Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

LLMs Generate Structurally Realistic Social Networks but Overestimate Political Homophily

Serina Chang[*1], **Alicja Chaszczewicz**[*1], **Emma Wang**[1], **Maya Josifovska**[1,2], **Emma Pierson**[3], **Jure Leskovec**[1]

[1]Department of Computer Science, Stanford University
[2]Department of Computer Science, University of California, Los Angeles
[3]Department of Computer Science, Cornell University

Abstract

Generating social networks is essential for many applications, such as epidemic modeling and social simulations. The emergence of generative AI, especially large language models (LLMs), offers new possibilities for social network generation: LLMs can generate networks without additional training or need to define network parameters, and users can flexibly define individuals in the network using natural language. However, this potential raises two critical questions: 1) are the social networks generated by LLMs realistic, and 2) what are risks of bias, given the importance of demographics in forming social ties? To answer these questions, we develop three prompting methods for network generation and compare the generated networks to a suite of real social networks. We find that more realistic networks are generated with "local" methods, where the LLM constructs relations for one persona at a time, compared to "global" methods that construct the entire network at once. We also find that the generated networks match real networks on many characteristics, including density, clustering, connectivity, and degree distribution. However, we find that LLMs emphasize political homophily over all other types of homophily and significantly *overestimate* political homophily compared to real social networks.

Code —
https://github.com/snap-stanford/llm-social-network

1 Introduction

The ability to generate realistic social networks is crucial for many applications, when the true social network cannot be observed (e.g., for privacy reasons) or a realistic network is desired between hypothetical individuals. For example, in epidemic modeling, synthetic social networks are frequently used so that researchers can model the spread of disease based on who has come into contact with whom (Barrett et al. 2009; Block et al. 2020). Synthetic networks are also useful for simulating and analyzing social media platforms (Pérez-Rosés and Sebé 2015; Sagduyu, Grushin, and Shi 2018) and social phenomena, such as polarization and opinion dynamics (Dandekar, Goel, and Lee 2013; Das, Gollapudi, and Munagala 2014).

Deep learning approaches to network generation typically require training on many domain-specific networks (You et al. 2018), making it difficult to generalize to new settings where networks are not yet observed. On the other hand, some classical network models require far less or no training, since they only have a few parameters, but in exchange for this simplicity, they make rigid and unrealistic assumptions about how social networks form. For example, Erdős–Rényi models only take in two inputs, n (the number of nodes) and p, and they assume each edge forms with uniform probability p (Erdős and Rényi 1959).

In contrast, LLMs balance these challenges: they can generate social networks in a zero-shot fashion, without any additional training or need to define network parameters, but they can also take in rich, flexible inputs describing individuals in natural language and use those inputs to generate a network between those individuals. A key question, however, is whether the social networks generated by LLMs are *realistic*. On one hand, LLMs have demonstrated capabilities to realistically simulate human responses and interactions (Aher, Arriaga, and Kalai 2023; Park et al. 2023; Argyle et al. 2023), suggesting that they may be able to generate realistic social networks as well. On the other hand, LLMs sometimes struggle with reasoning over graphs (Wang et al. 2023; Fatemi, Halcrow, and Perozzi 2024) and it is unclear if their language abilities generalize to structured objects like networks, so that they can reproduce structural characteristics of social networks such as low density and long-tailed degree distributions.

Furthermore, a central concern with using LLMs in social settings is bias. Prior works have shown that LLMs produce stereotyped descriptions of individuals based on their demographics (Cheng, Durmus, and Jurafsky 2023; Cheng, Piccardi, and Yang 2023) and skew towards the liberal opinions (Santurkar et al. 2023). These demographics, such as gender and political affiliation, play essential roles in the formation of real-world social networks, resulting in well-documented demographic homophily (McPherson, Smith-Lovin, and Cook 2001; Kossinets and Watts 2009; Halberstam and Knight 2016). Thus, we cannot evaluate whether LLMs' social network generation is realistic without incorporating demographics into our experiments; at the same time, we need to analyze how LLMs reason about these demographic features and investigate potential signs of bias.

[*]These authors contributed equally.

The goal of this work is to study these two issues—realism and bias—in the context of social network generation with LLMs. Our research questions are as follows:

- **RQ1:** Can LLM-generated networks match real-world social networks on structural characteristics? How do different prompting methods result in different structures?
- **RQ2:** Can LLMs capture demographic homophily? How do levels of homophily vary across demographic variables, and are there signs of bias?
- **RQ3:** How does incorporating interests, beyond demographics, affect LLM-generated networks?

To answer these questions, we propose three zero-shot prompting methods for social network generation (Figure 1). First, we find that "local" methods, where the LLM takes on the perspective of one person at a time, yield more realistic networks than "global" methods, where the LLM constructs the entire network at once, even though the LLM receives less information in the local setting. The LLM is able to generate networks that match real networks on many structural characteristics, including density, clustering, connectivity, and degree distribution. The LLM also exhibits clear demographic homophily, across gender, age, race/ethnicity, religion, and political affiliation. However, the LLM consistently emphasizes homophily in political affiliation above all other demographic variables, and overestimates levels of political homophily compared to reported levels in real online *and* offline social networks. Finally, we find that incorporating LLM-generated interests does not reduce political homophily, since the interests themselves encode political stereotypes. Overall, our work demonstrates the promise of using LLMs for social network generation while calling attention to challenges around integrating demographics.

2 Related Work

Social simulation with LLMs. Prior work has demonstrated LLMs' abilities to realistically simulate human responses and interactions (Aher, Arriaga, and Kalai 2023; Park et al. 2023; Argyle et al. 2023) and studied the dynamics of LLM agents interacting in a population, such as how conventions or consensus arise (Ashery, Aiello, and Baronchelli 2024; Marzo, Castellano, and Garcia 2024). However, while simulating interactions over social networks, existing work focuses less on using LLMs to generate the networks themselves, either making simplistic assumptions about the network structure such as sampling agents randomly to interact (Park et al. 2022; Chuang et al. 2023; Ashery, Aiello, and Baronchelli 2024) or assuming fully connected networks (Marzo, Castellano, and Garcia 2024), or requiring human involvement in building the network (Gao et al. 2023; Zhou et al. 2024). To improve the realism and usability of these simulations, it is essential to also explore LLMs' abilities to *generate* the network structure, a prerequisite to simulating interactions over networks.

A few contemporaneous works have explored LLMs for social network generation, with different focuses from ours. Marzo, Pietronero, and Garcia (2023) focus on degree distribution, showing that scale-free networks emerge from interactions between LLMs. He et al. (2023) focus on content homophily, analyzing a simulated society powered by LLM chatbots. The most similar work to ours is Papachristou and Yuan (2024), who analyze whether LLMs demonstrate network formation principles, such as preferential attachment and homophily. While their work establishes the existence of *general* network principles, our work compares generated networks to real social networks directly, computing many network metrics, and shows that all metrics can be matched at once, while their experiments primarily explore one principle at a time, with a different prompt for each principle. Also, to test homophily, they consider hobby, favorite color, and location, while we explore key demographic features.

LLM social biases. Using LLMs in social contexts raises concerns of biases and stereotyping (Cheng, Durmus, and Jurafsky 2023; Cheng, Piccardi, and Yang 2023; Wang, Morgenstern, and Dickerson 2024). When responding to public opinion or political questions, LLMs' answers often skew liberal (Santurkar et al. 2023; Hartmann, Schwenzow, and Witte 2023). When assigned a persona, LLMs show worse reasoning capabilities when assigned certain demographics (Gupta et al. 2024) and produce more toxic content under certain personas (Deshpande et al. 2023). However, bias in the context of social network generation remains unexplored. In this work, we investigate such biases by studying the effects of demographic variables on LLM-generated social networks.

Graph generation. Deep learning approaches aim to learn graph generation directly from observed data, resulting in realistic generated networks (You et al. 2018; Simonovsky and Komodakis 2018; Guo and Zhao 2023). However, these models require a set of training graphs, and typically have fixed feature sets, so they cannot be applied in situations without access to real-world network data, and cannot be easily extended to incorporate new features, while LLMs can be. In contrast, some simple classical models can generate networks with only 1-2 parameters, which means they can generate networks without requiring training and with minimal need to define parameters. However, their simplicity often results in unrealistic network structures. For example, Erdős–Rényi models assume that each edge forms with uniform probability (Erdős and Rényi 1959), Watts–Strogatz models (Watts and Strogatz 1998) generate small-world networks but struggle to produce realistic degree distributions, and Barabási–Albert models (Barabási and Albert 1999) capture power-law degree distributions but miss community structure and clustering.

LLMs & graphs. The use of LLMs has been explored for various graph tasks (Li et al. 2024; Jin et al. 2024), such as graph reasoning (Wang et al. 2023; Fatemi, Halcrow, and Perozzi 2024), node classification (Zhao et al. 2023; Chen et al. 2024; Ye et al. 2024), or tasks on knowledge graphs (Pan et al. 2024). These works have shown that LLMs possess preliminary graph reasoning capabilities, but struggle on larger graphs or harder tasks, such as finding a Hamilton path (Wang et al. 2023). Yao et al. (2024) is one of the first to explore LLMs for graph generation, but they explore it in the context of molecule generation, not social networks.

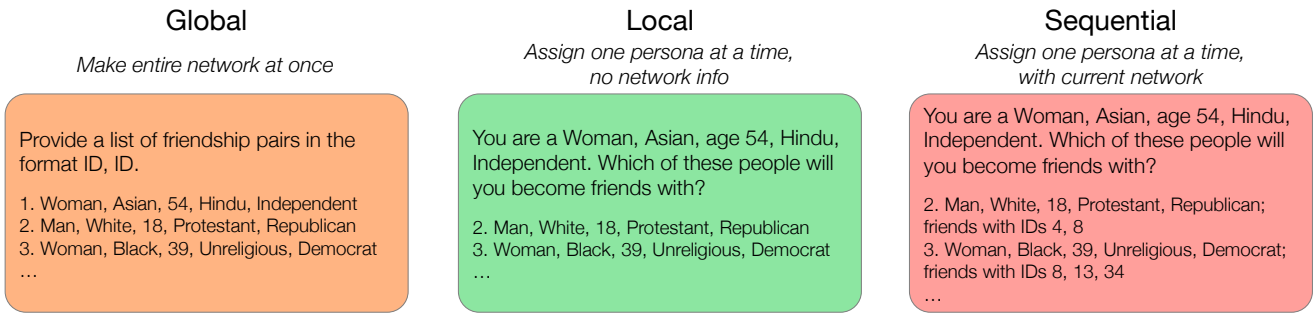

Figure 1: Our three prompting methods to generate social networks with LLMs. See full prompts in Figures C3-C5.

3 Generating Social Networks

Our process for generating social networks involves two steps: first, constructing a set of personas, and second, having the LLM generate networks over those personas. Here, we summarize the two steps, with details in Appendix C.

3.1 Persona Construction

For each persona, we include their gender, age, race/ethnicity, religion, and political affiliation, which are salient dimensions of homophily in real social networks (McPherson, Smith-Lovin, and Cook 2001; Halberstam and Knight 2016). We sample these characteristics based on the distribution of the US population. Using US Census data (US Census Bureau 2023), we acquire the joint distribution for gender, age, and race/ethnicity. Then, we sample the persona's religion, conditioned on their race/ethnicity (Statista 2016; PRRI Staff 2021), and political affiliation, conditioned on their gender and race/ethnicity (Pew Research Center 2024; Sanchez and Foxworth 2022). In Section 5.3, we also experiment with including interests for each persona, instead of only demographic variables.

3.2 Network Generation

We design three prompting methods for generating social networks, which we summarize in Figure 1.

Global. In our first method, which we call "Global", we provide the LLM with the entire list of personas, and prompt it to construct the network between them, in the form of edge pairs (referring to each persona's ID).

Local. In our second method, which we call "Local", we have the LLM take on the identity of one persona at a time, e.g., by saying, "You are a Woman, Asian, age 54, Hindu, Independent." We provide the LLM with the list of all other personas (in the same format) and prompt it to pick friends for the persona it is currently assigned. To construct the entire network, we iterate through all personas in a random order, and we keep an edge between personas A and B if the LLM selects B when acting as A or vice versa (so we do not require both to select each other). This method is inspired by techniques in machine learning that similarly model the graph generation process by iterating through nodes and selecting edges for each node at a time (You et al. 2018).

Sequential. In our third method, we also assign the LLM one persona at a time, but in addition to providing the list of all other personas, we also provide information about the constructed graph so far. We experiment with providing each persona's full list of current friends versus only their degree (i.e., total number of friends). These variations are similar to the preferential attachment experiments in Papachristou and Yuan (2024), where they also experimented with providing neighborhood information versus only degree, although their experiments only considered the network and no demographic features. In contrast, our experiments—and in particular, the comparison of the Local to Sequential methods—reveal how providing demographic and network information compare to only providing demographic information.

4 Comparison to Real Networks

To evaluate the realism of our generated networks, we gather a set of real social networks from the CASOS (CASOS 2024) and KONECT (KONECT 2024) repositories. We kept networks that described *friendships* between individuals, which filtered out other types of networks, such as work-related interactions or visiting ties between families. We included eight real networks, which capture friendships within diverse communities, such as among physicians, students, and prisoners (see Appendix C.3 for details).

We extract graph-level and node-level metrics from the real networks and our generated networks, and compare their distributions. For consistency, we treat all networks as undirected. Since the number of nodes varies across networks, we focus on network metrics that are comparable across graphs of different sizes, and scale those that are dependent on network size based on how they are expected to scale in an Erdös-Rényi random graph (Erdős and Rényi 1959). Below, we define and motivate the network metrics that we evaluate on.

Density. A basic property of a network is its density of edges, and social networks tend to be sparse, meaning lower density (Wong, Pattison, and Robins 2006). Density computed as the number of observed edges divided by the total number of possible edges in the network, which comes out to $\frac{2E}{N(N-1)}$, where N is the number of nodes and E is the number of edges in the network.

Average clustering coefficient. Social networks are known to exhibit clustering, where one's friends are likely to be friends with each other (Alizadeh 2017). For a node i, its clustering coefficient is $\frac{2E_i}{N_i(N_i-1)}$, where N_i is its number of neighbors and E_i is the number of edges between its neighbors. The average clustering coefficient computes the average over nodes.

Largest connected component (LCC). Social networks are known to be well-connected (Ugander et al. 2011), with the vast majority (over 99%) of the nodes in the largest connected component (LCC), i.e., the largest subgraph where all nodes within the subgraph are reachable by each other. Thus, as a metric, we compute the proportion of all nodes in the network that are in the LCC, $\frac{N_{LCC}}{N}$.

Average shortest path. Social networks are not only well-connected, meaning nodes can reach each other, but also they can reach each other in relatively short paths (Alizadeh 2017). So, we measure the average shortest path over all pairs of nodes in the LCC, divided by $\log N_{LCC}$, since shortest paths scale with $\log N$ in Erdös-Rényi graphs (Watts and Strogatz 1998). We compute shortest paths within the LCC instead of the entire network, since the distance between two disconnected nodes is infinite.

Community structure. Social networks exhibit strong community structure, with more edges within communities and fewer edges across communities (Newman 2004). To measure community structure, first we use the Louvain algorithm (Blondel et al. 2008) to partition the network into communities, then we assess the quality of the partition with modularity (Eq. 5). Higher levels of community structure correspond to higher modularity.

Degree distribution. Social networks are said to follow a power law degree distribution, where $P(k) \sim k^{-\gamma}$, for degree k and constant γ (Barabási and Albert 1999). This results in long-tailed degree distributions with a few people having far more friends than most others. To measure degree distribution, we compute the degree of each node in the network, and to make degree comparable across graphs, we divide degree by N. To summarize degree distribution from a set of networks, we pool all of the (normalized) degrees of nodes in the networks in the set, and compute the distribution over the pooled degrees in bins of 0.05, from 0 to 1.

Homophily. Finally, social networks are known to exhibit homophily, where "birds of a feather flock together", i.e., people with similar traits are likelier to be friends (McPherson, Smith-Lovin, and Cook 2001). To measure homophily, we use a common metric adopted in prior work (McPherson, Smith-Lovin, and Cook 2001; Easley and Kleinberg 2010; Smith, McPherson, and Smith-Lovin 2014): the ratio of observed-to-expected cross-group edges. Specifically, first we compute the *observed* proportion of edges that are cross-group (e.g., different gender), then we compute the *expected* proportion of edges that are cross-group (based on the number of nodes that belong to each group), then we take the ratio of these two proportions. We define this as

$$H = \frac{C_{\text{obs}}}{C_{\text{exp}}} = \frac{\frac{\sum_{ij} A_{ij} \cdot \mathbb{1}[g_i \neq g_j]}{E}}{\frac{\sum_g \sum_{g' \neq g} N_g N_{g'}}{N(N-1)}}, \quad (1)$$

where A_{ij}, as the adjacency matrix, is 1 if nodes i and j are connected and 0 otherwise; g_i indicates node i's group; and N_g is the number of nodes in group g.[1] If the ratio is below 1, this indicates homophily, since there are fewer cross-group edges than expected, while ratios above 1 indicate heterophily (e.g., this appears in heterosexual dating networks).

