Who follows Whom on Twitter? An analysis of homophily patterns

ALEXI QUINTANA MATHE
Northeastern University, USA

ZHEN GUO
Northeastern University, USA

DAVID LAZER
Northeastern University, USA

NIR GRINBERG
Ben-Gurion University, Israel

Introduction

From mental health concerns (Masciantonio et al., 2021; Meier et al., 2020) to information and misinformation spread (Bakshy et al., 2015; Bovet & Makse, 2019; Grinberg et al., 2019a), the impact of Social Media on society is broad. In one way or another, this impact is influenced by the relational structure among users in these platforms. For some of the most used Social Media sites, such as Facebook or Instagram, the connections between accounts almost completely control the content users see and experience. Conceptually, the social network behind each of these platforms can be thought as a backbone structure upon which dynamics, like sharing or accessing

information, take place. Hence, understanding and accurately describing this network of relationships is crucial to grasp the functioning and implications of these sites.

Over the last fifteen years, Twitter has consolidated as a major Social Media platform, playing an important role in the information ecosystem of western societies (Priya et al., 2019; Scharkow et al., 2020; Shearer & Matsa, 2018). 23% of US adults say they use Twitter (Odabaş, 2022), and 69% of US Twitter users say they get news from this platform (Mitchell et al., 2021). In addition, 46% of US Twitter users answer that Twitter is an important way or the most important way they get news. Twitter's influence is recognized for disparate domains such as politics (Bovet & Makse, 2019; Enli & Skogerbø, 2013; Stier et al., 2018), science dissemination (Cormier & Cushman, 2021), supply chain and organizational reputation (Etter et al., 2019; Schmidt et al., 2020), and consumer media relationship (Gibbs & Haynes, 2013; Wu et al., 2011), making it an important platform to study. However, while it has deserved a lot of attention from researchers, some aspects of its underlying social network remain unexplored.

A notable feature of social relationships, online and offline, is homophily, defined as the tendency of people to have ties with similar others (McPherson et al., 2001). Follower relationships on Twitter are not an exception. In particular, Twitter users show a tendency to follow other users with similar political attitudes (Eady et al., 2019; Garimella & Weber, 2017; Mosleh et al., 2021), interests (Romero et al., 2013), race (De Choudhury, 2011; Messias et al., 2017), and religion (Chen et al., 2014). In addition, physical proximity is strongly associated with following relationships (Stephens & Poorthuis, 2015; Takhteyev et al., 2012). Homophily on Twitter is linked to the appearance of echo chambers, and, in general, to heterogeneous and inefficient flows of information in the platform (Grabowicz et al., 2016; Halberstam & Knight, 2016; Sasahara et al., 2021; Tokita et al., 2021). High levels of homophily are also associated with concerns of polarization and segregation in the US (DiPrete et al., 2011a; Garimella & Weber, 2017; Gentzkow & Shapiro, 2011). While there is evidence in the literature that sharing attributes such as partisanship or physical location increases likelihood of

following someone on Twitter, very few studies compare the importance of these attributes (De Choudhury, 2011; Messias et al., 2017). Furthermore, they are often inferred using profile pictures or activity in the platform itself, which can lead to biased estimates of homophily (Berry et al., 2021).

Our work consists in a large-scale analysis of homophily in Twitter following relationships in the US. Our primary research question is: *how similar are US Twitter users to the people that they follow on Twitter?* We leverage a dataset of about 1.6 million US registered voters matched to Twitter accounts to analyze homophily patterns in terms of partisanship, race/ethnicity, sex, age, and residence. This dataset of accounts linked to public U.S. voter records, previously validated and fairly representative of US registered voters on Twitter (Hughes et al., 2021; Shugars et al., 2021), enables accessing demographic, location and partisanship information external to activity on the platform at a scale unprecedented in the literature. Our main contributions are as follows:

- We evaluate how much each variable is associated with following someone on Twitter. We find substantial levels of partisanship, race/ethnicity, and age homophily, but geographical proximity is the variable most strongly associated with following ties. Living in rural, urban, or suburban areas is also important.
- We use multivariate analysis to disentangle how the association between some variables and follower ties is driven by other variables. For example, the importance of partisanship diminishes in multivariate analysis, suggesting that some of the partisanship homophily observed on Twitter is driven by residential segregation.
- Because we have access to user attributes external to behavior on the platform, we are able to run a comprehensive analysis of homophily for US Twitter users. For example, we have access to precise age and residential data. For partisanship, we use a measure inferred for everyone in our sample without conditioning on political activity in the platform as most research on polarization on Twitter does.

Literature Review

We start this section by defining the various types of homophily found in the literature and discussing which ones we address. Afterwards, we delve into the research on homophily by the different variables we study, with a special focus on Social Media and Twitter. For partisanship, we review some of the research on the link between polarization and Social Media, because of its relationship to online political homophily. Then, we summarize the literature on homophily based on race/ethnicity, age, gender, and location, discussing the connection between Social Media, Twitter and geography.

Homophily

As defined in the work by McPherson, Smith-Lovin and Cook, homophily "is the principle that a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson et al., 2001). Homophily is a pervasive pattern of social relationships, robust over widely varying types of relations. While the notion behind homophily is simple, there are different types of homophily, substantial variation in its measurement, and inconsistencies in the terminology used in the literature (Bojanowski & Corten, 2011; Lawrence & Shah, 2020; McPherson & Smith-Lovin, 1987; Wimmer & Lewis, 2010). Here, we use the word *segregation* to describe the raw composition of the network, that is, the extent to which similar people are connected. This is often called absolute homophily or homogeneity in other work. A first basic distinction is the split of segregation in *baseline homophily*, the expected homophily level generated by the opportunity set in a given social context (that is, the number of members of each group), and *inbreeding homophily*, the homophily "over and above" this opportunity set (Blau, 1977; McPherson et al., 2001). Inbreeding homophily encompasses both segregation generated by "the individual-level propensity to choose similar others", termed *choice homophily* (McPherson & Smith-Lovin, 1987), and segregation generated by other confounding dimensions, by shared foci of activity, by network effects such as triadic

closure, and by the differential likelihood of groups to send or receive ties (Kossinets & Watts, 2009; Wimmer & Lewis, 2010).

Some of main impacts of homophily in Social Media, such as diffusion and information spread, depend on the extent that the network is segregated regardless of the factors leading to this segregation (Centola, 2011; Korkmaz et al., 2019; Stein et al., 2023; Yavaş & Yücel, 2014) Therefore, in our work, we closely examine network segregation levels by calculating the percentage of the follower relationships of the members of a group that are directed towards each group. We measure inbreeding homophily by calculating probabilities of a tie between members of two groups and use measures of homophily that control for the propensity of different groups to send and receive ties. This is especially relevant on Twitter, where some groups may have widely varying numbers of followers or followees, potentially driving homophily or heterophily patterns. In addition, homophily by one variable may drive homophily by another correlated variable. Therefore, we use multivariate analysis and model the probability of a follower tie with all of our variables as predictors, allowing us to examine when the inbreeding homophily we find for some variables is driven by other variables.

Taken together, our measures provide information on how the structure of relationships behind Twitter is related, in the US, to the partisanship, demographics, and location of its users. However, we do not aim at completely disentangling choice homophily from the influence of factors such as platform algorithms or triadic closure. Our goal is to provide a set of bivariate and multivariate descriptive findings relative to homophily on Twitter, and we do not offer full causal explanations relative to the role of homophilous preferences on tie formation and dissolution on Twitter. This type of analysis is hard given the opacity of the platforms' internal functioning, and needs experiments to be fully addressed, that, however, may have their own ecological validity issues (see Mosleh et al., 2021 for an example of such work).

Political orientation

A major concern associated with political and partisan homophily in the US, especially online, is polarization. One of the core theories behind these concerns, in the online context, is Selective Exposure theory (Sears & Freedman, 1967). This theory predicts a low diversity in people's media diets as a consequence of preferences for like-minded content and the possibility of choosing between a myriad of news providers (Gentzkow & Shapiro, 2011; Sunstein, 2001). When viewed through this lens, Social Media platforms, the main way people access news in the US (Shearer & Matsa, 2018), would be making matters worse due to homophilous tie-making and algorithmic influence. The argument is that the tendency to select news providers and build connections on Social Media of a similar political orientation, reinforced by algorithmic systems that feed users content and tie recommendations based on these preferences, would generate stark echo chambers, leading to a polarized society (Sunstein, 2018). This line of reasoning is supported by the growing levels of polarization in US society in the last 20 years, since the rise of Social Media, through various dimensions such as affect towards opposing party supporters or residential and dating preferences (Boxell et al., 2017; Huber & Malhotra, 2017; Hui, 2013; Iyengar & Westwood, 2015). In particular, research has found substantial and increasing ideological homogeneity in political discussion and confidant networks between 1992 and 2016 (Butters & Hare, 2022; Lee & Bearman, 2020). Findings on other types of tie, such as work ties or weaker ties, seem to point to somewhat lower levels of polarization (DiPrete et al., 2011b; Eveland et al., 2018; Mutz & Mondak, 2006).

While simulations and theoretical models support this relationship between Social Media and echo chambers, they are generally validated on specific online conversations, contexts, and outcomes (Halberstam & Knight, 2016; Rychwalska & Roszczyńska-Kurasińska, 2018; Sasahara et al., 2021; Tokita et al., 2021). Overall, empirical evidence regarding the impact of Social Media on polarization is mixed. Polarization levels have grown faster for older rather than younger age cohorts, while

younger cohorts use the Internet and Social Media more (Boxell et al., 2017). In opposition with Selective Exposure theory, Scharkow et al. found that social network sites increase the diversity of news domains visited in Germany (2020), and data from Facebook shows exposure to politically cross-cutting content to be common in this site (Bakshy et al., 2015).

Regarding Twitter, the high heterogeneity across types of users and countries and the importance of context (Urman, 2020; Vaccari et al., 2016) make it hard to reach a generalized conclusion. Furthermore, different methods and data may lead to conflicting conclusions. For example, a substantial number of studies on Twitter find segregated political conversations, in particular along party lines in the US (Barberá et al., 2015, 2015; Cinelli et al., 2021; Conover et al., 2011; Williams et al., 2015). However, by using tweets as their core unit of analysis, this research overstates the average levels of segregation of users interested in politics. The content generated on Social Media is highly concentrated in a relatively small set of users (Mcclain, 2021; Odabaş, 2022; Wu et al., 2011). For instance, 78% of all political tweets from U.S. adults are created by users aged 50 and older, and users who frequently tweet about politics are more politically engaged and partisan (Bestvater et al., 2022; Pew Research Center, 2019). This implies that political conversations on Twitter tend to be monopolized by particularly polarized individuals.

When looking at partisan segregation at the user level and going beyond political content, a different picture emerges. Shore et al. (2018) found that the average Twitter account tweets links from more moderate news sources than the links they are exposed to, while a small set of users, responsible for the majority of tweets received, show the opposite pattern. Research using representative survey samples from the US, Italy, and Germany finds significant average exposure to dissimilar political opinions on Twitter, but also substantial variation by ideology and slant level (Eady et al., 2019; Vaccari et al., 2016). Recent studies document that the majority of users of this platform do not follow any political elite, and that non-political opinion leaders attract considerably more attention

(Mukerjee et al., 2022; Wojcieszak et al., 2022). However, users who do follow political elites show sizable polarization levels across a range of behaviors on the platform, including following (Halberstam & Knight, 2016; Wojcieszak et al., 2022). Furthermore, polarization seems to have increased among this subset of users between 2009 and 2016 (Garimella & Weber, 2017). Finally, users who retweeted partisan media display a clear preference to follow in-party accounts in experimental settings (Mosleh et al., 2021).

One of the main takeaways from this literature is that users displaying some kind of political behavior on the platform tend to be polarized, while average Twitter users do not engage in politics and are not polarized. In particular, this seems to hold for the US and for following behavior, the focus of our work. Most of this literature studies users that display interest in politics in the platform, either by sharing tweets about politics or by interacting with political accounts (Boutyline & Willer, 2017; Colleoni et al., 2014; Garimella & Weber, 2017; Halberstam & Knight, 2016; Mosleh et al., 2021; Wojcieszak et al., 2022). Eady et al. (2019) use survey information on political attitudes external to platform activity for a representative set of Twitter users, but they infer the ideology of the accounts they follow using their following behavior (Barberá et al., 2015) and focus on the interactions with political elites. Therefore, current research has still not explored how strong is partisanship homophily for Twitter following relationships of general US users. We are able to fill this gap by using partisanship information independent of activity on the platform. By doing so, we can take into account the ideological orientation of users who do not engage in political behavior on the site, whether they are interested in politics or not. On top of that, our matching of users to voter file information ensures our sample is composed of real people on Twitter, and not bots or organizational accounts.

