NBER WORKING PAPER SERIES

HOMOPHILY, GROUP SIZE, AND THE DIFFUSION OF POLITICAL INFORMATION IN SOCIAL NETWORKS: EVIDENCE FROM TWITTER

Yosh Halberstam
Brian Knight

Working Paper 20681
http://www.nber.org/papers/w20681

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2014

We are particularly indebted to Zack Hayat for getting this project off the ground and providing continual advice. We thank seminar participants at UC-Berkeley, CU-Boulder, Michigan State, Stanford, Toronto, the National University of Rosario, the Central Bank of Colombia and the 2014 Media and Communications Conference at Chicago-Booth. Ashwin Balamohan, Max Fowler, Kristopher Kivutha and Somang Nam jointly created the infrastructure to obtain the Twitter data used in this paper, and Michael Boutros helped design the MTurk surveys we used to analyze the content in tweets. Dylan Moore provided outstanding research assistance. Special thanks to Darko Gavrilovic, the IT consultant at Toronto, who facilitated the data work for this project, and Pooya Saadatpanah for providing computing support. We gratefully acknowledge financial support from the Social Sciences and Humanities Research Council of Canada. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter

Yosh Halberstam and Brian Knight

NBER Working Paper No. 20681
November 2014
JEL No. D7,D8

ABSTRACT

In this paper, we investigate political communications in social networks characterized both by homophily–a tendency to associate with similar individuals–and group size. To generate testable hypotheses, we develop a simple theory of information diffusion in social networks with homophily and two groups: conservatives and liberals. The model predicts that, with homophily, members of the majority group have more network connections and are exposed to more information than the minority group. We also use the model to show that, with homophily and a tendency to produce like-minded information, groups are disproportionately exposed to like-minded information and the information reaches like-minded individuals more quickly than it reaches individuals of opposing ideologies. To test the hypotheses of our model, we analyze nearly 500,000 communications during the 2012 US elections in a social network of 2.2 million politically-engaged Twitter users. Consistent with the model, we find that members of the majority group in each state-level network have more connections and are exposed to more tweets than members of the minority group. Likewise, we find that groups are disproportionately exposed to like-minded information and that information reaches like-minded users more quickly than users of the opposing ideology.

Yosh Halberstam
Department of Economics
150 St. George Street
University of Toronto
Toronto, ON M5S 3G7, Canada
yosh.halberstam@utoronto.ca

Brian Knight
Brown University
Department of Economics, Box B
64 Waterman Street
Providence, RI 02912
and NBER
Brian_Knight@brown.edu

1 Introduction

A long tradition of scholarship, starting with Black (1958), Downs (1957), and Becker (1958), has highlighted the key role of providing voters with a significant degree of information that is useful in the selection of high quality candidates and in the monitoring the behavior of politicians once in office. Further, it is often argued that information is most valuable to voters when it comes from different sources (Gentzkow and Shapiro, 2008). Naturally, if information is highly correlated across sources, then it will provide little value to voters. Taken together, these two points emphasize the importance of providing voters with an ideologically diverse set of high-quality information for a well-functioning democracy (Putnam et al., 1994).

At the same time, a literature investigating relationships between individuals has documented (a) a significant degree of homophily–a tendency to develop relationships with similar individuals– and (b) the fact that members of larger groups tend to have more relationships than members of smaller groups (Currarini et al., 2009; Marsden, 1987). To the extent that communication of political opinions and beliefs occurs within these social networks, then it is natural to hypothesize that members of larger groups will be exposed to more information than members of smaller groups. Moreover, if different types of individuals have different opinions and beliefs, then individuals may also be disproportionately exposed to like-minded information. For example, if conservatives tend to associate with other conservatives and have relatively negative opinions about Democratic politicians, then conservatives will be disproportionately exposed to negative opinions about Democratic politicians.

In this paper, we investigate the role of ideological homophily and group size in determining exposure to information on social media. Use of social media has grown dramatically during the past decade, with 60 percent of American adults and over 20 percent of the worldwide population currently using social networking sites (Rainie et al., 2012).[1] In terms of exposure to information on social media, 19 percent of all American adults reported regularly using social media as a source of news in 2012, a substantial increase from 2 percent just four years earlier. In addition to using social media to gather information, Americans also use social media to produce and transmit information. Indeed, new survey data released by the Pew Research Center show that half of social network users share or repost news stories, images or videos, while nearly as many discuss news issues or events on social network sites.[2] In particular, two thirds of American social media users, or 39 percent of all American adults, have engaged in some form of civic or political activity using social media, and 22 percent of registered US voters used social media to let others know how they

[1]The worldwide statistic comes from a June 18, 2013, report by eMarketer Inc. The link to the report is: http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976

[2]For more details, see: Pew Research Center, March, 2014, "State of the News Media 2014: Overview". The link to this report is: http://www.journalism.org/packages/state-of-the-news-media-2014/

voted in the 2012 elections.

Three key features distinguish social media from other forms of communication, including traditional media outlets and social interactions. First, social media allow users to not only consume but to also produce information, whereas the supply of information in traditional media markets is typically concentrated in the hands of a small number of outlets. Second, the information to which users are exposed depends upon self-chosen links among users. That is, users may be exposed to significantly different information depending on the set of individuals with whom they are connected and the content created by this set of individuals. Third, information on social media travels more rapidly and broadly than in other forms of social interactions. For example, a tweet from a user on Twitter is simultaneously transmitted to all of that user's followers, and each time one of these followers retweets this tweet, another set of followers is exposed to the information.

Given these differences between social media and other forms of communication and its growing role, we are motivated to examine how ideological homophily and group size influence the degree of exposure individuals have to information and its partisan composition. To this end, we develop a simple model of the diffusion of political information in a social network characterized by homophily and two groups: conservatives and liberals. The model predicts that, with homophily, members of the majority group have more network connections and are exposed to more information than the minority group. We also use the model to show that, with homophily and a tendency to produce like-minded information, groups are disproportionately exposed to like-minded information and information reaches like-minded individuals more quickly than it reaches individuals of opposing ideology.

