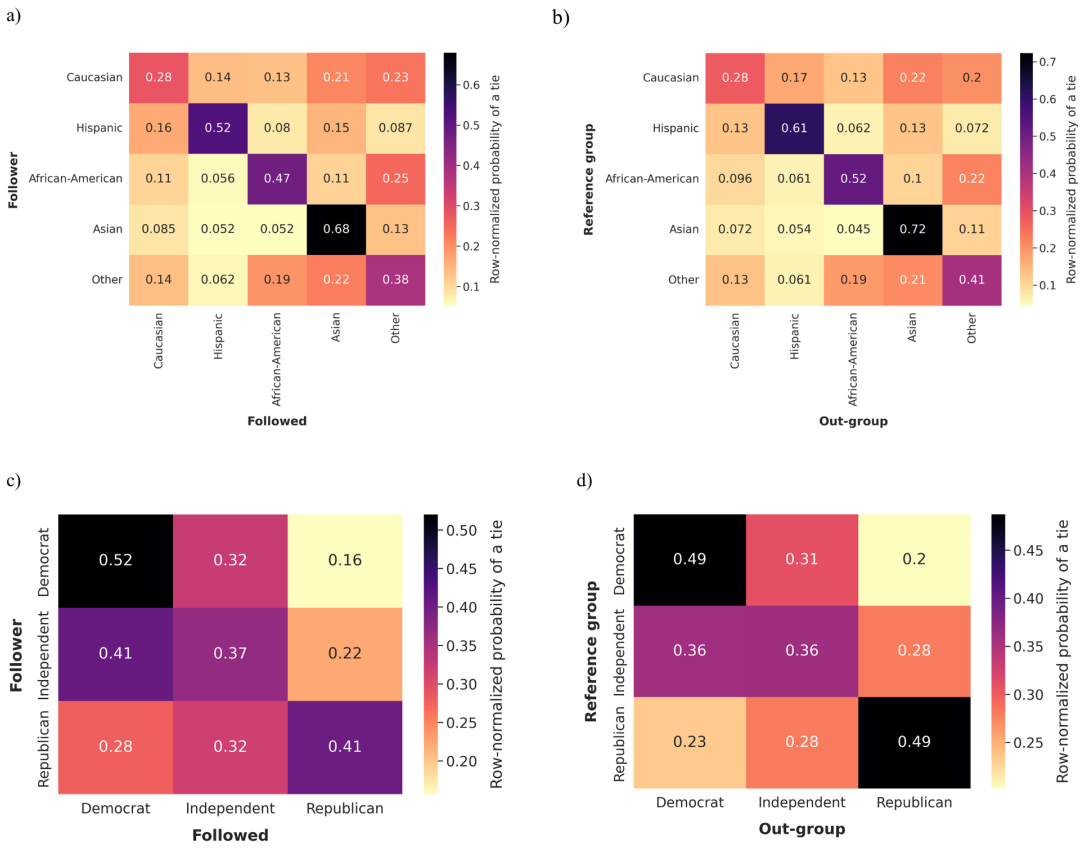
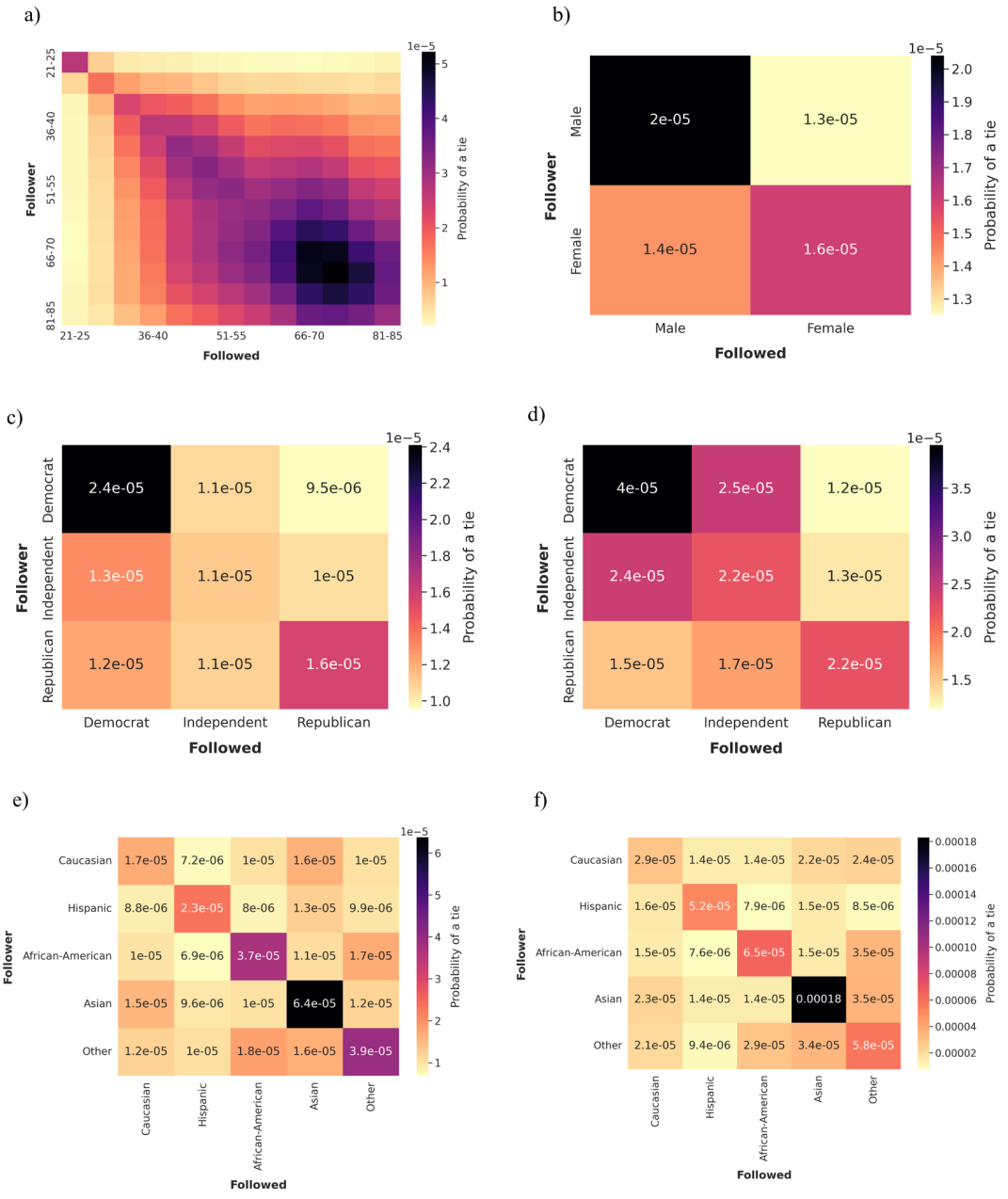
Who follows Whom on Twitter? An analysis of homophily patterns