On the other hand, almost all the literature to date has studied ideological homophily on Twitter without comparing it to other possible attributes influencing following behavior. The only exception is De Choudhury's study (2011), which compares different types of homophily and finds substantially more partisanship than location, ethnicity or gender homophily. However, this study uses profile descriptions to match users to party labels,

probably leading to matches for only highly politicized users, and infers gender and ethnicity using names. Hence, another way we add to the literature is by putting into context partisanship homophily with other types of homophily such as race or location, for a large and diverse sample of US voters on Twitter. Furthermore, no research has explored how partisanship homophily may be driven by other variables. Partisanship in the US is strongly associated with age, race/ethnicity, and residential patterns (Brown & Enos, 2021), variables also found to be related with following behavior on Twitter (see next subsections). Some reviewed studies implicitly assume the associations found between partisanship similarity and following reveal preferences for ideological congruity (Boutyline & Willer, 2017; Colleoni et al., 2014; Garimella & Weber, 2017). However, we find that some partisanship homophily for US Twitter users is driven by geographical proximity and the type of residential area of users.

Race/Ethnicity

While homophily tends to be high in the US for the sociodemographic characteristics we use in this work (race/ethnicity, sex, age), race is the demographic generally found to segregate social networks the most (McPherson et al., 2001; Smith et al., 2014). Race and ethnicity homophily have been thoroughly studied, especially in contexts like schools, finding high levels of homophily (Currarini et al., 2010; Moody, 2001; Shrum et al., 1988; Wimmer & Lewis, 2010). If we focus on network confidants, a canonical measure of core ties, the frequency of cross racial ties was, in 1985, less than one seventh the expected if ties were random (Marsden, 1988; McPherson et al., 2001), and stayed similar in 2004 (Smith et al., 2014). African-Americans, followed by Asians, Hispanics, and finally Whites, show the highest levels of inbreeding homophily. Research with Facebook data from US college campuses and high schools find similar racial and ethnic homophily patterns as for offline relationships, although only a portion of this homophily is attributed to choice homophily (Lewis et al., 2012; Mayer & Puller, 2008; Wimmer & Lewis, 2010). Race homophily has also been found in other online environments, such as

MySpace or Grindr, with some variation in the groups displaying more homophily (Mazur & Richards, 2011; Salamanca et al., 2019; Thelwall, 2009). Twitter follower networks also present homophily for this attribute, however, it is generally lower than for offline relationships (Cesare et al., 2017; De Choudhury, 2011; Messias et al., 2017). These studies also find higher African-American homophily than for Whites or Asians. In general, research on racial and ethnic homophily in Social Media has focused on specific populations (Lewis et al., 2012; Mayer & Puller, 2008; Salamanca et al., 2019; Wimmer & Lewis, 2010) and suffered from small samples (Cesare et al., 2017; Mazur & Richards, 2011; Thelwall, 2009) or biased race/ethnicity inference methods based on names and profile pictures (De Choudhury, 2011; Golder et al., 2022; Kozlowski et al., 2022; Messias et al., 2017). We are the first to quantify racial and ethnic homophily on Twitter for a large-scale sample of US voters, using either self-reported information or inferences that we are able to compare against self-reports (see Data section for more details)

Gender

Regarding gender and sex, the homophily patterns across different contexts and types of relationships are more complex. Gender homophily is generally significant before adulthood and then decreases, mainly due to kin ties (McPherson et al., 2001; Shrum et al., 1988; Smith et al., 2014). Non-kin confidant ties remain significantly homogeneous by gender and sex, though less than for race/ethnicity or age. Other types of ties, such as work or political discussion ties, tend to be more sex segregated, especially for men and often due to opportunity structures. The role of sex and gender in online spaces is mixed, and generally minor. Early research on MySpace, Facebook, and online videogames found small or no gender/sex homophily (Huang et al., 2013; Mayer & Puller, 2008; Mazur & Richards, 2011; Thelwall, 2009; Ugander et al., 2011; Utz & Jankowski, 2016), and similarly on Twitter (De Choudhury, 2011; Messias et al., 2017). However, a large-scale study with Tuenti, a social network platform from Spain similar to Facebook, found substantial homophily, especially for women, and points that younger users show

more preferences for gender heterogeneity (Laniado et al., 2016). Even if most networks in these arenas are sex and gender heterogenous, some research documents inequalities in the probability of a tie by these attributes (Utz & Jankowski, 2016). In particular, both men and women are more likely to follow men on Twitter than expected by chance (Messias et al., 2017). Our work contributes to this literature by measuring how sex is associated with following behavior on Twitter. We use administrative information on sex, avoiding inferences based on name or profile pictures (Santamaría & Mihaljević, 2018). However, a drawback of our data is that it assumes a binary gender categorization and is only accurate for cisgender users.

Age

Focusing on age, homophily for this attribute is generally high for offline ties. However, there is variation by type of tie, namely, kin ties are heterogenous regarding age (Marsden, 1988; McPherson et al., 2001). Online relationships tend to involve substantial homophily for this demographic, especially in Social Media platforms that most closely mimic offline ties, such as Myspace or Facebook, but also for anonymous spaces like virtual worlds (Kang & Chung, 2017; Liao et al., 2014; Mazur & Richards, 2011; Thelwall, 2009; Ugander et al., 2011; Utz & Jankowski, 2016). Interestingly, Ugander et al. (2011) find a strong decrease in the age homogeneity of ties in Facebook as age increases, similarly as for offline ties (Marsden, 1988; McPherson et al., 2001). Only two studies explore age homophily on Twitter, pointing to its importance for ties in this platform (Liao et al., 2014; Zamal et al., 2012). However, they are focused on the task of inferring age from neighbors and use small samples and birthday wishes as a measure of age, leading to a strong bias for younger and 18-years old users. In our work, we draw a complete picture of how age is associated with follower relationships on Twitter.

Location

The introduction of communication technologies and in particular the widespread usage of the internet have generated a myriad of new possibilities for the creation and maintenance of social ties. This has led some scholars to argue that geographical constraints now play a secondary role for social relationships, a claim part of the "death of distance" hypothesis (Cairncross, 1997; Rainie & Wellman, 2012). However, research through the last 20 years still finds a strong impact of location on different types of social interaction, such as mobile phone calls and messages, and on social networks in general (Mesch et al., 2012; Mok et al., 2010; Onnela et al., 2011; Phithakkitnukoon et al., 2012). An interesting question is how much geography influences relationships in online spaces where it is theoretically possible to establish connections with anyone, regardless of preexisting social ties. The evidence in this regard is clear: shorter distances or spatial co-occurrence are strongly associated with having a tie in a variety of platforms, such as LiveJournal (Liben-Nowell et al., 2005), Facebook (Backstrom et al., 2010; Bailey et al., 2018; Spiro et al., 2016), Flickr (Crandall et al., 2010), Twitter (Grabowicz et al., 2014; Stephens & Poorthuis, 2015; Takhteyev et al., 2012) and even virtual worlds (Huang et al., 2013). Furthermore, the probability of a tie tends to decrease linearly with distance in a log-log scale, so that users living at a short distance (such as less than 10 km or within the same ZIP code) are orders of magnitude more likely to be connected than users at a distance of 1000 km or more (Backstrom et al., 2010; Grabowicz et al., 2014; Huang et al., 2013; Liben-Nowell et al., 2005). Still, across different countries and platforms, the probability of a tie at distances above 1000 km stays roughly similar.

Another important insight from this literature is that distance alone cannot explain the impact of geography on ties because users are not evenly distributed in space. In particular, the number of users living closer to each other than to another user has been found to predict ties better than distance for Facebook and LiveJournal (Backstrom et al., 2010; Liben-Nowell et al., 2005), similarly to how population density in between two locations is associated with commuting, migration, phone calls, and commodity flows

(Simini et al., 2012). In other words, what impacts the magnitude of different types of relationships between two users or locations is not so much the distance between them, but the amount of other opportunities at a similar or lesser distance. This implies that people living in low density areas are more likely to establish ties, work or migrate at larger distances. In addition, other factors, such as state lines, country borders, language, or number of air flights, also impact the likelihood of a tie on social media (Bailey et al., 2018; Kulshrestha et al., 2012; Takhteyev et al., 2012).

The case of relationships on Twitter and their connection to users' location is particularly interesting, due to the informational component of this platform. Follower relationships on Twitter are based both on information consumption and on pre-existing ties, and researchers have characterized the platform as both a social network and as an information network (Kwak et al., 2010; Myers et al., 2014). Hence, the norms and functioning of the platform should imply a lower influence of geography than for other Social Media platforms such as Facebook. Although some evidence points in this direction (Grabowicz et al., 2014), several studies find a strong relationship between physical location and different types of interactions on Twitter. On the platform, networks of mentions between locations map well to existing sociocultural and political regions when clustered with community detection methods (Arthur & Williams, 2019; Hedayatifar et al., 2019). Follower relationships are tied to physical distance and offline connections: Takhteyev et al. (2012) find 39% of ties to fall within regional clusters of 100 km or less, and that number of flights, language, and country borders between locations are good predictors of Twitter ties. Users at shorter distances are also more likely to reciprocate ties and to be embedded in smaller and denser networks (Quercia et al., 2012; Stephens & Poorthuis, 2015). Still, long distance and transnational ties are also common (De Choudhury, 2011; Kulshrestha et al., 2012; Takhteyev et al., 2012).

Most of this literature uses small samples of ego networks; only Grabowicz et al. (2014) take into account the distribution of distances among random dyads of Twitter users by calculating probabilities of a tie. In addition, the impact of the population density

between users, discussed in the previous paragraph, has not been measured on Twitter. In our study, we calculate the probability of following someone on Twitter as distance grows, and use a measure that takes into account the population density in between two users. We also compare the importance of this measure with state lines in multivariate analysis. Our administrative data includes detailed residence information for our sample of Twitter users, an advantage over previous research which relied on user defined location or geolocated tweets, leading to noisy, sparse, and biased location data.

Missing: literature on reciprocated ties on twitter

Data and Methods

Data Description

Our main dataset is a panel consisting of about 1.6 million twitter users, who are paired to U.S. voter files compiled by TargetSmart (Grinberg et al., 2019b; Hughes et al., 2021; Shugars et al., 2021). The panel is built from a 10% sample of Twitter collected between January 2014 and March 2017, signifying a near exhaustive collection of profiles that were active during that period. The Twitter profiles of these accounts with identifiable names and US locations were matched to public voter records accumulated by the vendor TargetSmart in October 2017. Individuals in the voter file were matched to Twitter users when a single user matched the name and city of the individual. Voter file records were also matched to users with unique state and name combinations. A detailed description of the matching procedure can be found in Hughes et al. (2021). The linkage between public voter record and the Twitter profile allows us to obtain more reliable data of Twitter users, including location, age, sex, race/ethnicity, and party registration than if inferring them from Twitter profiles. In addition, the matching procedure ensures that

members of our sample on Twitter correspond to real people offline, avoiding the presence of bots and organizations, and provides a sampling frame for our study, that is, US registered voters on Twitter (Tufekci, 2014). Our panel has been found to notably overrepresent white users and to include slightly more females than a sample representative of US voters on Twitter collected by Pew Research Center (Hughes et al., 2021). Hispanics and Asians are underrepresented, while African-Americans are correctly represented. We also miss Twitter users below 18 years old when the panel was built in 2017.

Between September 2020 and January 2021, we collected the lists of users followed by each member of this panel of Twitter users. We construct a network by incorporating directed edges when one panel member follows another. We eliminate isolated nodes without any incoming or outcoming nodes, ensuring that all nodes in the follower network have at least one connection with other nodes and therefore restricting our analysis to users with some degree of following activity on Twitter. The resulting network comprises 1,146,313 million nodes and about 20 million connections, and represents a subgraph of the whole Twitter follower network. In particular, we use it as a sample of the network of relationships among US registered voters on Twitter. We also build a network of reciprocal ties from there, keeping connections between nodes that shared a bidirectional follower tie and dropping nodes with no edges in this reciprocal network. This undirected network has 799,808 nodes and 4,349,484 edges.

In addition to the previously described panel dataset, we also use a dataset of survey respondents who provided their Twitter handle. This data comes from the Covid States project[1], an online survey with regular waves roughly every two months since April 2020 yielding a total of 27 waves. Each wave of the survey includes a sample of more than 20,000 respondents per wave, recruited through PureSpectrum, a professional survey company. As part of this survey, respondents who use Twitter are invited to provide their Twitter handle. We run a series of steps to authenticate the handles that we use, removing

[1] https://www.covidstates.org/

the ones that consisted of common names and that corresponded to different survey respondents (Joseph et al., 2021). We also removed the respondents providing different handles in more than one survey wave. After this process, we end up with a dataset of 31,653 Twitter users from which we have self-reported survey data on gender, race/ethnicity, age, partisanship, voting preferences, education, and household income. In Appendix Section 2, we detail how we build a network from the follower relationships among these users, and apply an additional cleaning step, namely, removing users with a high number of followers. The analysis of this network provides a complementary view to the results from the panel network; we discuss the advantages and disadvantages of both samples in the Results and in Appendix Section 2.

