Online Appendix:

Political Homophily in a Large-Scale Online Communication Network

Further Validation with Author Flair

In the main text we describe the use of author flair to validate the ideological measure. To attempt to further validate the scale, we defined three subsets of users: (1) those who used any liberal term in their author flair (e.g., “liberal”, “democrat”, “progressive”), (2) those who used any conservative term in their author flair (e.g., “conservative”, “republican”, “trump”), and (3) those who used any moderate or centrist term in their author flair (e.g., “moderate”, “centrist”, “independent”). Figure A1 shows the distribution of ideological estimates for each of the three groups. The three distributions closely map onto expectations of where the three groups should be in ideological space – liberal users tend to be on the left, conservative users tend to be on the right, and moderate users are relatively central. Interestingly, conservative users show strong evidence of bimodality. This suggests some degree of fracturing within conservative politics, which makes sense given the nature of the 2016 U.S. presidential election (Smeltz et al., 2017). The results of our validation are similar to previous work using Twitter and Facebook to estimate ideological preferences using social media (Barberá, 2015; citation masked for review). In those instances, as here, the ideological estimates generally conform to expectations, but such data contain enough uncertainty that there is considerable overlap in the ideologies of liberals and conservatives.

Cross-Validation of Ideology Measure

To test whether the correlation in ideologies of those who interact in the subreddits we use for estimating ideology is sensitive to the fact that users must select into the subreddits used for estimation in order to interact, we used a cross validation technique. To do so, we randomly split the set of 101 subreddits into two groups (subreddit set one consisting of 50 subreddits and subreddit set two consisting of 51 subreddits). We then estimated the ideology of users using

each subset alone. First, this enables us to test the extent to which the ideology estimate was consistent when using different subsets of the data for estimation. Among users who contributed to a subreddit in both set one and set two (n = 106,613), the correlation in the estimated ideology is strongly positive (r = .28, p < .01). As would be expected, the ideology estimates from the subsets of the data are also highly correlated with the measure using the full set of subreddits reported in the main text. Among users who contributed to the first set of subreddits (n = 220,440), the correlation in ideology as estimated using the subset and ideology using the full set of subreddits is strongly positive (r = .37, p < .01). Similarly, among users who contributed to the second set of subreddits (n = 579,671), the correlation in ideology as estimated using the subset of subreddits and ideology using the full set of subreddits is strongly positive (r = 0.93, p < .01)[1]. The strong correlation in ideology using separate subsets of subreddits for estimation suggests that a common underlying process – likely selecting subreddits based on ideology – accounts for choices of where to submit on the site.

Second, we use the ideology estimates from one set of subreddits to estimate the correlation in ideology in author and commenter ideology in the other set. That is, we use set one to estimate ideology and observe the correlation in ideology between interacting users only in the second set, and vice versa. In this way, the data used for estimation is not the same as that used to observe the correlation in ideology between interacting users. When we estimate ideology using set one of subreddits, the correlation in ideology between interacting users in set two of the subreddits is positive and significant (r = .35, p < .01). When we estimate ideology using set two

[1] The substantial difference in the size of the set of users is driven in large part by the fact that one subreddit, The_Donald, has a very substantial user base (n = 443,741). Whichever half of the set of subreddits The_Donald is randomly assigned to has a much higher number of users than the other. This also likely explains the stronger correlation in ideology observed between the first and second sets of subreddits and the overall measure.

of subreddits, the correlation in ideology between interacting users in set one of the subreddits is positive and significant ($r = .46$, $p < .01$). Although the correlation coefficients differ from one another, and from the correlation reported in the main text ($r = 0.58$, $p < .01$), the substantive interpretation across all three estimates is similar. For all three, there is a positive and significant correlation indicating that there is substantial positive selection in conversation partners on Reddit. However, in all three cases the correlation is not so strong as to indicate that users do not interact with others with different political views.

As described below in the next section, some of the correlation we observe in the main text is due to the way in which the ideology measure is estimated. However, these results give us confidence that the correlation is, in substantial part, due to users selecting to interact with others who are relatively similar politically.