Appendix

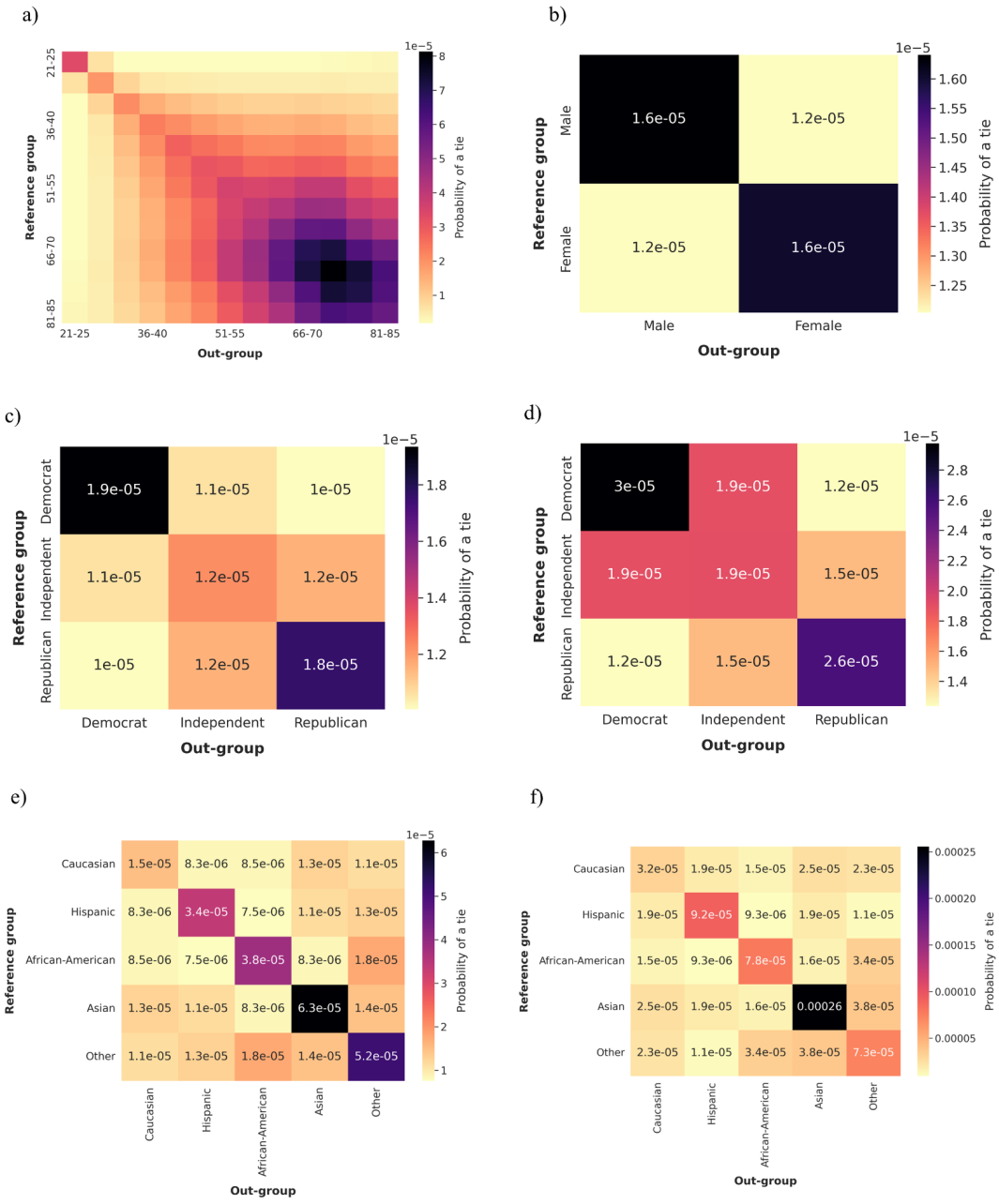
Appendix figures



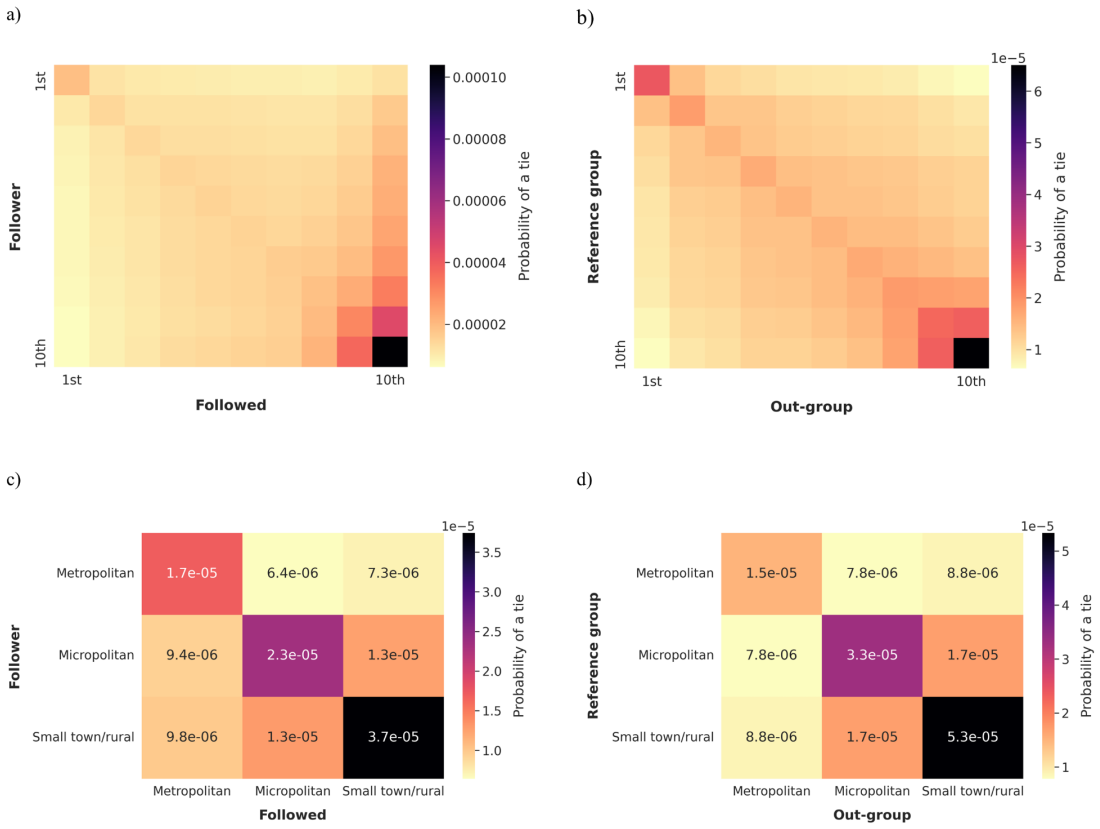
Appendix figure 1. Heatmaps with row-normalized probabilities of a tie for race/ethnicity in VRA states and party registration (panels a) and c), respectively), and with row-normalized probabilities of a reciprocated following tie for race/ethnicity in VRA states and party registration (panels b) and d), respectively). Color scales change for each panel.