Regarding the variables from the voter file, we have access to administrative data on age and sex. We add three years to the voter file age to match the period through which we collected the follower data. Our sex variable is a binary measure, self-reported during voter registration. Therefore, it is unable to describe gender beyond the gender binary and may not account for transgender individuals. The race/ethnicity variable provided in the voter file is self-reported for states affected by the Voting Rights Act (VRA)[2], and is inferred by TargetSmart for the other states. In addition, the voter file provides partisanship information as party registration and as the inferred probability of supporting the democratic party. We bin this probability as recommended by TargetSmart into republican (0-0.35), independent (0.35-0.65), and democrat (0.65-1). Regarding party registration, we consider as independents individuals with party registration listed as either “independent”, “no party”, or “unaffiliated” in the voter file records. Party registration is included in 31 of the 50 US states, however, there is substantial variation in how it is registered by state (Ansolabehere & Hersh, 2012; Hughes et al., 2021). In contrast, the inferred measure of partisanship provides coverage across all of the US and a more consistent measurement across states. It is largely based on party registration for the states that include it, therefore, we use it as our main partisanship measure.

[2] The list of these states are: Alabama, Georgia, Louisiana, Mississippi, South Carolina, Virginia, Alaska, Arizona, Texas, and North Carolina

The voter file data on race/ethnicity and partisanship has been validated in previous work, at the county level for partisanship and against a different voter file based inference for race/ethnicity, with overall good results (Shugars et al., 2021). In addition, individual level validation against survey data was realized for a small sample of 182 panel members, finding good matching for gender, but more significant mismatch for race/ethnicity minorities and partisanship (Hughes et al., 2021). However, this small sample size prevented extracting reliable conclusions. Here, we use our sample of Twitter users from the Covid States survey data for further validation, using the set of 799 Twitter handles we have in both the panel and the survey data. The comparison of the voter file to survey information for gender, age, race/ethnicity, and partisanship is included in Appendix Section 1. We find excellent matching for gender and age, good matching for partisanship and whites, decent matching for african americans, and significant discrepancies for hispanics and whites. Through this comparison, we also observe how party registration matches better the survey responses than inferred partisanship, therefore, we provide some of our results for both measures as a robustness check. The matching for race/ethnicity is not higher in states under VRA preclearance, however, the number of users from these states in both survey and panel datasets is small. Hence, we calculate some of our race/ethnicity results on homophily for users in these states as a comparison point.

In addition to demographic and partisanship information, we use a set of geographical variables derived from the census tract and the latitude and longitude where each panel member resides. We expect this data to be of high accuracy, except for the small fraction of users who changed residence between 2017, the year of collection of the voter records, and 2020, the year of collection of the follower data. Using the latitude and longitude information, we calculate the distance in km between all the pairs of panel members with a tie between them in our network. We also calculate it for a random set of about 110M dyads, that we use to approximate the probabilities of a tie at different distance bins (see *Measuring Homophily* subsection). We exclude panel members residing in Hawaii and

Alaska from this calculation. In addition, we use another measure of geographical separation between users, inspired by the work of Simini et al. (2012). In particular, we compute, for a given pair of panel members *a* and *b*, the total population residing in a circle around *a* of radius the distance between *a* and *b*, and call this value the *population in radius* between two users. This measure provides an estimation of the opportunities that *a* has to build a tie to someone else residing closer or as close to them as *b*. In addition, it deals with one of the major issues when using distance to measure geographical separation between two users, namely, the effect of residing in more or less dense areas. Someone residing in a dense metropolitan area has a lot of opportunities for contact at short distances, while someone residing in a rural area has very few opportunities at short distances. In particular, the number of alternative options to build a tie on Twitter may be more strongly associated with having a following relationship than the distance between two users, and our *population in radius* measure approximates the former. To calculate it, we geolocate the census tract of panel member *a*, and sum the population of all the census tracts intersecting with a radius *r* around the centroid of *a*'s census tract, where *r* is the distance between *a* and *b*. We run this calculation for all pairs of panel members with a tie between them and for a random subsample of about 20M dyads, also excluding users residing in Hawaii and Alaska.

Finally, we are also interested in how the rural or urban status of the residential area of users is associated with their following patterns. We use the Rural-Urban Commuting area (RUCA) codes of the census tract of users as a measure of the rural status of their tract. This categorization of census tract takes into account commuting flows in addition to population density and urbanization[3]. For example, low density census tracts with large commuting flows to a metropolitan area are categorized as metropolitan areas. We group codes 1-3 as Metropolitan areas, codes 4-6 as Micropolitan areas, and codes 7-10 as Small Town/Rural areas. In addition to the RUCA codes, we also use the population density of the users' census tract, binned in deciles calculated from the users in the network of following relationships. While the RUCA codes are effective at detecting

[3] https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation/

rural areas isolated from metropolitan areas, it categorizes most census tracts as metropolitan. Using the census tract population density allows differentiating highly dense metropolitan areas, such as city centers, from other lower density metropolitan areas.

Table 1 provides the number of users of the follower and reciprocal network in each category, together with average degrees.

	Follower Network			**Reciprocal Network**	
	N	**Avg. in-degree**	**Avg. out-degree**	**N**	**Avg. degree**
Party score: Democrat	608,958	21.4	20.0	431,781	56.2
Party score: Independent	159,038	12.4	13.7	107,073	36.6
Party score: Republican	378,317	13.4	15.1	260,954	39.2
Party reg: Democrat	269,379	25.3	22.6	195,546	63.7
Party reg: Independent	168,805	19.6	17.2	116,813	51.0
Party reg: Republican	156,857	13.7	13.7	106,760	40.6
Male	490,485	19.3	18.2	419,560	50.1
Female	605,464	16.3	17.3	345,765	46.5
African-American	81,883	14	13.8	54,252	39.9
Asian	21,954	19.2	18.3	15,266	51.7
Caucasian	961,639	18.1	18.1	679,325	49.2
Hispanic	46,516	9.2	11.0	27,317	31.8
VRA state: African-American	34,793	5.2	5.4	23,139	15
VRA state: T.Asian	3,034	5.8	5.9	2,087	16.3
VRA state: Caucasian	193,588	6.6	6.5	133,317	18.1

VRA state: Hispanic	12,925	3.9	4.2	7,787	12.6
Age: 18-29	272,666	8.0	9.1	198,094	21.9
Age: 30-49	525,080	19.8	19.2	369,588	53.3
Age: 50-64	237,515	21.7	22.2	161,786	61.9
Age: 65+	82,932	21.9	21.9	51,860	67.0
RUCA: Small Town/Rural	56,276	10.4	13.1	38,648	32.3
RUCA: Micropolitan	75,368	9.0	12.0	50,803	29.1
RUCA: Metropolitan	1,013,939	18.5	18.2	709,888	50.2

Table 1. Number of respondents and average degree by inferred partisanship, party registration, sex, inferred race/ethnicity, race/ethnicity for the subnetwork of respondents in VRA states, age, and RUCA category of census tract.

Measuring homophily

As detailed in the literature review, there are various processes behind the observed amount of segregation in a network, that is, the percentage of ties that are directed towards same group members. While we do not aim to disentangle "pure" choice homophily from the complex set of interrelated processes that lead to a segregated network, we use descriptive measures that allow us to sort out some of these processes. We enumerate below the different ways we quantify homophily in our network. We use the following notation: *A, B, C, …* stand for each of the attributes we work with, and a_1, a_2, ..., a_n, b_1, b_2, ..., $b_{m, \ldots}$ for the different groups of each attribute. We use attribute A as an example to explain each measure, but the same reasoning applies to the other attributes. In addition, we term as *followee* a user that is followed by another user. We describe our different measures below for the directed network of following relationships, but apply very similar measures to the undirected network of reciprocal connections as well.

Before describing our measures, we discuss some of the factors leading to a segregated network that we address in our work. A first reason why a network may be segregated for a given attribute is simply the unequal sizes of the groups within this attribute. In the stylized case of an attribute with two groups, the larger group will generally send and receive more ties, in aggregate, while the smaller groups will send and receive less ties. This may cause members of the smaller group to have a higher average fraction of ties to the larger group than to other members of their group, even if members of the smaller group have a preference for same group ties. Similarly, members of the larger group may have a higher average fraction of intragroup than outgroup ties even without the presence of homophily preferences. We use the term baseline homophily for this segregation generated by the sizes of groups, also called opportunity structures. To deal with the effect of uneven group sizes, we calculate probabilities of a tie at the dyad level and fit logistic regressions with having a tie as outcome. These methods are described in points 2 and 3 below. In general, we say that a group displays inbreeding homophily if the probability of an intra-group tie is higher than the probability of a non-homophilous tie.

Another factor that may lead to segregation and also inbreeding homophily in a directed network is the differential propensity of groups to send and receive ties. In other words, a group may disproportionately attract ties, causing this group to display inbreeding homophily without any preference for same group ties, that is, just because of the tendency to send ties to nodes with high in-degree. Likewise, groups with low average in-degree may appear to not have choice homophily when looking at their inbreeding homophily levels, because their members are likely to send ties to high in-degree nodes that also happen to be out-group members. Comparable effects can also appear due to the different propensity to send ties of each group. A similar effect can exist as well for undirected networks, in this case due to the differential tendency of groups to establish bidirectional ties. We deal with these effects, often called *activity* and *popularity* (for directed networks) or *sociality* (for undirected networks) in the literature (Goodreau et al., 2009; Wimmer & Lewis, 2010), with univariate logistic regressions, described in point 2 below. We also deal with differential activity by normalizing the probabilities of a tie by

the sum of the probabilities of the group sending the tie, as explained in point 3. Finally, it is also possible that inbreeding homophily or heterophily by one attribute drives inbreeding homophily or heterophily by another attribute. For example, if two attributes are correlated, like, in our case, partisanship and race/ethnicity, a preference for same group ties for one attribute will generate inbreeding homophily for the other attribute even without choice homophily for this second attribute. We take into account this possibility, for the attributes that we have data on, with multivariate logistic regressions, described in point 4.

1. We measure the extent that the network is *segregated*, by calculating, for all users, what percentage of their followees are in each group a_1, ..., a_n. We plot the distribution of these values for each group and compute their averages, allowing a close look at the percentage of followees of users of group a_i that are members of each group. To reduce noise, we only include in the results the users who follow at least 10 other users in our panel, resulting in 400,000 users. With the goal of correcting for the lack of representativity of our sample regarding sex and race, we weight the calculation using the comparison to representative survey samples of Twitter users in Hughes et al. (2021). For distance, population in radius, and age, we group them in four bins and look at how the percentages of followees falling in each bin are distributed.
2. We fit simple logistic regressions with having a tie in the network as outcome (y) and the following variable as predictor: given attribute A, a categorical variable H taking value 0 when the group of the sender and the receiver of the tie are different, and the value of the matching group when they are equal. The resulting coefficients for H are measures of *inbreeding homophily*. We also fit additional models, with variables for the group of the sender of the tie (a_s) and for the group of the receiver of the tie (a_r), to control for the differential *activity* and *popularity* of each group. The equation would take this form:

$$y \sim \alpha a_s + \beta a_r + \gamma H,$$

where γ stands for a set of as many parameters as the number of groups in *A*. This model can be quickly fitted after calculating the number of ties and the number of dyads without a tie for each combination of the groups (that is, with the data in a frequency format). The resulting parameters α, β, and γ approximate relative risks after exponentiation, given that ties are orders of magnitude more likely than no-ties (King & Zeng, 2002; Kossinets & Watts, 2009). The estimates exp(γ) corresponding to each category $a_1, a_2, ..., a_n$ of *H* are interpreted as the number of times an homophilous tie for the category is more likely than a non-homophilous tie. Therefore, we interpret that there is inbreeding homophily for a group when the corresponding exponentiated parameter is larger than 1. The α and β parameters model how the fact that the sender and receiver of the tie are members of each group associates with the probability of a tie, regardless of homophily, encoded in γ. Hence, when a_s and/or a_r are included in the model, the exp(γ) estimates are measures of inbreeding homophily when controlling for the differential propensity of each group to send and/or receive ties. For reciprocal ties, we control for the sociality of each group by including variables a_i, one for each category in *A*, with the number of dyad members that are of the category (so with values ranging from 0 to 2). We model age homophily as the absolute difference in age between members of the dyad. We include the age of the sender and the age of the receiver of the tie as controls for differential activity and popularity by age, and the sum of ages of the two members of the dyad as a control for the differential sociality by age in the case of reciprocal ties.

3. We calculate the probability that a user in group a_i follows another user in group a_j, by counting the number of ties between users of groups a_i and a_j and dividing it by the number of dyads (that is, possible ties) of users of groups a_i and a_j. We display these *probabilities of a tie* in heatmaps that allow visualizing the likelihood that there is a tie between two random members of each pair of groups. We also row-normalize the previous heatmaps, by dividing the probability of a tie from group a_i to group a_j by the sum of probabilities of a tie from group a_i to groups $a_1, ..., a_n$,. By row-normalizing, we control for the differential activity of

each group and visualize, from the point of view of one group, the relative probability of a tie with each other group. We display the row-normalized heatmaps in the main text, and the ones with raw probabilities in the appendix. In the case of the distance variables, we calculate the probability of a tie for a large set of small-sized bins. Given the huge amount of dyads in the network, we approximate the number of dyads in each distance bin from a random sample of dyads.

