(1.189)
Observations	10,300	10,300	8,560
R^2	0.025	0.030	0.038
Adjusted R^2	0.024	0.029	0.035

Residual Std. Error	17.823 (df = 10287)	17.780 (df = 10289)	17.737 (df = 8531)
F Statistic	22.211*** (df = 12; 10287)	31.498*** (df = 10; 10289)	12.073*** (df = 28; 8531)

Note: Standard errors are two-way clustered at the network and participant level.

*p < 0.10, **p < 0.05, ***p<0.01

Supplemental Table 3

	Dependent variable:			
	Starting Propensity to Donate	Contribution in Round 1	Perceived Descriptive Norm	Bias in Norm perception
	(1)	(2)	(3)	(4)
Eigenvector Centrality (z-scored)				0.403
				(0.422)
Similarity Attraction (1 = high, -1 = low)				0.777*
				(0.443)
Pre-Experiment Propensity to Donate to Public Good (z-scored)		28.459***	4.919***	1.149***
		(0.315)	(0.374)	(0.443)
Animal Moniker: Brown Bear	0.018	-0.780	0.037	2.670***
	(0.033)	(1.888)	(2.271)	(0.461)
Animal Moniker: Cheetah	0.008	-1.548	-1.551	0.796
	(0.033)	(1.888)	(2.292)	(2.258)
Animal Moniker: Coyote	0.023	-0.709	-0.649	-0.905
	(0.033)	(1.888)	(2.298)	(2.328)
Animal Moniker: Deer	-0.010	0.066	-1.131	0.004
	(0.033)	(1.888)	(2.261)	(2.188)
Animal Moniker: Elephant	0.030	-1.447	0.605	-0.582
	(0.033)	(1.888)	(2.271)	(1.964)
Animal Moniker: Fox	0.017	-0.596	0.664	1.558
	(0.033)	(1.888)	(2.292)	(2.321)
				1.367
				(2.126)

Animal Moniker: Giraffe	0.030	0.428	-3.141	-2.749
	(0.033)	(1.888)	(2.293)	(2.243)
Animal Moniker: Gorilla	0.025	2.648	-4.731**	-3.949*
	(0.033)	(1.888)	(2.303)	(2.271)
Animal Moniker: Grizzly Bear	0.022	0.934	-0.741	-0.082
	(0.033)	(1.888)	(2.287)	(2.183)
Animal Moniker: Kangaroo	-0.003	1.411	-2.923	-2.059
	(0.033)	(1.888)	(2.266)	(2.339)
Animal Moniker: Koala	-0.018	-4.011**	-1.235	-0.800
	(0.033)	(1.888)	(2.287)	(2.071)
Animal Moniker: Lion	0.003	0.256	-3.239	-2.813
	(0.033)	(1.888)	(2.292)	(2.389)
Animal Moniker: Panda	0.008	-1.548	-3.813*	-3.230
	(0.033)	(1.888)	(2.271)	(1.976)
Animal Moniker: Polar Bear	0.028	-0.638	-1.074	-0.325
	(0.033)	(1.888)	(2.267)	(2.183)
Animal Moniker: Tiger	0.005	-2.012	-4.173*	-3.577
	(0.033)	(1.888)	(2.276)	(2.185)
Animal Moniker: Wolf	0.000	-1.042	-1.269	-0.594
	(0.033)	(1.888)	(2.261)	(2.127)
Animal Moniker: Zebra	0.008	-0.923	-3.275	-2.114
	(0.033)	(1.888)	(2.257)	(2.228)
Eigenvector * Similarity				0.806**
				(0.402)
Eigenvector * Main-Effect				0.421
				(0.404)

Similarity * Main-Effect				0.580
				(0.443)
Eigenvector * Similarity * Main-Effect				1.072***
				(0.399)
Constant	0.543***	-3.921***	47.445***	-0.711
	(0.024)	(1.492)	(1.808)	(1.678)
Observations	2,160	2,160	1,969	1,969
R^2	0.003	0.793	0.090	0.060
Adjusted R^2	-0.005	0.792	0.081	0.047
Note:				*p < 0.10, **p < 0.05, ***p<0.01