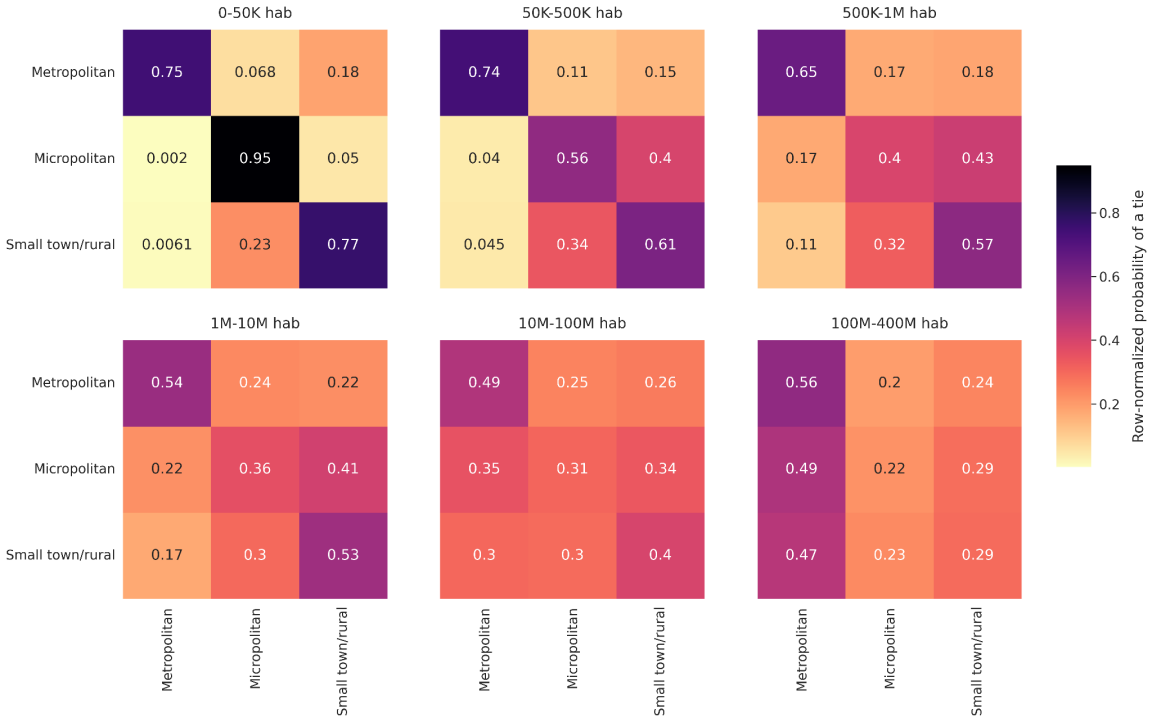
Appendix Figure 2. Heatmaps with probabilities of a tie for a) Age, b) Gender, c) Inferred partisanship, d) Party registration, e) Inferred Race/ethnicity, f) Race/ethnicity in VRA states. Color scales change for each panel.



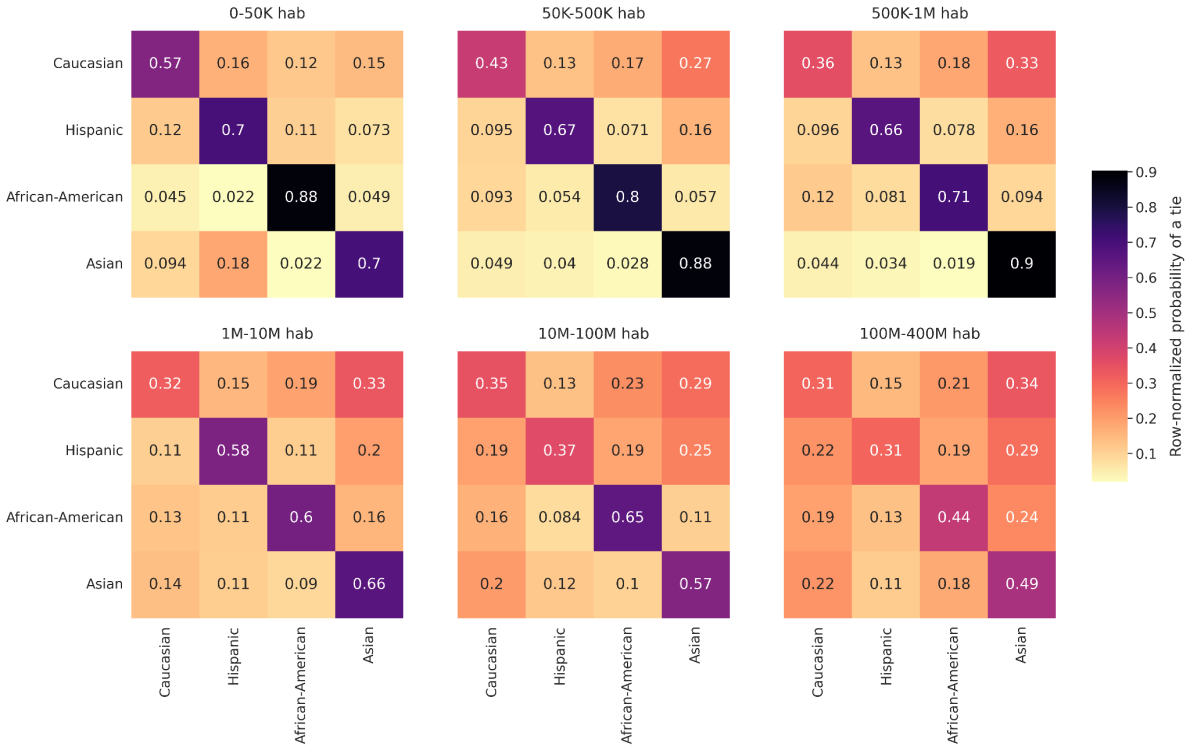
Appendix Figure 3. Heatmaps with probabilities of a reciprocated following tie for a) Age, b) Gender, c) Inferred partisanship, d) Party registration, e) Inferred Race/ethnicity, f) Race/ethnicity in VRA states. Color scales change for each panel.