We investigate these hypotheses using data from Twitter, one of the leading social media websites. As the model makes clear, measuring exposure to information on social networks requires data on both links between users and the content produced and transmitted by them. To measure links on Twitter, we begin by selecting politically-engaged users, defined as users who followed at least one account associated with a candidate for the US House during the 2012 election period. Among this population of over 2.2 million users, we identified roughly 90 million links. Using these links, we construct a single national network and 50 state subnetworks comprising users who follow candidates from the same state. To infer the ideology of these users, we use the political party of the candidates they follow. We also collected and analyzed nearly 500,000 retweets of candidate tweets as well as tweets that mention candidates. By combining the data on links and content, we are able to measure whether or not users are exposed to a given candidate tweet or mention via these political networks. Further, using the time associated with these retweets and the information on network connections, we measure the speed of information transmission.

We find that patterns of homophily in the political networks are similar to those documented in other social networks, such as the offline high-school friendship networks analyzed by Currarini

et al. (2009). We also find that that followers of the majority party have more connections than followers of the minority party. As predicted by our model, we next show that members of larger groups have more connections and are exposed to more tweets on a per-capita basis than members of smaller groups. Turning to exposure to like-minded information, we first show that a key condition of the model–production of like-minded information–is satisfied. Given this, we then show that groups are indeed disproportionately exposed to like-minded tweets, and that retweets of candidate tweets flow through the national network more quickly to like-minded users than to users of opposing ideology.

The paper proceeds as follows: After reviewing the relevant literature, we provide a simple model that yields the key hypotheses for our empirical investigation. Section 4 describes the data, Section 5 develops the empirical framework for measuring ideological homophily and ideological segregation, Section 6 presents the empirical results on network structure, and Section 7 examines communications within the network. Section 8 concludes, and discusses the implications of our findings.

2 Literature Review

This paper contributes to the literatures on homophily and group size in social networks, communication in social networks, and the role of the media sector in political mobilization and ideological segregation.

The first literature has documented that pairs of individuals with common characteristics are more likely to associate with one another than pairs of individuals with differing characteristics. This empirical regularity has come to be known as homophily, and applies to many different individual characteristics, including racial identity, gender, age, religion, and education (McPherson et al., 2001). Recent contributions focusing on group size and homophily include Currarini et al. (2009), who develop a theoretical model of network formation in which homophily can arise from both biases in preferences and biases in opportunities for meetings. They use their model to explain three empirical findings from analyses of friendship networks in high schools: larger groups have a larger fraction of same-type links, larger groups have more per-capita links, and the most extreme bias towards own-type relationships occurs for medium-sized groups. Marsden (1987) investigates similar issues in the context of advice networks, also finding that members of larger groups tend to have more connections. We build upon these studies by examining, in addition to network connections, the role of group size and homophily in the diffusion of information, in terms of both overall exposure and exposure to like-minded information.

Our paper is also related to a literature that has examined communications within networks, typically taking network structure as given. This literature is summarized in Jackson and Yariv

(2010). Most relevant to our research are two key papers on the role of homophily in communications. In particular, Golub and Jackson (2012) examine how network structure, and homophily in particular, impacts the speed of learning. The authors show that, in a model with average-based updating (DeGroot, 1974), homophily tends to slow convergence in beliefs across groups since it increases interactions within groups but decreases interactions across groups. By contrast, in a model of direct diffusion, homophily does not impact the speed of convergence since the average distance between individuals in the network is unaffected. Turning to exposure, in Jackson and Lopez-Pintado (2013) the authors explore how homophily influences whether or not an idea can spread throughout an entire network. Relative to these studies, our analysis combines the role of homophily in exposure to information and the tendency for individuals to be exposed to like-minded information.[3]

Third, there is also evidence that communications within social networks, and social media in particular, have real effects on offline behavior and collective action in particular. Most recently, communications on social media have been shown to precipitate protests during the Arab Spring (Acemoglu et al., 2014). Relatedly, previous research has shown that political mobilization is weaker among minority groups (Oberholzer-Gee and Waldfogel, 2005). At the same time, media outlets are less likely to cover issues for which demand among its consumer base is low (Oberholzer-Gee and Waldfogel, 2009). As a result, it has been suggested that politicians are more likely to target larger groups because the infrastructure for transmitting information to them has already been created by the media. We offer a new possible explanation that is based on the nature of interactions in social networks: political mobilization among minority groups is lower than among majority groups because they are exposed to less information through social interactions, independent of media markets.

Finally, our study is related to a literature that investigates the role of the media sector in ideological segregation.[4] Campante and Hojman (2013), for example, provide evidence that the introduction of television in the United States led to a decline in political polarization. In contrast, some have argued that the Internet is detrimental to democracy because it allows citizens to isolate

[3] Although not examining homophily specifically, Lerman and Ghosh (2010) find that information spreads on Twitter slower but farther compared to a social news site that is denser and more interconnected, suggesting a connection between the flow of information and network structure.

[4] More generally, there is also a related literature in economics on media bias and voter exposure to partisan information in the media. This literature has examined both the causes and consequences of media bias. Possible causes of media bias include consumer preferences (Gentzkow and Shapiro, 2010) and media ownership (Durante and Knight, 2012). Studies on the consequences of media bias have tended to focus on voting outcomes. DellaVigna and Kaplan (2007) document that the introduction of Fox News increased support for Republican candidates. George and Waldfogel (2003) document the impact of the entry of the New York Times on local political outcomes. Chiang and Knight (2011) document that surprising newspaper endorsements (e.g., those for Republican candidates from left-leaning papers) are more influential than unsurprising endorsements. Enikolopov et al. (2011) show that access to a partisan television station in Russia increased support for the party affiliated with the station.

themselves within "echo chambers", groups that share similar views and experiences (Sunstein, 2001). In a challenge to this view, Gentzkow and Shapiro (2011) document that users of online media (e.g., *www.nytimes.com*) are as segregated as users of traditional (i.e., offline) media and are less segregated than face-to-face (offline) social networks.[5] Our paper contributes to this literature by examining the quickly growing role of social media in contributing to ideological segregation.

3 Theoretical Model

This section develops a theoretical model of network structure under homophily and the diffusion of partisan information through this network. In particular, we consider the canonical Bass model of the diffusion of information but with two groups, conservatives and liberals, and biased interactions between these groups.