Correlation due to Measurement Strategy

As explained in the main text, we measure ideology using data on which subreddits users have posted to. We then use this measure to examine the similarity between interacting users. However, because of the way that Reddit is structured, users must post to the same subreddit in order to have an interaction. Therefore, it is possible that the similarity in ideology between interacting users is in part due to the measurement strategy rather than due to the selection of ideologically similar other users and subreddits to interact with.

To more fully understand the extent to which the measurement strategy might account for the observed correlation in ideological posting, we simulated a null model and compared our observed correlation to the distribution of null correlation estimates. In the null model, we kept all user interactions fixed. That is, if user *i* and user *j* interacted in the observed data, they would still interact in the null simulation. However, we randomly reassigned the subreddits in which

users interacted, keeping the number of interactions that took place in each subreddit fixed. In doing so, we kept fixed (a) the number of interactions each user had, (b) the number of interactions that took place in each subreddit, and (c) that each interacting pair of users would have to have posted to the same subreddit to have an interaction. In doing so, this model simulates a world very similar to the observed data, but in which users do not select subreddits based on ideology. We then estimated the "ideology" of users based on this data in the same way that we did on the observed data and calculated the correlation in ideology between interacting users using the observed set of interactions. The resulting correlation values represent the degree of correlation that is inherent to the measurement process.

We repeated the above process 1,000 times, creating a distribution of observed correlations in ideology in the null simulation. In all cases the correlations were positive and significant, with values between .01 and .07 (in all cases, $p < .01$). The correlation in ideology in the real data was .58 ($p < .01$). Figure A2 visually displays the distribution of the null correlation values from the simulated data and the observed correlation in the real data. Clearly, the observed correlation is substantially higher than the correlations from the null model. We tested for the statistical difference between each of the simulated null correlations and the observed correlation using the procedures outlined by Diedenhofen and Musch (2015). In all cases, we found that the correlation in the real world data was significantly greater than in the simulated null (in all cases, $p < .01$). This suggests that the measurement strategy does indeed account for some of the correlation in ideology between interacting users, but only a small proportion. To a much greater degree, users' selection into subreddits is based on ideology.

Auto-Correlation of Political Homophily

As mentioned in the main text, to further test H2 we performed an autoregression analysis of the correlation coefficients between a user's ideology and a commenter's ideology on each day. To do so, we created a daily time series of the correlation values throughout the 18-month period. To investigate the effect that events have on the correlation observed on a given day, we identified 62 days on which at least one significant event occurred during the time period. To identify events, we read Wikipedia's lists of significant events in the United States in 2016 and 2017 (Wikipedia, n.d.). These lists include events of many types, including politics, sports, music and entertainment, weather, business, crime, among others. We identified from the list events that were of a political nature and that we felt were likely to spur a national conversation about the event. The full set of events we identified are presented in Table A1. Using this data, we created two indicator variables. The first simply indicates whether an event occurred on a given date (*Political event*). The second indicates whether an event occurred on a given date or on the previous day (*Political event 2*), as the conversation surrounding an event often substantially occurs on the day after the event takes place.

We then tested autoregression models with lags of up to 8 days to identify which lags would provide the best model fit. The model with a one-day lag and a seven-day lag represented the best tradeoff between model parsimony and model fit. The results of the model with the inclusion of *political event* (measuring whether an event occurred on that day) included are presented in Table A2. The coefficient for a political event occurring on a day is negative and significant (B = -0.01, p = .01). This indicates that on days in which a political event occurs, the correlation is significantly lower than on days in which political events do not occur, controlling for the autoregression trend in the data. Similarly, the results of the model with the inclusion of

the *political event 2* (measuring whether an event occurred on that day or on the previous day) included are presented in Table A3. The coefficient for a political event occurring on a given day or the day before is negative and significant (B = -0.01, p < .01). This indicates that on days in which a political event occurs or has occurred on the previous day, the correlation is significantly lower than on days in which political events do not occur, controlling for the autoregression trend in the data.

As shown in the models, both the one-day lag and the seven-day lag are positive and significant predictors of the correlation in ideology. This makes sense, as there is a clear auto-regression in the data. Many processes are likely responsible for the autocorrelation from day to day. These include trends, such as the approaching election or holiday periods. However, processes specific to the site may also account for the day-to-day correlation, such as how the commenting patterns on a given day will influence the prominence of particular threads on the next day. Popular threads are displayed in more prominent places within a subreddit and, in exceptional circumstances, a thread may be displayed on the front page of the site due to its popularity. Therefore, if a thread becomes popular on one day due to high levels of commenting, it may be more likely to attract new commenters on subsequent days if its position on the site is due to its popularity. Future research may wish to investigate these trends in more detail.