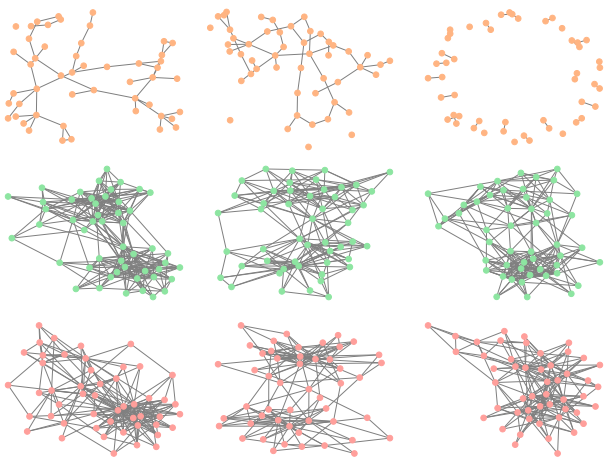

Figure 2: Examples of social networks generated by our three prompting methods: Global (top), Local (middle), and Sequential (bottom).

5 Results

Experimental set-up. We experiment with the following LLMs: OpenAI's GPT-3.5 Turbo and GPT-4o (Brown et al. 2020; OpenAI et al. 2023), Meta's Llama 3.1 (8B and 70B) (Touvron et al. 2023), and Google's Gemma 2 (9B and 27B) (Gemma Team et al. 2024). These six models represent a range across companies, different model sizes, and proprietary (the GPT models) versus open-source (the Llama and Gemma models). We find that GPT-3.5 Turbo performs the best at matching the real social networks, so we report results from GPT-3.5 Turbo in the main text, but we report results from all models in Appendix A. We show that our main results about political homophily being most emphasized and overestimated hold for all six models. We also include sensitivity analyses, with different temperatures (Figure B5) and minor changes to the prompt (Figure B6), and show that results are stable.

We sample $N = 50$ personas and we use the same set of personas for all LLM experiments. In Table C1, we report the demographic make-up of these 50 sampled personas. For each prompting method, we generate 30 networks, to

[1]For age, we use the average age difference in observed edges divided by the expected age difference (i.e., average difference over all possible pairs of nodes).

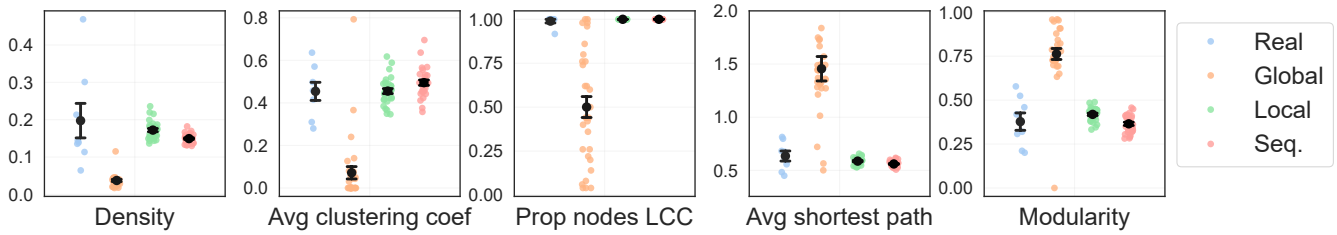

Figure 3: Graph-level metrics over real and generated social networks. We visualize mean and standard error (in black) and individual data points corresponding to each network.

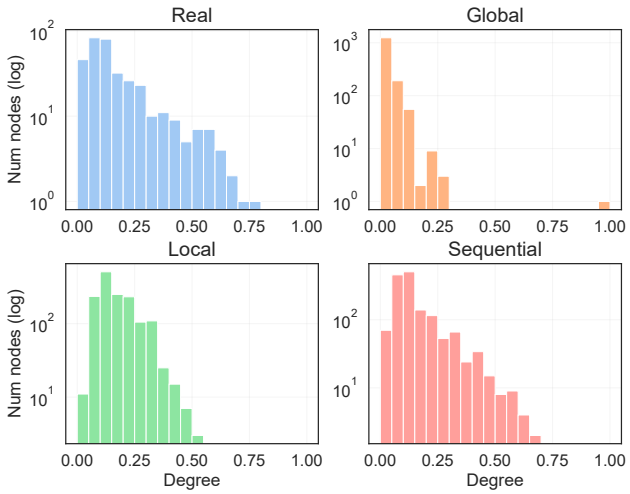

Figure 4: Degree distributions over real and generated social networks. For each set of networks, we pool degrees over nodes in the networks (Section 4).

capture variation in prompting (e.g., order of personas) and model response. For the graph-level metrics (all metrics in Section 4 besides degree distribution), we visualize their mean and standard error, along with individual data points (Figures 3 and 5). Visualizing both the standard error and individual data points capture inferential uncertainty and outcome variability, as recommended by Zhang et al. (2023). For degree (Figure 4), we visualize the pooled degree distribution over nodes in the networks, as described in Section 4.

5.1 Evaluating Network Structure

Here, we describe our main results from evaluating the structure of the generated networks.

Local and Sequential are more realistic than Global. First, we find that the prompting methods produce visually different network structures, as shown in Figure 2. Furthermore, the networks produced by the Global method are far less realistic than those produced by Local and Sequential. As shown in Figures 3-4, Global has unrealistically low density, clustering, and connectivity, too much community structure, and misses the long tail of the degree distribution. In comparison, Local and Sequential overlap with the real distributions for all graph-level metrics and show much greater variation in node degrees.

Thus, LLMs produce more realistic social networks when we assign the LLM to act as one persona at a time, instead of prompting it to produce the entire network at once. This is interesting, since the LLM has strictly less information under the one-persona setting: in the Local setting, it has no access to any network information, only making local decisions per persona based on demographics, and in the Sequential setting, it only knows the network based on previous personas' choices without any ability to see into the future. In comparison, the Global method allows the LLM to take into account the entire network at once, along with all personas' demographic information, so that it can theoretically consider dependencies between all these pieces of information. However, the LLM is not able to effectively leverage all of these dependencies and instead produces far less realistic networks.

Sequential captures long-tail degree distribution. By comparing Local and Sequential, we can isolate the impact of incorporating network information, on top of providing demographic information. The main difference between the two methods is the degree distribution: as shown in Figure 4, Sequential gets much closer to the real degree distribution, in terms of exhibiting a long tail, while Local approximately matches the mode without matching the tail. We also find that Sequential exhibits slightly less community structure and less homophily than Local, which makes sense since Local only matches on demographic similarity while Sequential also takes into account network information.

Thus, the Sequential method is able to match real social networks on many structural characteristics. In Table B1, we quantify how well each LLM method matches the real networks, and compare to classical network models with few parameters, including random graph models (Erdős and Rényi 1959), small-world models (Watts and Strogatz 1998), and preferential attachment models (Barabási and Albert 1999). Even in the best case—when we allow the classical models to choose parameters based on the real networks—these models cannot match all of the characteristics as well as our Local or Sequential methods can. For example, when using the Kolmogorov-Smirnov statistic (Eq. 7) to measure the distance between the generated versus real networks' distributions per structural characteristic, Sequential achieves an average distance of 0.330, while

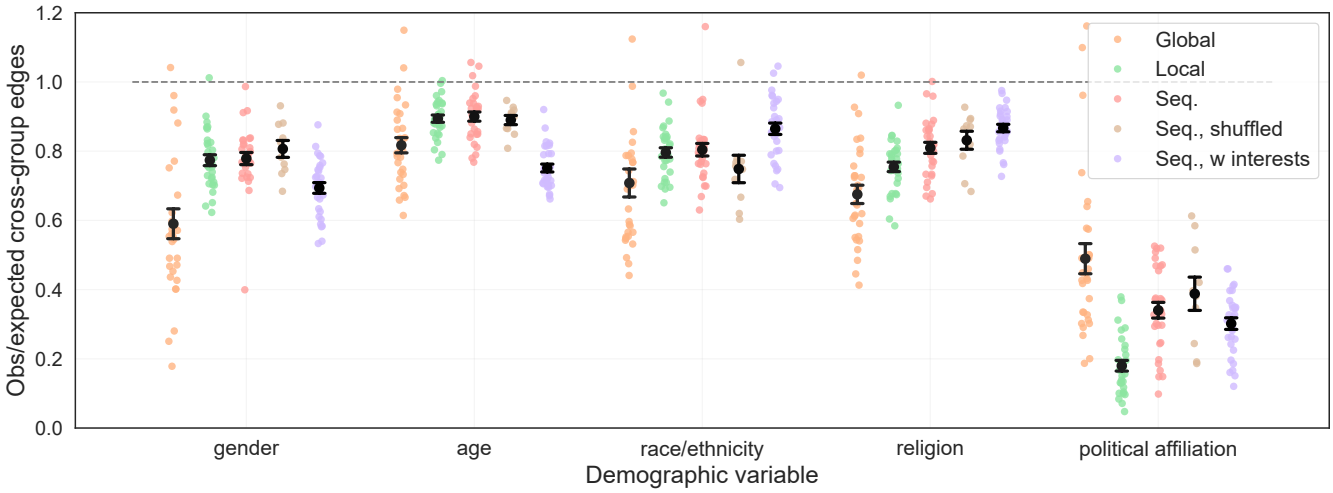

Figure 5: Rates of homophily in our generated networks, per demographic variable. Ratios below 1 (marked by the grey line) indicate homophily, with lower ratios indicating more homophily. We visualize mean and standard error (in black) and individual data points corresponding to each network.

the best classical model (small-world) yields an average distance of 0.499, 51% higher than Sequential. Thus, the LLM achieves better realism than these classical network models when generating social networks, while maintaining their advantages of requiring no additional training or minimal need to define parameters.

5.2 Evaluating Homophily

We measure homophily for gender, age, race/ethnicity, religion, and political affiliation, using the ratio of observed-to-expected proportion of cross-group edges (Eq. 1).

LLMs capture homophily, with greatest emphasis on politics. In Figure 5, we show that, across all prompting methods and demographic variables, the ratio is significantly below 1, indicating that the generated networks clearly exhibit homophily. Furthermore, we see different levels of homophily for different demographic variables. For the more realistic Local and Sequential methods, homophily is by far the strongest for political affiliation: observed cross-party relations are 82% less frequent than expected under Local and 66% less frequent than expected under Sequential. Rates of political homophily are even stronger for the other LLMs that we test, most extremely for GPT-4o and Llama 3.1 70B, where none of the edges are cross-party and the network fractures into two disconnected components (Appendix A).

Does this mean that LLMs actually pay the most attention to political affiliation when choosing social ties? Despite political homophily being the strongest, this could be due to correlations between political affiliation and other demographics; for example, hypothetically, if all Democrats had the same gender and race, and all Republicans had the same gender and race, then apparent homophily in political affiliation could actually be due to similarity in other demographics. To test this, we try shuffling the demographics, so that, while maintaining the same numbers of each group per demographic, each persona is randomly assigned to a group, thus removing correlations between demographics. When we run Sequential with these shuffled personas, we find that political homophily remains by far the strongest (Figure 5), demonstrating that the LLM is, in fact, paying most attention to political affiliation when choosing social ties. As an additional test, we also try ablations of the demographics, where we present the LLM with only one demographic variable at a time, or with two variables, one being political affiliation and the other being one of the four others. Here, we also find that political affiliation continues to be the dominant factor: when only one variable is presented, levels of homophily increase for *all* demographics but it remains the highest for political affiliation; when two variables are presented, political homophily is always stronger than the other demographic's homophily (Figure B2).

Finally, we directly test what the LLM is paying attention to by prompting it to generate a short reason for each friend that it selects. Then, given the reason, we use GPT-4o to classify the reason, e.g., "I'm a woman too and we share the same religion and political affiliation" is classified as [gender, religion, political affiliation]. We allow the LLM to generate free-text reasons during network generation since we do not want to constrain its response (e.g., if it is using other information, such as degree or ID). We find that political affiliation strongly dominates here as well: it is

Demographic	**Reason** %
Political affiliation	86.7%
Religion	43.0%
Age	21.8%
Race/ethnicity	12.1%
Gender	7.3%

Table 1: Frequency that each demographic is part of the LLM's reason for choosing a persona as a friend.

Measure	Local	Sequential	Source	Description	Value
Cross-group ratio (Eq. 1) ↓	0.180 (0.015)	0.340 (0.022)	Halberstam and Knight (2016)	Twitter	0.528
Same-group ratio (Eq. 8) ↑	1.851 (0.016)	1.685 (0.023)	Halberstam and Knight (2016)	Twitter	1.404
Isolation index (Eq. 9) ↑	0.729 (0.020)	0.530 (0.027)	Halberstam and Knight (2016)	Twitter	0.403
			Gentzkow and Shapiro (2011)	Voluntary associations	0.145
				Work	0.168
				Neighborhood	0.187
				Family	0.243
				People you trust	0.303
				Political discussants	0.394
Polarization (Eq. 10) ↑	0.639 (0.037)	0.515 (0.041)	Garimella and Weber (2017)	Twitter, follow	0.33-0.42
				Twitter, retweet	0.37-0.41

Table 2: Comparing political homophily in our generated networks to real-world networks. We consider different measures of homophily and indicate with ↓ or ↑ the direction that indicates greater homophily for that measure. For Local and Sequential, we report the mean and standard error (in parentheses) over each method's 30 generated networks.

part of the reason 86.7% of the time, while the next most-mentioned demographic, religion, is only mentioned 43.0% of the time (Table 1). As a caveat, prior work has shown that we cannot always trust an LLM's own explanation for its choices (Agarwal, Tanneru, and Lakkaraju 2024). However, given the alignment of these results with our other results, there is strong evidence that LLMs pay the most attention to political affiliation when generating social relations.

LLMs *overestimate* political homophily. Given the LLM's emphasis on political homophily, we seek to compare its level of political homophily to reported levels from real-world social networks. We are not able to compare to the eight social networks from Section 5.1 here since we do not have demographic features per node. However, we are able to find reported political homophily in several papers, covering both online and offline social networks. In Table 2, we summarize these comparisons, showing that Local and Sequential consistently overestimate political homophily across different measures of homophily. For example, Halberstam and Knight (2016) analyze political homophily on Twitter. In their data, cross-party relations appear 47% less often than expected, which indicates homophily, but not as strong as what the LLM predicts. In addition to the cross-group ratio, we can also compute a *same-group* ratio, using the ratio of observed-to-expected proportion of same-group edges, where a higher ratio indicates more homophily (Eq. 8). Using this measure, we find that same-party relations only appear 40% more often than expected in the Twitter data, while same-party relations are 85% and 68% more frequent than expected for Local and Sequential, respectively.

Halberstam and Knight (2016) also compute the isolation index, which is the difference in average conservative exposure between conservatives and liberals (Eq. 9), with larger indices indicating greater isolation. They find an isolation index of 0.403, while the LLM's is far higher, at 0.720 for Local and 0.530 for Sequential. Furthermore, the authors note that homophily could be overestimated in their data, since they selected users who follow politicians, which "may tend to disproportionately include individuals with strong preferences for linking to like-minded users." Even though their homophily results could be overestimates, the LLM's estimates still significantly exceed theirs.

Isolation indices are even lower in Gentzkow and Shapiro (2011), who study ideological segregation in social networks. In face-to-face interactions, the highest isolation indices they report are 0.243 (family), 0.303 (people you trust) and 0.394 (among people who discuss politics). Finally, we also compare to Garimella and Weber (2017), who study political polarization on Twitter and define it as $p_i = 2 \cdot |0.5 - \alpha/(\alpha + \beta)|$, where α and β indicate how many left-leaning and right-learning users, respectively, are followed by user i, and they take the average p_i over users (Eq. 10). Their measure captures the difference between observed leaning and a balanced leaning of 0.5, with higher numbers indicating greater polarization. They report polarization levels of 0.33-0.42 for following relations and 0.37-0.41 for retweets (increasing over time), while we find significantly higher levels in the LLM's generated networks, with 0.639 for Local and 0.515 for Sequential.