3.1 Network Structure

We first define the network and examine the role of homophily in terms of interactions. More formally, suppose individuals can be partitioned into two types, or groups, conservatives and liberals ($t \in \{C, L\}$). Group sizes are given by w_t such that $w_C + w_L = 1$, and, without loss of generality, assume that conservatives are the majority group and that liberals are the minority group ($w_C \geq 0.5$).

In any given period, two randomly-selected individuals of the same group interact with probability π_s and two randomly-selected members of different groups interact with probability π_d, and it will be natural to assume a bias in these interaction probabilities (i.e. $\pi_s > \pi_d$). Then, in any given period, a typical member of group t will have $\pi_s w_t$ same-type interactions and $\pi_d(1-w_t)$ different-type interactions. Then defining homophily for group t as the fraction of interactions with same type individuals, we have that:

$$H_t = \frac{\pi_s w_t}{\pi_s w_t + \pi_d(1-w_t)}$$

Note that this basic index does not account for the distribution of types in the populations. Specifically, if conservatives dominate the population and links are formed at random, then liberals would appear to be homophilous and conservatives would appear heterophilous. To address this issue, the literature has also focused on *relative homophily*. In particular, if the majority group has a higher degree of homophily, then the network is said to satisfy relative homophily. Also,

[5] In a recent working paper, Flaxman et al. (2013) use data on the browsing histories of Internet Explorer users to show that these individuals are more ideologically fragmented when they read articles, most of which are opinion articles, returned by search engines or on social media than when they read descriptive news on online media.

inbreeding homophily for group t is satisfied when $H_t > w_t$, and *heterophily* for group t is satisfied when $H_t < w_t$.

Given all of this, we have the following result with respect to group size and network structure.

Proposition 1: With biased interactions ($\pi_s > \pi_d$), an increase in the size of group t increases total network interactions for group t. Moreover, an increase in group size increases homophily for group t and thus relative homophily is satisfied. Finally, inbreeding homophily is satisfied.

To see the result regarding total interactions, note that total interactions are given by $\pi_s w_t + \pi_d(1-w_t)$, which is increasing in w_t so long as $\pi_s > \pi_d$. That is, since interactions are biased towards the own-group, an increase in group size leads to more total interactions. To see the result regarding homophily, note that an increase in group size increases same-type interactions but deceases interactions with the other group, leading to an increase in homophily. Finally, one can show that inbreeding homophily is satisfied when $\pi_s > \pi_d$.

Using these results, the relationship between group size and homophily is presented in Figure 1a, under the assumption of biased interactions. As shown, homophily is increasing in group size. Further, all groups experience inbreeding homophily as homophily is greater than baseline homophily for all groups.

3.2 Homophily, Group Size and the Diffusion of Information

Given these results with respect to network structure, we next consider the role of homophily in terms of how information flows through the network. We begin by considering the role of group size in exposure to information and then extend the model to two types of information, liberal and conservative, to examine the role of homophily in exposure to like-minded information.

In terms of the production of information, we consider a case in which each individual produces information with probability ε at time $\tau = 0$. Given our empirical application to the spread of information via retweets through Twitter, we abstract from the subsequent production of information after $\tau = 0$, coined the rate of innovation (p) in the original Bass model, and thus set $p = 0$ after $\tau = 0$.

We then consider how this information, once produced, spreads through the network. In particular, following the Bass model, we assume that, conditional on an interaction, previously exposed individuals transmit information to previously unexposed individuals with probability q. Following the Bass model, we define F_t^τ as the fraction of group t exposed to information at time τ. This is then linked to the fraction exposed at time $\tau - 1$ as follows:

$$F_t^\tau = F_t^{\tau-1} + (1 - F_t^{\tau-1}) f_t^\tau$$

where, $f_t^\tau$is the hazard rate, or the probability of group t exposure at time τ, conditional on not

being exposed at time $\tau - 1$:

$$f_t^\tau = qw_t\pi_s F_t^{\tau-1} + q(1-w_t)\pi_d F_{-t}^{\tau-1} - q^2 w_t(1-w_t)\pi_s\pi_d F_t^{\tau-1} F_{-t}^{\tau-1}$$

where $-t$ refers to the other group. In this expression, the first term represents the likelihood of being exposed to the information via the own group, the second term represents the likelihood of being exposed to the information via the other group, and the third term represents the likelihood of being exposed by both groups.

Then, we have the following result with respect to group size and exposure to information.

Proposition 2: With biased interactions ($\pi_s > \pi_d$), members of the majority group are exposed to more information than the minority group. That is, $F_C^\tau > F_L^\tau$for all times τ. In the absence of biased interactions ($\pi_s > \pi_d$), there will be no difference between majority and minority groups in exposure to information. Further, in the absence of differences in group size ($w_C = 0.5$), there will be no group-level difference in exposure to information.

While the proof is relegated to an appendix, we provide an overview of the basic intuition here. In particular, in the first period, total exposure to information for group t is given by:

$$F_t^1 = w_t\pi_s\varepsilon + (1-w_t)\pi_d\varepsilon$$

That is, a typical conservative is exposed to a fraction of other conservatives equal to $w_C\pi_s$ and to a fraction of liberals equal to $(1-w_C)\pi_d$. A similar logic applies to a typical liberal, and a comparison of these two groups shows that $F_C^1 > F_L^1$ so long as $w_C > 0.5$ and $\pi_s > \pi_d$. Having shown that the majority has higher initial exposure, the proof follows by induction, demonstrating that $F_C^{\tau-1} > F_L^{\tau-1}$ implies that $F_C^\tau > F_L^\tau$.

The logic behind Proposition 2 is presented in Figure 1b. As shown, when group sizes are equal, the relationship between the fraction of group t exposed to the information at time τ is the same and is given by the solid line for both groups, conservatives and liberals. The shape of the curve is identical to that in the standard Bass model, with an initial slow rise due to a small fraction of the population being exposed to the information, and thus a small fraction able to transmit, followed by a steep rise, and finally a tapering off as most of the population has already been exposed. Increasing the size of the conservative group and reducing the size of the liberal group leads to an upward shift in exposure for conservatives, due to the fact that they have more network interactions, and a downward shift in exposure for liberals, due to the fact that they have fewer network interactions. This leads to a disparity in exposure levels between the two groups for all times τ.