4. Finally, we fit logistic regressions adding multiple variables in the same model in a stepwise fashion. In this manner, we are able to detect when the association between one variable and the probability of a tie is driven by another variable. As detailed before, we can approximate the exponentiated parameters from the model to relative risks because of the minimal fraction of dyads with a tie in the graph. This allows interpreting them as the factor by which the probability of a tie is multiplied when increasing a continuous variable by one unit or when a categorical variable takes a given value instead of the reference value. When including each of the categorical variables in the model, we incorporate the activity and sociality controls as in 2. We also add in these models the population in radius variable, which need to be calculated for each observation (that is, dyad) we fit the model to. Thus, it would be computationally intractable to use all dyads in the network, and we apply the case-control methodology by randomly sampling as many dyads without a tie as the number of edges (King & Zeng, 2001, 2002; Smith et al., 2014). For similar reasons, using ERGM's or QAP with a dyadic variable like that is nearly impossible. Recent advances in the efficiency of ERGM estimation (Stivala et al., 2020) may allow fitting an ERGM to our network, but without measures of geographical distance (because they cannot be derived from nodal attributes). These are the variables with the largest explanatory power in our results, and they have previously been studied only in isolation with respect to other predictors of following behavior on Twitter. Hence, we prioritized including them against using the (otherwise) optimal statistical framework. In addition, using a logistic regression with our network is not likely

to incur in heavily biased estimates: its very low density should yield to few violations on the assumption of independence of observations.

Results

First, we analyze the bivariate homophily patterns for the political and demographic variables: inferred partisanship, party registration, sex, race/ethnicity, and age. Then, we focus on the different location-based variables: distance and population in radius between two users, RUCA code, and population density of the census tract. Finally, we turn our attention to the multivariate analysis.

Political and Demographic homophily

Table 2 shows the inbreeding homophily measure, as exponentiated coefficients of the logistic regressions discussed in point 2 of the *Measuring Homophily* subsection. Hence, each coefficient can be interpreted as the number of times an homophilous tie for a given category is more likely than a non homophilous tie for this variable. Overall, we see some degree of homophily for all the variables considered, but especially for race/ethnicity, for democrats, and for age. Figures 1 and 2 include the heatmaps with row-normalized probabilities of a tie, allowing a closer look at the following patterns among the categories of each variable. In Appendix Figure 1 we display the row-normalized heatmaps for race in VRA states and for party registration, while Appendix Figures 2 and 3 include the heatmaps with probabilities of a tie without normalization. We show in Figures 3 to 6 how the followees of users in our sample are distributed by different variables (point 1 of the *Measuring Homophily* subsection). Finally, we analyze the homophily patterns in the following behavior of the Covid States users in section 2 of the Appendix. The variables from this data are self-reported, and hence generally of higher

quality than the voter file data, especially regarding race and partisanship. However, the sample is built from handles provided by a small percentage of the survey respondents who use twitter, and we identified some as not likely to correspond to respondents. In addition, it is a much smaller sample, yielding a network with potentially more noise. Overall, it is a network less representative of the following ties among US Twitter users than the panel network. Given these strengths and weaknesses, we use it to check for the robustness of some of the findings from the panel data, and also explore the homophily patterns by SES status, only available in the survey data.

Focusing first on partisanship, we find that democrats display substantially more inbreeding homophily than republicans for both inferred partisanship and party registration. In addition, republicans are more likely to follow democrats than democrats to follow republicans (Figure 1 and Appendix Figure 1). However, with the controls for popularity and activity, the homophily levels of democrats and republicans become closer for inferred partisanship and even lower for registered democrats compared to registered republicans. This is in line with the higher average number of followers and followees of democrats (Table 1): Democrats tend to have more followers and also to follow more users, leading to more ties among them regardless of any preference for same partisanship relationships. This group is also, on average, the majority partisanship in the followee sets of users of all three parties (Figure 3), and is substantially segregated: 87% of democrats have at least 50% of the people they follow that are also democrats. There is substantial variation in the composition of followee sets, that is, they are politically very homogenous for some users, while very diverse for others.

	Directed ties				**Reciprocal ties**	
	Homophily	**Homophily + Activity**	**Homophily + Popularity**	**Homophily + Activity + Popularity**	**Homophily**	**Homophily + Sociality**

Party score: Democrat	2.17	2.44	1.93	2.03	1.84	2.08
Party score: Independent	1.01	0.93	1.04	0.96	1.17	1.00
Party score: Republican	1.41	1.31	1.61	1.60	1.68	1.60
Party reg: Democrat	2.14	2.12	1.95	1.75	1.90	1.82
Party reg: Independent	1.20	1.09	1.01	0.83	1.21	0.87
Party reg: Republican	1.20	1.38	1.78	2.65	1.64	2.66
Same Sex	1.32	1.34	1.35	1.35	1.34	1.34
African-American	3.59	3.67	3.58	4.21	4.17	5.19
Asian	6.20	4.35	4.13	3.21	7.22	4.20
Caucasian	1.62	1.64	1.59	1.39	1.64	1.51
Hispanic	2.27	2.64	3.21	4.32	3.63	4.76
VRA state: African-American	4.31	4.47	4.75	6.91	4.71	8.38
VRA state: T.Asian	12.18	8.63	8.94	8.06	15.42	9.34
VRA state: Caucasian	1.93	1.96	1.84	1.32	1.96	1.36
VRA state: Hispanic	3.43	3.49	3.86	5.16	5.51	6.28
Age	0.97	0.96	0.96	0.95	0.96	0.94

Table 2. Exponentiated logistic regression coefficients of group homophily with respect to non-homophilous ties. The first column shows coefficient for the model with only the homophily variable, the second incorporates the group of the sender of the tie, the third the group of the receiver of the tie, and the fourth both sender and receiver. The fifth and sixth columns are for models on reciprocal ties, with only the homophily variable in column five and with the different sociality controls in column 6. In the case of age, the coefficients are for the absolute difference in age between sender and receiver of the tie, incorporating controls for age of sender (activity), age of receiver (popularity), and the sum of ages of the two nodes (sociality)

For reciprocal ties, the differences between republicans and democrats vanish: republicans show somewhat more homophily than for directed ties, while democrats less (Table 2), and both are similarly likely to have a reciprocal tie with members of the opposing party (Figure 3, Appendix Figure 1). Hence, the following patterns of

democrats and republicans on Twitter are different, maybe stemming from how dominant democrats are on the platform. An interpretation fitting the patterns we observe is that democrats tend to be influential, gathering a lot of non-reciprocal followers, while reciprocal relationships tend to be established among users with a similar status that are less likely to be democrats. Democrats and republicans behave similarly for both inferred partisanship and party registration, except when controlling for sociality and activity or by sociality in the case of reciprocal ties. Partisanship homophily is generally lower in the Covid States data (Appendix Section 2), especially for democrats, which seems to imply that our main measure of partisanship, inferred from the voter file data, is not leading to an underestimation of homophily.

Finally, to check if partisanship homophily is higher for Twitter users that tweet about politics, we look at the following patterns of users by our two measures of partisanship broken down by the number of tweets about the 2020 election they posted (Appendix Figures 11 and 12, see Appendix Section 3 for an explanation of how we made these heatmaps). For both measures, homophily is significantly higher for democrats who tweeted about the elections. A democrat-democrat tie using inferred partisanship is 2.5 times more likely than a democrat-republican tie for users without tweets about the elections, while it is 5.2 times more likely for users with more than 10 election tweets. These factors are 2.0 and 3.6, respectively, for party registration. In contrast, republican homophily only increases for users who posted more than 50 tweets about the elections, and to lower levels than for democrats. Furthermore, for both measures, the probability of a republican-republican tie is lower for users with 1 to 50 election tweets compared to users with 0 election tweets and the probability of a republican-democrat tie is higher for users with 1 to 50 election tweets.

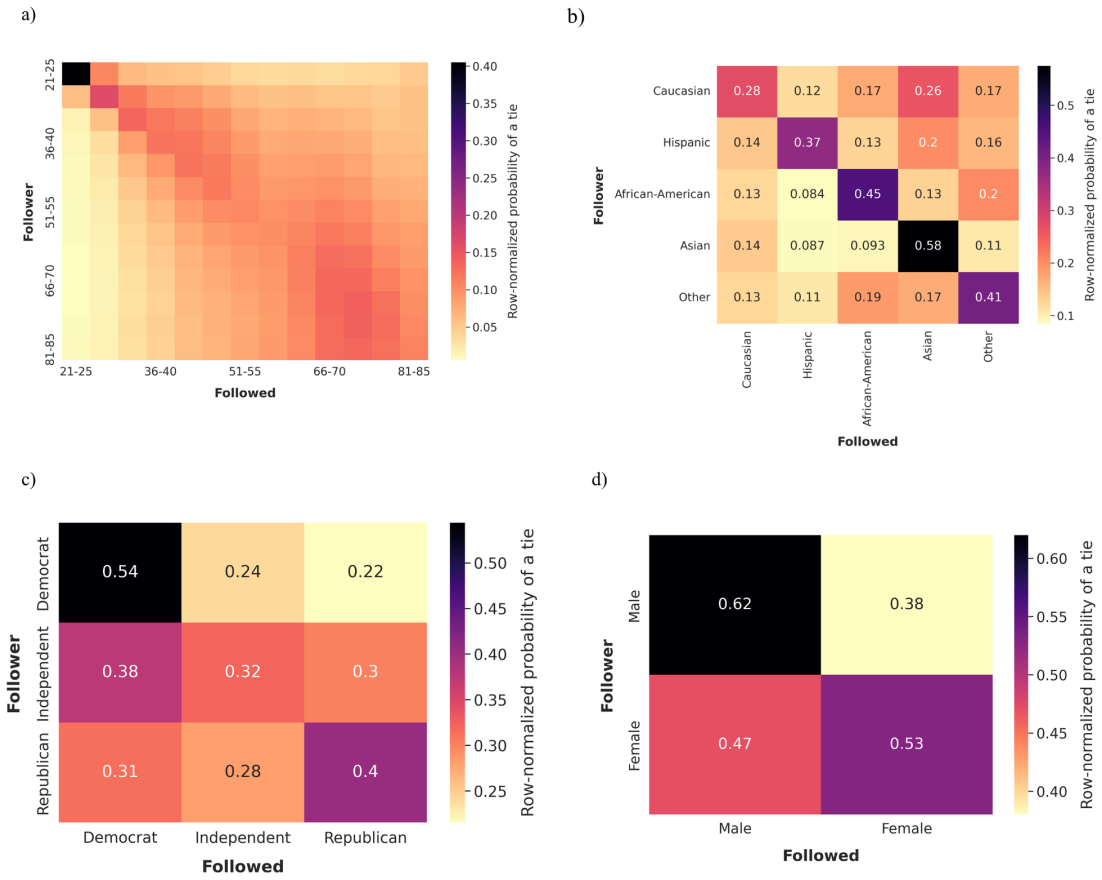

Figure 1. Heatmaps with row-normalized probabilities of a tie by age, race/ethnicity, inferred partisanship, and sex, in panels a), b), c), and d), respectively. Color scales change for each panel.

We find moderate homophily for sex in Table 2, with similar values when controlling for differential likelihood to send and receive ties and for reciprocal ties. However, the regression framework with these controls for a bivariate attribute does not allow estimating distinct homophily parameters for male and female, because there are not enough degrees of freedom. When looking at their following patterns separately, we find an asymmetric pattern: there is substantial homophily for male users, but not for females (Figure 1), and females are more likely to follow males than vice versa (Appendix Figure 2). As reported in previous literature (cite), men generally have more followers on the