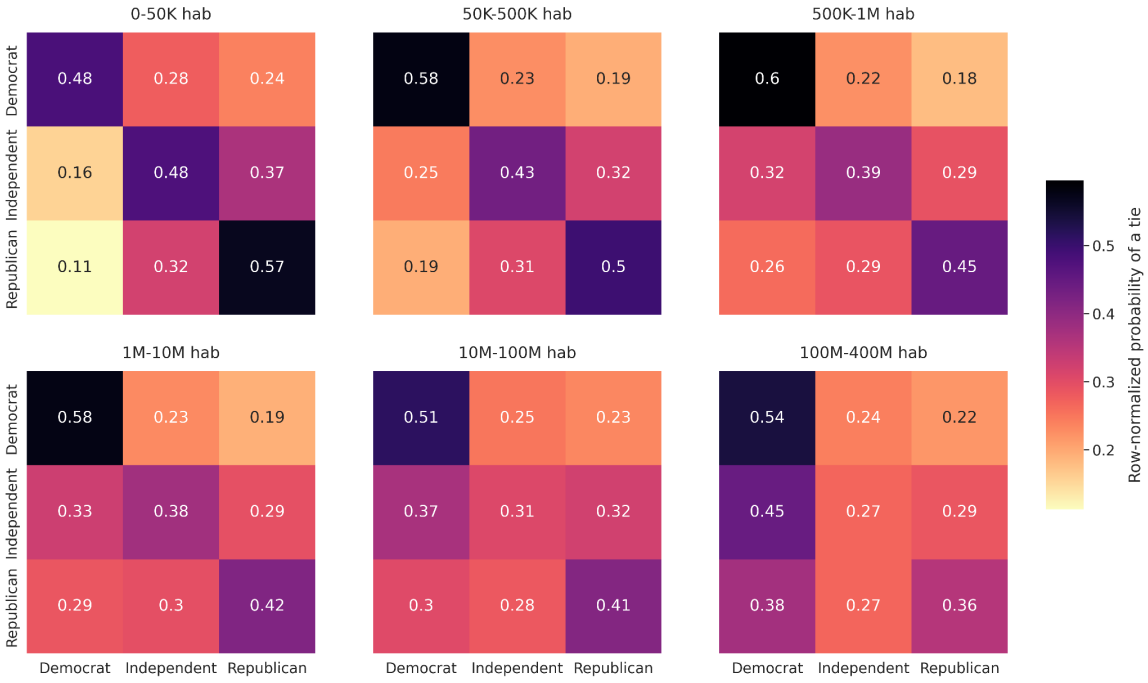
Appendix Figure 4. Heatmaps with row-normalized probabilities of a tie for population density of census tract and RUCA category of tract (panels a) and c), respectively), and with row-normalized probabilities of a reciprocated following tie for population density of census tract and RUCA category of tract (panels b) and d), respectively). Color scales change for each panel.



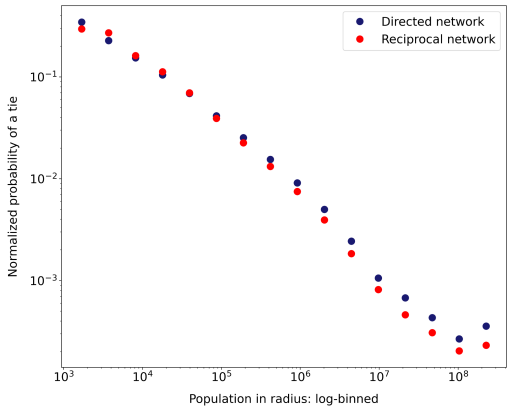
Appendix Figure 5. Heatmaps of row-normalized probability of a tie by RUCA category, broken down by the population in radius between members of the dyad. The probabilities in a tie in each cell are calculated by summing the number of ties between users in each combination of RUCA categories that are within a given range of population in radius, and dividing this value by the corresponding number of dyads.



Appendix Figure 6. Heatmaps of row-normalized probability of a tie by race/ethnicity, broken down by the population in radius between members of the dyad. The probabilities in a tie in each cell are calculated by summing the number of ties between users in each combination of race/ethnicity values that are within a given range of population in radius, and dividing this value by the corresponding number of dyads.



Appendix Figure 7. Heatmaps of row-normalized probability of a tie by inferred partisanship, broken down by the population in radius between members of the dyad. The probabilities in a tie in each cell are calculated by summing the number of ties between users in each combination of partisanships that are within a given range of population in radius, and dividing this value by the corresponding number of dyads.



Appendix Figure 8. Normalized probability of following ties by population in radius compared to reciprocal ties

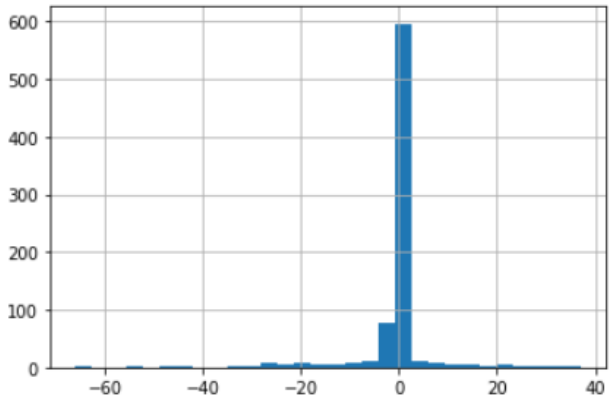
Appendix Section 1: Comparison to Covid States

In this section, we compare the Twitter panel data and the Covid State data for the users present in both datasets. This comparison allows us to assess the reliability of the panel data and, in turn, confirm the validity of the results derived from this data. First, we examine in Appendix Table 1 the alignment of binary gender between voter file and survey. 96% of users have the same gender in the voter file and the survey. Overall, this validates the sex data of the panel.