Mechanisms for Decreased Correlation Surrounding Events

In the main text, we mention two possible pathways through which political events could engender increased diversity in conversation. First, it is possible that when political events occur, users who already regularly contribute to the subreddits we study diversify the areas on the site that they contribute to. For example, users who are primarily dedicated to contributing to a particular political subreddit may seek out a wider set of other subreddits to contribute to when

an event occurs, perhaps because conversations about the event will be taking place in multiple areas of the site.

H_4: On days in which political events take place, users contribute to a wider set of political subreddits than on days in which political events do not take place.

Second, it is possible that political events draw new users to participate in politics who have previously not participated. For many users, participation in political talk is likely not a primary goal of the use of Reddit, but when politically focusing events occur, such users may be drawn to political subreddits to make sense of the news events of the day. If so, these users may increase the diversity of conversation through their participation.

H_5: On days in which political events take place, users who have not previously posted in political threads are more likely to post in political threads.

To test these hypotheses, we performed additional analyses. To investigate H_4 we first took a subset of the users who had posted to our set of subreddits both on days in which a political event took place or the following day and on at least once on non-event dates. We then calculated the average number of unique subreddits a user posts to per day on days in which a political event took place or the day after a political event took place ($M = 0.011$, $SD = 0.008$) and the average number of unique subreddits a user posts to per day on all other days ($M = 0.005$, $SD = 0.004$). A difference of means test shows that on days in which a political event took place or the day after, the number of subreddits posted to per day is significantly higher than on days in which no political event took place ($t = 360.14$, $p < .01$). This suggests that when political events occur, users are more likely to post to a wider array of political subreddits than on days in which political events do not take place.

To investigate H_5 we calculated the number of users who were first-time contributors to the set of political subreddits on each day. We then compared the proportion of users who were first-time posters on days in which an event took place or the day after (.101) to the proportion of users who were first-time posters on all other days (.089). A difference of proportions test shows that the proportion of first-time posters is higher on days in which a political event took place or the day after ($\chi^2 = 2524.5$, $p < .01$). Similarly, we calculated the proportion for each day. That is, if on a given day 1,000 users posted and 300 of them had never before posted in the political subreddits, the proportion of first-time users would be .30. We then investigated the average proportion of first-time posters for days in which a political event took place, excluding January, 2016, as a very high proportion of these users are new users as we define them. That is, by definition on January 1, 2016 all users were first-time posters, and on subsequent days many users were first-time posters simply because it is the beginning of the period we study. We found that the average proportion of new users on days in which a political event took place of the day after ($M = 0.103$, $SD = 0.052$) was significantly higher than the average proportion of new users on days in which no political events took place ($M = 0.090$, $SD = 0.044$, $t = 2.41$, $p = .02$). These results suggest that on days surrounding political events, new users are drawn to converse about politics on Reddit.

Another possible explanation for the change in the ideological diversity of conversation surrounding events is that the ideological distribution of users who post on high correlation days is different than the ideological distribution of users who post on low-correlation days. In order to investigate this, we tested for the difference in the distribution of ideology for users who post on days with the lowest quartile of correlation values against the distribution of ideology of users who post on the highest quartile of correlations. These distributions are presented in Figure A3.

A Kolmogorov-Smirnov test was used to assess differences between the two distributions D = 0.146, $p < .01$. We note that Kolmogorov-Smirnov tests are sensitive to sample size, so with the large sample we have, it is likely that the test will identify even small differences between distributions. Although the distributions are significantly different, as Figure A1 shows the differences between them are subtle. It does not seem to be the case that a largely different set of users post on high homophily days. On high homophily days, there is a slightly higher proportion ideologically extreme users who post, but the difference is substantively small. This suggests that the users on these days have different behavior in terms of whom they interact with, rather than that they are substantially different in ideology.