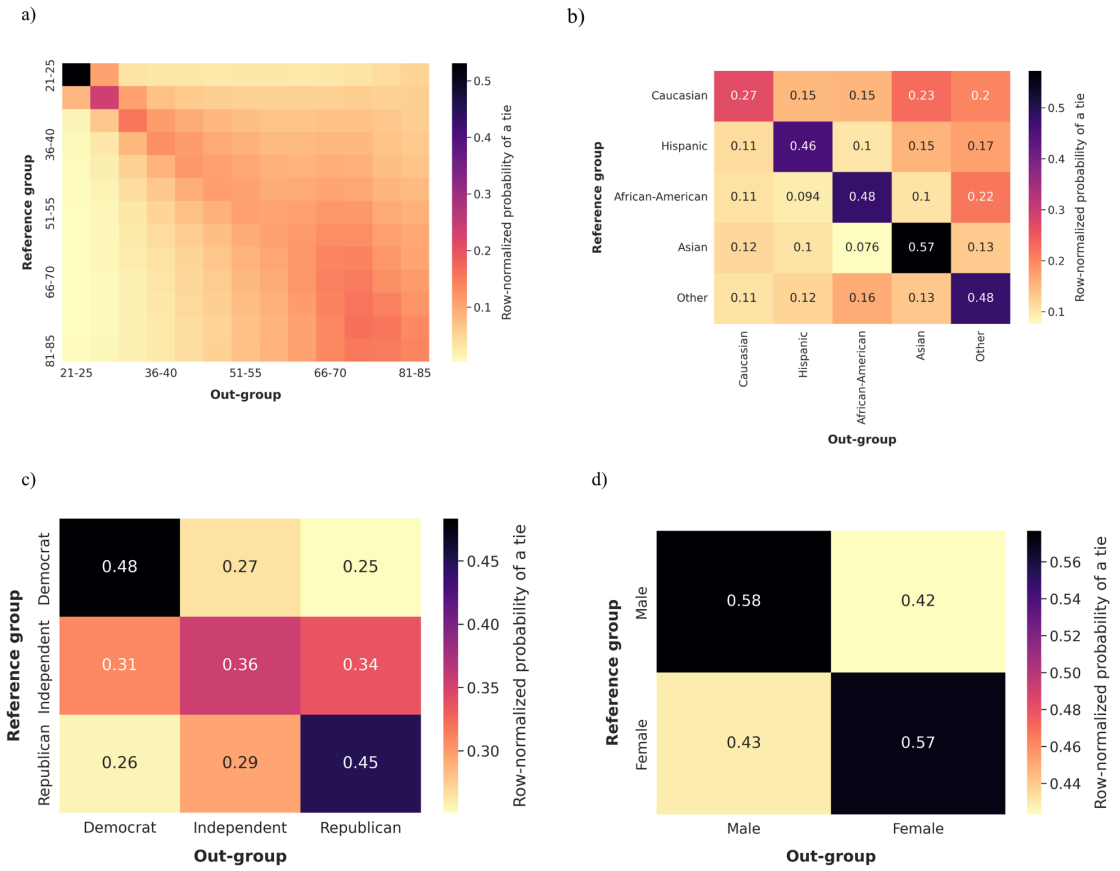

Figure 2. Heatmaps with row-normalized probabilities of a tie for reciprocated following relationships only, by age, race/ethnicity, inferred partisanship, and sex, in panels a), b), c), and d), respectively. Color scales change for each panel.

platform (Table 1) and both males and females are more likely to follow males. This leads to both groups having, on average, more males in the users they follow (Figure 4): 87% of males have a majority of males in their followees, while only 42% of females have a majority of females in their followees. These patterns change for reciprocal ties: both groups have very similar levels of homophily, and males have less than for directed ties (Figure 2). We find that males tend to dominate unreciprocated relationships between

individuals with different status on the platform, while females display homophily for reciprocal, same status ties.

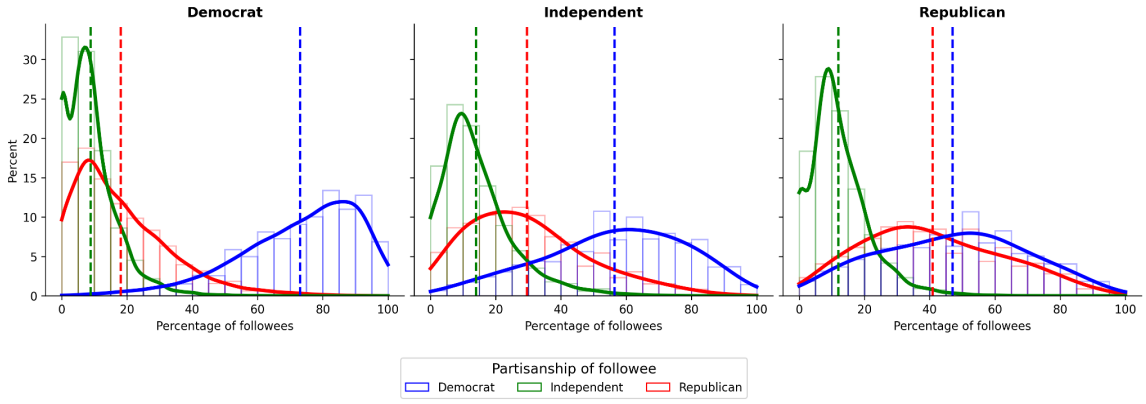

Figure 3. Distribution of followees by inferred partisanship. Vertical lines correspond to average values.

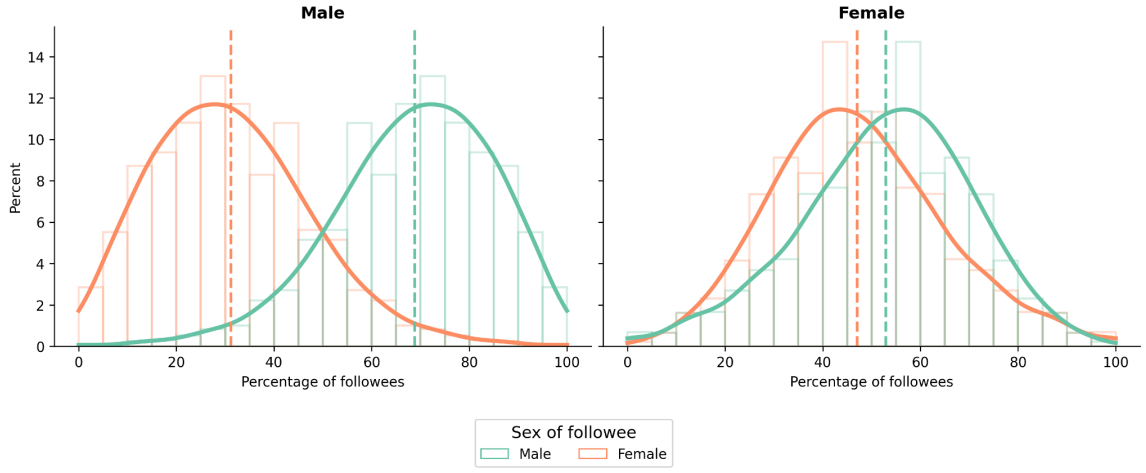

Figure 4. Distribution of followees by sex, weighted to representative samples of Twitter users. Vertical lines correspond to average values.

Turning now our attention to race/ethnicity, we find high levels of homophily for minorities, and moderate homophily for whites (Table 2). When taking into account the

higher likelihood of whites and asians to send and receive ties (Table 1), we find decreased homophily for these groups and higher for african americans and hispanics. Homophily is higher in VRA preclearance states, where the voter file race/ethnicity measures should be more accurate. However, these are also states with a distinctive history of racial discrimination, where it is possible that Twitter users tend to also be more homophilous in their following relationships.

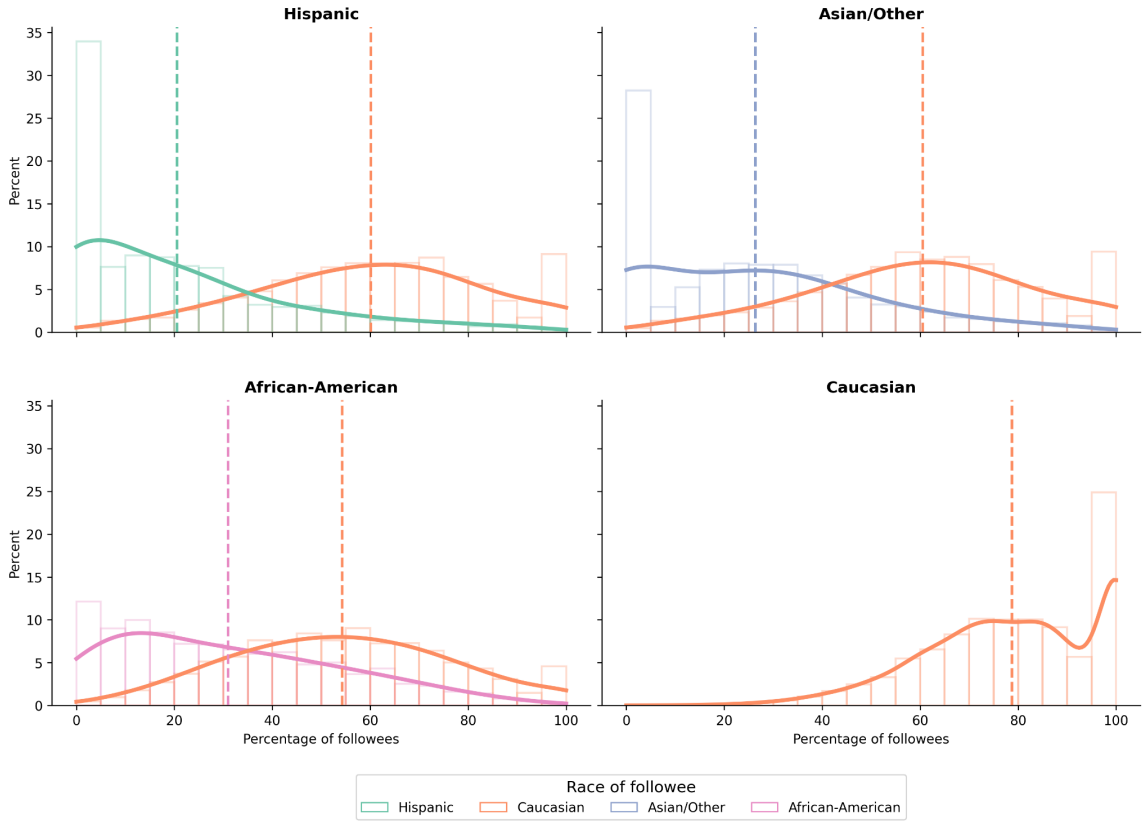

Figure 5. Distribution of followees by inferred race/ethnicity , weighted by representative samples of US voters on Twitter. Vertical lines correspond to average values. Only groups composing an average of at least 10% of the followee sets of focal group members are displayed.

Whites dominate the followee sets of all groups (Figure 5), due to the majority of US Twitter users being white (Table 1), though the dispersion in the distribution of percentage of followed users of the same race is high. While an important portion of the members of racial and ethnic minorities follow few users of their race/ethnicity, a relatively significant amount have followee sets where their race/ethnicity is a majority or at least whites are not the majority. This is especially true for African-Americans: 21% of users of this racial subgroup have 10% or less of African-Americans in the users they follow, while 22% have 50% or more. All groups are more homophilous for reciprocal ties (Table 2). Race/ethnicity homophily for the Covid States users is generally lower than for panel members (Appendix Section 2). Hence, our imperfect measure of race/ethnicity may be leading to an overestimation of homophily for minorities. Assuming that the inferred measure of race/ethnicity from the voter file is based on census tract or bloc composition, minorities in geographically segregated areas may be more likely to be correctly classified in this measure, while minorities in low segregation areas are likely to be misclassified. If residential segregation correlates with segregation in Twitter following relationships, it is possible that we detect higher homophily on Twitter by race/ethnicity with TargetSmarts' inference than we would with a self-reported measure.

Regarding age, we observe strong homophily patterns. Assuming a linear relationship between difference in age and probability of a tie, and controlling for differential activity and popularity also assuming linearity (Table 2), a tie between two users of the same age is about 3 times more likely than between two users 20 years apart, and about 8 times more likely if the two users are 40 years apart. However, as Table 1, Appendix Figure 2 and Figure 1 show, homophily and following patterns vary significantly by age cohort. Younger users, especially users less than 30 years old, have a clearly lower number of followers and followees on average (Table 1, Appendix Figure 2), and display very high in-breeding homophily when taking into account this lower propensity of sending a tie (Figure 1). Users between 30 and 60 display progressively less homophily, and a significant propensity to follow users between 65 and 75 years old, an age cohort that

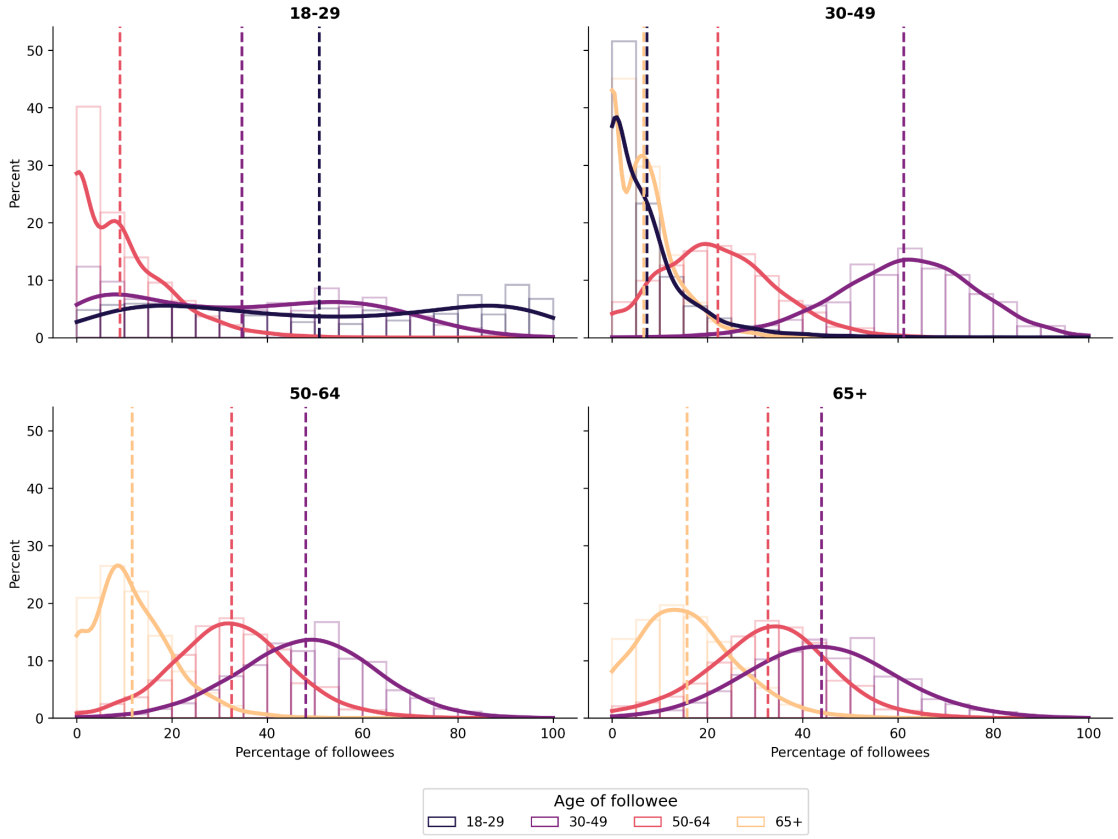

Figure 6. Distribution of followees by age. Vertical lines correspond to average values. Only groups composing an average of at least 6% of the followee sets of focal group members are displayed.

gets a disproportionate number of followers (see Appendix Figure 2) . There is significant age homophily for users older than 50, but it is less concentrated on users of a very similar age and more on following users that are 5 or 10 years older. These patterns are similar for reciprocal ties (Figure 2 and Appendix Figure 2), but with more homophily in general. While the older age cohorts dominate in terms of probabilities of a tie, the composition of the followee sets is dominated by users between 30-49 years old, for all age cohorts except 18-29 years old (Figure 6). The variation in the age composition of followed users for this age cohort is huge: 14% of 18-29 year olds have more than 90% of their followees in their age cohort, while a similar percentage, 10%, have less than 10% of followees in their age cohort.