We hypothesize that the LLM overestimates political homophily partially due to high levels of polarization in its online pretraining data and frequent discussions of such polarization, although future work is needed to carefully study this phenomenon. These results also have important implications if one seeks to run experiments over social networks generated by LLMs. For example, overestimated political homophily may result in unrealistically high levels of polarization, potentially leading to incorrect conclusions and interventions.

Homophily between pairs of groups. As described in Eq. 1, we measure homophily as the ratio of observed-to-expected proportion of cross-group edges. Now, we extend this definition to compute an observed-to-expected ratio for

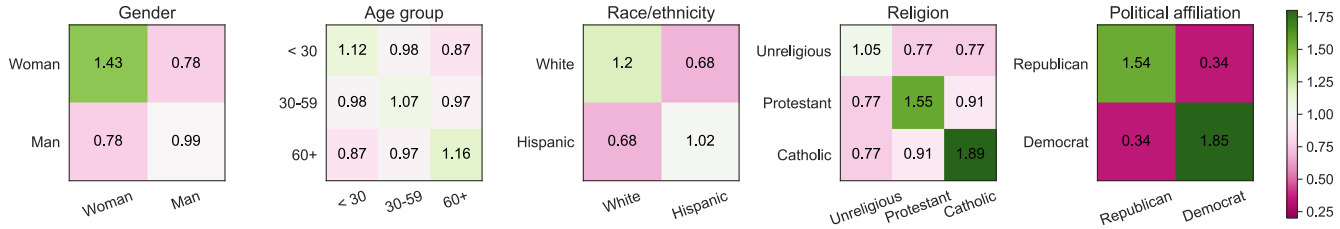

Figure 6: Ratio of observed-to-expected proportion of edges, for all pairs of demographic groups (Eq. 2). All subfigures share the same colormap (right). Groups with at least 10 nodes are kept.

any pair of groups, A and B (where A could equal B):

$$H_{AB} = \frac{\frac{\sum_{ij} A_{ij} \cdot \mathbb{1}[g_i = A] \cdot \mathbb{1}[g_j = B]}{E}}{\frac{N_A(N_B - \mathbb{1}[A = B])}{N(N-1)}}. \quad (2)$$

This measures the observed proportion of edges that are between nodes in groups A and B divided by the expected proportion of edges between nodes in these groups. In Figure 6, we visualize H_{AB} for all pairs, for each of the five demographic variables. For each variable, we keep all groups with at least 10 personas in our set of 50 personas (Table C1), and compute H_{AB} for each pair of groups, reporting the mean ratio over the 30 generated networks from Sequential.

We find, as expected, that the diagonal (i.e., same-group ratios) tends to be above 1, although notably it is not for relations between men, with a ratio of 0.99, while relations between women have a ratio of 1.43. Thus, even within one demographic variable, the LLM's levels of homophily vary for different groups, such as homophily within women being stronger than homophily within men. Variability across groups is present for all demographic variables: most extremely, for religion, relations between Catholics occur 89% more frequently than expected, while relations are only 5% more frequent for personas that identify as Unreligious. This plot also reveals that not all cross-group relations are equally unlikely: for example, when we divide age into three age groups, the adjacent age groups have higher cross-group ratios (0.98 for under 30 and 30-59 and 0.97 for 30-59 and 60+) compared to the non-adjacent age groups (0.87 for under 30 and 60+). Finally, this plot reiterates that political affiliation has the lowest cross-group ratio, at 0.34, and reveals that it also has among the highest same-group ratios, although the LLM gives Democrats stronger same-group preferences, with a same-group ratio of 1.85, than Republicans, with a same-group ratio of 1.54.

5.3 Incorporating Interests

A natural question is whether demographic homophily is exaggerated because we only give the LLM demographic information, without other important details such as the person's interests. Thus, we run an additional set of experiments where we allow the LLM to also generate interests for each persona. To generate interests, we prompt the LLM with, "In 8-12 words, describe the interests of someone with the following demographics" (full prompt in Figure C2). In Table 3, we provide examples of the generated interests, with the full list of personas with interests available online.

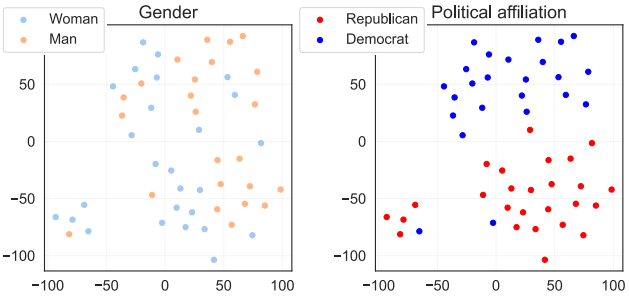

Figure 7: Embeddings of interests, with T-SNE projection to 2D. Each dot is a persona, colored by its gender (left) or its political affiliation (right).

Effect of interests on networks. As shown in Figure 5, by comparing Sequential vs. Sequential with interests, we find that incorporating interests strengthens homophily for some demographics (clearly for gender and age and slightly for political affiliation) and weakens homophily for the others (race/ethnicity and religion). Notably, after these adjustments, political homophily remains the strongest across the demographic variables. Thus, our finding that LLMs prioritize political homophily is robust to incorporating interests.

In fact, the interests themselves strongly encode political homophily, with evidence of political stereotyping. For example, among the Democrat personas, the most common interests are "social justice" (62.5% of Democrats vs. 0.2% of Republicans), "community service" (29.3% vs. 13.9%), and "progressive policies" (18.6% vs. 0%). In contrast, among the Republican personas, the most common interests are "conservative politics" (41.6% of Republicans vs. 0% of Democrats), "church activities" (32.1% vs. 13.2%), and "gardening" (23.2% vs 16.4%). We provide the complete list of top 10 interests per demographic group in Table B4. We also analyze interests by mapping them to text embeddings, using OpenAI's text-embedding-3-small model. In Figure 7, we visualize the embeddings, coloring them by their persona's gender (left) and political affiliation (right). We compare these two demographic variables, since both have two, approximately equally sized groups in our sam-

Gender	Race / Ethnicity	Age	Religion	Political Affiliation	Interests
Man	White	47	Protestant	Republican	Hunting, fishing, classic rock, church activities, patriotic events, home improvement
Woman	Black	69	Unreligious	Republican	History, gardening, community service, classic jazz music, financial news, travel
Man	Hispanic	75	Unreligious	Democrat	Historical documentaries, community events, family gatherings, literature on social justice
Man	American Indian / Alaska Native	30	Protestant	Republican	Outdoor activities, traditional crafts, conservative politics, music, community service, history
Woman	Asian	58	Catholic	Democrat	Volunteering, social justice, culinary arts, family activities, church community involvement

Table 3: Examples of LLM-generated interests for personas with different demographics.

ple of 50 personas. From this comparison, we can see how much more distinct the political groups are than the gender groups, demonstrating the level of political homophily encoded in the interests. Finally, we conduct a supplementary experiment where we try network generation with *only* interests and no demographics. Homophily decreases across all demographic variables, but political homophily remains by far the strongest (Figure B4).

These results demonstrate how the LLM's emphasis on political affiliation also appears in interest generation, and, as a result, incorporating LLM-generated interests cannot help to reduce overestimated political homophily. These results build on prior results showing that LLM sometimes exhibit bias in political settings (Cheng, Piccardi, and Yang 2023; Santurkar et al. 2023; Wang, Morgenstern, and Dickerson 2024), exploring these issues through a novel lens of homophily and social network generation.

6 Discussion

Our work has established several findings. First, with the right prompting method, the LLM is able to simultaneously match many structural characteristics of real social networks, outperforming classical network models with few parameters. Second, "local" prompting methods produce more realistic networks than "global" methods and, within local methods, adding network information (i.e., Sequential) helps the LLM capture long-tailed degree distributions. Third, the LLM exhibits clear homophily across five key demographic variables, but political homophily dominates, to the extent that it is overestimated relative to real-world measures. Finally, incorporating LLM-generated interests does not reduce these overestimates, since the interests themselves encode strong political homophily.

Future directions. Our findings demonstrate the promise of generating social networks with LLMs, as they are zero-shot, flexible, and structurally realistic. However, more needs to be done to address potential biases, especially with regards to political homophily. One possibility could be to incorporate more information per persona: we found in our ablations that the LLM tends to produce stronger homophily per variable when fewer variables are provided (Figure B2), so we might see political homophily reduce if more variables or richer descriptions (such as interests) are provided per personas. However, given our findings on LLM-generated interests, researchers may want to handcraft interests with fewer political stereotypes, although other methods would be needed to scale such efforts to large social networks. Furthermore, while we focus in this work on how LLMs overestimate political homophily compared to real-world measures, in the Appendix we discuss real-world measures of homophily for the other demographic variables and find that the LLM, in fact, seems to *underestimate* homophily for race/ethnicity and religion (Table B2), calling for future work that thoroughly investigates potential biases along the other demographic dimensions.

There are also other limitations to LLM social network generation: for example, while we have shown that Sequential can approximately match the *means* of the real networks' metrics, its generated networks consistently demonstrate less *variance* than the real networks (Figure 3), reflecting known issues of LLMs to flatten demographic groups (Wang, Morgenstern, and Dickerson 2024) and lack output diversity (Kirk et al. 2023). Furthermore, all of our methods require listing N personas per prompt, which becomes infeasible with larger networks, due to context windows and cost. In Table C2, we conduct a big-O analysis of how the number of tokens scales with network size, revealing a tradeoff between the improved realism of Sequential and Local versus the lower costs of Global. To address this tradeoff, future work could explore how to make the Sequential or Local methods more scalable: for example, instead of listing N personas per prompt, where each persona chooses from the full list of all other personas, one could provide a *subset* of the other personas. In Appendix C.2, we discuss a simple implementation of this idea where the subsets are sampled uniformly at random, and show how this extension enables us to generate networks with 300 personas (sampling 30 personas per prompt). These larger networks are similar to our smaller networks in terms of homophily and some structural characteristics, although density and clustering drop due to the subsampling (Figure C6).

In future work, we hope to further extend these methods to make them both scalable and realistic, as well as generate networks with greater variance. We also hope to study whether our results on political homophily generalize outside of the US, and to explore methods to mitigate the political biases we observe in this work.

Acknowledgments

We thank Joon Sung Park and Marios Papachristou for helpful comments. S.C. was supported in part by a Meta PhD Fellowship and NSF Graduate Research Fellowship. We also gratefully acknowledge the support of NSF under Nos. OAC-1835598 (CINES), CCF-1918940 (Expeditions), DMS-2327709 (IHBEM); Stanford Data Applications Initiative, Wu Tsai Neurosciences Institute, Stanford Institute for Human-Centered AI, Chan Zuckerberg Initiative, Amazon, Genentech, GSK, Hitachi, SAP, and UCB.

References

Agarwal, C.; Tanneru, S. H.; and Lakkaraju, H. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *arXiv preprint arXiv:2402.04614*.

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*.

Alizadeh, M. 2017. Generating and analyzing spatial social networks. *Computational and Mathematical Organization Theory*, 23: 362–390.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2024. The Dynamics of Social Conventions in LLM populations: Spontaneous Emergence, Collective Biases and Tipping Points. *arXiv preprint arXiv:2410.08948*.

Barabási, A.-L.; and Albert, R. 1999. Emergence of scaling in random networks. *Science*, 286(5439): 509–512.

Barrett, C. L.; Beckman, R. J.; Khan, M.; Kumar, V. S. A.; Marathe, M. V.; Stretz, P. E.; Dutta, T.; and Lewis, B. 2009. Generation and analysis of large synthetic social contact networks. In *Proceedings of the 2009 Winter Simulation Conference (WSC'09)*, 1003–1014.

Block, P.; Hoffman, M.; Raabe, I. J.; Dowd, J. B.; Rahal, C.; Kashyap, R.; and Mills, M. C. 2020. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behaviour*, 4: 588–596.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Brown, A. 2022. About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth. *Pew Research Center*. https://pewrsr.ch/3Qi2Ejd.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

CASOS. 2024. Public Datasets. http://www.casos.cs.cmu.edu/tools/datasets/external/index.php. Accessed June 30, 2024.

Chen, Z.; Mao, H.; Li, H.; Jin, W.; Wen, H.; Wei, X.; Wang, S.; Yin, D.; Fan, W.; Liu, H.; et al. 2024. Exploring the potential of large language models (LLMs) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2): 42–61.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, 1504–1532.

Cheng, M.; Piccardi, T.; and Yang, D. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*, 10853–10875.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2023. Simulating Opinion Dynamics with Networks of LLM-based Agents. *arXiv preprint arXiv:2311.09618*.

Coleman, J.; Katz, E.; and Menzel, H. 1957. The Diffusion of an Innovation Among Physicians. *Sociometry*, 20(4): 253–270.

Dandekar, P.; Goel, A.; and Lee, D. T. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences (PNAS)*, 110(15): 5791–5796.

Das, A.; Gollapudi, S.; and Munagala, K. 2014. Modeling Opinion Dynamics in Social Networks. In *Proceedings of the 7th ACM international conference on Web search and data mining (WSDM'14)*, 403–412.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270.

Easley, D.; and Kleinberg, J. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge university press Cambridge. Available at https://www.cs.cornell.edu/home/kleinber/networks-book/.

Erdős, P.; and Rényi, A. 1959. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297.

Fatemi, B.; Halcrow, J.; and Perozzi, B. 2024. Talk like a Graph: Encoding Graphs for Large Language Models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Garimella, K.; and Weber, I. 2017. A Long-Term Analysis of Polarization on Twitter. In *ICWSM'17: Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 528–531.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gemma Team; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

Gentzkow, M.; and Shapiro, J. M. 2011. Ideological Segregation Online and Offline. *The Quarterly Journal of Economics*, 126: 1799–1839.

Guo, X.; and Zhao, L. 2023. A Systematic Survey on Deep Generative Models for Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5370–5390.

Gupta, S.; Shrivastava, V.; Ameet Deshpande, A. K.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*.

Halberstam, Y.; and Knight, B. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*, 143: 73–88.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

He, J.; Wallis, F.; Gvirtz, A.; and Rathje, S. 2023. Artificial Intelligence Chatbots Mimic Human Collective Behaviour. *Research Square*.

Hodges Jr., J. L. 1958. The Significance Probability of the Smirnov Two-Sample Test. *Arkiv fiur Matematik*, 3(43): 469–486.

Holland, P. W.; Laskey, K. B.; and Leinhardt, S. 1983. Stochastic blockmodels: First steps. *Social Networks*, 5(2): 109–137.

Jin, B.; Liu, G.; Han, C.; Jiang, M.; Ji, H.; and Han, J. 2024. Large Language Models on Graphs: A Comprehensive Survey. *arXiv preprint arXiv:2312.02783*.

Kapferer, B. 1972. *Strategy and transaction in an African factory*. Manchester: Manchester University Press.

Kirk, R.; Mediratta, I.; Nalmpantis, C.; Luketina, J.; Hambro, E.; Grefenstette, E.; and Raileanu, R. 2023. Understanding the Effects of RLHF on LLM Generalisation and Diversity. *arXiv preprint arXiv:2310.06452*.

KONECT. 2024. Networks. http://konect.cc/networks/. Accessed February 27, 2024.

Kossinets, G.; and Watts, D. J. 2009. Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, 115(2).

Krackhardt, D. 1999. The Ties That Torture: Simmelian Tie Analysis in Organizations. *Research in the Sociology of Organizations*, 16: 183–210.

Laniado, D.; Volkovich, Y.; Kappler, K.; and Kaltenbrunner, A. 2016. Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, 5(1): 19.

Li, Y.; Li, Z.; Wang, P.; Li, J.; Sun, X.; Cheng, H.; and Yu, J. X. 2024. A Survey of Graph Meets Large Language Model: Progress and Future Directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI'24)*.

MacRae Jr., D. 1960. Direct Factor Analysis of Sociometric Data. *Sociometry*, 23(4): 360–371.

Marzo, G. D.; Castellano, C.; and Garcia, D. 2024. Large Language Model agents can coordinate beyond human scale. *arXiv preprint arXiv:2409.02822*.

Marzo, G. D.; Pietronero, L.; and Garcia, D. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. *arXiv preprint arXiv:2312.06619*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1): 415–444.

Newman, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69: 066133.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3580–3599.

Papachristou, M.; and Yuan, Y. 2024. Network Formation and Dynamics Among Multi-LLMs. *arXiv preprint arXiv:2402.10659*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST'23)*, 1–22.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Pérez-Rosés, H.; and Sebé, F. 2015. Synthetic generation of social network data with endorsements. *Journal of Simulation*, 9(4): 279–286.