3.3 Homophily and Exposure to Like-Minded Information

In order to examine the role of homophily in exposure to information, we next extend the model to allow for two types of information, conservative and liberal. Let L_t^τ and C_t^τ denote the fraction of group t exposed to conservative and liberal information, respectively, at time τ and, as above, l_t^τ and c_t^τ represent the group t hazard rates for liberal and conservative information, respectively. In terms of the production of information of two types, we consider a case in which each individual produces like-minded information with probability ε_s and produces opposing information with probability ε_d at time $\tau = 0$.[6] That is, conservatives produce conservative information with probability ε_s and liberal information with probability ε_d. To the extent that partisan information is disproportionately produced by like-minded individuals, then it will be natural to assume that $\varepsilon_s > \varepsilon_d$. Given the focus on the overall role of homophily and our extension to two types of information, we simplify the model by abstracting from majority and minority differences and focus on a special case of the model with equally sized groups ($w_C = 0.5$). Then, we have the following result.

Proposition 3: With biased interactions ($\pi_s > \pi_d$) and the production of like-minded information ($\varepsilon_s > \varepsilon_d$), groups are disproportionately exposed to like-minded information. That is, $C_C^\tau > L_C^\tau$ and $L_L^\tau > C_L^\tau$ for all times τ. In the absence of either biased interactions or the production of like-minded information, groups are equally likely to be exposed to both conservative and liberal information at any point in time τ.

While the reader is referred to Appendix for a proof, we begin by showing that both groups are exposed to like-minded information in the first period:

$$C_C^1 - L_C^1 = L_L^1 - C_L^1 = 0.5(\pi_s - \pi_d)(\varepsilon_s - \varepsilon_d) > 0$$

Given this, we also show that a tendency to associate with similar members tends to reinforce these initial differences in exposure to like-minded information. If either $\pi_s = \pi_d$ or $\varepsilon_s = \varepsilon_d$, it is clear that there will not be initial differences in exposure rates.

Finally, we consider the implication of Proposition 3 for the speed of the transmission of information through the network.

Proposition 4: With biased interactions ($\pi_s > \pi_d$) and the production of like-minded information ($\varepsilon_s > \varepsilon_d$), average time to exposure is lower for like-minded information than for opposing information.

Since, as shown in Proposition 3, groups are more likely to be exposed to like-minded in-

[6] We have also considered an extension in which individuals may be more likely to transmit like-minded information at higher rates. That is, for the case of conservative information, it may be the case that transmission rates for conservatives (q_s) exceed transmission rates for liberals (q_d). This will tend to reinforce homophily, in the sense that own-type transmissions now occur with probability $w_t q_s \pi_s$ and different-type transmissions occur with probability $(1\text{-}w_t) q_d \pi_d$.

formation at any given time period, it then follows that average time to exposure to conservative information will be lower for same-type information than for opposing information.

To summarize, the model predicts that members of the majority group will have more network interactions, will have higher homophily, and will be exposed to more information on a per-capita basis. Extending the model to conservative and liberal information, we have that groups are disproportionately exposed to like-minded information and receive like-minded information more quickly than opposing information.

4 Data

To test these hypotheses, our study uses data from Twitter, an internet platform through which users connect and communicate with each other. We describe below the data on the political network, voter ideology, and political communications.

4.1 The Political Network

Our goal is to construct a network of politically engaged users of social media. Given this and lacking a direct measure of the ideology of Twitter users, we focus on Twitter users who follow politicians, defined here as candidates for the House of Representatives in 2012, and we use the party affiliation of these politicians to infer the ideology of the Twitter user. In November 2012, there were 825 candidates for the House, and we found 751 candidates with at least one Twitter account for a total of 976 candidate accounts.[7]

A comprehensive list of these candidate accounts was used to retrieve the set of Twitter users who followed at least one of the accounts on the list. In particular, on November 5th, one day before the 2012 election, we downloaded information on all 2.2 million Twitter users who followed a House candidate (henceforth, *voters*). These voters comprise our sample of Twitter users.

To construct the network, we use information on links among voters, and this process is depicted in Figure 2. In particular, we downloaded the list of followers of each of the 2.2 million voters.[8] Using these links, we construct a national network of politically-engaged Twitter users and, in some specifications, state-level networks based upon the state associated with candidates.

To provide a sense of the geographic distribution of these voters in the network, we examine

[7]Multiple accounts are especially common among incumbents, with one account serving as the official account and another serving as the campaign account. In addition, some politicians have personal accounts that are followed by voters.

[8]Following is unlike friendship or connections on other social media sites because the connection is not necessarily mutual. Except for protected accounts, users do not approve who follow them, and they do not need approval to follow other individuals.

user-supplied locations, which are provided by roughly one-quarter of voters.[9] Figure 3 plots the percent of Twitter voters from a given user-supplied state against the state's percent of US population. Remarkably, all states line up near the 45 degree line except for California, which has a lower share of voters relative to its share in the US population.[10] This finding suggests that our set of Twitter voters closely reflect the distribution of actual voters in the United States.

4.2 Voter Ideology

We further characterize voters as either liberal or conservative based upon the party affiliation of the candidates that they follow, and this process is depicted in Figure 4. In particular, voters who follow more Democratic than Republican candidate are coded as liberal, and voters that follow more Republican than Democratic candidates are coded as conservative. Given our desire to focus on two groups of voters, conservatives and liberals, we exclude voters who follow an equal number of candidates from the two parties. Among liberals and conservatives, we further distinguish is some specifications between extremists, voters who only follow candidates from one party, and moderates, voters who follow candidates from both parties.

To shed light on the validity of these measures of voter ideology and geography, we correlated our measures with survey responses from the latest Gallup State of the States political survey. In Figure 5a, we compare our estimate of the share of liberals in each state, using both the user-supplied location and the inferred ideology measures, to the share of liberals in each state in the Gallup survey. As shown, our estimates for the liberal share of voters in each state are positively correlated with the Gallup measure, and most states line up close to the 45 degree line.