Summing up the results from this subsection, we found high homophily in follower relationships for racial and ethnic minorities and age, lower but still significant homophily for partisanship, especially for democrats, moderate homophily for males, and no homophily for females. As an example of the lower homophily by partisanship compared to race/ethnicity and age, we find that the maximum gap in probability of a tie generated by our partisanship categories, a democrat-democrat tie compared to a democrat-republican tie, consists in multiplying this probability by 2.5. On the other hand, an african-american to african-american edge is 3.5 times more likely than an edge from an african-american to a white user, and this value is 4.1 for asians. Regarding age, the probability of a 21-29 to 21-29 years old tie is 3.6 times higher than a 21-29 to 50-64 years old tie, and the probability of a 30-49 to 30-49 tie is 4.6 times higher than a 30-49 to 21-29 tie. These patterns stay roughly similarly for reciprocal ties, which tend to be more homophilous in general, with some exceptions like democrats or males. Overall, homophily in the Covid States data (see Appendix Section 2) is lower for all attributes, but confirms the higher homophily found with the panel for race/ethnicity minorities and age compared to partisanship. In terms of segregation in the users followed, we find high segregation for democrats, but low for republicans. White users also have, on average, homogeneous sets of followees by race and ethnicity, and males are somewhat segregated on Twitter, as well as users between 30 and 49 years old.

Location homophily

We plot in Figure 7 how the probability of a following relationship decreases with distance and population in radius. Both variables are very strongly associated with having a tie on Twitter: the plots display heavy-tailed distributions and can be reasonably well approximated with straight lines in log-log scale. This is especially the case for population in radius and distance up to 1,000 km. The association between distance and probability of a tie vanishes for distances larger between 1,000 km 3,500 km and becomes positive for larger distances.

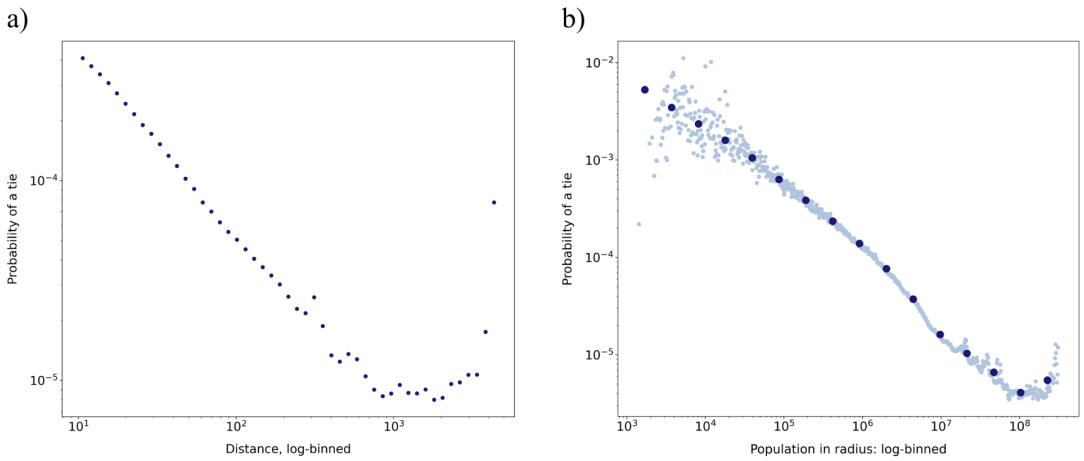

Figure 7. Probability of a following tie by distance (in km), in panel a), and by population in radius, in panel b). Both plots are in log-log scale.

This finding is consistent with research across different platforms and countries, where distance also ceases to be negatively associated with online relationships beyond 1000 km (Backstrom et al., 2010; Grabowicz et al., 2014; Liben-Nowell et al., 2005). The distance between California and the Northeast regions in the US is roughly 3,900 km, which may be behind the high probability of a tie at distances larger than 3,500 km. In contrast to distance, population in the radius is predictive of Twitter ties across almost its full range of values, pointing to a stronger association for this variable than distance. To confirm this finding, we run logistic regressions on having a tie, in a similar fashion as detailed in point 5 of the *Measuring homophily* section, first with each of the variables separately as the only predictors and after with both of them together. We include the variables after taking their base 10 logarithm. In the first regressions, the exponentiated coefficient is 0.35 for distance and 0.37 for population in radius. When including both variables in the regression, this value increases to 0.97 for distance and stays similar, at 0.38, for population in radius, clearly showing the larger explanatory power of population in radius over distance.

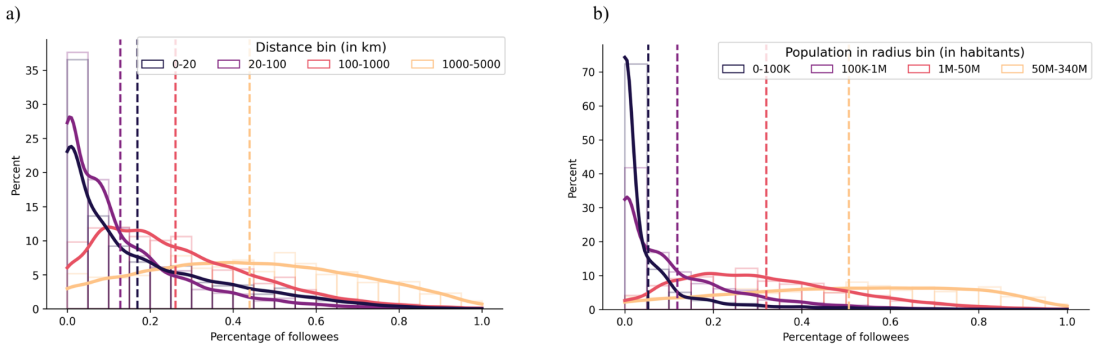

Figure 8. Distribution of followees by Distance and Population in radius, in four bins, in panels a) and b), respectively. Vertical lines correspond to average values.

Turning now our attention to the composition of networks (Figure 8), we observe a different picture: while the probability of a tie at low distances is high, the median tie spans 941 km and 77M population in radius, and users that reside far tend to dominate followee sets. Even if ties to close users are much more likely, most other users reside far away: we estimate that only 2% of pairs of users in our network are within 100 km. There is significant dispersion in the distributions of Figure 8, and a substantial portion of users have relatively local followee networks: 21% of panel members who follow at least 10 users have 50% or more of these users at a distance of 100 km or lower.

In relative terms, reciprocated ties among users residing closer are more likely than following ties (Appendix Figure 8). The population in radius value at which following ties start to be more likely than reciprocal ties, in proportion to the corresponding probability of a tie, is roughly 100,000. In particular, a reciprocated tie below this value is 191 times more likely than a tie above it, while this factor is reduced to 136 for following ties. In Appendix Section 4, we build an index of social connectedness between US counties based on the relative probability of a reciprocal following tie between users in each pair of counties, inspired by Facebook's Social Connectedness Index (Bailey et al., 2018). We compare the association between our Twitter connectedness Index and the distance between counties to this association for Facebook's index, and find a significant

association for Twitter, but lower than for Facebook (see Appendix Figure 13). Overall, these findings match the complex nature of Twitter documented in the literature (cites) as a social media platform with both informational and social components, and highlight how local environments, even if less important than for traditional Social Media like Facebook, may still play a significant role in its structure of relationships.

We also evaluate if homophilous ties by race/ethnicity and partisanship are more likely among users who reside closer, given the strong residential segregation in the US by these two variables and the high likelihood of a tie between users living close. We plot race/ethnicity and partisanship heatmaps broken down by different population in radius bins (Appendix Figures 6 and 7). We find very high homophily for nearby users across both whites and minorities, while following relationships through high population in radius values are less homophilous by race/ethnicity than the average tie. Appendix Figure 7 depicts a similar pattern for republicans, with small differences for democrats. Ties spanning large population in radius values tend to be directed towards democrats, regardless of the partisanship of the sender of the tie. In particular, a republican-democrat tie at 100 M or more population in radius is more likely than a republican-republican tie within the same range of population in radius. This points to differential patterns of partisanship homophily for republicans and democrats. We conclude that following ties to individuals residing closer, more likely to be offline relationships, are more homophilous by race/ethnicity and for republicans.

Regarding homophily by the type of residential area of the user, we observe high homophily by RUCA code, especially for users in tracts in small towns or rural areas (Figure 9c, Appendix Figure 4c), and very high homophily for users in the highest density decile (Figure 9a, Appendix Figure 4a). Panel members in census tracts in the lowest decile of population density, only 33% of which are small town/rural tracts by RUCA code, also display significant homophily. To check whether the homophily patterns by RUCA code are driven by the high tendency of users to follow nearby users, we plot Figure 9c broken down by population in radius in Appendix Figure 5.

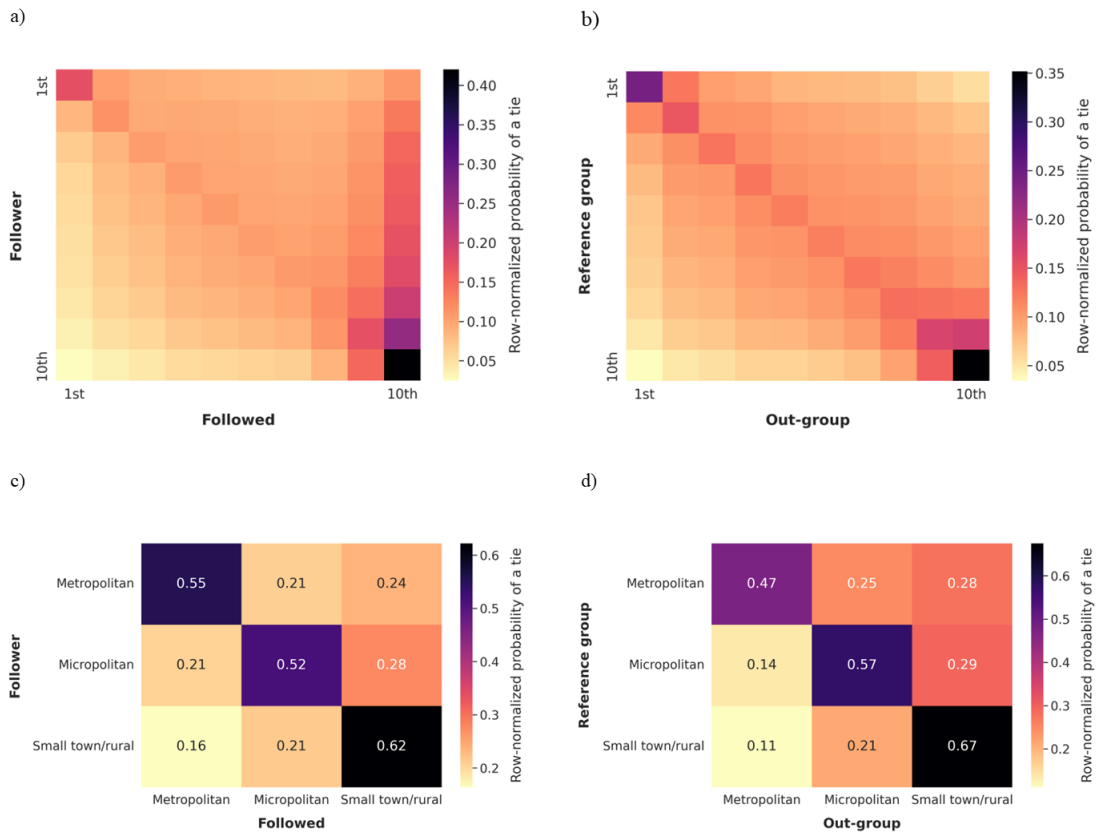

Figure 9. Heatmaps with row-normalized probabilities of a tie by population density of census tract in deciles and grouped RUCA category (panels a) and c), respectively), and with row-normalized probabilities of a reciprocated following tie for the same two variables (panels b) and c), respectively). Color scales change for each panel.