Pew Research Center. 2024. Partisanship by race, ethnicity and education. https://www.pewresearch.org/politics/2024/04/09/partisanship-by-race-ethnicity-and-education/#partisanship-by-race-and-gender.

PRRI Staff. 2021. 2020 PRRI Census of American Religion: County-Level Data on Religious Identity and Diversity. https://www.prri.org/research/2020-census-of-american-religion/.

Sagduyu, Y. E.; Grushin, A.; and Shi, Y. 2018. Synthetic Social Media Data Generation. *IEEE Transactions on Computational Social Systems*, 5(3): 605–620.

Sanchez, G. R.; and Foxworth, R. 2022. Native Americans support Democrats over Republicans across House and Senate races. *Brookings*. https://www.brookings.edu/articles/native-americans-support-democrats-over-republicans-across-house-and-senate-races/.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, 29971–30004.

Simonovsky, M.; and Komodakis, N. 2018. GraphVAE: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, 412–422. Springer.

Smith, J. A.; McPherson, M.; and Smith-Lovin, L. 2014. Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004. *American Sociological Review*, 79.

Statista. 2016. Religious identity of adults in the United States in 2016, by race and ethnicity. https://www.statista.com/statistics/749128/religious-identity-of-adults-in-the-us-by-race-and-ethnicity/.

Thelwall, M. 2009. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2): 219–231.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The Anatomy of the Facebook Social Graph.

US Census Bureau. 2023. National Population by Characteristics: 2020-2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Wang, H.; Feng, S.; He, T.; Tan, Z.; Han, X.; and Tsvetkov, Y. 2023. Can Language Models Solve Graph Problems in Natural Language? In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684): 440–442.

Wong, L. H.; Pattison, P.; and Robins, G. 2006. A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360.

Yao, Y.; Wang, X.; Zhang, Z.; Qin, Y.; Zhang, Z.; Chu, X.; Yang, Y.; Zhu, W.; and Mei, H. 2024. Exploring the Potential of Large Language Models in Graph Generation. *arXiv preprint arXiv:2403.14358*.

Ye, R.; Zhang, C.; Wang, R.; Xu, S.; and Zhang, Y. 2024. Language is All a Graph Needs. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1955–1973.

You, J.; Ying, R.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. GraphRNN: Generating realistic graphs with deep auto-regressive models. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, 5708–5717.

Zachary, W. W. 1977. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4): 452–473.

Zhang, S.; Heck, P. R.; Meyer, M. N.; Chabris, C. F.; Goldstein, D. G.; and Hofman, J. 2023. An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences (PNAS)*, 120(33).

Zhao, J.; Zhuo, L.; Shen, Y.; Qu, M.; Liu, K.; Bronstein, M.; Zhu, Z.; and Tang, J. 2023. GraphText: Graph Reasoning in Text Space. *arXiv preprint arXiv:2310.01089*.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; and Sap, M. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our work generates synthetic networks over hypothetical individuals, so it does not violate privacy norms, and we specifically analyze potential social harms, such as stereotyping or exacerbating segregation.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the claims in the abstract and introduction are supported by our main results in Section 5, which are further tested with robustness checks in our Appendices.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, to analyze LLMs' social network generation, we need to both provide methods for generating social networks with LLMs (Section 3) and an evaluation framework (Section 4).

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we specify how we sample persona demographics following the US distribution (Section 3 and Appendix C.1), and we describe which populations are captured in the real social networks (Section 4 and Appendix C.3). The rest of the data is generated by LLMs, i.e., the generated social networks, which are our object of study, so we study their artifacts closely.

   (e) Did you describe the limitations of your work? Yes, we discuss limitations in the Discussion (Section 6),

with additional results and discussion referred to in the Appendices.

(f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss potential negative impacts of generating social networks with LLMs, particularly related to bias and stereotyping. We discuss related work on bias in social settings (Section 2) and investigate such biases in our work, documenting political biases in the generated networks and generated interests (Section 5).

(g) Did you discuss any potential misuse of your work? Yes, we point out that it would be risky to simulate politically-related social phenomena, such as polarization or opinion dynamics, over networks generated by LLMs, since we find that they overestimate political homophily.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we have carefully documented the risks of generating social networks with LLMs, such as overestimated political homophily, underestimated variance, and LLM costs. Our code, data, and generated networks are available at https://github.com/snap-stanford/llm-social-network. This includes all code to run our experiments and reproduce our results.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA, we do not have theoretical results.

(b) Have you provided justifications for all theoretical results? Yes, we have provided evidence to support all claims, see Section 5 and Appendices.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes. For example, we test whether the fact that political homophily is the highest can be explained by correlations between demographics, so we try shuffling demographics, and we find that political homophily is still the highest. We also test six different LLMs, to test how broadly our results generalize. Finally, we test minor changes to the experimental set-up (e.g., temperature and prompt), and we show that our main results are not sensitive to these changes.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see above.

(e) Did you address potential biases or limitations in your theoretical framework? NA, we do not have a theoretical framework.

(f) Have you related your theoretical results to the existing literature in social science? Yes, we discuss related work from sociology (Watts and Strogatz 1998; Krackhardt 1999; McPherson, Smith-Lovin, and Cook 2001; Kossinets and Watts 2009; Smith, McPherson, and Smith-Lovin 2014), sociometry (Coleman, Katz, and Menzel 1957; MacRae Jr. 1960), economics (Gentzkow and Shapiro 2011; Halberstam and Knight 2016), and political science (Argyle et al. 2023).

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discuss the implications in Section 6 and the Appendices.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA, we do not have theoretical proofs.

(b) Did you include complete proofs of all theoretical results? NA, we do not have theoretical proofs.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, all code, data, and instructions are available at https://github.com/snap-stanford/llm-social-network, which is included on the first page.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? We do not train models, but we do describe experimental parameters, such as LLM temperature, the number of personas sampled, the number of networks sampled, randomization in the prompts, etc.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we generate 30 networks per prompting method and report mean, standard error, and individual data points.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we describe the APIs we used—OpenAI API and Llama API—that we used to call the models. In our GitHub repository, we also describe our version of Python (3.10) and list the required Python packages and versions.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in Section 4, we describe our evaluation framework, including the real networks that we compare against and the network characteristics that we measure.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA, we do not have a classification problem.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets,

(a) If your work uses existing assets, did you cite the creators? Yes, we cited our data sources, which included statistics to sample demographics following the US population and the real social networks.

(b) Did you mention the license of the assets? NA, these data sources are all publicly available.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we provide our code (with methods to generate social networks) and generated networks in our GitHub repository.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA, we are generating synthetic networks over hypothetical individuals.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we describe that the real networks do not contain node features, so they are not personally identifiable, and the population statistics are aggregated over the entire US population.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA, we are not releasing new datasets.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA, we are not releasing new datasets.

6. Additionally, if you used crowdsourcing or conducted research with human subjects,

(a) Did you include the full text of instructions given to participants and screenshots? NA, our work does not involve crowdsourcing or research with human subjects.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA, our work does not involve crowdsourcing or research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA, our work does not involve crowdsourcing or research with human subjects.

(d) Did you discuss how data is stored, shared, and deidentified? NA, our work does not involve crowdsourcing or research with human subjects.

Appendix

In Appendix A, we compare results across GPT, Llama, and Gemma models. In Appendix B, we describe additional experiments and findings. In Appendix C, we provide details on methods and experimental set-up.

A Comparison Between LLMs

We experiment with the following LLMs: OpenAI's GPT-3.5 Turbo and GPT-4o (Brown et al. 2020; OpenAI et al. 2023), Meta's Llama 3.1 (8B and 70B) (Touvron et al. 2023), and Google's Gemma 2 (9B and 27B) (Gemma Team et al. 2024). These six models represent a range across companies, different model sizes, and proprietary (the GPT models) versus open-source (the Llama and Gemma models). Llama 3.1 in particular includes some of the best-performing open-source models on holistic benchmarks.[2]. For the OpenAI models, we use the OpenAI API,[3] and for the other models, we use the Llama API,[4] which also includes other open-source models. We report our main results on GPT-3.5 Turbo since we find that it best matches the structure of the real-world social networks (discussed below), but here, we discuss results from all other models. For these experiments, we generate networks with the Sequential method, using the same experimental settings as before (same set of 50 personas, prompt, temperature of 0.8, etc.). For these experiments, we generate 10 instead of 30 networks per model, but we find that standard errors are small. We visualize results for GPT-4o in Figure A1, Llama 3.1 8B and 70B in Figure A2, and Gemma 2 9B and 27B in Figure A3.

Structural characteristics. We find that GPT-3.5 Turbo best matches the structure of the real networks, most notably matching the real-world density. All of the other models have much higher densities, which also contributes to unrealistically high clustering and low shortest paths. Due to density's outsized effect, we wanted to see if providing the model a bit of help on density might be all that it needs to match the other characteristics as well. Thus, we try a variant of the Sequential method where we specify n, i.e., how many friends should be chosen (instead of only asking "Which of these people will you become friends with?", see full prompt in Figure C5). We sample n from Exponential($\lambda = 0.2$), with mean $1/\lambda = 5$, independently for each persona. Note that specifying each persona's number of choices does not predetermine the exact density or degree distribution of the network, since a persona's total set of friends at the end of the network generating process is the union of its chosen friends along with anyone who chose it. However, specifying these numbers can help to guide the models to lower densities.

With this variant, which we call "$+\lambda$", all models' densities are brought down to a reasonable range. Llama 3.1 8B $+\lambda$ is also able to approximately match the real networks and GPT-3.5 Turbo on all other characteristics now (Figure A2). However, even with $+\lambda$, GPT-4o and Llama 3.1 70B still generate networks that are often disconnected into two components; this is due to extreme political segregation, which we discuss below. With Gemma 2 27B, the networks are almost always fully connected, but even with $+\lambda$, clustering and modularity remain slightly too high.[5]

Homophily. We find that all six models exhibit clear homophily, with ratios above 1 for gender, age, race/ethnicity, religion, and political affiliation. We also find, consistently across the models, that political homophily remains by far the strongest form of homophily and it is always overestimated compared to real-world measures. In fact, the models we test in this section all produce levels of political homophily even higher than GPT-3.5 Turbo, which we discussed were overestimates in the main text (Table 2).

The emphasis on political homophily is particularly ex-

[2]https://crfm.stanford.edu/helm/lite/latest/#/leaderboard

[3]https://platform.openai.com/docs/api-reference

[4]https://www.llama-api.com/

[5]We were not able to add $+\lambda$ to Gemma 2 9B, since it could not consistently follow instructions to choose exactly n friends.

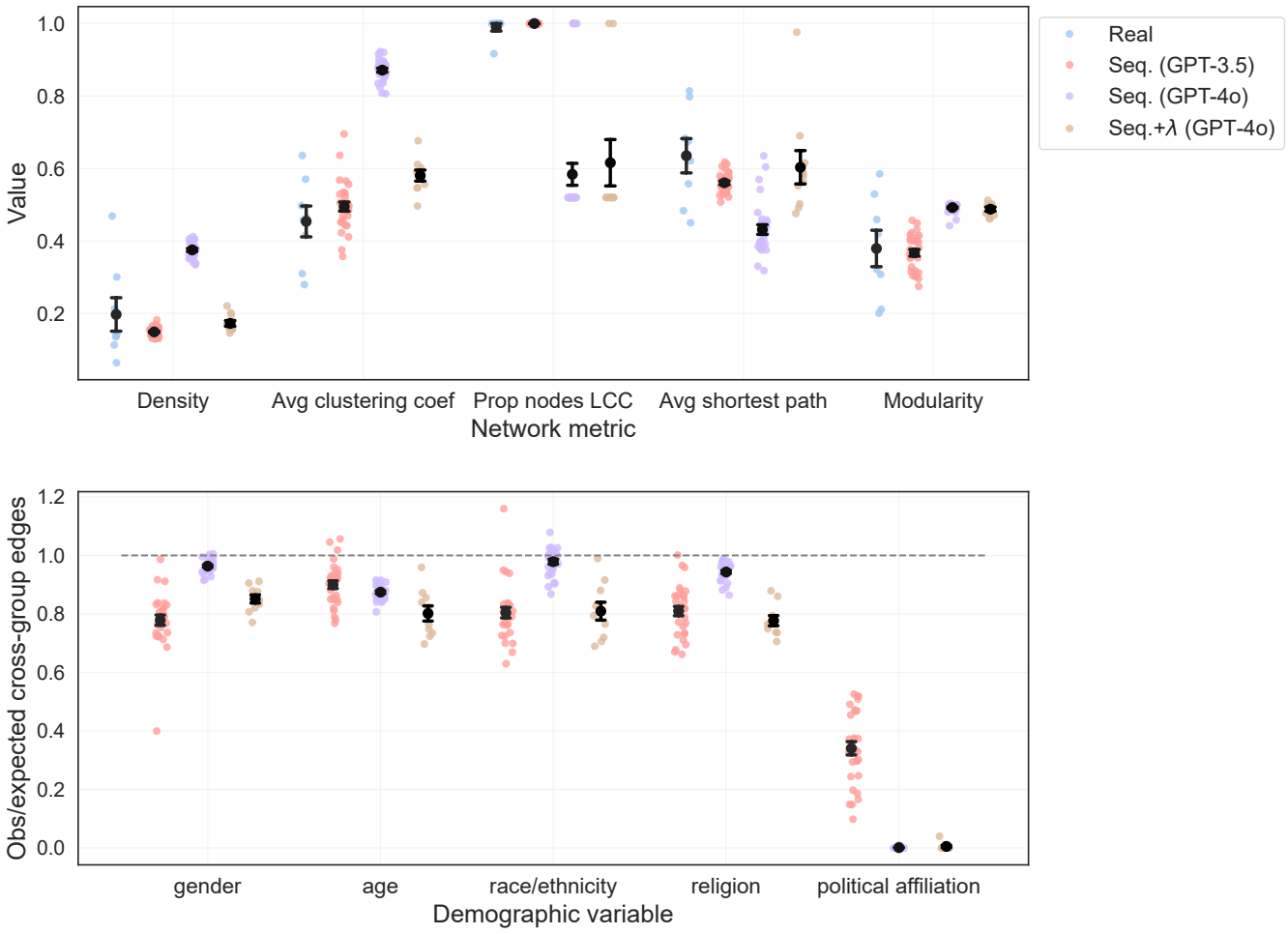

Figure A1: Evaluating results from GPT-4o, compared to GPT-3.5 Turbo. **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network.

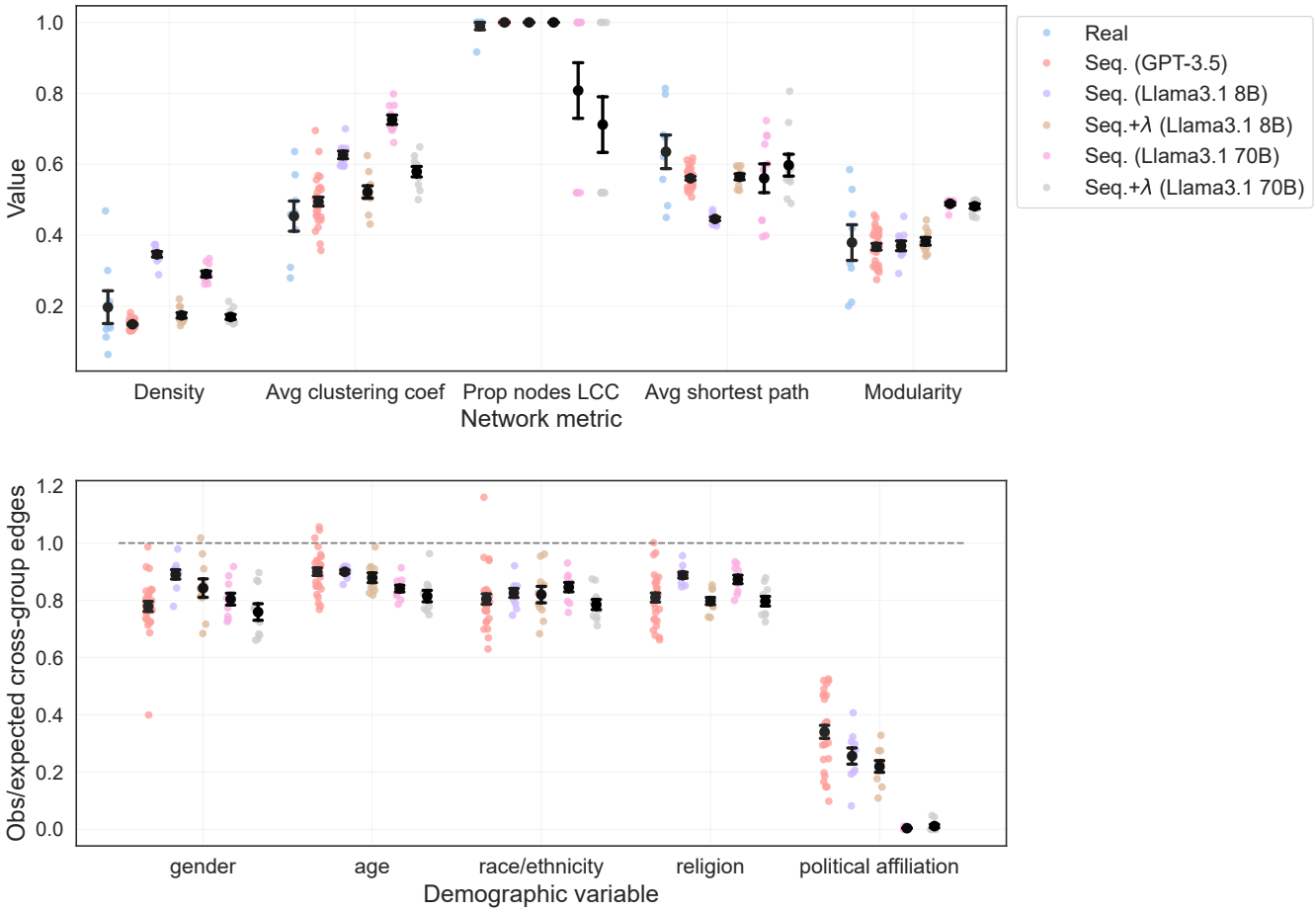

Figure A2: Evaluating results from Llama 3.1 8B and 70B, compared to GPT-3.5 Turbo. **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network.