As further evidence on our proxies for ideology, we have also downloaded information on Twitter accounts associated with significant media outlets and computed the fraction of liberal voters following each media outlet.[11] Using this information, Figure 5b plots, for the 25 outlets with the most followers in our sample of voters, the likelihood that a liberal voter follows a given outlet, relative to the likelihood that a conservative voter follows the same outlet. As shown, media outlets and programs traditionally considered to be right leaning, such as Rush Limbaugh, The Hannity Show, and Fox News, have very low likelihood ratios. On the other hand, media outlets and programs traditionally considered to be left-leaning, such as the New York Times and the Rachel Maddow show, have a likelihood ratio in excess of one. These results are also broadly

[9] While these location entries vary in specificity and format, we have used a simple procedure for inferring a user's state from the information he or she supplies, with a focus on two letter postal codes or full state names.

[10] The point above the reference line accounting for nearly zero percent of US population is Washington D.C.

[11] In particular, we downloaded followers of Twitter accounts associated with significant network television outlets and shows (as defined by journalism.org), significant cable television outlets and shows (as defined by journalism.org), the top 10 newspapers in terms of national circulation (as defined by www.stateofthemedia.org), the top 10 talk radio hosts in terms of the number of listeners (as defined by www.stateofthemedia.org), and the top six political blogs (as defined by http://technorati.com/blogs/directory/politics/ (accessed September 19, 2012).

consistent with the measures of media bias developed by Groseclose and Milyo (2005), who find the New York Times as one of the most left-leaning outlets and Fox News as one of the most right-leaning. In summary, these results suggest that our measures of voter ideology are reasonable and do capture some underlying measure of political preferences.

Finally, there is also support at the individual level for the validity of our ideology measure. Using information on voter registration history, Barberá (2013) matches a sample of voters from Ohio, a state that requires party registration to vote in a primary election, to their Twitter accounts and finds a strong correlation between party registration and the parties these voters follow on Twitter.

4.3 Political Communications

To examine how partisan information flows through the network, we have collected information on tweets associated with candidate accounts and retweets of these candidate tweets by voters. We also collected information on mentions of candidates by voters. We focus on the candidate tweets and mentions produced during a six-week window centered around the 2012 Election Day: October 15 through November 28.

During this time period, House candidates produced over 22,000 unique tweets, with roughly 64 percent coming from Republican accounts and 36 percent from accounts associated with Democratic candidates. These candidate tweets were retweeted over 167,000 times by over 70,000 different voters. For mentions, we have over 308,000 mentions of candidates by voters, with 74 percent mentioning Republicans and 26 percent mentioning Democrats.[12]

Turning to the speed of information transmission, we calculate the time associated with a given voter being exposed to a given candidate tweet, and time is normalized so that it equals zero for the first retweet. Using these measures, the average time to exposure is 102 minutes.

5 Empirical Framework

Based upon these Twitter data, we use the network structure to develop measures of the degree of homophily and ideological segregation. Then, using network structure and communications within the network, we develop measures of the exposure of voters to information.

[12]For mentions of multiple candidates, we focus on the party with the most candidates mentioned and exclude cases in which a mention focused on an equal number of candidates from the two parties.

5.1 Measures of Homophily in Social Networks

For measures of homophily, we follow Currarini et al. (2009). Let I be the total number of voters and I_t be the total number of type t voters. With two groups, conservatives and liberals, we have that $I = I_C + I_L$. Then, $w_t = \frac{I_t}{I}$ is the fraction of type t in the voter population. Let v_{it} denote the number of type t voters followed by voter i. Then $s_t = \frac{1}{I_t}\sum_{i \in I_t} v_{it}$ denotes the average number of type t voters followed by type t voters (same) and $d_t = \frac{1}{I_t}\sum_{i \in I_t} v_{i-t}$ denotes the average number of non-type t voters followed by type t voters (different). With these in hand, we can then define the homophily index for type t voters is as follows:

$$H_t = \frac{s_t}{s_t + d_t}.$$

This index measures the proportion of type t connections that are with voters of the same type t. We then compare this to baseline homophily ($H_t = w_t$), which occurs under the assumption of random links between voters. To examine the relationship between group size and overall connections, we will also use the measure of connections per capita, $s_t + d_t$, for group t.

5.2 Measuring Ideological Segregation

For comparison with existing measures of ideological isolation in different settings, we also compute the isolation index following Gentzkow and Shapiro (2011). This measure has been developed by White (1986) and Cutler et al. (1999), and widely applied to study ethnic and urban segregation.

For each voter $j \in J$, let v_{jC} denote the number of conservative followers and v_{jL} the number of liberal followers. We can then define the *share conservative* of voter j as the fraction of his or her followers who are conservative:

$$share\ conservative_j = \frac{v_{jC}}{v_{jC} + v_{jL}}.$$

We can then define conservative exposure for each voter i as follows:

$$conservative\ exposure_i = \frac{1}{\sum_{j \in J} \phi_{ij}} \sum_{j \in J} \phi_{ij} \times share\ conservative_j,$$

where $\phi_{ij} \in \{0, 1\}$ as an indicator equal to one if voter i follows voter j. Taking averages across voters within groups, we then have conservative exposure for conservatives and conservative exposure among liberals. With these in hand, the isolation index is given by:

$$isolation = conservative\ exposure_C - conservative\ exposure_L,$$

where $conservative\ exposure_t = \frac{1}{I_t}\sum_{i\in I_t} conservative\ exposure_i$.

This index varies between 0 and 1 and captures the degree to which conservatives, relative to liberals, have a greater tendency to follow voters whose other followers are conservative. As the index increases, both groups become increasingly isolated from each other, as measured by a shrinking share of voters who have both conservative and liberal followers.

Comparing these two measures, homophily captures within-group attributes, while segregation measures captures cross-group attributes. That is, homophily captures the tendency of voters to link to voters with the same ideology. Segregation captures the degree to which voters of different types have distinct patterns of exposure, in terms of the set of voters they follow. The more these sets of followers are distinct from one another, the more isolated are the voter types from each other.