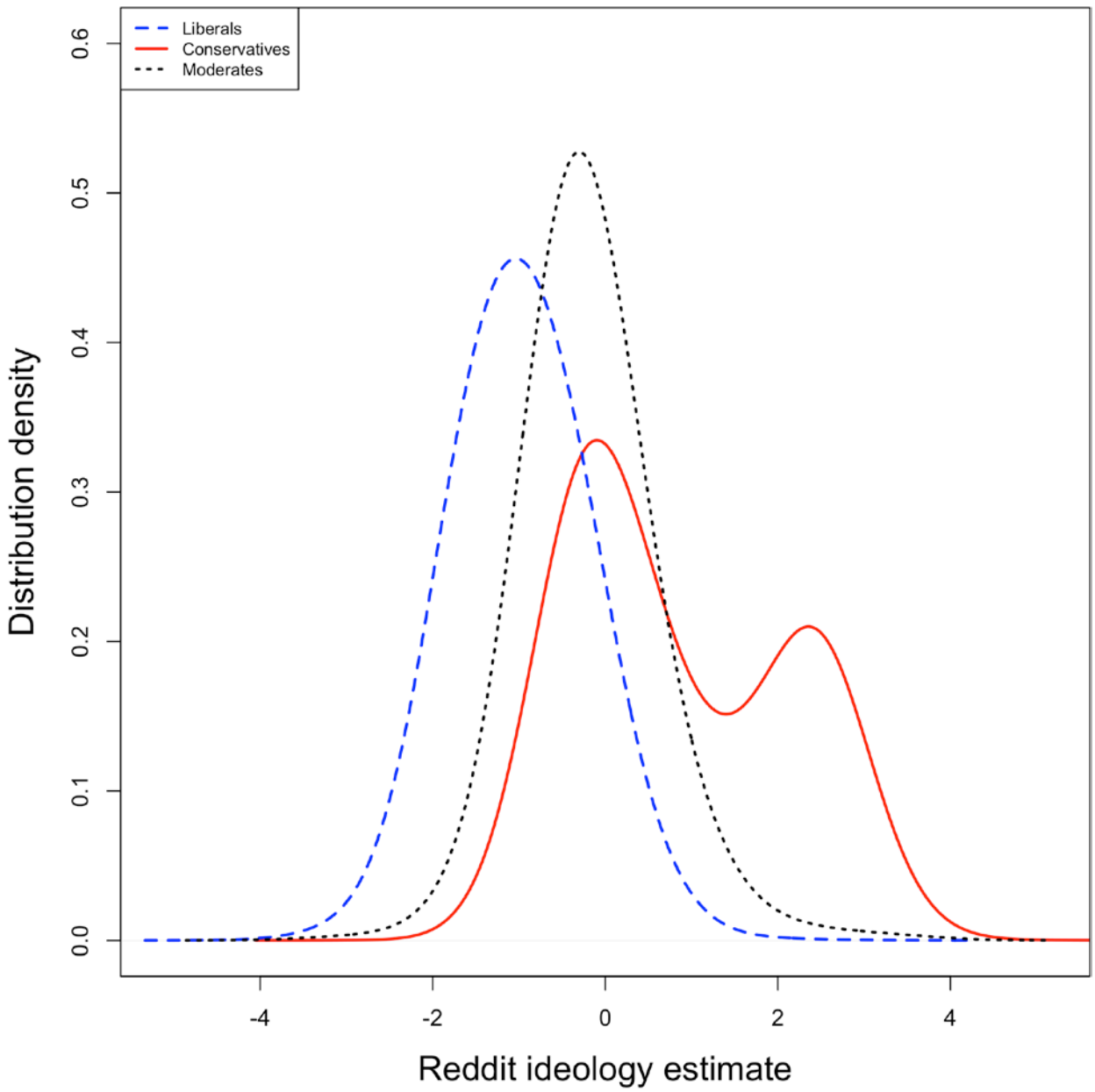

Figure A1. Distributions of estimated Reddit user ideology by self-defined political preference category (author flair).