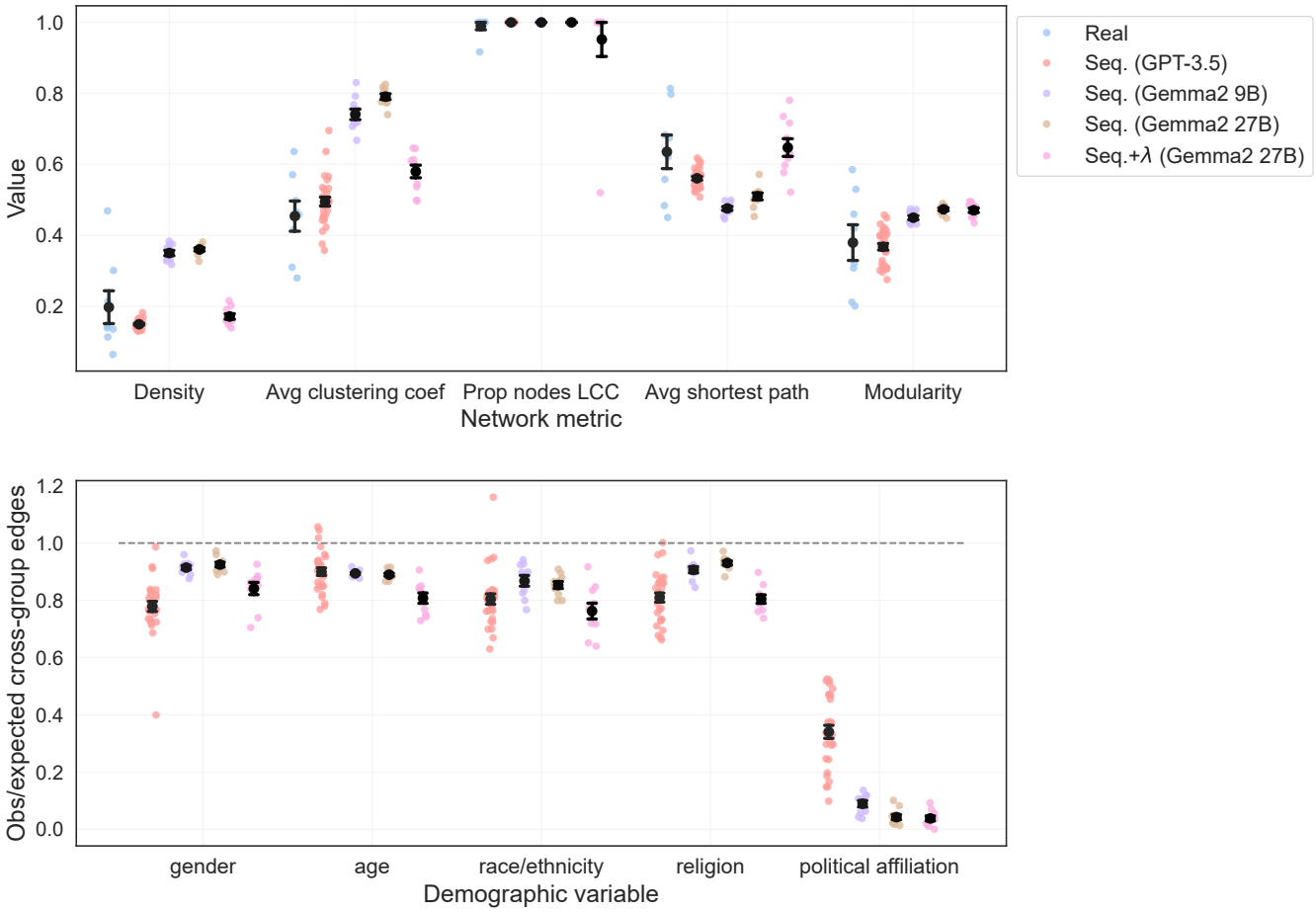

Figure A3: Evaluating results from Gemma 2 9B and 27B, compared to GPT-3.5 Turbo. **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network.

treme for GPT-4o and Llama 3.1 70B: they show complete segregation between Republicans and Democrats, such that there are *no* cross-group edges and the cross-group ratio is around 0 (Figures A1 and A2). Furthermore, as a result of the complete segregation, their networks fracture into two disconnected components, one for Republicans and one for Democrats, which is why we see that the proportion of nodes in the largest connected component is often around 50% for these two models. This is highly unrealistic, as it is well-known that social networks, despite having strong community structure and homophily, are also characterized by having a giant connected component that contains the vast majority of the nodes (Ugander et al. 2011). Splitting a synthetic network into two disconnected components has important implications for downstream modeling and use cases of these generated networks: for example, an epidemic outbreak started in one component would never reach the other component, or opinions could not spread, greatly altering the trajectory of dynamic network processes.

B Additional Results

In this section, all experiments are run with GPT-3.5 Turbo, unless otherwise specified.

Comparing to classical network models. In the main text, we showed that the LLM can match many structural characteristics of real social networks, including density, clustering, connectivity, and degree distribution (Figures 3 and 4). However, how does this compare to existing models for network generation? Here, we consider three classical network models: (1) Erdős–Rényi random graph models (Erdős and Rényi 1959), (2) Barabási-Albert preferential attachment models (Barabási and Albert 1999), and (3) Watts-Strogatz small-world models (Watts and Strogatz 1998). We choose these models to compare to since they only have 1-2 parameters, and we are interested in LLM's capabilities to generate networks without additional training and with minimal need to defined parameters. In contrast, we do not compare to stochastic block models (Holland, Laskey, and Leinhardt 1983), which require edge probabilities between all pairs of blocks and block assignments, or machine learning models for graph generation, since they have many parameters and require a substantial set of observed graphs for training to fit those parameters (You et al. 2018; Simonovsky and Komodakis 2018; Guo and Zhao 2023).

To quantify how well a network metric is matched, we extract the metric from each real social network, using the same eight social networks as in our main experiments (Sections 5.1 and C.3), and from each generated network, with 30 generated networks per model. To compare the two distributions of the metric, we report both the difference in the means, normalized by the real networks' standard deviation (Eq. 6), and the two-sample Kolmogorov–Smirnov statistic (Eq. 7), which measures the distance between two empirical distributions. As we show in Table B1, even in the best case—when we allow the models to choose parameters based on the real social networks—these models cannot match all of the real network metrics as well as our Local or Sequential methods can. Thus, being able to match the structural characteristics of real social networks is nontrivial, adding significance to our finding that LLMs can match many structural characteristics at once.

Real-world homophily for other demographics. In the main text, we showed that levels of political homophily predicted by the LLM are unrealistically high, compared to reported levels of homophily in real-world social networks (Table 2). What about real-world levels of homophily for the other demographic variables? Data from Thelwall (2009), who studies connections on MySpace, suggests cross-group ratios of 0.44 for race/ethnicity, 0.69 for religion, and 1.04 for gender. Data from Laniado et al. (2016), who study connections on Tuenti (a Spanish social network platform), suggests cross-gender ratios of 0.96 when measuring friendship networks and 0.88 when measuring interaction networks. Smith, McPherson, and Smith-Lovin (2014), using data from the US General Social Survey (GSS), finds cross-group ratios of 0.17 for race/ethnicity, 0.45 for religion, 0.59 for age, and 0.81 for gender in 1985 and ratios of 0.25 for race/ethnicity, 0.44 for religion, 0.60 for age, and 0.88 for gender in 2004.

These numbers, which we summarize in Table B2, reveal how much variation there is in levels of homophily across studies and over time, making it difficult to evaluate whether the LLM's levels of homophily are realistic or not. However, some trends emerge: gender homophily tends to be weaker, race/ethnicity homophily tends to be stronger, and religion and age are somewhere in between. Quotes from summary papers on homophily support this ordering: McPherson, Smith-Lovin, and Cook (2001) say, "By the time that they are adults, people have friendship and confidant networks that are relatively sex-integrated (at least when compared to other dimensions like race, age, and education)," and Thelwall (2009) says, "For U.S. friendship in the last century, the key factors, in decreasing order, seem to be race and ethnicity, age, religion, educational level, occupation, and gender."

In contrast, in our LLM experiments, the Sequential method predicts similar levels of homophily for gender, race/ethnicity, and religion—all around 0.8—and less homophily for age (0.9) (Table B2). The Local method predicts slightly more homophily for all variables, but a similar ranking: gender, race/ethnicity, and religion have cross-group ratios around 0.75 and age still has less homophily (0.89). While these levels of homophily seem realistic for gender, it seems that the LLM seriously underestimates homophily for race/ethnicity and religion (and likely for age as well, although this is inconclusive since we only have age homophily from one study and age homophily is less comparable to the others since it is not a categorical variable). It is possible that LLMs, due to their guardrails and instruction fine-tuning, have been guided away from using sensitive attributes like race or religion to influence their generation; however, future work is needed to thoroughly investigate evidence of underestimation and probe for possible reasons.

Age homophily. Age is our only non-categorical demographic variable, which slightly complicates our definition of homophily, which assumes well-defined groups. We are able to measure an approximately analogous cross-group ra-

Model	Density	Avg CC	% LCC	Avg SP	Mod.	Degree	Avg
Random graph	0.013†	2.286	**0.378**	1.156	1.149	0.154	1.025
Preferential attachment	0.113†	1.363	**0.378**	1.081	1.160	0.053	0.807
Small world	0.054†	0.041†	**0.378**	0.808	0.404	0.188	0.444
GPT-3.5 Turbo, Global	1.311	3.399	17.740	6.547	2.959	0.914	5.478
GPT-3.5 Turbo, Local	**0.207**	**0.012**	**0.378**	**0.383**	0.316	**0.023**	**0.220**
GPT-3.5 Turbo, Sequential	0.394	0.363	**0.378**	0.596	**0.088**	0.174	0.332
Random graph	0.625	1.000	**0.125**	0.750	0.750	0.465	0.619
Preferential attachment	0.625	0.750	**0.125**	0.750	0.750	0.416	0.569
Small world	0.625	0.375	**0.125**	0.750	0.500	0.617	0.499
GPT-3.5 Turbo, Global	0.967	0.933	0.833	0.900	0.967	0.740	0.890
GPT-3.5 Turbo, Local	0.525	**0.250**	**0.125**	**0.500**	0.500	0.265	0.361
GPT-3.5 Turbo, Sequential	**0.375**	0.325	**0.125**	0.625	**0.342**	**0.190**	**0.330**

Table B1: Quantitative results on structural characteristics. The top six rows indicate difference in means, normalized by the real networks' standard deviation (Eq. 6). † indicates that the model parameters were fitted on the real-world mean value for this characteristic, so the mean difference should be ignored, and it is left out of the average. The bottom six rows represent the two-sample Kolmogorov–Smirnov statistic, which measures the distance between two empirical distributions (Eq. 7). For both measures, lower is better.

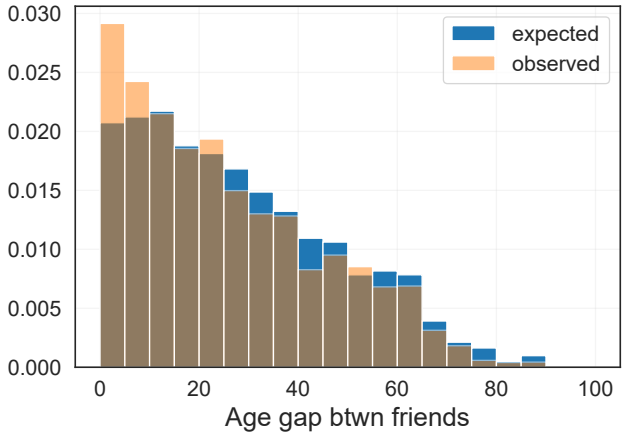

Figure B1: Observed (left) vs. expected (right) distribution of age gaps in the generated social networks, under the Sequential method.

tio (Eq. 1) for age by comparing the observed-to-expected average age gaps between connected nodes, as described in the main text. As a more detailed measure, we can also plot the distribution of observed age gaps vs. expected age gaps, which we show in Figure B1. The observed distribution consists of the age gaps for all observed edges pooled over the 30 networks generated by the Sequential method. The expected distribution consists of the age gaps between all possible pairs of nodes among the 50 personas. Compared to the expected distribution, the observed distribution is clearly shifted to the left (i.e., smaller gaps), with more edges than expected with age gaps of 0-10.

Ablations of demographic variables. Here, we explore how the LLM's generated networks change when we provide subsets of the original set of five demographic variables (gender, age, race/ethnicity, religion, and political affiliation). First, we try providing one variable at a time. When only one variable is provided, we find that homophily for that variable gets stronger, compared to when all five variables are provided (Figure B2, top). This is not entirely surprising: when the LLM is only provided this variable, then it forms the social network based entirely on that variable. For example, if the LLM is only given gender, the LLM is unlikely to connect a woman and a man since they do not have any listed demographics in common, but if more demographics were provided, they might now have a demographic in common. However, this result is not guaranteed: the LLM could have learned the joint distribution of these variables during pretraining, so theoretically it could have added the missing demographics to each persona by sampling from that joint distribution. Given that the LLM does not seem to do this on its own, it would be interesting in future work to explore this capability: adding greater variance in the generated networks by augmenting each persona with additional traits, generated by the LLM. However, one would have to be careful about exacerbating biases with this technique, as we saw with LLM-generated interests.

Second, we try providing two variables at a time: political affiliation and one of the four other variables. For each variable, its level of homophily when in a pair tends to be between its level of homophily when alone versus with all five variables (Figure B2, bottom). We also find in both of these experiments that political affiliation continues to be the dominant factor. In the single-variable experiments, political homophily is the strongest when political affiliation is provided, compared to homophily for any other variable when it is the only variable provided. In the two-variable experiments where political affiliation and one other variable is provided, political homophily is always stronger than the other variable's homophily.