	Female	**Male**
Female	535	15
Male	16	201

Appendix Table 1: Alignment of binary sex between panel and survey. Columns correspond to panel sex, while rows denote survey gender.

Then, we look at the matching of age between both data sources. Appendix Figure 9 presents the distribution of age differences between the survey and the voter file. Most of the users (82%) have their age in both sources within a maximum of 1 year of absolute difference, and 84% have it within two years. Still, 14% of the users have absolute differences of 5 years or more. Given that both the voter file and the survey data on age should be of high accuracy, this mismatch indicates potential errors either in the survey to Twitter user or in the voter file to Twitter user linkages.



Appendix Figure 9: Histogram of age differences between survey and voter file. All the survey ages are dated back to the panel built date (2017) based on the year the survey was filled. The x-axis is the result of survey age minus panel age. The y-axis corresponds to the number of users.

Regarding partisanship, we compare the two measures of partisanship from the voter file, inferred partisanship and party registration, to three different measures of political orientation from the survey responses: partisanship identification in three categories, partisanship identification in seven categories, and candidate voted or intended to vote in the 2020 elections. Appendix Table 2 illustrates the alignment between panel partisanship and survey partisanship in three categories. A majority or close to a majority of voter file identified Republicans (47%-55%) and Democrats (62%-74%) validate their partisanship in the survey. Party registration matches the survey responses better than inferred partisanship across all partisanships, with larger differences in voter file percentage matching survey answers for Independents and Democrats (14% and 12%, respectively) than for Republicans (7%).

	Voter file: inferred partisanship(N=796)	**Voter file: party registration (N=431)**

Survey	Republican	Independent	Democrat	Republican	Independent	Democrat
Republican	47%	19%	12%	55%	18%	10%
Independent/Other	32%	38%	26%	24%	52%	15%
Democrat	20%	43%	62%	21%	30%	74%
Total	238	118	439	112	123	196

Appendix Table 2: Column normalized distribution of voter file partisanship against the answers of the survey question, "Generally speaking, do you think of yourself as a...Republican, Independent, Democrat, or Other.", grouping the Independent and Other categories. Columns represent voter file partisanship through inferred partisanship (left) or registration (right), while rows signify partisan categories from the survey.

Next, Appendix Table 3 displays a similar comparison but using survey partisanship in 7 categories. This variable is made by asking follow-up questions after the main partisanship question: "Generally speaking, do you think of yourself as a… Republican, Independent, Democrat, or Other." Respondents who identify as either Republican or Democrat are asked if they consider themselves to be a strong Republican/Democrat, or a not very strong Republican/Democrat. Respondents who do not identify as Republicans or Democrats are asked if they are closer to the republican party, the democratic party, or neither. We observe how independents in the voter file measures tend to lean Democrat, and that a small percentage of voter file Democrats are Republicans or lean Republican in the survey data (15% for inferred partisanship, and 12% for party registration). However, a relatively high percentage of voter file Republicans identify as Democrats or Democrat leaning in the survey (29% for inferred partisanship and 27% for party registration)

	Voter file: inferred partisanship(N=796)			Voter file: party registration (N=431)		
Survey	**Republican**	**Independent**	**Democrat**	**Republican**	**Independent**	**Democrat**

Strong Republican	28%	11%	7%	31%	10%	7%
Republican	19%	8%	5%	23%	8%	3%
Leaning Republican	11%	8%	3%	11%	7%	2%
Independent	13%	18%	12%	8%	19%	6%
Leaning Democrat	9%	12%	11%	6%	26%	8%
Democrat	7%	21%	20%	9%	15%	23%
Strong Democrat	13%	22%	43%	12%	15%	52%
Total	237	118	439	111	123	196

Appendix Table 3: Column normalized distribution of voter file partisanship against survey partisanship in 7 categories. The survey variable is made from answers to 4 different survey questions, where respondents identifying as partisans are asked about the strength of their partisan identity, and respondents identifying as independents or from another party are asked if they lean Republican or Democrat.

Finally, we also compare the voter file partisanship to the candidate voted or intended to vote in the 2020 election in Appendix Table 4. For responses after the elections and respondents who voted, we use the answers to the question "Who did you vote for in the 2020 U.S. presidential election?". For respondents who did not vote, we use the question “Who did you support in the 2020 U.S. presidential election?”. For responses before the election, we use the question "If the 2020 U.S. presidential election were held today, who would you vote for?". This table displays higher agreement between the voter file inferred partisanship and the survey data than Table X: a majority of each partisanship voted or intended to vote for their corresponding candidate. However, a significant fraction of voter file Republicans voted or intended to vote Biden. In addition, most voter file independents voted for Biden.

	Party ID Panel(N=785)			**Party Registration Panel (N=432)**		
Survey	**Republican**	**Independent**	**Democrat**	**Republican**	**Independent**	**Democrat**

Trump	52%	27%	16%	53%	24%	15%
Other/Not sure	13%	15%	11%	15%	20%	7%
Biden	35%	58%	73%	32%	55%	78%
Total	239	118	438	113	123	196

Appendix Table 4: Column normalized distribution of voter file partisanship against candidate voted, supported, or intended to vote in the 2020 presidential election : "Who did you support in the 2020 U.S. presidential election?", "If the 2020 U.S. presidential election were held today, who would you vote for?", and "Who did you vote for in the 2020 U.S. presidential election?".

Summing up, we find that Democrats identified from the voter file measures are generally well aligned to the survey responses, while Republicans are less so. In addition, party registration matches the survey data slightly better than inferred partisanship. The reasons for the mismatches found may be years of difference between the voter file data, from 2017, and the survey data, recorded between 2020 and 2022. In some cases, it is also possible that the Twitter user linked does not correspond to the survey respondent or the voter record. Finally, the voter file data may misclassify users, an option probably more likely for inferred partisanship than for party registration.