We find persisting homophily for panel members in Small town/rural areas up to ties at 10 M population in radius, roughly the population of 1 large US state or a few smaller states, and significant homophily for Metropolitan areas across all population in radius values. Long range following relationships are more likely when the receiver of the tie resides in Metropolitan areas than ties to same RUCA code areas, while users in Micropolitan census tracts are only homophilous for ties to their surrounding users.

Users in almost all deciles of population density are more likely to follow someone in the highest density decile than in their own decile (Figure 9a). The highest decile of tract population density corresponds to densities higher than 14,622 inhabitants per square miles. 62% of the census tract in this decile are from Los Angeles and San Francisco counties in California, from New York city, from Cook County in Illinois, or from Suffolk county in Massachusetts. Interestingly, this pattern is not present for reciprocated ties (Figure 9b), which are generally homophilous, showing how follower ties to these high density areas tend to go to users with higher status on the platform. We conclude that users residing in city centers and central metropolitan areas in the US gather disproportionate followers on Twitter.

Multivariate analysis

As explained in point 4 of the *Measuring Homophily* subsection, we run nested logit models with our different variables for following ties, and a single model with all variables for reciprocal ties (Table 3). Partisanship, age, race/ethnicity and sex are included as homophily variables, similarly as in Table 2. We include census tract population density and RUCA code as categorical variables, with the combinations of values found as having high probabilities of a tie in bivariate analysis (see *Location homophily* subsection). The regressions include population in radius instead of distance, because of the higher predictive power of this variable identified in the Results. In addition, we add a dummy variable when both members of the dyad reside in the same state, to evaluate whether state boundaries play a role beyond population in radius. The first model we use includes only partisanship, because we are especially interested in detecting how other variables might be driving its homophily levels as this is the attribute generally gathering the most attention from researchers.

	Model 1	Model 2	Model 3	Model 4	Model 5	Rec. ties
McFadden's Pseudo R^2	0.025	0.118	0.131	0.211	0.226	0.264
Inferred partisanship						
Democrat homophily	2.03	1.76	1.75	1.78	1.65	1.64
Independent homophily	0.96	0.88	0.87	0.86	0.87	0.88
Republican homophily	1.60	1.51	1.51	1.47	1.46	1.41
Log_{10}(population in radius)		0.37	0.60	0.56	0.56	0.45
Same State			4.05	4.07	3.92	3.56
Age difference				0.95	0.95	0.95
Race/ethnicity						
African-American homophily				2.99	3.17	3.80
Asian homophily				1.91	1.99	2.47
Caucasian homophily				1.24	1.23	1.25
Hispanic homophily				2.54	2.61	2.72
Same Sex				1.37	1.38	1.35
Tract Population Density						
10th dec. to 10th dec.					3.96	2.37
1st-9th dec. to 10th dec.					1.97	1.15
1st dec. to 1st dec.					1.21	1.40
RUCA code						
Small town/rural to small town/rural					1.36	1.27
Micropolitan to Micropolitan					1.20	1.22
Metropolitan to Metropolitan					1.74	1.71

Table 3. Logistic regression models goodness of fit and exponentiated coefficients. The first five columns correspond to nested models with probability of a following tie as outcome, while the last column is for a single model with probability of a reciprocated tie as outcome. Regressions include controls for differential activity and popularity/sociality for partisanship, age, race and sex All coefficients are significant at the 0.01 significance level.

Overall, we find that most homophily effects identified in bivariate analysis hold, albeit with generally smaller effect sizes. Partisanship homophily decreases when including population in radius in the model, suggesting that some of this homophily was driven by residential segregation and the higher likelihood of Twitter users to follow nearby users. The coefficient for democrat homophily is reduced further when census tract population density and RUCA code are added, probably because democrats are more likely to reside in metropolitan areas that we find have considerable homophily on Twitter. Race/ethnicity homophily is also lower in multivariate analysis, which is likely due to residential segregation as we identify in the *Location homophily* subsection. Homophily by sex stays similar, and even slightly higher, than in bivariate analysis, similarly as for age homophily. Regarding population in radius, we find a relatively significant drop in its coefficient as the variable for residing in the same state is included, showing that some of the preference for nearby users on Twitter is driven by a preference for users of the same state. The homophily of metropolitan and high density areas holds in the multivariate regressions, as well as the preference for users in high density census tracts, while rural and lower density areas homophily is still present, but significantly reduced. Finally, we find similar deviations from the bivariate associations previously found for reciprocal ties across all variables.

In addition to evaluating confounding effects, Model 5 in Table 3 is useful to compare the relative effects of each variable on the probability of a tie. First of all, the importance of the different location variables is striking: the model predicts that residing in the same state multiplies the probability of a tie by roughly four, and that a follower relationship among users with around 10,000 habitants between them is roughly 6 and 10 times more likely than among users with 1M and 10M habitants between them, respectively. These effect sizes dwarf the importance of partisanship: a democrat-democrat tie is less than two times more likely than an independent-independent tie, the largest partisanship difference in the model. The measure of model fit, McFadden Pseudo R^2, is multiplied by six when adding population in the radius and same state in the model, confirming the importance of these variables over partisanship in terms of structuring follower

relationships on Twitter. Age also plays a significant role, with a difference in age of only 12 years generating the same change in probability as going from a democrat-democrat tie to an independent-independent tie. In addition, african-american and hispanic homophily is still substantial and clearly higher than for partisanship. Introducing age, sex, and race/ethnicity to the model significantly increase its Pseudo R^2, pointing at the importance of demographics for ties on Twitter. Comparing Model 5 to the model for reciprocal ties, we find how the differences identified in bivariate analysis are generally confirmed: homophily effects for partisanship, age, and sex are similar, while race homophily is higher for reciprocal ties (see Table 2). In addition, population in radius is more predictive of reciprocal ties, similarly as in the bivariate analysis, while residing in the same state is more predictive of follower ties. Interestingly, the high coefficient for follower ties among users in the highest decile of population density is almost halved for reciprocated ties, pointing at this homophily being driven by non-reciprocated ties.

Discussion

References

Ansolabehere, S., & Hersh, E. (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis*, *20*(4), 437–459.

Arthur, R., & Williams, H. T. P. (2019). The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales. *PLOS ONE*, *14*(4), e0214466. https://doi.org/10.1371/journal.pone.0214466

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. *In WWW '10: Proceedings of the 19th International Conference on World Wide Web,* 61–70.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, *32*(3), 259–280. https://doi.org/10.1257/jep.32.3.259

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Berry, G., Sirianni, A., Weber, I., An, J., & Macy, M. (2021). Estimating Homophily in Social Networks Using Dyadic Predictions. *Sociological Science*, *8*, 285–307. https://doi.org/10.15195/v8.a14

Bestvater, S., Shah, S., Rivero, G., & Smith, A. (2022, June 16). Politics on Twitter: One-Third of Tweets From U.S. Adults Are Political. *Pew Research Center - U.S. Politics & Policy*. https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/

Blau, P. M. (1977). A Macrosociological Theory of Social Structure. *The American Journal of Sociology*, *83*(1), 26–54.

Bojanowski, M., & Corten, R. (2011). *Measuring Segregation in Social Networks* (SSRN Scholarly Paper No. 1873465). https://doi.org/10.2139/ssrn.1873465

Boutyline, A., & Willer, R. (2017). The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks: Political Echo Chambers. *Political Psychology*, *38*(3), 551–569. https://doi.org/10.1111/pops.12337

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-018-07761-2

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated

with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617. https://doi.org/10.1073/pnas.1706588114

Brown, J. R., & Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*, *5*(8), Article 8. https://doi.org/10.1038/s41562-021-01066-z

Butters, R., & Hare, C. (2022). Polarized Networks? New Evidence on American Voters' Political Discussion Networks. *Political Behavior*, *44*(3), 1079–1103. https://doi.org/10.1007/s11109-020-09647-w

Cairncross, F. (1997). *The Death of Distance: How the Communications Revolution Will Change Our Lives*. Harvard Business Review Press.

Centola, D. (2011). An Experimental Study of Homophily in the Adoption of Health Behavior. *Science*, *334*(6060), 1269–1272. https://doi.org/10.1126/science.1207055

Cesare, N., Lee, H., McCormick, T., & Spiro, E. S. (2017). *Redrawing the "Color Line": Examining Racial Segregation in Associative Networks on Twitter* (arXiv:1705.04401). arXiv. https://doi.org/10.48550/arXiv.1705.04401

Chen, L., Weber, I., & Okulicz-Kozaryn, A. (2014). U.S. Religious Landscape on Twitter. *ArXiv:1409.8578 [Physics]*. http://arxiv.org/abs/1409.8578

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M.

(2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), Article 1.

Cormier, M., & Cushman, M. (2021). Innovation via social media – The importance of Twitter to science. *Research and Practice in Thrombosis and Haemostasis*, *5*(3), 373–375. https://doi.org/10.1002/rth2.12493

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 22436–22441. https://doi.org/10.1073/pnas.1006155107

Currarini, S., Jackson, M. O., & Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, *107*(11), 4857–4861.

https://doi.org/10.1073/pnas.0911793107

De Choudhury, M. (2011). Tie Formation on Twitter: Homophily and Structure of Egocentric Networks. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 465–470. https://doi.org/10.1109/PASSAT/SocialCom.2011.177

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., & Zheng, T. (2011a). Segregation in Social Networks Based on Acquaintanceship and Trust. *American Journal of Sociology*, *116*(4), 1234–1283. https://doi.org/10.1086/659100

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., & Zheng, T. (2011b). Segregation in Social Networks Based on Acquaintanceship and Trust. *American Journal of Sociology*, *116*(4), 1234–1283. https://doi.org/10.1086/659100

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open*, *9*(1), 2158244019832705. https://doi.org/10.1177/2158244019832705

Enli, G. S., & Skogerbø, E. (2013). Personalized Campaigns in Party-Centred Politics. *Information, Communication & Society*, *16*(5), 757–774. https://doi.org/10.1080/1369118X.2013.782330

Etter, M., Ravasi, D., & Colleoni, E. (2019). Social Media and the Formation of Organizational Reputation. *Academy of Management Review*, *44*(1), 28–52.

https://doi.org/10.5465/amr.2014.0280

Eveland, W. P., Appiah, O., & Beck, P. A. (2018). Americans are more exposed to difference than we think: Capturing hidden exposure to political and racial difference. *Social Networks*, *52*, 192–200. https://doi.org/10.1016/j.socnet.2017.08.002

Friedkin, N. E. (2004). Social Cohesion. *Annual Review of Sociology*, *30*(1), 409–425. https://doi.org/10.1146/annurev.soc.30.012703.110625

Garimella, V. R. K., & Weber, I. (2017). A Long-Term Analysis of Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1.

Gentzkow, M., & Shapiro, J. M. (2011). Ideological Segregation Online and Offline *. *The Quarterly Journal of Economics*, *126*(4), 1799–1839. https://doi.org/10.1093/qje/qjr044

Gibbs, C., & Haynes, R. (2013). A Phenomenological Investigation Into How Twitter Has Changed the Nature of Sport Media Relations. *International Journal of Sport Communication*, *6*(4), 394–408. https://doi.org/10.1123/ijsc.6.4.394

Golder, S., Stevens, R., O'Connor, K., James, R., & Gonzalez-Hernandez, G. (2022). Methods to Establish Race or Ethnicity of Twitter Users: Scoping Review. *Journal of Medical Internet Research*, *24*(4), e35788. https://doi.org/10.2196/35788

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks*. *Demography*, *46*(1), 103–125. https://doi.org/10.1353/dem.0.0045

Grabowicz, P. A., Ganguly, N., & Gummadi, K. P. (2016, March 31). Distinguishing between Topical and Non-Topical Information Diffusion Mechanisms in Social Media. *Tenth International AAAI Conference on Web and Social Media*. Tenth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13130

Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., & Eguíluz, V. M. (2014). Entangling Mobility and Interactions in Social Media. *PLOS ONE*, *9*(3), e92196. https://doi.org/10.1371/journal.pone.0092196

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019a). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019b). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Halberstam, Y., & 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. https://doi.org/10.1016/j.jpubeco.2016.08.011

Hedayatifar, L., Rigg, R. A., Bar-Yam, Y., & Morales, A. J. (2019). US social fragmentation at multiple scales. *Journal of The Royal Society Interface*, *16*(159), 20190509. https://doi.org/10.1098/rsif.2019.0509

Huang, Y., Shen, C., & Contractor, N. S. (2013). Distance matters: Exploring proximity and homophily in virtual world networks. *Decision Support Systems*, *55*(4), 969–977. https://doi.org/10.1016/j.dss.2013.01.006

Huber, G. A., & Malhotra, N. (2017). Political Homophily in Social Relationships: Evidence from Online Dating Behavior. *The Journal of Politics*, *79*(1), 269–283. https://doi.org/10.1086/687533

Hughes, A. G., McCabe, S. D., Hobbs, W. R., Remy, E., Shah, S., & Lazer, D. M. J. (2021). Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets. *Public Opinion Quarterly*, *85*(S1), 323–346. https://doi.org/10.1093/poq/nfab020

Hui, I. (2013). Who is Your Preferred Neighbor? Partisan Residential Preferences and Neighborhood Satisfaction. *American Politics Research*, *41*(6), 997–1021. https://doi.org/10.1177/1532673X13482573

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Joseph, K., Shugars, S., Gallagher, R., Green, J., Quintana Mathé, A., An, Z., & Lazer, D.