Degree per demographic group. In the main text, we discussed and visualized degree distributions over all nodes in the network (Figure 4), but degree could also differ

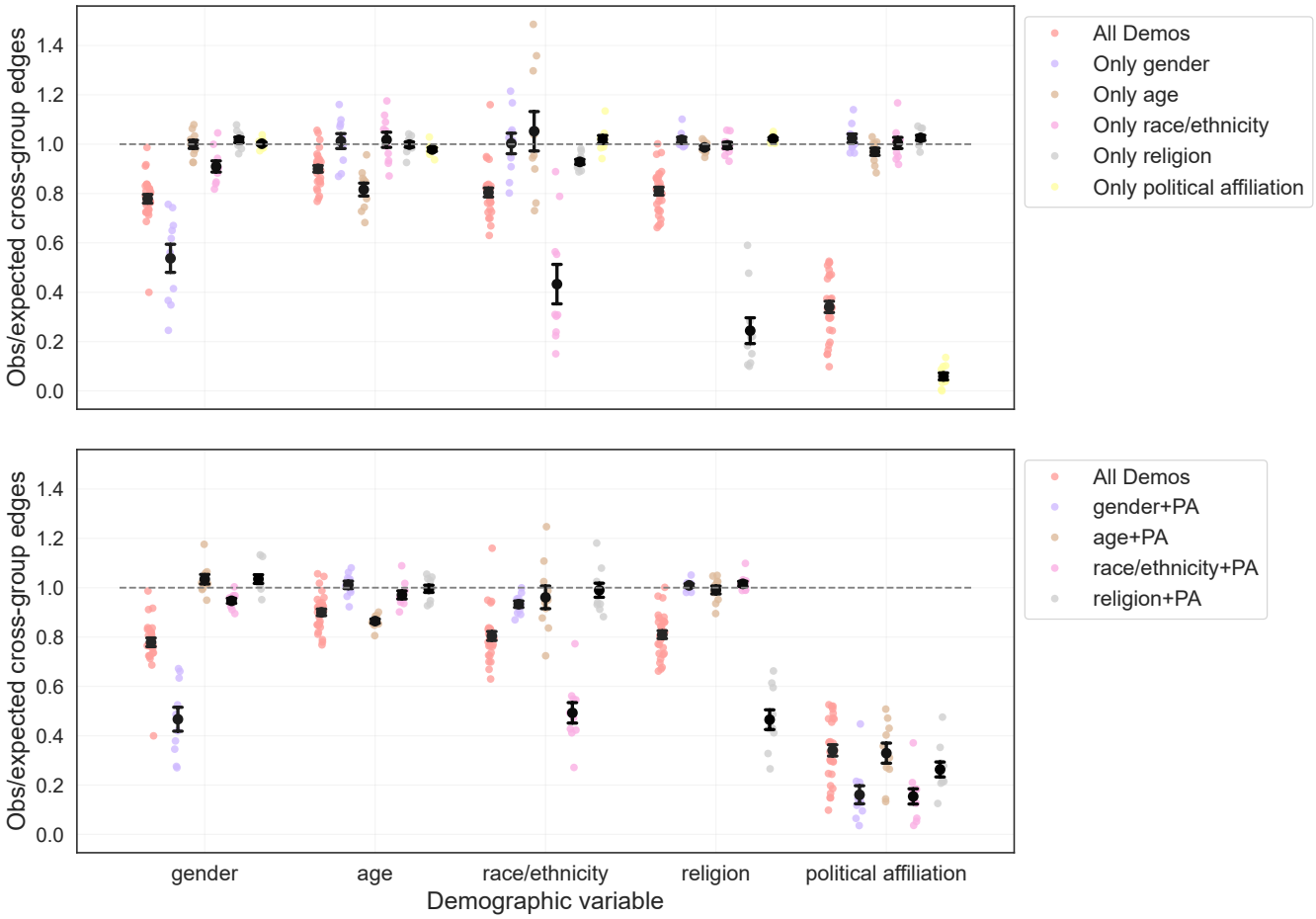

Figure B2: Evaluating homophily per demographic variable, when providing subsets of the demographics. **Top:** providing one variable at a time. **Bottom:** providing two variables at a time, political affiliation (PA) and one of the other four. Ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network. All model results shown here use the Sequential method and GPT-3.5-Turbo.

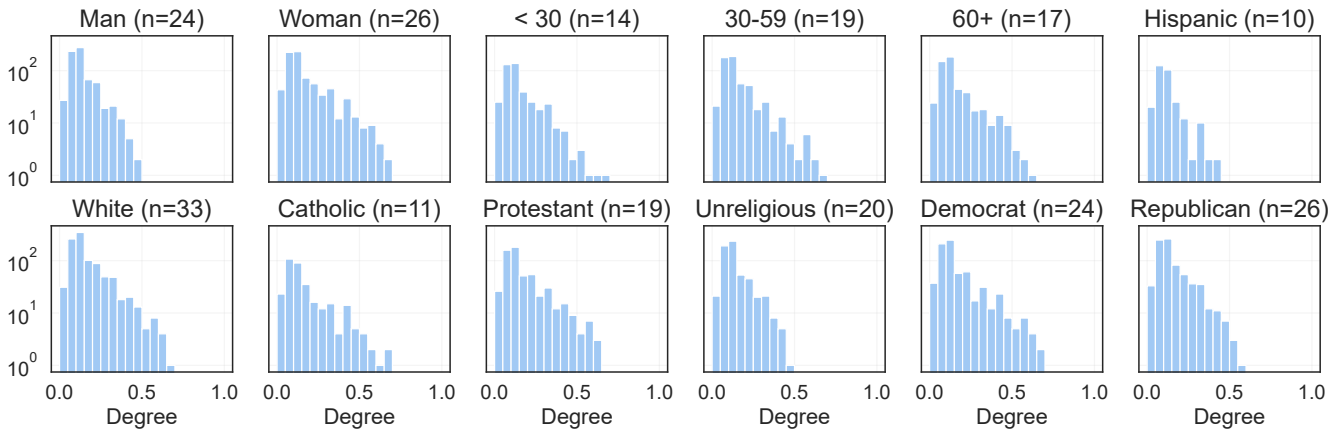

Figure B3: Degree distribution per demographic group, pooled over the 30 networks generated by the Sequential method.

Study / LLM Method	Demographic	Cross-group ratio (Eq. 1) ↓
Thelwall (2009), MySpace	Gender	1.04
	Race/ethnicity	0.44
	Religion	0.69
Laniado et al. (2016), Tuenti, friendship	Gender	0.96
Laniado et al. (2016), Tuenti, interaction	Gender	0.88
Smith, McPherson, and Smith-Lovin (2014), GSS in 1985	Gender	0.81
	Age	0.59
	Race/ethnicity	0.17
	Religion	0.45
Smith, McPherson, and Smith-Lovin (2014), GSS in 2004	Gender	0.88
	Age	0.60
	Race/ethnicity	0.25
	Religion	0.44
Local	Gender	0.774 (0.015)
	Age	0.894 (0.010)
	Race/ethnicity	0.796 (0.013)
	Religion	0.755 (0.013)
Sequential	Gender	0.779 (0.017)
	Age	0.900 (0.013)
	Race/ethnicity	0.804 (0.018)
	Religion	0.810 (0.016)

Table B2: Comparing levels of homophily in real-world vs. LLM-generated networks, for the demographic variables besides political affiliation. For Local and Sequential, we report the mean and standard error (in parentheses) over each method's 30 generated networks.

	n	Mean degree	Gini coef.	GS Index
Overall	50	0.149 (0.002)	0.347 (0.005)	–
Man	24	0.131 (0.003)	0.281 (0.009)	0.368 (0.011)
Woman	26	0.166 (0.004)	0.370 (0.008)	0.299 (0.009)
Under 30	14	0.147 (0.005)	0.321 (0.010)	0.524 (0.009)
30-59	19	0.150 (0.004)	0.336 (0.009)	0.543 (0.007)
60+	17	0.150 (0.005)	0.328 (0.009)	0.541 (0.006)
Hispanic	10	0.114 (0.004)	0.259 (0.013)	0.426 (0.017)
White	33	0.160 (0.003)	0.335 (0.008)	0.330 (0.008)
Catholic	11	0.157 (0.006)	0.354 (0.013)	0.511 (0.009)
Protestant	19	0.163 (0.004)	0.352 (0.010)	0.475 (0.011)
Unreligious	20	0.132 (0.004)	0.276 (0.008)	0.487 (0.009)
Democrat	24	0.157 (0.004)	0.355 (0.009)	0.189 (0.011)
Republican	26	0.143 (0.005)	0.313 (0.007)	0.210 (0.014)

Table B3: Degree distribution and Gini-Simpson (GS) diversity index per demographic group, for all groups with at least 10 personas in our set of 50 personas. For each statistic, we compute its value per network, and report the mean value and standard error (in parentheses) over the 30 networks generated by the Sequential method.

across demographic groups. In Figure B3, we visualize each group's degree distribution, for all groups with at least 10 personas in our set of 50 personas. As in Figure 4, the distribution consists of degrees for all nodes in the group, pooled over the 30 networks generated by the Sequential method. We see minor differences between groups: for example, in the generated networks, women have slightly higher degree on average than men (0.166 vs. 0.131, Table B3). This could be explained by two factors: first, there are slightly more women (n=26) than men (n=24) in our set of 50 personas, and second, women show higher levels of same-group preference, as discussed in the main text (same-group ratios of 1.43 vs 0.99, Figure 6). We also find that White personas have slightly higher average degree than Hispanic personas (0.160 vs 0.114), which could also be explained by there being more White personas (n=33) vs. Hispanic personas (n=10), and higher same-group ratios among White personas compared to Hispanic personas (1.20 vs. 1.02).

Beyond using the mean to describe the degree distribution, we can also use the Gini coefficient as a measure of the inequality in degree. For a set of values $x_1, x_2, \cdots, x_n$, the Gini coefficient is defined as

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}, \tag{3}$$

where $\bar{x}$ is the mean of the values. First, when we compute the Gini coeffient over all nodes, we find a higher coefficient (more inequality) when using the Sequential method ($G = 0.347$), compared to the Local method ($G = 0.261$) or Global method ($G = 0.255$). This aligns with our earlier results showing that Sequential can better capture the

long-tailed degree distribution of real social networks. Per group, we find similar results to what we found for degree distribution: higher average degree tends to correspond to higher Gini coefficient, since the higher average tends to be driven by longer tails, which results in greater inequality. As described in Table B3, women compared to men, White compared to Hispanic, Protestant and Catholic compared to Unreligious, and Democrat compared to Republican all have slightly higher average degrees and Gini coefficients.

Diversity per demographic group. We can also measure rates of homophily and diversity per group. In Figure 6, we visualized observed-to-expected ratios to capture same-group preferences (the diagonal) and cross-group preferences (the off-diagonal). This analysis revealed that not all cross-group relations are equally unlikely; for example, adjacent age groups were likelier to have cross-group relations. It also revealed that not all same-group preferences were equally strong; for example, we saw stronger preferences within women compared to men or within Catholic compared to Unreligious.

Related to homophily, we can also measure the *diversity* in a persona's 1-hop neighbors (i.e., their friends). We use the Gini-Simpson (GS) index to measure diversity, which is defined in terms of a set of groups, $\mathcal{G} = \{g_1, g_2, \cdots, g_k\}$, and the proportion p_g of values that belong to each group:

$$GS = 1 - \sum_{g \in \mathcal{G}} p_g. \quad (4)$$

For a given node i, we define its 1-hop diversity with respect to a demographic variable by computing the GS index over their friends' group identities for that variable. In Table B3, we report the average 1-hop diversity for members in a group, with respect to that group's demographic variable (e.g., the average 1-hop gender diversity for men). As expected, groups with higher diversity correspond to those with lower same-group ratios in Figure 6: men compared to women, Hispanic compared to White, Catholic compared to Protestant and Unreligious, and Republican compared to Democrat. The three age groups are similar to each other for both measures.

Individuals and intersections of group identities. We found that LLMs overemphasize political homophily overall and that each political group, Democrats and Republicans, shows strong same-group preferences (Figure 6) and low diversity in 1-hop neighbors (Table B3). However, how does political preference and diversity vary over individuals, as their other group identities intersect with their political affiliation? To analyze this, we use individual-level metrics: the GS index to measure individual-level diversity in 1-hop neighbors and an individual's"Democrat lean", i.e., the proportion of the individual's 1-hop neighbors who are Democrats, as an easily interpretable measure.

We find large within-group variability over individuals: for example, the persona with the lowest 1-hop political diversity (measured by GS index) is a woman, age 75, White, Unreligious, and Democrat (mean=0.111, SE=0.032), while the persona with the highest 1-hop political diversity is a woman, age 38, White, Protestant, and Democrat (mean=0.353, SE=0.028). Relatedly, we find that the lowest-diversity persona has a very high Democrat lean (mean=0.908, SE=0.030), while the highest-diversity persona's lean is less extreme (mean=0.722, SE=0.028). Since both individuals are women, White, and Democrat, we can attribute their differences to their age and religion. From pure numbers, we would not expect religion to play a major role in their differences, since the share of Democrats vs. Republicans is about the same for personas that identify as Unreligious (10 vs. 10) and as Protestant (10 vs. 9). However, it is possible that the LLM expects that an Unreligious Democrat's preference for befriending Democrats is stronger than a Protestant Democrat's preference. Age could also partially explain their differences: among personas within 10 years of 75 (65-85), there are 6 Democrats and 3 Republicans, but among personas within 10 years of 38 (28-48), there are 8 Republicans and 5 Democrats, which could help to explain why the former persona has a stronger Democrat lean and less political diversity in friends.

These results reveal that, even though we see strong political homophily and low political diversity overall, the LLM generates some variability across individuals, due to the intersection of their political affiliation with other demographic identities, such as religion or age.

Incorporating interests. In Figure B4, we visualize demographic homophily under three versions of the Sequential method: when only demographic information is provided, when demographic information and interests are provided, and when only interests are provided. In all cases, the LLM places the largest emphasis on political homophily.

In Table B4, we report the top interests per demographic group. The LLM generates interests as a comma-separated list (see examples in Table 3), which makes it straightforward to separate the list into individual interests and compute, for each group and interest, what percentage of personas in that group have that interest. For this experiment, we use 1,000 personas, instead to our usual set of 50 personas, so that we can estimate percentages over larger populations. We report percentages for all groups with at least 30 personas (see the counts per group in Table C1). We find that "social justice" is a common interest across groups since it dominates Democrats' interests (62.5%) and Democrats account for a large portion of every other group besides Republicans. We can also see the result of correlations between demographic identities: for example, men, White, and religious populations are likelier to be Republicans, resulting in higher percentages for "conservative politics".

Sensitivity analyses. In our experiments, we use a default temperature of 0.8. In Figure B5, we show that our main results do not significantly change if we use a temperature of 0.6 or 1.0 instead. Since the LLM overemphasizes political homophily, we also try adding "Pay attention to all demographics" to the system prompt. We include GPT-4o in this study as well due to its extreme levels of political homophily (Figure A1). In Figure B6, we show that adding this prompt does not significantly change results for GPT-3.5 Turbo or GPT-4o. Finally, as mentioned before, we also find that results do not significantly change if we prompt the LLM to

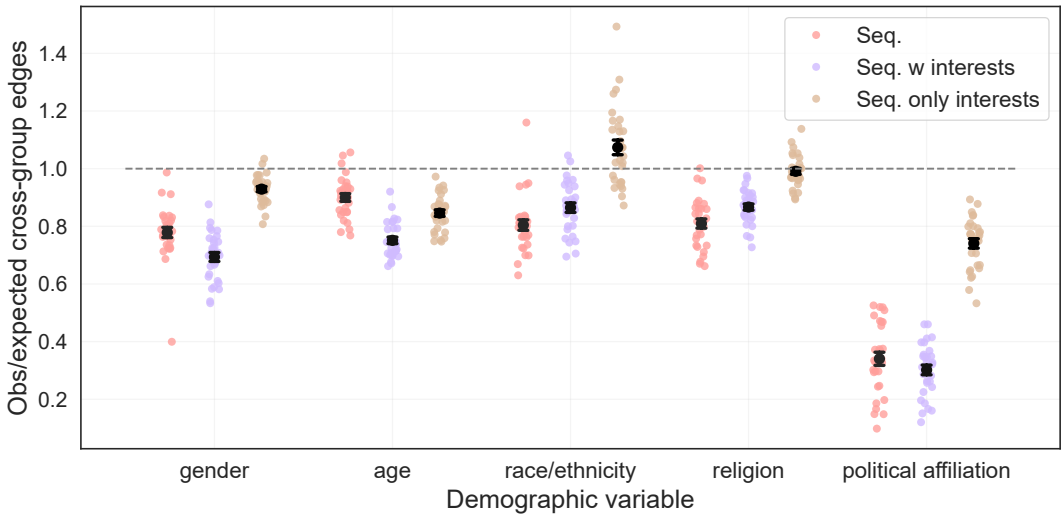

Figure B4: Visualizing demographic homophily under three methods: "Seq." (only demographic information), "Seq. w interests" (demographic information and interests), and "Seq. only interests" (only interests). The LLM places the largest emphasis on political homophily in all cases. Ratios below 1 (marked by the grey line) indicate homophily, with lower ratios indicating more homophily. We visualize mean and standard error (in black) and individual data points corresponding to each network. All model results shown here use the Sequential method and GPT-3.5-Turbo.

generate a short reason for each friend that it selects. Thus, our results are robust to these perturbations in temperature and prompt.