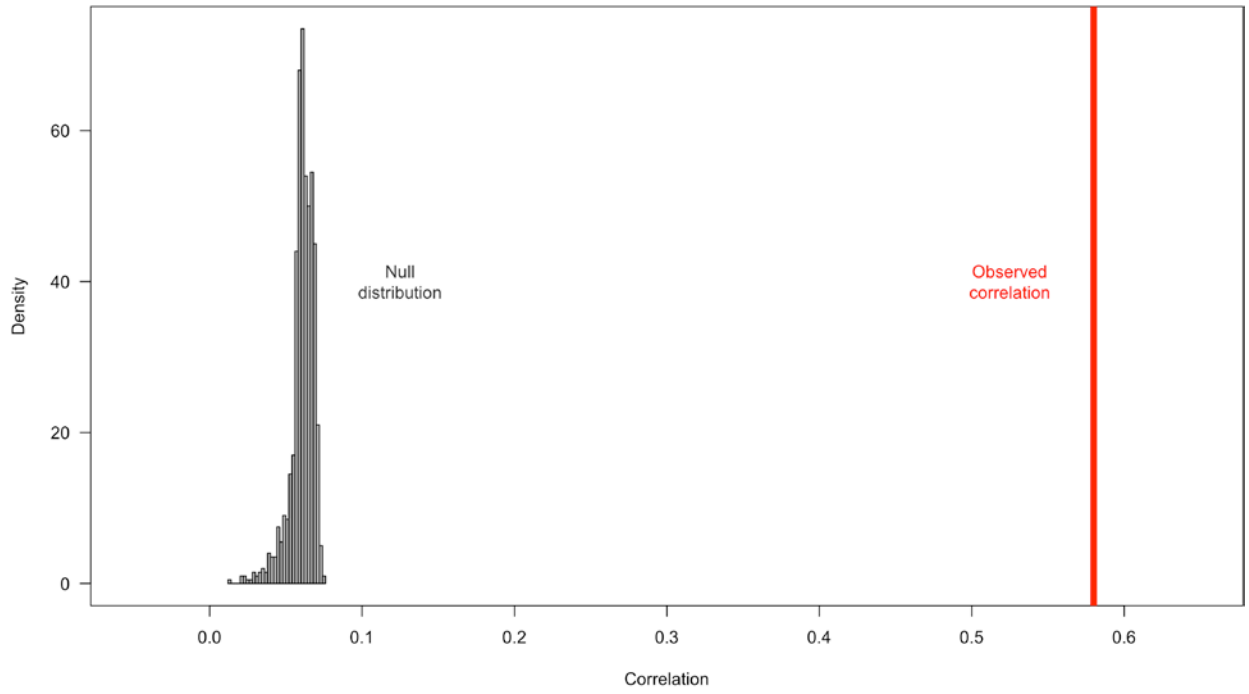

Figure A2. Distribution of correlation in ideology between interacting users in null simulation (grey) and observed correlation in ideology between interacting users in the real data (red).