Regarding race/ethnicity, we first convert the survey question, which is a multiple choice question, into a single answer variable. We do this by selecting the minority race/ethnicity when a respondent identifies as white and another race/ethnicity, and, for the few cases where respondents identify as hispanic and african-american or asian, we categorize them as hispanics. Appendix Table 5 demonstrates the alignment between the voter file race/ethnicity variable and this grouped race/ethnicity variable from the survey, for all users, and for users in states affected by the Voting Right Act. It reveals that most voter file whites and african-americans self-report as the corresponding race in the survey, while a majority of hispanics and asians also self-report as their race (50% and 54%, respectively). However, the sample sizes of panel asians and hispanics are small, and recall is low for these categories: only 19% of survey asians are categorized as asians in the panel, and only 16% of survey hispanics are categorized as hispanics in the panel.

Surprisingly, the overall percentage of users with matching voter file to survey race/ethnicity is 88%, while it is 81% for users in VRA states. However, the sample of users in both the survey and the voter file is small.

	All States(N=794)				VRA States (N=155)			
	White	**Hispanic**	**African-American**	**Asian**	**White**	**Hispanic**	**African-American**	**Asian**
White	89%	25%	13%	31%	81%	0%	15%	100%
Hispanic	3%	50%	0%	8%	8%	67%	0%	0%
African American	3%	0%	82%	0%	7%	0%	79%	0%
Asian	3%	25%	1%	54%	3%	33%	3%	0%
Total	672	20	68	13	113	6	33	1

Appendix Table 5: Column normalized distribution of voter file race/ethnicity against and recoded survey responses to multiple choice questions on race/ethnicity.. Columns correspond to panel-identified races, categorized through all states (left) and VRA states(right). In contrast, rows denote races derived from survey answers. We do not display the values for the "Other" category from the survey responses, or from the "Unknown" category of the voter file.

Appendix Section 2: Covid States homophily

In this section, we discuss the homophily results for the Covid States survey data and compare them to the panel data. We build a network of following relationships on Twitter among users who provided their Twitter handles similarly as we do for the panel members: for each of them, we collected the users they follow through January 2023, and filtered these followee lists to include only users that also provided handles in the survey. A majority of the users in the Covid States data do not follow and are not followed by

any other Covid States Twitter user, however, we keep these isolates in the network because they can influence the homophily estimates. This subgraph of the whole Twitter follower network has 31,653 nodes and 63,329 edges. However, after inspecting the network, we found that a small portion of users provided clearly false handles, such as the twitter handles of famous singers or actors. These users, with a very high number of followers, make up for a large fraction of the edges of the network. To deal with this issue, we filter the network to nodes with in-degree below a certain threshold. Since it is hard to know the right threshold to filter for, given that some survey respondents may be relatively popular on Twitter, we use four different thresholds and provide our results for the four resulting networks. Appendix table 6 provides the descriptives and filtering thresholds for each network. Filtering more nodes with high degree leads to substantially less total edges, however, it also reduces the risk that a few high degree users, with maybe a false handle not associated to the corresponding survey respondent, drive most of the results.

Network	**Maximum degree**	**Number of nodes**	**Number of Edges**
G1	99	31,555	9,717
G2	49	31,505	5,949
G3	24	31,460	4,360
G4	14	31,395	2,945

Appendix Table 6. Descriptives for the Covid States Twitter networks.

We show the results of the simple logistic regressions from point 2 of the *Measuring homophily* subsection in Appendix Table 7. We acknowledge the issues with this data, however, we think it is valuable as a comparison point to the panel network.

	Directed ties

	Homophily				Homophily + Activity + Popularity			
	G1	**G2**	**G3**	**G4**	**G1**	**G2**	**G3**	**G4**
Party: Democrat	1.46 (0.07)	1.55 (0.09)	1.60 (0.11)	1.58 (0.14)	1.45 (0.14)	1.43 (0.18)	1.43 (0.21)	1.49 (0.27)
Party: Independent/Other	0.98 (0.07)	0.90 (0.09)	0.96 (0.11)	1.04 (0.14)	0.87 (0.10)	0.86 (0.13)	0.83 (0.15)	0.89 (0.19
Party: Republican	1.31 (0.11)	1.48 (0.16)	1.47 (0.19)	1.27 (0.21)	1.54 (0.19)	1.78 (0.28)	2.14 (0.40)	1.70 (0.40)
Vote: Biden	1.35 (0.06)	1.49 (0.08)	1.69 (0.11)	1.81 (0.14)	1.36 (0.14)	1.46 (0.21)	1.70 (0.29)	1.87 (0.40)
Vote: Other/Not Sure	1.33 (0.17)	0.97 (0.20)	0.74 (0.21)	0.84 (0.29)	1.14 (0.19)	1.00 (0.25)	0.82 (0.28)	0.83 (0.34)
Vote: Trump	1.40 (0.10)	1.58 (0.15)	1.78 (0.19)	1.74 (0.23)	1.55 (0.19)	1.61 (0.26)	1.68 (0.33)	1.69 (0.41)
Same Sex	1.17 (0.05)	1.28 (0.07)	1.27 (0.08)	1.25 (0.1)	1.27 (0.05)	1.38 (0.07)	1.35 (0.08)	1.31 (0.1)
African-American	2.5 (0.27)	2.75 (0.37)	3.05 (0.49)	3.47 (0.66)	2.41 (0.37)	2.35 (0.47)	3.16 (0.80)	2.92 (0.89)
Asian	1.85 (0.41)	2.42 (0.63)	2.3 (0.78)	1.81 (0.90)	3.00 (0.82)	2.73 (0.89)	3.17 (1.37)	2.82 (1.75)
Caucasian	1.30 (0.05)	1.45 (0.08)	1.80 (0.12)	1.83 (0.14)	1.55 (0.16)	1.69 (0.24)	1.70 (0.24)	1.95 (0.43)
Hispanic	1.74 (0.30)	1.32 (0.36)	1.09 (0.43)	1.31 (0.59)	1.08 (0.23)	1.20 (0.39)	1.06 (0.49)	1.10 (0.59)
Age	0.98 (0.00)	0.98 (0.00)	0.98 (0.00)	0.98(0.00)	0.98(0.00)	0.98(0.00)	0.98(0.00)	0.98(0.00)