C Methodological Details

C.1 Persona Construction

As described in the main text, we include gender, age, race/ethnicity, religion, and political affiliation. In Table C1, we list the number of personas in each demographic group for the sample of 50 personas we used in most of our experiments, as well as the sample of 1,000 personas we used for evaluating top interests per demographic (Table B4). In Figure C1, we visualize the distribution of ages in the sample of 1,000 personas. Below, we explain how we sampled demographic variables per persona.

Gender, race/ethnicity, and age. First, we use data from the US Census (US Census Bureau 2023), who provide monthly population estimates for sex, race/ethnicity, and age (individual years, from 0 to 100 years old). Specifically, we downloaded `nc-est2023-alldata-r-file07.csv` from US Census datasets archived online[6] and used the data for June 2023. We use these estimates to calculate joint distributions of gender, race/ethnicity, and age. Additionally, using data from Pew Research Center (Brown 2022), we sample from the age-dependent distribution of those who identify as non-binary.

Religion. We sample religion conditioned on the persona's race/ethnicity. Statista (Statista 2016) provides distributions of religious identity for adults in the US in 2016, for most race/ethnicities. Additionally, using data from 2020 PRRI Census of American Religion (PRRI Staff 2021), we acquire the distribution for Native Americans.

Political affiliation. Finally, we sample political affiliation conditioned on the persona's race/ethnicity and gender. We primarily use data from Pew Research Center (Pew Research Center 2024), using the 2023 numbers from their figure, "Partisan identification by gender among racial and ethnic groups," which cover most race/ethnicities. Additionally, we use data from Brookings (Sanchez and Foxworth 2022), who report Native Americans' distribution of political support in 2022.

Interests. In Figure C2, we provide the prompt that we use to generate interests. We randomize the order of demographics provided, since we find that the LLM seems to pay special attention to the first listed demographic when generating interests. We use GPT-4o to generate interests, since we find that it follows the required format a little better.

C.2 Network Generation

In Figures C3-C5, we provide the full basic prompts for each of our network generation methods: Global, Local, and Sequential. These are the prompts that are used when only demographic variables are provided per persona. When interests are provided, we add "interests include: ..." per persona. In variants of the Sequential prompt, discussed in Appendices A and B, we experiment with specifying the number of friends that should be chosen; prompting the LLM to generate a short reason for each selected friend; and adding "Pay attention to all demographics" to the prompt.

For the Global method, the entire list of personas is given in one prompt, while in the Local and Sequential methods,

[6]https://www2.census.gov/programs-surveys/popest/datasets/2020-2023/national/asrh/

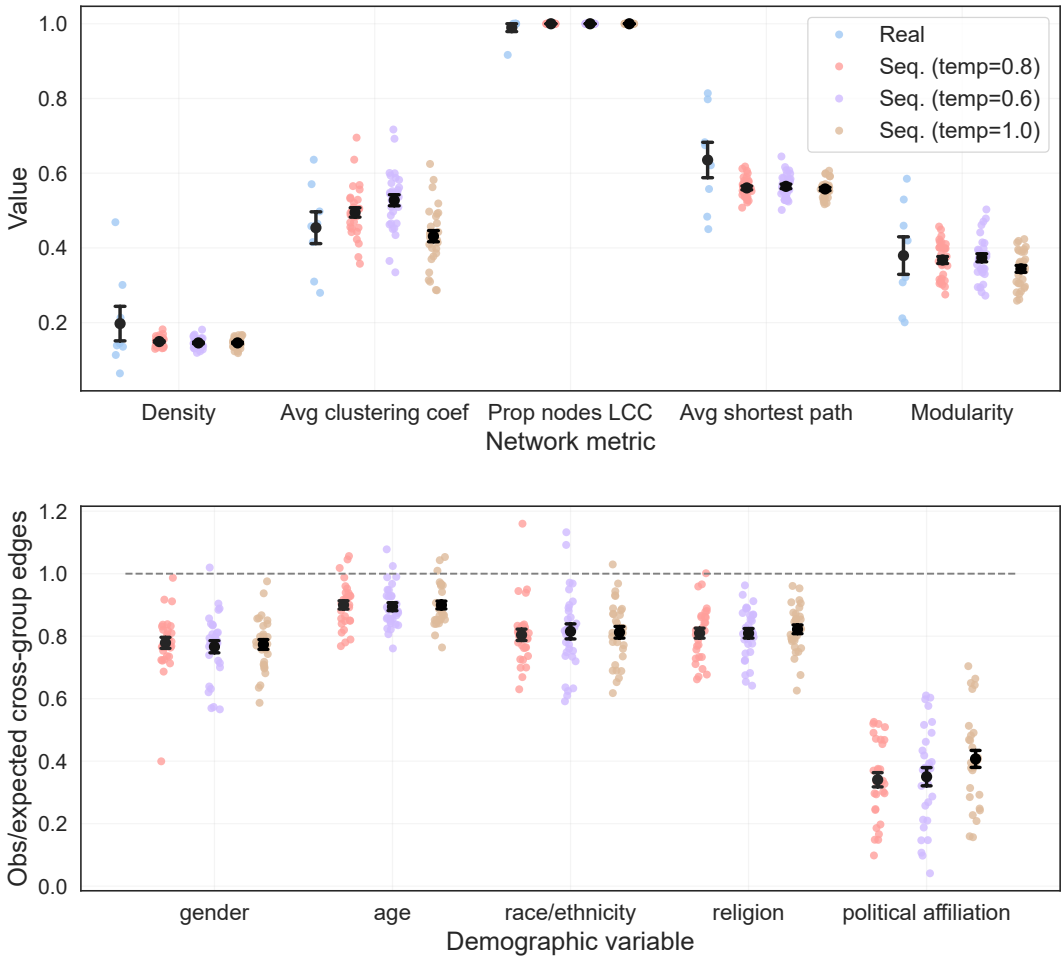

Figure B5: In our experiments, we use a default temperature of 0.8. Our main results do not change significantly if we use a temperature of 0.6 or 1.0 instead. **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network. All model results shown here use the Sequential method and GPT-3.5-Turbo.

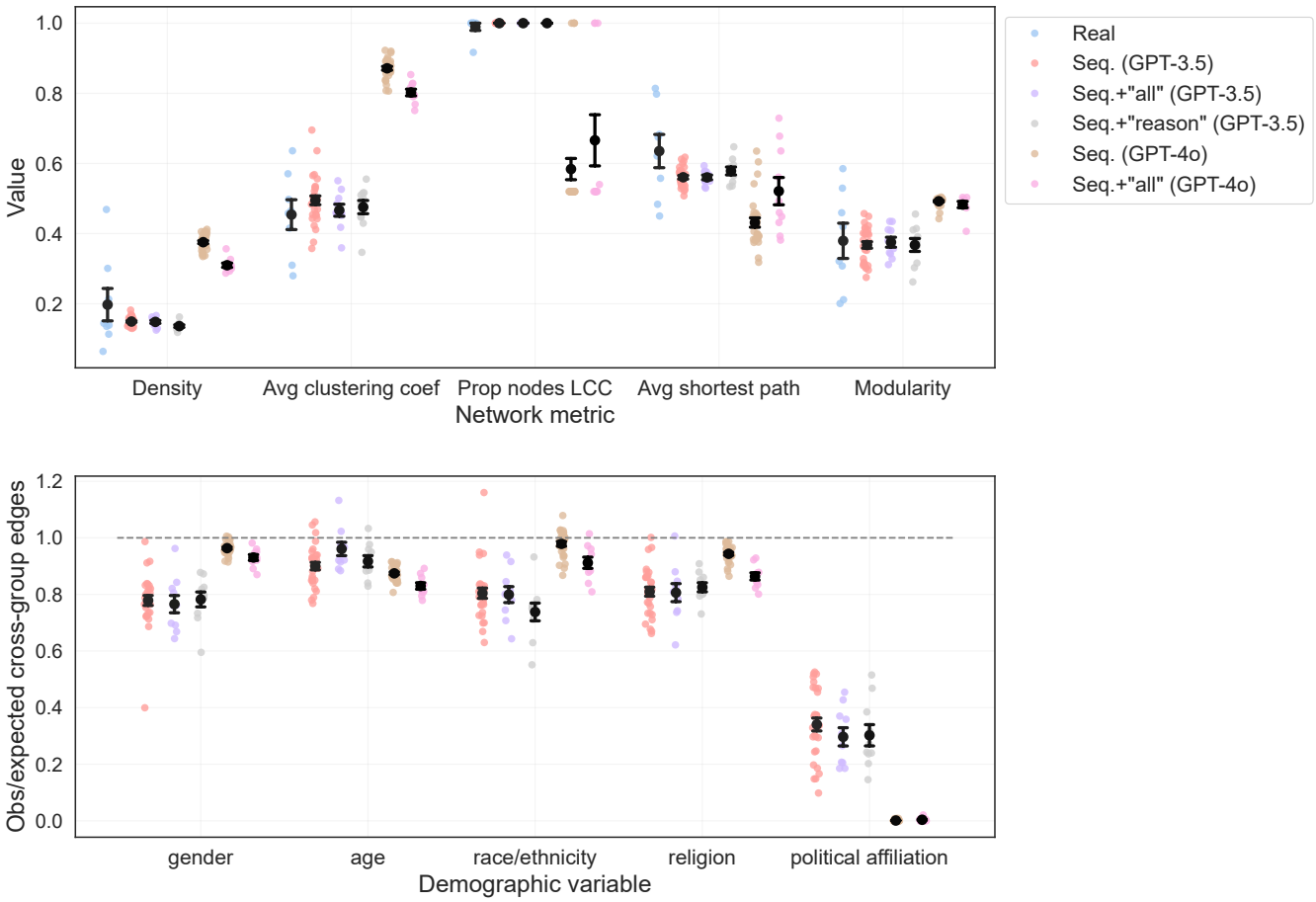

Figure B6: Our results do not change siginificantly with minor changes to the prompt. We try adding "Pay attention to all demographics" (+"all") and prompting the LLM to give a short reason for each friend that it selects (+"reason"). **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network. All model results shown here use the Sequential method.

Demographic group	Top interests
Man	social justice (29.1%), church activities (23.8%), conservative politics (23.4%), sports (17.1%), community service (16.7%), technology (15.7%), golf (14.0%), history (13.6%), video games (12.6%), progressive policies (11.2%)
Woman	social justice (35.3%), gardening (31.5%), community service (27.7%), reading (21.6%), church activities (20.8%), conservative politics (16.2%), volunteering (12.8%), family gatherings (11.6%), travel (10.8%), cooking (9.2%)
White	gardening (27.4%), conservative politics (25.8%), church activities (24.0%), social justice (23.0%), community service (15.9%), reading (15.6%), outdoor activities (12.0%), history (11.0%), volunteering (11.0%), golf (10.7%)
Hispanic	social justice (46.0%), community service (35.8%), family gatherings (27.3%), church activities (18.2%), cultural heritage (16.0%), music (13.4%), progressive policies (12.3%), conservative politics (12.3%), family (10.7%), soccer (10.7%)
Black	social justice (58.6%), community service (33.1%), church activities (21.1%), gospel music (17.3%), progressive policies (15.8%), community activism (14.3%), music (14.3%), reading (13.5%), jazz music (11.3%), sports (11.3%)
Asian	social justice (27.9%), travel (24.6%), technology (23.0%), church activities (16.4%), community service (14.8%), conservative politics (14.8%), yoga (13.1%), fitness (11.5%), entrepreneurship (11.5%), gardening (11.5%)
Protestant	church activities (36.1%), social justice (33.2%), community service (25.2%), gardening (22.0%), conservative politics (21.4%), reading (13.9%), sports (10.9%), volunteering (9.5%), history (9.3%), outdoor activities (8.9%)
Catholic	community service (37.7%), social justice (35.2%), family gatherings (25.5%), church activities (24.3%), conservative politics (21.5%), gardening (16.2%), reading (10.1%), sports (9.3%), volunteering (8.9%), progressive policies (8.5%)
Unreligious	social justice (28.3%), technology (22.8%), gardening (21.0%), travel (18.8%), fitness (16.3%), conservative politics (15.9%), reading (13.0%), outdoor activities (13.0%), progressive policies (12.7%), entrepreneurship (9.4%)
Republican	conservative politics (41.6%), church activities (32.1%), gardening (23.2%), outdoor activities (15.0%), golf (14.3%), community service (13.9%), sports (13.5%), family gatherings (12.7%), history (12.2%), hunting (11.4%)
Democrat	social justice (62.5%), community service (29.3%), progressive policies (18.6%), gardening (16.4%), reading (15.6%), church activities (13.2%), volunteering (12.4%), music (11.0%), travel (10.4%), technology (10.0%)

Table B4: Top 10 interests per demographic group.

the LLM is assigned one persona at the time and all other personas are listed. We generate 30 networks per method where, for each generated network, we randomize the order that the personas are listed, and, for the Local and Sequential methods, we also randomize the order in which personas are assigned (using different orders for listing and assignment). For the Sequential method, we experiment with providing each persona's list of friends versus only their degree. We find that the model performs better with only degree, while listing friends results in unrealistically high densities. Furthermore, fewer tokens are used with only degree, so we use this version of Sequential.

In Figure 2, we visualize examples of networks generated by each of the three methods. The networks are visualized using `networkx`, with a spring layout and fixed seed for the visualization. The visualizations primarily reveal differences in the generated networks across methods, but they also reveal differences between networks generated by the same method. Since the visualization method is fixed, these differences should be attributed to variation in actual networks, driven by the randomization in persona ordering (for listing and assignment) and randomness in LLM outputs (using a default temperature of 0.8).

Model costs and comprehension. In this work, we focus primarily on realism and bias to evaluate the generated networks. However, when using LLMs, measures such as token costs and model comprehension are also very important. We measure token cost as the number of input tokens and output tokens, summed over generating the entire network. We also conduct a big-O analysis of how the input tokens scale with network size, based on N (the number of nodes in the network), E (the number of edges), D (the number of tokens to describe a persona's demographics), and I (the number of tokens to describe a persona's interests). To measure model comprehension, we consider the expected number of turns (prompt and response) to generate the entire network, compared to the actual number of turns. The actual exceeds the expected if we could not parse the model's response (e.g., it gave an ID that does not exist in the network). So that we can measure the concepts of cost and model comprehension independently, we only add to the number of input/output tokens on *successful* turns.

We report results in Table C2, with all measures averaged over the 30 networks generated per method. The Global method requires, by far, the fewest number of input tokens, since we only need to list all personas once. How-

Demographic	Group	Count in 50	Count in 1,000
Gender	Woman	26	499
	Man	24	492
	Nonbinary	0	9
Race/ethnicity	White	33	609
	Hispanic	10	187
	Black	4	133
	Asian	2	61
	American Indian/Alaska Native	1	9
	Native Hawaiian/Pacific Islander	0	1
Religion	Protestant	19	440
	Unreligious	20	276
	Catholic	11	247
	Jewish	0	12
	Hindu	0	9
	Buddhist	0	8
	Muslim	0	6
	Other Christian	0	2
Political affiliation	Democrat	24	501
	Republican	26	474
	Independent	0	25

Table C1: Marginal distributions of personas' demographics. Count in 50 indicates the number of personas belonging to each group in the sample of 50 personas we used for most experiments. Count in 1,000 indicates their number in the sample of 1,000 personas we used for evaluating top interests per demographic group (Table B4)

Method	O(input)	# input tokens	# output tokens	Expected # turns	Actual # turns
Global, demos	$O(ND)$	607	95.267	1	1.133
Local, demos	$O(N^2D)$	30450	196.767	50	50
Sequential (friend list), demos	$O(N^2D + NE)$	42601.233	335.633	50	50
Sequential (degree), demos	$O(N^2D)$	37735	200.300	50	50.033
Sequential (degree), interests	$O(N^2I)$	37035	257.967	50	50.033
Sequential (degree), demos + interests	$O(N^2(D + I))$	57385	190.733	50	50.067

Table C2: Comparing model costs and comprehension over different LLM methods for social network generation. For Sequential, we consider two versions: listing each persona's current friend list vs. only their current degree. N is the number of nodes in the network, E is the number of edges, and D and I are the number of tokens needed to describe all demographics and interests, respectively.