5.3 Measuring Exposure to Information

We next develop measures of exposure to like-minded information and isolation in exposure to information. Let ε_{is} denote the total number of same-type tweets (or mentions) to which voter i is exposed. Then $\varepsilon_{ts} = \frac{1}{I_t}\sum_{i\in I_t}\varepsilon_{is}$ denotes the average number of same-type tweets to which voters of type t are exposed (same) and $\varepsilon_{td} = \frac{1}{I_t}\sum_{i\in I_t}\varepsilon_{id}$ the average number of different-type tweets to which they are exposed (different). We next define the exposure index paralleling the homophily index. In particular, the *exposure index* for type t voters is as follows:

$$E_t = \frac{\varepsilon_{ts}}{\varepsilon_{ts}+\varepsilon_{td}}.$$

For comparison purposes, we next define *baseline exposure* as follows:

$$\varepsilon_t = \frac{\sum_{i\in I}\varepsilon_{it}}{\sum_{i\in I}\varepsilon_{it}+\sum_{i\in I}\varepsilon_{i-t}}$$

This is equal to the share of type t tweets to which all voters are exposed.

Recall that, in the absence of homophily, the production shares $\frac{\varepsilon_s}{\varepsilon_s+\varepsilon_d}$ determine the composition of partisan exposure, which is group invariant. We approximate these shares using our baseline measure, ε_t. Thus, if $E_t > \varepsilon_t$ then this would be evidence that homophily plays a role in partisan exposure. The larger the exposure index is relative to baseline exposure, the greater the bias in exposure to same-type information due to homophily. Finally, to measure the relationship between group size and total exposure to information, we will use the measure of tweets per capita, $\varepsilon_{ts}+\varepsilon_{td}$, for group t.

6 Results on Network Structure

Using the data described in Section 4 and the measures developed in Section 5, we next present our results on network structure. We begin by describing our results on homophily and segregation at the national level before turning to results at the candidate state level.

6.1 National Political Network

In Table 1, we first display the ideological composition of voter followees as a function of the ideology of the voter. While liberals account for 36 percent of voters, 67 percent of their followees are liberal, with just 33 percent conservative. Likewise, conservative voters make up 64 percent of the sample, and 80 percent of their followees are also conservative, with just 20 percent liberal.

Using these measures, Table 2 provides estimates of homophily at the national level. The shares of each group in the population are identical to those in Table 1. As shown in the first row, 69 percent of followees of liberals are also liberal. This follows from Table 1, which shows that liberals have 40 same-type links and 59 combined links, both on a per-capita basis. For conservatives, homophily equals 84 percent, as they have, on average, 58 links to conservatives out of 68 links across both conservatives and liberals. Relative homophily thus holds since homophily is higher for the larger group, conservatives in this case. Likewise, inbreeding homophily is satisfied for both groups since the homophily index, as shown in the final column, exceeds the population share for both groups.

We next examine how isolation in our network compares to that in other settings, including face-to-face social interaction and traditional media outlets. In the final three columns of Table 2, we report conservative exposure estimates for liberals and conservatives at the national level. As shown, conservative exposure among conservatives is 0.776, and conservative exposure among liberals is 0.372, implying an isolation index of 0.403.

Note that this result differs from those in Gentzkow and Shapiro (2011), who find that the internet is surprising unsegregated along ideological lines, with a baseline estimate of segregation equal to 0.075. While their estimate of ideological segregation on the internet is close to their estimated ideological segregation for traditional media outlets, our estimated segregation places social media on par with those associated with face-to-face interactions with political discussants, the second most segregated environment studied in Gentzkow and Shapiro (2011).

To attempt to reconcile these two sets of findings, high segregation when examining links on Twitter and low segregation when examining consumption of news on the internet, we next examine two differences between these studies. First, it is plausible that our sample, constructed by selecting users who follow politicians, may tend to disproportionately include individuals with strong preferences for linking to like-minded users. To investigate this issue, we split our sample

of voters into extremists, those who follow candidates from only one party, and moderates, those who follow both parties. As shown in Table 2, and consistent with the view that extremists have stronger preferences for linking to like-minded users, we find that both homophily and segregation are higher for extremists than for moderates. Second, we use information on the followers of our sample of media outlets described above to compute segregation in media consumption on Twitter. Note that this measure does not use any information on links between voters, and, instead, we treat users as if they are only consuming information from media outlets on Twitter. As shown in Table 3, isolation in media consumption (0.241) for our sample of voters is significantly higher than the measures in Gentzkow and Shapiro (2011) but is significantly lower than our network-based measure of isolation, which equals 0.394 in this subsample of voters. Thus, these same Twitter users experience lower segregation when consuming news from media outlets on Twitter than when using Twitter as a social network. Finally, we combine these two approaches by computing isolation in media consumption for moderates. As shown, segregation in media consumption for moderates equals 0.067, which is on par with the measure in Gentzkow and Shapiro (2011). Taken together, these results suggest that the differences in results between these two studies may be driven by differences between both the activities and types of individuals under consideration.

6.2 State Political Networks

Moving next to sub-networks at the state level, we investigate several characteristics of the political network we have constructed. In particular, using our state-level homophily estimates, we investigate whether our data support the predictions in Proposition 1. In particular, we investigate whether: (a) larger groups form a larger share of their friendships with people of their own type, (b) groups inbreed and (c) larger groups form more friendships per capita.

Using variation in group size across candidate states, Figure 6a plots the homophily index for each type against their share in the population. Each point in this figure is an ideological group at the state level. As shown, almost all observations lie above the 45 degree line, implying that inbreeding homophily is satisfied. Thus, our results support the prediction that groups inbreed. Also, consistent with the prediction of the model, homophily is broadly increasing in group size. We have also verified that, in every state, homophily is larger for the majority group; thus relative homophily is also satisfied. Again using state-level variation, Figure 6b presents scatter plots of followees per capita for each group against the group's share in the population. The linear fit is presented as well to show the general trend. As shown, a move from 0 to 1 in the share of the population increases the number of followees per capita from about 40 to 60. Thus, our data are also consistent with the prediction that larger groups have more followees per capita.