(2021). (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 312–324. https://doi.org/10.18653/v1/2021.emnlp-main.27

Kang, J., & Chung, D. Y. (2017). Homophily in an Anonymous Online Community: Sociodemographic Versus Personality Traits. *Cyberpsychology, Behavior, and Social Networking*, *20*(6), 376–381. https://doi.org/10.1089/cyber.2016.0227

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

King, G., & Zeng, L. (2002). Estimating risk and rate levels, ratios and differences in case-control studies. *Statistics in Medicine*, *21*(10), 1409–1427. https://doi.org/10.1002/sim.1032

Korkmaz, G., Kuhlman, C. J., Goldstein, J., & Vega-Redondo, F. (2019). A computational study of homophily and diffusion of common knowledge on social networks based on a model of Facebook. *Social Network Analysis and Mining*, *10*(1), 5. https://doi.org/10.1007/s13278-019-0615-5

Kossinets, G., & Watts, D. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, *115*(2), 405–450. https://doi.org/10.1086/599247

Kozlowski, D., Murray, D. S., Bell, A., Hulsey, W., Larivière, V., Monroe-White, T., &

Sugimoto, C. R. (2022). Avoiding bias when inferring race using name-based approaches. *PLOS ONE*, *17*(3), e0264270. https://doi.org/10.1371/journal.pone.0264270

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. (2012). Geographic Dissection of the Twitter Network. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), Article 1.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 591. https://doi.org/10.1145/1772690.1772751

Laniado, D., Volkovich, Y., Kappler, K., & Kaltenbrunner, A. (2016). Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, *5*(1), 19. https://doi.org/10.1140/epjds/s13688-016-0080-6

Lawrence, B. S., & Shah, N. P. (2020). Homophily: Measures and Meaning. *Academy of Management Annals*, *14*(2), 513–597. https://doi.org/10.5465/annals.2018.0147

Lee, B., & Bearman, P. (2020). Political isolation in America. *Network Science*, *8*(3), 333–355. https://doi.org/10.1017/nws.2020.9

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, *109*(1), 68–72. https://doi.org/10.1073/pnas.1109739109

Liao, L., Jiang, J., Lim, E.-P., & Huang, H. (2014). A study of age gaps between online

friends. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 98–106. https://doi.org/10.1145/2631775.2631800

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, *102*(33), 11623–11628. https://doi.org/10.1073/pnas.0503018102

Marsden, P. V. (1988). Homogeneity in confiding relations. *Social Networks*, *10*(1), 57–76. https://doi.org/10.1016/0378-8733(88)90010-X

Masciantonio, A., Bourguignon, D., Bouchat, P., Balty, M., & Rimé, B. (2021). Don't put all social network sites in one basket: Facebook, Instagram, Twitter, TikTok, and their relations with well-being during the COVID-19 pandemic. *PLOS ONE*, *16*(3), e0248384. https://doi.org/10.1371/journal.pone.0248384

Mayer, A., & Puller, S. L. (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, *92*(1), 329–347. https://doi.org/10.1016/j.jpubeco.2007.09.001

Mazur, E., & Richards, L. (2011). Adolescents' and emerging adults' social networking online: Homophily or diversity? *Journal of Applied Developmental Psychology*, *32*(4), 180–188. https://doi.org/10.1016/j.appdev.2011.03.001

Mcclain, C. (2021, May 4). 70% of U.S. social media users never or rarely post or share about political, social issues. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2021/05/04/70-of-u-s-social-media-users-

never-or-rarely-post-or-share-about-political-social-issues/

McPherson, M., & Smith-Lovin, L. (1987). Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups. *American Sociological Review*, *52*(3), 370–379. https://doi.org/10.2307/2095356

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

Meer, T. van der, & Tolsma, J. (2014). Ethnic Diversity and Its Effects on Social Cohesion. *Annual Review of Sociology*, *40*(1), 459–478. https://doi.org/10.1146/annurev-soc-071913-043309

Meier, A., Gilbert, A., Börner, S., & Possler, D. (2020). Instagram Inspiration: How Upward Comparison on Social Network Sites Can Contribute to Well-Being. *Journal of Communication*, *70*(5), 721–743. https://doi.org/10.1093/joc/jqaa025

Mesch, G. S., Talmud, I., & Quan-Haase, A. (2012). Instant messaging social networks: Individual, relational, and cultural characteristics. *Journal of Social and Personal Relationships*, *29*(6), 736–759. https://doi.org/10.1177/0265407512448263

Messias, J., Vikatos, P., & Benevenuto, F. (2017). White, Man, and Highly Followed: Gender and Race Inequalities in Twitter. *Proceedings of the International Conference on Web Intelligence*, 266–274. https://doi.org/10.1145/3106426.3106472

Mitchell, A., Shearer, E., & Stocking, G. (2021, November 15). News on Twitter: Consumed by Most Users and Trusted by Many. *Pew Research Center's Journalism Project*. https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/

Mok, D., Wellman, B., & Carrasco, J. (2010). Does Distance Matter in the Age of the Internet? *Urban Studies*, *47*(13), 2747–2783. https://doi.org/10.1177/0042098010377363

Moody, J. (2001). Race, School Integration, and Friendship Segregation in America. *American Journal of Sociology*, *107*(3), 679–716. https://doi.org/10.1086/338954

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021). Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences*, *118*(7), e2022761118. https://doi.org/10.1073/pnas.2022761118

Mukerjee, S., Jaidka, K., & Lelkes, Y. (2022). The Political Landscape of the U.S. Twitterverse. *Political Communication*, *0*(0), 1–24. https://doi.org/10.1080/10584609.2022.2075061

Mutz, D. C., & Mondak, J. J. (2006). The Workplace as a Context for Cross-Cutting Political Discourse. *The Journal of Politics*, *68*(1), 140–155. https://doi.org/10.1111/j.1468-2508.2006.00376.x

Myers, S. A., Sharma, A., Gupta, P., & Lin, J. (2014). Information network or social network?: The structure of the twitter follow graph. *Proceedings of the 23rd International Conference on World Wide Web*, 493–498. https://doi.org/10.1145/2567948.2576939

Odabaş, M. (2022, 05). 10 facts about Americans and Twitter. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2022/05/05/10-facts-about-americans-and-twitter/

Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L., & Christakis, N. A. (2011). Geographic Constraints on Social Network Groups. *PLoS ONE*, *6*(4), e16939. https://doi.org/10.1371/journal.pone.0016939

Pew Research Center. (2019, October 23). National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets. *Pew Research Center - U.S. Politics & Policy*. https://www.pewresearch.org/politics/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. *PLOS ONE*, *7*(6), e39253. https://doi.org/10.1371/journal.pone.0039253

Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media*, *9*, 17–29.

https://doi.org/10.1016/j.osnem.2018.11.001

Quercia, D., Capra, L., & Crowcroft, J. (2012). The Social World of Twitter: Topics, Geography, and Emotions. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), Article 1. https://doi.org/10.1609/icwsm.v6i1.14254

Rainie, L., & Wellman, B. (2012). *Networked: The New Social Operating System*. https://doi.org/10.7551/mitpress/8358.001.0001

Romero, D., Tan, C., & Ugander, J. (2013). On the Interplay between Social and Topical Structure. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), Article 1.

Rychwalska, A., & Roszczyńska-Kurasińska, M. (2018). *Polarization on Social Media: When Group Dynamics Leads to Societal Divides*. http://hdl.handle.net/10125/50150

Salamanca, P., Janulis, P., Elliott, M., Birkett, M., Mustanski, B., & Phillips, G. (2019). An Investigation of Racial and Ethnic Homophily on Grindr Among an Ongoing Cohort Study of YMSM. *AIDS and Behavior*, *23*(1), 302–311. https://doi.org/10.1007/s10461-018-2262-7

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156. https://doi.org/10.7717/peerj-cs.156

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F.

(2021). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, *4*(1), 381–402. https://doi.org/10.1007/s42001-020-00084-7

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Schmidt, C. G., Wuttke, D. A., Ball, G. P., & Heese, H. S. (2020). Does social media elevate supply chain importance? An empirical examination of supply chain glitches, Twitter reactions, and stock market returns. *Journal of Operations Management*, *66*(6), 646–669. https://doi.org/10.1002/joom.1087

Sears, D. O., & Freedman, J. L. (1967). SELECTIVE EXPOSURE TO INFORMATION: A CRITICAL REVIEW*. *Public Opinion Quarterly*, *31*(2), 194–213. https://doi.org/10.1086/267513

Shearer, E., & Matsa, K. E. (2018, September 10). News Use Across Social Media Platforms 2018. *Pew Research Center's Journalism Project*. https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media-platforms-2018/

Shore, J., Baek, J., & Dellarocas, C. (2018). Network Structure and Patterns of Information Diversity on Twitter. *Management Information Systems Quarterly*,

42(3), 849–972.

Shrum, W., Cheek, N. H., & Hunter, S. MacD. (1988). Friendship in School: Gender and Racial Homophily. *Sociology of Education*, *61*(4), 227–239. https://doi.org/10.2307/2112441

Shugars, S., Gitomer, A., McCabe, S., Gallagher, R. J., Joseph, K., Grinberg, N., Doroshenko, L., Foucault Welles, B., & Lazer, D. (2021). Pandemics, Protests, and Publics: Demographic Activity and Engagement on Twitter in 2020. *Journal of Quantitative Description: Digital Media*, *1*. https://doi.org/10.51685/jqd.2021.002

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*(7392), 96–100. https://doi.org/10.1038/nature10856

Smith, J. A., McPherson, M., & 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*(3), 432–456. https://doi.org/10.1177/0003122414531776

Spiro, E. S., Almquist, Z. W., & Butts, C. T. (2016). The Persistence of Division: Geography, Institutions, and Online Friendship Ties. *Socius*, *2*, 2378023116634340. https://doi.org/10.1177/2378023116634340

Stein, J., Keuschnigg, M., & van de Rijt, A. (2023). Network segregation and the

propagation of misinformation. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-022-26913-5

Stephens, M., & Poorthuis, A. (2015). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*, *53*, 87–95. https://doi.org/10.1016/j.compenvurbsys.2014.07.002

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, *35*(1), 50–74. https://doi.org/10.1080/10584609.2017.1334728

Stivala, A., Robins, G., & Lomi, A. (2020). Exponential random graph model parameter estimation for very large directed networks. *PLOS ONE*, *15*(1), e0227804. https://doi.org/10.1371/journal.pone.0227804

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media* (NED-New edition). Princeton University Press. https://doi.org/10.2307/j.ctv8xnhtd

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81. https://doi.org/10.1016/j.socnet.2011.05.006

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

https://doi.org/10.1002/asi.20978

Tokita, C. K., Guess, A. M., & Tarnita, C. E. (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, *118*(50). https://doi.org/10.1073/pnas.2102147118

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14517

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). *The Anatomy of the Facebook Social Graph* (arXiv:1111.4503). arXiv. http://arxiv.org/abs/1111.4503

Urman, A. (2020). Context matters: Political polarization on Twitter from a comparative perspective. *Media, Culture & Society*, *42*(6), 857–879. https://doi.org/10.1177/0163443719876541

Utz, S., & Jankowski, J. (2016). Making "Friends" in a Virtual World: The Role of Preferential Attachment, Homophily, and Status. *Social Science Computer Review*, *34*(5), 546–566. https://doi.org/10.1177/0894439315605476

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement Among German and Italian Users of Twitter. *Social Media + Society*, *2*(3), 2056305116664221. https://doi.org/10.1177/2056305116664221

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Wimmer, A., & Lewis, K. (2010). Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. *American Journal of Sociology*, *116*(2), 583–642. https://doi.org/10.1086/653658

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., & Tucker, J. A. (2022). Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science Advances*, *8*(39), eabn9418. https://doi.org/10.1126/sciadv.abn9418

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. *Proceedings of the 20th International Conference on World Wide Web*, 705–714. https://doi.org/10.1145/1963405.1963504

Yavaş, M., & Yücel, G. (2014). Impact of Homophily on Diffusion Dynamics Over Social Networks. *Social Science Computer Review*, *32*(3), 354–372. https://doi.org/10.1177/0894439313512464

Zamal, F. A., Liu, W., & Ruths, D. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), Article 1.