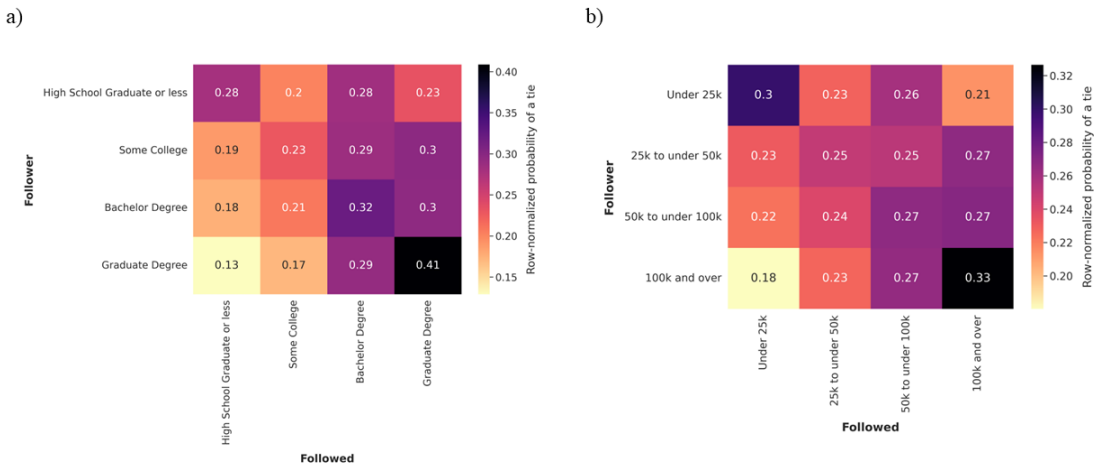
Appendix Table 7. Exponentiated logistic regression coefficients of group homophily with respect to non-homophilous ties, with and without controls for differential activity and popularity. G1, G2, G3 and G4 denote the networks from Covid States respondents after filtering nodes with degree higher than 99, higher than 49, higher than 24, and higher than 14, respectively. 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). 95% confidence interval are provided in parentheses.

We find some degree of homophily for all attributes, but generally lower than for the panel. This is somewhat surprising, given that the survey self-reported attributes should be more accurate than the voter file data, especially for partisanship and race/ethnicity, and that misclassification should, in general, lead to an underestimation of homophily. However, while the survey data is of higher quality, the sample is arguably of less quality

than for the panel, given that we rely on a small fraction of survey users who choose to provide us their Twitter handle. In addition, the results we show effectively focus on users that tend to have a high number of followers and followees, because these are more likely to follow or be followed by at least one other survey user. Therefore, this sample is probably not as representative of the population of US registered voters as the panel sample. However, the lower homophily by race/ethnicity is especially surprising, given the issues with race/ethnicity inference methods and the relatively high mismatch between survey and panel data (see Appendix Section 1). It is possible that the voter file inferred measure of race/ethnicity is more accurate for minorities that live in segregated neighborhoods, because of the usage of residential information. These populations may be more segregated on Twitter, leading to an overestimation of homophily patterns.

Focusing on the political information, homophily by party identification is relatively low, but it is higher when classifying users by candidate voted or intended to vote in the 2020 elections. In particular, homophily for Trump voters is higher than for republicans, as well as inferred republicans and registered republicans in the panel network (see Table 2). This points to some potential underestimation of republican voter homophily in the panel data, because the voter file data is from 2017, so a bit outdated. Even with this measure of political orientation, ideological homophily is still lower than homophily of african-americans and asians, or than homophily by age (a difference in age of 30 years multiplies the probability of a tie by 0.55, the same value as when going from a non-homophilous tie by voting preference to a Biden voter to Biden voter tie for the G4 network).

Finally, we provide in Appendix figure 10 row-normalized heatmaps of probabilities of a tie by education level and income, for the G3 network (they are qualitatively similar to the ones for the G4 network). We have not found any analysis of association between following behavior on Twitter and socioeconomic status in the literature, thus, we consider it valuable to provide our results here. We find a tendency of users to follow higher SES users, and relatively strong homophily for high SES individuals.



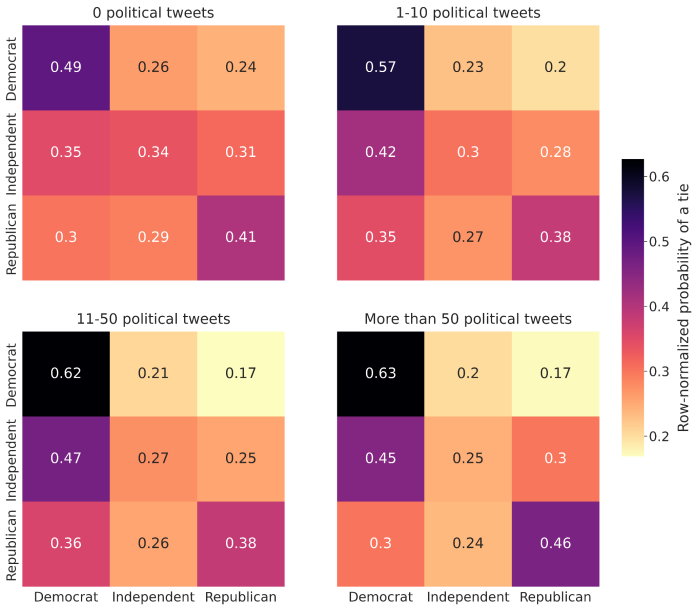
Appendix Figure 10. Heatmaps with row-normalized probabilities of a tie for education level and income in four bins, from the network of Covid States respondents with maximum degree 24. Color scales change for each panel.