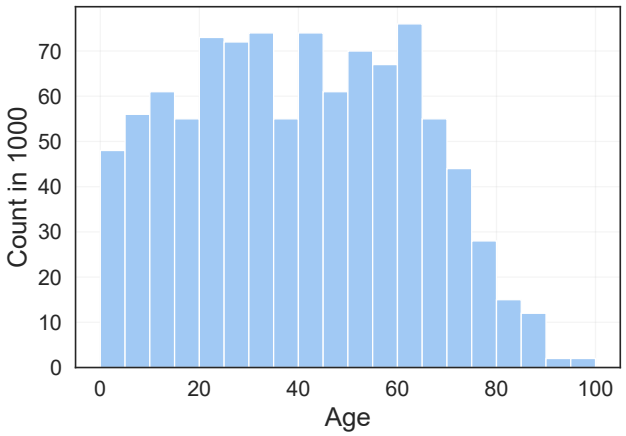

Figure C1: Age distribution in sample of 1,000 personas.

ever, Global has the highest rate of errors, resulting in an actual number of turns that is 13% higher than the expected number of turns. The Local method and Sequential method scale similarly, since they both require iterating through all N personas, and for each persona, presenting them with information about the other $N-1$ personas. Their main difference is what information is provided about each persona: Local only provides D demographics, while Sequential also provides network information, which can be each persona's friend list (resulting in an additional E per iteration) or only each persona's degree (resulting in $D+1$ per persona). Both Local and Sequential have good model comprehension, with the models rarely failing and actual turns being only 0-6.7% above expected turns. Overall, this table lays out tradeoffs, since the Global method requires far fewer tokens, but as we saw in the main text (Figures 2-4), it produces much less realistic networks, and it is more prone to invalid responses.

Scalability. To address these tradeoffs, we propose a simple extension of the Local and Sequential methods to make them more scalable. Previously, when we queried a given persona, we presented it with information about the $N-1$ other personas. Instead, we could provide it a *subset* of the other personas, with a fixed subset size k. As a result, instead of scaling on the order of $O(N^2D)$, Local and Sequential would scale on the order of $O(ND)$, which is the same as the Global method.

To demonstrate how this could be done, we implement a simple version of this extension, where the subset is sampled uniformly at random. Then, using the Sequential method, we generate networks with 300 personas, sampling 30 random personas per query. Compared to the smaller 50-node networks we generated before, these larger networks show similar levels of homophily across the five demographic variables, and political homophily remains by far the strongest (Figure C6, bottom). In terms of structural characteristics, the larger networks match the smaller ones (and the real networks) on average shortest paths and modularity, but the larger networks have lower density and clustering (Figure C6, top). This is expected, due to the sampling: previously, each persona would select friends from 100% of all other personas, but now each persona selects friends from only 10% of other personas. Future work could explore methods of non-uniform sampling that can correct for these lower levels of density and clustering, such as using a recommender system-like model (potentially powered by a graph neural network) to choose the subset of other personas that each persona sees.

Network	# nodes	# edges
Galesburg	31	63
Hi-tech	36	91
Karate	34	78
Prison	67	142
Tailor 1	39	158
Tailor 2	39	223
Moreno freshmen	31	218
Moreno high school	70	274

Table C3: Summary statistics of the eight real social networks that we use.

C.3 Evaluating Network Structure

In this section, we provide more details about the real social networks that we compared against and define various measures we used to characterize the networks. In Table C3, we provide basic statistics about the real networks.

Real networks. We use the following six networks from the CASOS repository (CASOS 2024):

Galesburg (Coleman, Katz, and Menzel 1957). This network describes friendship ties between physicians, where they were asked to name three doctors whom they considered personal friends and to nominate three doctors with whom they would discuss medical matters. The goal of this study was to analyze the diffusion of a new drug in terms of when physicians first prescribed it, studied in the context of their social network.

Hi-tech (Krackhardt 1999). This network describes friendship ties between employees of a small hi-tech firm. In a survey, they answered the question, "Who do you consider to be a personal friend?" Most friendship nominations were reciprocated, and an edge is kept only if both people nominated each other.

Karate (Zachary 1977). This network describes friendships between members of a karate club at a US university. Due to a schism where the club split into two, this network has often been used to study community structure.

Prison (MacRae Jr. 1960). This network describes friendship ties between prison inmates. All were asked, "What fellows on the tier are you closest friends with?" Each respondent could choose as many or as few friends as he desired.

Tailor (Kapferer 1972). This network describes relations between workers at a tailor shop in Zambia (then Northern Rhodesia). The dataset includes both "instrumental" ties (work-related), which we leave out, and "sociational" ties (friendship, socioemotional), which we include. Networks were recorded twice, seven months apart, so we have two

User: In 8-12 words, describe the interests of someone with the following demographics:
race/ethnicity: White
age: 72
gender: Man
political affiliation: Republican
religion: Catholic
Answer by providing ONLY their interests. Do not include filler like "She enjoys" or "He has a keen interest in".

Figure C2: Prompt to generate interests for persona.

System: Your task is to create a realistic social network. You will be provided a list of people in the network, where each person is described as "ID. Gender, Age, Race/ethnicity, Religion, Political affiliation". Provide a list of friendship pairs in the format ID, ID with each pair separated by a newline. Do not include any other text in your response. Do not include any people who are not listed below.

User: 28. Man, age 48, Hispanic, Protestant, Democrat
11. Man, age 31, White, Protestant, Democrat
10. Man, age 58, Hispanic, Catholic, Democrat
41. Woman, age 41, White, Catholic, Republican
...

Figure C3: Prompt for Global method.

System: You are a Man, age 48, Hispanic, Protestant, Democrat. You are joining a social network. You will be provided a list of people in the network, where each person is described as "ID. Gender, Age, Race/ethnicity, Religion, Political affiliation". Which of these people will you become friends with? Provide a list of *YOUR* friends in the format ID, ID, ID, etc. Do not include any other text in your response. Do not include any people who are not listed below.

User: 11. Man, age 31, White, Protestant, Democrat
10. Man, age 58, Hispanic, Catholic, Democrat
41. Woman, age 41, White, Catholic, Republican
2. Woman, age 20, White, Catholic, Republican
...

Figure C4: Prompt for Local method.

System: You are a Man, age 48, Hispanic, Protestant, Democrat. You are joining a social network. You will be provided a list of people in the network, where each person is described as "ID. Gender, Age, Race/ethnicity, Religion, Political affiliation", followed by their current number of friends. Which of these people will you become friends with? Provide a list of *YOUR* friends in the format ID, ID, ID, etc. Do not include any other text in your response. Do not include any people who are not listed below.

User: 11. Man, age 31, White, Protestant, Democrat; has 4 friends
10. Man, age 58, Hispanic, Catholic, Democrat; has 2 friends
41. Woman, age 41, White, Catholic, Republican; has 0 friends
2. Woman, age 20, White, Catholic, Republican; has 7 friends
...

Figure C5: Prompt for Sequential method, when only degree is provided. We also have a version where each persona's current list of friends (in the form of ID) is provided.

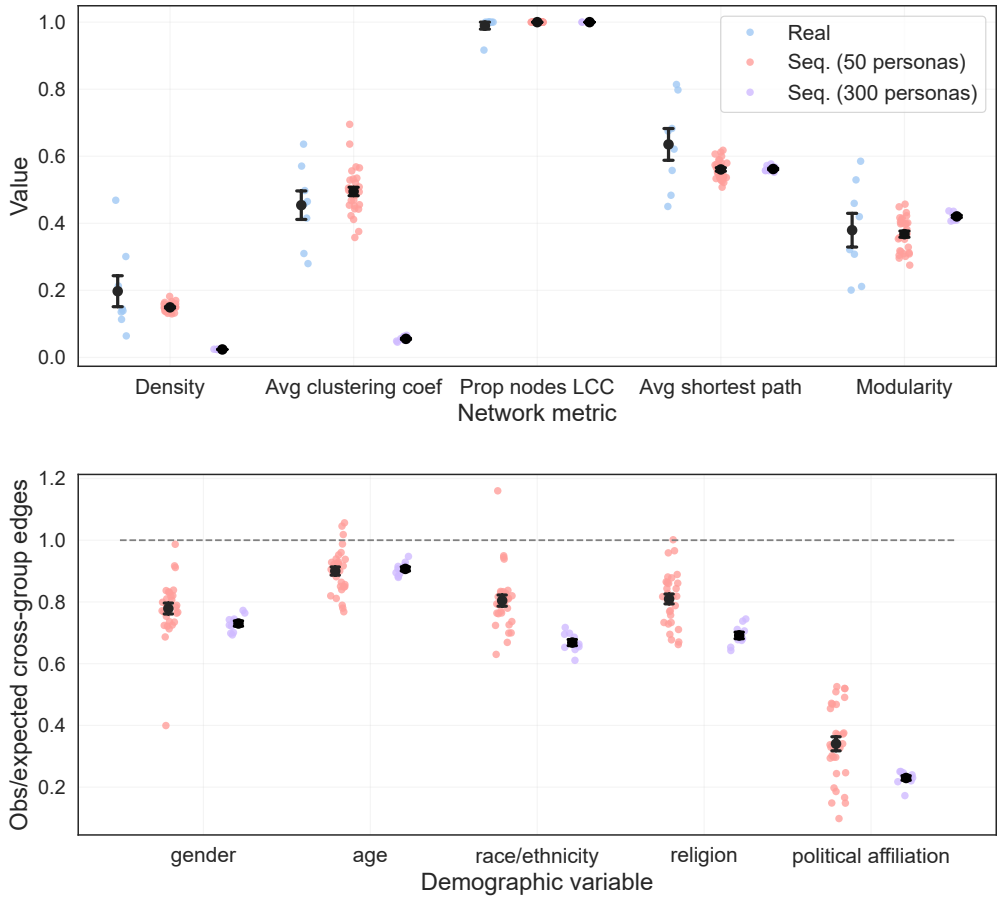

Figure C6: Comparing original generated networks with 50 personas to generated networks with 300 personas, where 30 personas are sampled uniformly at random per query. **Top:** structural network metrics. **Bottom:** homophily, where ratios below 1 (marked by the grey line) indicate homophily and lower ratios indicate more homophily. In both plots, we visualize mean and standard error (in black) and individual data points corresponding to each network. All model results shown here use the Sequential method and GPT-3.5-Turbo.

networks from this dataset.

We also include two networks from the KONECT repository (KONECT 2024):

Moreno freshmen. This network describes friendship ratings between university freshmen. The edge weights range from -1 (risk of getting into conflict) to +3 (best friend). We keep all edges with strictly positive weight.

Moreno high school. This network describes friendship ratings between high school boys. The edge weights range from -1 (risk of getting into conflict) to +3 (best friend). We keep all edges with strictly positive weight.

Network metrics. Most of the network metrics that we compare against are straightforward, such as density or average clustering coefficient, which are defined in the main text (Section 4). The one more involved metric is modularity, which assesses the quality of a community partition. Modularity measures the number of edges within the community, compared to how many edges are expected, and it is defined as

$$Q = \frac{1}{2E}\sum_{ij}(A_{ij} - \gamma\frac{N_i N_j}{2E})\mathbb{1}[c_i = c_j], \tag{5}$$

where γ is the resolution parameter (set by default to 1) and c_i indicates node i's community in the partition. As in the main text, E is the total number of edges in the network; A_{ij}, as the adjacency matrix, is 1 if nodes i and j are connected and 0 otherwise; and N_i is i's number of neighbors.

We also need to define how we quantified the distance between the generated networks and real networks in Table B1. Let $x_1, \cdots, x_m$ represent the values of a metric (e.g., density) from the real networks (where $m = 8$), and let $y_1, \cdots, y_n$ represent the values of the metric from the generated networks (where $n = 30$). First, we compute the difference in their mean, divided by the standard deviation of the real network distribution:

$$D = \frac{|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{n}\sum_{j=1}^{n} y_j|}{\sigma_{\text{real}}}. \tag{6}$$

We normalize by standard deviation to make differences comparable across metrics. We only normalize by the real networks' standard deviation since normalizing by the generated networks' standard deviation would arbitrarily reward higher variance methods. Second, we use the two-sample Kolmogorov–Smirnov (KS) statistic, which measures the distance between two empirical distributions by comparing their cumulative distribution functions (Hodges Jr. 1958):

$$F_{\text{real}}(u) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[x_i \le u]$$

$$F_{\text{gen}}(u) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[y_j \le u]$$

$$D_{KS} = \sup_u |F_{\text{real}}(u) - F_{\text{gen}}(u)|. \quad (7)$$

Comparing to classical models. We compare to several classical models for social network generation, which we describe below, along with how we chose their parameters. For all models, we set the number of nodes, N, to 50, to mimic our LLM experiments with 50 personas. We use the `networkx` implementation for all three models.[7]

Erdős–Rényi random graph (Erdős and Rényi 1959). We use the $G_{N,p}$ random graph model, which has N nodes and each edge is included with independent probability p. In our experiments, we simply set p to the average density of the eight real social networks.

Barabási-Albert preferential attachment (Barabási and Albert 1999). In this model, one node is added to the graph at each step, and it forms m edges with existing nodes, where each neighbor is sampled with probability proportional to its current degree. The `networkx` implementation starts by default with a star graph on $m+1$ nodes, then adds the remaining $N-m-1$ nodes one at a time. Thus, the number of edges in the graph is always $m + ((N-m-1)\cdot m)$. We choose $m = 5$, which minimizes the difference between the generated graph's density and the average density of the real social networks.

Watts-Strogatz small world (Watts and Strogatz 1998). In this model, first a ring is created over N nodes, then each node is joined to its k nearest neighbors, forming a lattice. Then, with independent probability p, each edge (i, j) is rewired, meaning it is replaced with (i, j'), where j' is selected from all nodes (aside from i and i's existing neighbors) uniformly at random. Since each node is joined to its k nearest neighbors, the number of edges in this graph is always $\frac{nk}{2}$. So, we choose $k = 10$, which also minimizes the difference between the generated and real networks' average density. Then, with $N = 50$ and $k = 10$, we sweep over possible values of p in $\{0.01, 0.02, \cdots, 0.5\}$ to minimize the difference between the generated and real networks' average clustering coefficient, resulting in $p = 0.15$.

C.4 Evaluating Homophily

To evaluate the LLM's level of homophily, we typically use the cross-group ratio in this work (Eq. 1), which measures the ratio of observed-to-expected proportion of cross-group edges. We also use a closely related measure, the same-group ratio, which measures the ratio of observed-to-expected proportion of same-group edges:

$$H = \frac{S_{\text{obs}}}{S_{\text{exp}}} = \frac{\frac{\sum_{ij} A_{ij}\cdot \mathbb{1}[g_i = g_j]}{E}}{\frac{\sum_g N_g(N_g-1)}{N(N-1)}}. \quad (8)$$

To compare the LLM's level of political homophily to prior work, we needed to use their measures of homophily. First, we define the isolation index, used in Halberstam and Knight (2016) and Gentzkow and Shapiro (2011). We define it following Halberstam and Knight (2016) (Appendix, p. 5). First, for voter $j \in J$ (they refer to all nodes in their network as voters), let v_{jC} and v_{jL} indicate their number of conservative and liberal followers, respectively. Then, isolation is defined as

$$\text{share-C}_j = \frac{v_{jC}}{v_{jC} + v_{jL}}$$

$$\text{C-exposure}_i = \frac{1}{\sum_{j\in J} A_{ij}} \sum_{j\in J} A_{ij} \cdot \text{share-C}_j$$

$$\text{C-exposure}_t = \frac{1}{N_t} \sum_{i \in I_t} \text{C-exposure}_i$$

$$\text{isolation} = \text{C-exposure}_C - \text{C-exposure}_L \quad (9)$$

Thus, C-exposure$_i$ for a voter i is the average share-C$_j$ over the voters that they follow, C-exposure$_t$ is the average conservative exposure for voters in group t, and isolation is the difference in average conservative exposure for conservative versus liberal voters.

We also use the polarization measure from Garimella and Weber (2017). First, they compute the user's leaning, l, which is

$$l = \frac{\alpha}{\alpha + \beta},$$

where α and β indicate how many left-leaning and right-leaning users, respectively, are followed by this user. Specifically, they begin with a uniform prior, $\alpha = \beta = 1$, then every follow/retweet of a user on each side adds one to that side's parameter (α for left, β for right). Note that their definition of leaning is very similar to the definition of share-C$_j$ in Halberstam and Knight (2016), with the addition of the uniform prior. Then, their definition of polarization p is

$$p = 2 \cdot |0.5 - l|, \quad (10)$$

which lies between 0 and 1, representing how much the user deviates from a balanced leaning of 0.5. They report average p over users, which is what we measure on our networks.

[7] https://networkx.org/documentation/stable/reference/generators.html