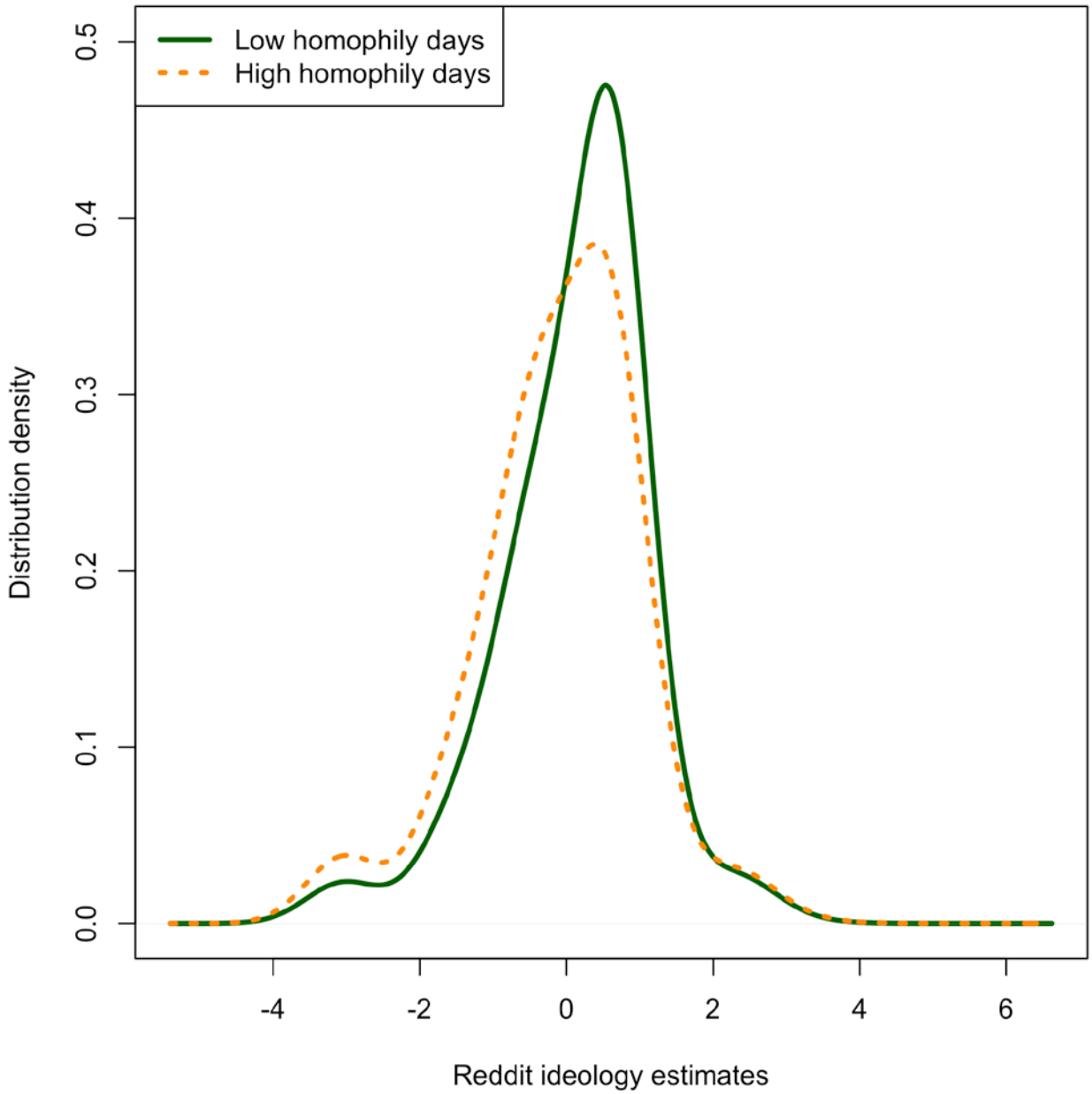

Figure A3. Distributions of ideology of posters on low- and high-homophily days.

Table A1

Political Events of Significant National Importance from January 2016 to June 2017

Date	Event
1/12/16	Obama's final State of the Union Address
1/16/16	Obama signs executive order to help Flint, MI with water crisis
2/1/16	Iowa Caucuses
2/9/16	New Hampshire primary
2/13/16	Supreme Court Justice Antonin Scalia dies
3/1/16	Super Tuesday
3/11/16	Trump cancels Chicago rally after fight breaks out between supporters and detractors
3/15/16	Super Tuesday 2
3/16/16	Merrick Garland nominated to Supreme Court
3/21/16	Obama flies to Cuba to talk with Castro
4/26/16	Super Tuesday 3
4/30/16	Obama's final correspondents' dinner
5/4/16	Kasich suspends campaign, making Trump presumptive nominee
5/25/16	State Dept Inspector General announces result of email server investigation
6/12/16	Orlando Pulse nightclub shooting
7/1/16	US military lifts ban on transgender people serving in the military
7/1/16	Attorney General Loretta Lynch announces that she will leave it up to the FBI to decide whether to bring charges against Hillary Clinton for her use of a private email server as Secretary of State
7/6/16	After FBI Director James Comey recommends against indicting Hillary Clinton, Attorney General Lynch announces that the federal investigation of Clinton will be closed with no charges
7/7/16	Sniper kills five police officers in Dallas during a Black Lives Matter protest
7/8/16	State Dept reopens investigation into Hillary Clinton's use of email server
7/18/16	Republican National Convention
7/19/16	Republican National Convention
7/20/16	Republican National Convention
7/21/16	Republican National Convention
7/25/16	Democratic National Convention
7/26/16	Democratic National Convention
7/27/16	Democratic National Convention
7/28/16	Democratic National Convention
9/11/16	Hillary Clinton faints at a 9/11 memorial service
9/26/16	First presidential debate
10/1/16	New York Times releases part of Trump's 1995 tax returns
10/4/16	Vice presidential debate
10/7/16	Obama administration accuses Russia of hacking the Democratic National Committee
10/7/16	Washington Post releases Access Hollywood tape of Trump
10/9/16	Second presidential debate