Appendix Section 3: Partisanship homophily by election tweets

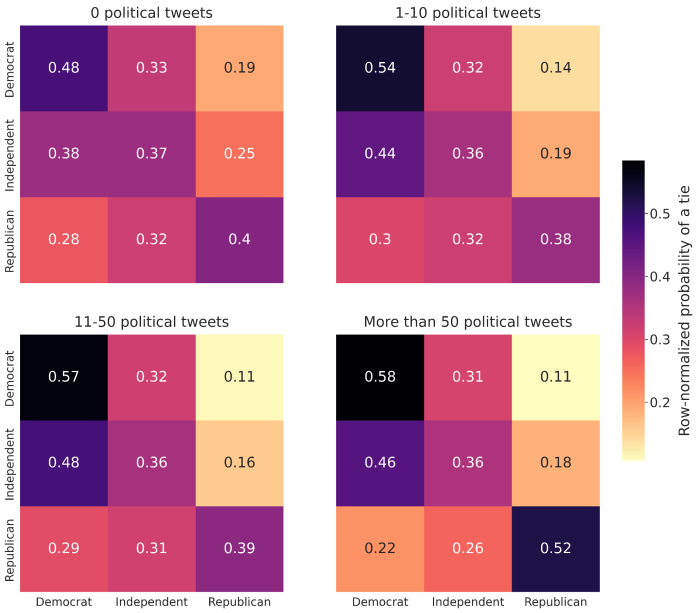
With the goal of examining if users interested in politics on Twitter display higher partisanship homophily levels, we built a classifier to detect political tweets. In particular, the classifier works very well to detect tweets about the 2020 U.S. presidential elections. We used the same keyword expansion approach utilized by prior work (Bakshy et al., 2015; Eady et al., 2019; Grinberg et al., 2019), and updated it for the 2020 U.S. Presidential election. The updating involved selecting high-specificity seed keywords that are likely to identify tweets about the U.S. election or politics more generally. Like prior work, we used a combination of general political keywords, hashtags, and candidate names to form our seed list. Then, we trained our classifier daily on a balanced set of political and non-political tweets (identified through a seed keyword list) to enable the classifier to identify additional words that co-occur with known political terms or figures.

We evaluated the political classifier using a stratified sample of 2065 tweets covering the entire study period and manual labeling by two raters on Amazon Mechanical Turk. Crowdworkers assigned the tweets into one of the following categories: (1) U.S. Presidential Election, (2) U.S. Politics, (3) Non-U.S. politics, (4) Other, or (5) I don't know. One of the authors resolved conflicts whenever they occurred. We find that the classifier can retrieve nearly all U.S. Presidential Election tweets with a recall of 96.4%. When collapsing categories (1) and (2) into one class of political content, we find that the classifier has a precision of 88.8% and a recall of 80.0%. These results are on par with the performance reported by Grinberg et al. (2019).

Using this classifier, we categorize users into 4 different bins following the number of political tweets they posted through 2020. Then, we calculate the probability of a tie from users with a given range of political tweets posted and of a given partisanship to each partisanship, regardless of the number of political tweets posted by the receiver of the tie. In other words, we compute inbreeding homophily levels by partisanship broken down by the number of political tweets posted by the sender of the tie. We display the result in row-normalized heatmaps in Appendix Figure 11 for inferred partisanship, and in Appendix Figure 12 for party registration.



Appendix Figure 11. Heatmaps with row-normalized probabilities of a tie by inferred partisanship, broken down by number of election tweets posted by the sender of the tie.



Appendix Figure 12. Heatmaps with row-normalized probabilities of a tie by party registration, broken down by number of election tweets posted by the sender of the tie.

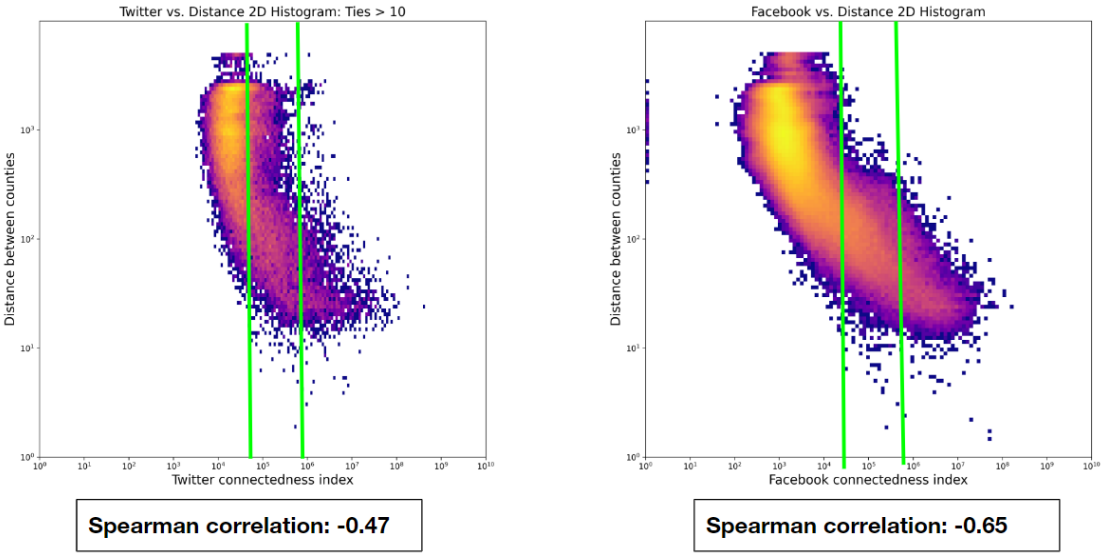
Appendix Section 4: Social Connectivity Index

Facebook measures the intensity of connectedness between locations, called connectedness index, using the friendship network. The facebook connectedness index of two locations (i,j) are derived from three variables: number of users in i, number of users in j, and friendship counts between i and j (cite **Facebook Data For Good Social Connectedness Index Methodology**). Similarly, we calculated the connectedness index for the Twitter reciprocal network. Our primary measure of Twitter Connectedness between two counties i and j is:

$$TwitterConnectednessIndex_{i,j} = \frac{Reciprocal\_Ties_{i,j}}{num\_users_i * num\_users_j}$$

Here, num_users i and num_users j are the number of Twitter users in counties i and j, and Reciprocal_Ties i,j is the number of reciprocal ties, meaning ties between users who follow each other, between the two counties.

After we derived Twitter connectedness for every two pairs of US counties, we made a scatterplot with connectedness as x-axis and distance between the two counties as y-axis, displayed in Appendix Figure 13. For comparison, we made a similar plot for Facebook connectedness which is available for download from Facebook. For both plots, we only show county pairs with at least 10 reciprocal ties in the Twitter data. We also calculated Spearman correlation for the two sets. As shown in the plots, both connectedness indexes are negatively correlated with distance, meaning that the closer the counties, the more friendship between them. Not surprisingly, Twitter is less stronger than Facebook in the correlation between connectedness and distance.



Appendix Figure 13: Heat scatter plots of Twitter connectedness index vs. Distance between counties (left) compared to Facebook connectedness index vs. Distance between counties (right). Note that we only show county pairs with at least 10 reciprocal ties in the Twitter network.