7 Results on Network Communications

Having documented evidence of network structure consistent with the model and the existing literature on homophily, we next examine how information flows through this political network. That is, as a result and homophily in the network, do members of larger groups receive more information, are voters disproportionately exposed to like-minded content, and, conditional on exposure, does political content reach like-minded voters more quickly?

7.1 Production of Information

Before turning to exposure to information, we first examine the degree to which users disproportionately produce like-minded information, a key condition in the model for exposure to like-minded information. As shown in Table 4, there is strong correlation between voter ideology and candidate party in the production of retweets. In particular, 91 percent of retweets of tweets by Democratic candidates are produced by liberal voters, and almost 99 percent of retweets of tweets by Republicans candidates are produced by conservative voters. While this may reflect a preference for producing like-minded information, it may also reflect the transmission mechanism, through which voters retweeted the tweet after being exposed via another voter. That is, due to homophily, it may be that liberal voters are disproportionately exposed to tweets from Democratic candidates via other liberal voters and likewise for conservative voters and Republican candidates. To address this issue, we next focus on the first retweet of a candidate tweet by a voter in our network. In this case, voters could not have been previously exposed to the tweet via another voter. As shown, a strong correlation between voter ideology and candidate party remains in the production of first retweets, with 86 of retweets of tweets by Democratic candidates produced by liberal voters, and almost 98 percent of retweets of tweets by Republicans candidates produced by conservative voters. Finally, we examine the production of mentions, and, as shown, 66 of retweets of mentions of Democratic candidates are produced by liberal voters, and 77 percent of mentions of Republicans tweets are produced by conservative voters. One possible difference between retweets of candidate tweets and candidate mentions involves sentiment. In particular, since candidates control the sentiment of tweets but voters control the sentiment of mentions, it is possible that some mentions of Democrats by conservative voters have negative sentiment and hence can be considered to have conservative content and likewise for mentions of Republicans by liberal voters. We return to this issue of sentiment later in this section.

7.2 Communications in the National Political Network

Having established that production of information is like-minded in nature, we next present our measures of exposure to like-minded information in terms of tweet exposure, retweet exposure, and exposure to mentions, all at the national level. In particular, we develop analogues to our homophily measures based upon the exposure to tweets and retweets from like-minded sources (i.e conservative voters and Republican candidates and liberal voters and Democratic candidates). As shown in Table 5, among voters exposed to at least one tweet, liberal voters are exposed to around 58 tweets on average, and 52 of those, or roughly 90 percent, originate from Democratic accounts. Likewise, exposure to like-minded information for conservative voters is also 90 percent, with 63 out of 70 tweets originating from Republican accounts. Were voters exposed randomly to tweets, liberal voters would have a like-minded exposure index of 48 percent, and conservatives would have a like-minded exposure index of 51 percent. Note also that these exposure measures of 90 percent are even larger than those in Table 2, which are based upon links between voters, suggesting that communication serves to amplify an already significant degree of homophily in the network structure.

We next turn to exposure to like-minded information based upon retweets, which account for multiple exposures to the same candidate tweet. Given that the Twitter interface separately identifies all of the retweeters of a single tweet, it is natural that a candidate tweet may be more influential when a voter is exposed to retweets from multiple accounts. As shown in Table 5, exposure to like-minded information is even larger (92 percent for liberal voters and 93 percent for conservative voters) when measured using exposure to retweets. Were voters exposed randomly to retweets, liberal voters would have an index of exposure to like-minded information of 31 percent and conservative voters would have an index of 69 percent. Comparing the index based upon the tweets to the index based upon retweets, the measures based upon retweets are somewhat larger. This is presumably due to the fact that, conditional on being exposed to a tweet, the number of retweet exposures is higher for tweets from like-minded sources (i.e. liberal voters and Democratic candidates and conservative voters and Republican candidates).

Results using data from candidate mentions are provided at the bottom of Table 5. As shown, among exposure to mentions for liberal voters, 39 percent are mentions of Democratic candidates, and, among exposure to mentions for conservative voters, 84 percent are mentions of Republican candidates. While these results are also consistent with voters being exposed to like-minded information, the patterns are less strong than those regarding candidate tweets and retweets. One natural explanation for this difference, as noted above, is that the production of mentions is less like-minded in nature than the production of tweets and retweets.

Finally, we consider measures of speed, or time to exposure, in the flow of information through the network at the national level. In particular, the model predicts that information may reach like-

minded users more quickly. As noted above, we measure speed as, conditional on exposure, the number of minutes that it takes for a voter to be exposed to a tweet, where the time associated with the first retweet is normalized to zero, and the unit of observation in this analysis is the at the level of the candidate tweet and exposed voter. To test this hypothesis, we first run a linear regression with minutes to voter exposure to a given candidate tweet as the dependent variable. In this regression, we control for a set of tweet fixed effects, which incorporates candidate party, an indicator for liberal voters, and an indicator for a mismatch between voter ideology and candidate party (i.e. indicating either a Republican candidate tweet and a liberal voter or a Democratic candidate tweet and a conservative voter). As shown in Table 6, liberal voters are exposed to tweets more slowly than conservative voters on average and, more interestingly, a mismatch between voter ideology and candidate party is associated with an increase in time to exposure of almost 10 minutes, representing a roughly 10 percent increase when compared to the sample average of 102 minutes. To provide results in percentage terms, we next run a similar regression but with the natural log of minutes as the dependent variable.[13] As shown, a mismatch between voter ideology and candidate party is associated with a 14 percent increase in time to exposure. Finally, we estimate a Cox survival model, again with candidate tweet fixed effects. As shown, a mismatch between voter ideology and candidate party is associated with a decrease in the likelihood of exposure, conditional on not being previously exposed, in any given time period. Note that a decrease in the likelihood of exposure is associated with an increase in expected time to exposure, and thus the results are consistent with those using linear regressions. In summary, and consistent with the predictions of the theoretical model, this section provides evidence that, in social networks characterized by homophily and the production of like-minded information, users are exposed to like-minded information more quickly than they are exposed to information of opposing ideology.