10/19/16	Third presidential debate
10/28/16	FBI Director Comey informs Congress that investigation into Hillary Clinton's use of email server
11/6/16	FBI Director Comey letter to Congress states newest investigation has not changed conclusions from July
11/8/16	US presidential election
11/18/16	Settlement announced in Trump University lawsuit
11/25/16	State election commission of Wisconsin announces recount
12/4/16	Man opens fire at Comet Ping Pong, a DC pizzeria due to conspiracy theory
12/9/16	CIA tells Congress it has "high confidence" Russia conducted operations to assist Trump to win presidency
12/19/16	Electoral college elects Trump
12/29/16	Obama administration announces sanctions against Russia in response to interference in 2016 presidential election. Trump urges country to move on from issue.
1/6/17	Declassified version of intelligence on Russian interference released
1/17/17	Obama commutes Chelsea Manning's sentence
1/20/17	Trump inaugurated
1/21/17	Women's march in opposition to Trump inauguration
1/27/17	Travel ban announced
1/31/17	Trump nominates Gorsuch to Supreme Court
2/3/17	Restraining order issued in the enforcement of travel ban
2/14/17	New York Times reports Trump campaign aides had repeated contacts with Russians
3/15/17	Revised travel ban blocked by federal judge in Hawaii
3/20/17	US House Permanent Select Committee on Intelligence holds hearing on Russian interference in 2016 election
3/30/17	Michael Flynn testifies to Congress about ties to Russia
4/5/17	Steve Bannon fired
4/22/17	March for Science
5/9/17	Comey fired by Trump
5/16/17	Trump accused of asking Comey to drop investigation of Flynn
5/16/17	Trump reported to have shared highly classified information with Russia
5/17/17	Mueller appointed as special counsel to investigate Russian interference in the 2016 election
6/8/16	Comey testifies before Congress about conversations with Trump concerning Flynn investigation
6/12/17	Ninth US Circuit Court of Appeals upholds decision blocking travel ban
6/14/17	House of Representatives Majority Whip Steve Scalise and his aides are hit by gunfire during a baseball practice in Virginia

Table A2

Regression of Daily Correlation in Ideology Between Interacting Users

	DV: Correlation in Ideology on Day *t*		
	B	*SE*	*T*
Intercept	0.03	0.01	3.60***
Political event	-0.01	<0.01	-2.50*
Correlation in Ideology *t*-1	0.72	0.03	24.37**
Correlation in Ideology *t*-7	0.22	0.03	7.62***
F		1,196***	
R^2		0.87	
N		539	

Note. * $p < .05$, ** $p < .01$, ***$p < .001$.

Table A3

Regression of Daily Correlation in Ideology Between Interacting Users

	DV: Correlation in Ideology on Day *t*		
	B	*SE*	*T*
Intercept	0.04	0.01	3.92***
Political event 2	-0.01	<0.01	-3.94***
Correlation in Ideology *t*-1	0.72	0.03	24.24**
Correlation in Ideology *t*-7	0.22	0.03	7.83***
F		1,219***	
R^2		0.87	
N		539	

Note. * $p < .05$, ** $p < .01$, ***$p < .001$.

References

2016_in_the_United_States. (n.d.). In *Wikipedia*. Retrieved August 21, 1018, from https://en.wikipedia.org/wiki/2016_in_the_United_States

2017_in_the_United_States. (n.d.). In *Wikipedia*. Retrieved August 21, 1018, from https://en.wikipedia.org/wiki/2017_in_the_United_States

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis, 23*, 76-91. doi:10.1093/pan/mpu011

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*(4): e0121945. doi:10.1371/journal.pone.0121945

Smeltz, D., Daalder, I. H., Friedhoff, K., & Kafura, C. (2017, October 2). What Americans think about America first. *The Chicago Council on Global Affairs.* Retrieved from https://www.thechicagocouncil.org/publication/what-americans-think-about-america-first