7.3 Communications in the State Political Networks

Turning to political communications within state-level networks, we present our findings on the role of group size in overall exposure to information on a per-capita basis and exposure to like-minded information. In the former, we investigate whether group size influences how much information voters obtain. That is, given that members of larger groups have more connections per capita, do these members of larger groups also receive more information on a per-capita basis? In the latter, we examine whether our findings on ideological homophily in connections extend to the ideological composition of communications to which voters are exposed.

In Figure 7, we investigate how exposure to total information varies with group size. In panel a), we show the relationship between per-capita retweets and group size, and, in panel b), the

[13] In this specification, we add one minute to all times in order to address the issue of immediate exposure, or zero minutes.

relationship between per-capita mentions and group size. Consistent with the model, exposure to information, in terms of both retweets and mentions, increases with group size. For example, a one standard deviation increase in group size is associated with a 10 percent increase in exposure to retweets and 19 percent increase in exposure to mentions.

The measure of exposure to like-minded information is similar to the one used at the national level. We examine communications within a state network using data on tweets produced and received by voters in the state network. For the liberal group, for example, exposure is measured by the share of retweets received that originate from (or mention) Democratic candidate accounts. Baseline exposure is then defined as exposure for a voter that is randomly exposed to tweets produced in his state network. We illustrate the connection between exposure to like-minded information and this baseline measure in Figure 8. In panel a), we show this relationship using retweets. As shown, in all states, and for both conservative and liberal voters, exposure to like-minded information exceeds baseline exposure. A second notable pattern is the positive relationship between exposure to like-minded information and baseline exposure. In particular, increasing the production of like-minded information results in higher exposure to like-minded information, as predicted by the model. This relationship is analogous to the relationship between H and w, and we find that *relative* exposure holds in the same sense that relative homophily holds. In panel b), we plot the same relationship using mentions data. We find that exposure to like-minded mentions increases in their share produced and exceeds baseline exposure. Yet, unlike retweets, exposure to mentions is significantly less biased towards like-minded information. It is tempting to interpret the difference between mentions and retweets as resulting from a more limited effect of homophily on the production rather than transmission of information. However, production bias in mentions is also narrower that in retweets, suggesting that other differences between mentions and retweets may driving the wedge in the exposure index.

Finally, in Figure 9, we examine the relationship between group size and the ratio between exposure and homophily (E/H). Focusing on retweets, we first note that the ratio E/H is strictly decreasing in group size. In other words, a marginal increase in group size has a diminishing effect on voter exposure to like-minded information relative to same-type connections. The trend for mentions is similar but less pronounced than for retweets. In general, rates of homophily and exposure to like-minded information are highly correlated as implied by the narrow range of values that E/H takes around one, and this is particularly true for retweets.

To summarize, our results suggest that group size influences both the degree and type of communications within social networks characterized by homophily. Importantly, majority and minority group members have distinct patterns of interactions and communications. The majority is more homophilous and has higher exposure to information in general and to like-minded information in particular.

7.4 Content Analysis

In this section we examine heterogeneity in our data on communications according to the content of the tweet. For retweets of candidate tweets, we distinguish between political and non-political information and investigate whether the patterns of exposure to like-minded information and the speed of diffusion differs between these two types of information. For mentions, we investigate differences between positive mentions of candidates and negative mentions since, as noted above, it is natural that mentions by conservative voters, for example, of Republican candidates might tend to be positive and mentions of Democratic candidates might tend to be negative. We categorized the full set of candidate tweets and, given the large sample, a 10% random sample of candidate mentions.[14]

Starting with the production of information, we find some evidence in Table 7 that production of political information tends to be more like-minded than the production of non-political information, although some of the differences are small in magnitude. In particular, while liberal voters account for over 92 percent of retweets of political tweets by Democratic candidates, they account for less than 85 percent of retweets of non-political tweets. Differences for retweets of tweets by Republican candidates are small, with conservative voters accounting for almost 99 percent of political information and 98 percent of non-political information. Turning to mentions, we do find significant differences in the like-minded production of mentions depending upon whether the mention was positive or negative. For example, while liberal voters are responsible for a majority of positive mentions of Democratic candidates, conservative voters are responsible for a majority of negative mentions.

In Table 8, we next investigate whether these differences in production translate into differences in exposure. In panel a), we do find that exposure to like-minded information is higher for political tweets than for non-political tweets. In terms of magnitudes, however, the differences between political and non-political retweets are relatively small. The fact that differences are small is not surprising given, as noted above, that differences between political and non-political production of retweets is relatively small.

Turning to mentions, as shown in panel b), we find significant differences in homophily and segregation between positive and negative sentiment, with high exposure to like-minded information for positive sentiment mentions and low exposure for negative sentiment mentions. This is consistent with the fact, as documented in Table 8, that the production of mentions is more like-minded in nature for positive mentions, whereas, for negative mentions, voter ideology and

[14]To classify candidate tweets, we designed two surveys on MTurk that asked workers to categorize our sample. The workers were asked to choose one of three responses to each tweet that we presented, where indifference was the third category. In the survey aimed to distinguish between political and non-political retweets we asked "Is the content of this tweet related to politics?". Each retweet was rated by two separate workers and the ratings are correlated at 0.683. In the survey on sentiment for mentions we asked "What is the sentiment expressed in this tweet?".

candidate party are less correlated.

Finally, in Table 9, we present regression results analogous to those we presented earlier on the speed of information diffusion. As documented above, the production of political retweets is somewhat more like-minded in nature than the production of non-political retweets. Given this, we investigate whether political tweets reach like-minded users more quickly than non-political tweets. To do so, we estimate augmented versions of the previously-discussed regression models with time to exposure as the dependent variable and also estimate Cox survival models. Most importantly, we now allow the coefficient on mismatch between candidate party and voter ideology to vary depending upon whether the tweet is political or non-political in nature. As shown, in all three specifications, we find that non-political tweets do reach like-minded users more quickly but that the difference in time to exposure is larger for political tweets. That is, non-political tweets reach like-minded users 7 minutes faster and political tweets reach like-minded users almost 11 minutes faster, a difference of roughly 4 minutes, and this difference is statistically significant at conventional levels.

8 Conclusion

While scholars have long argued that voters should have access to high quality information from a diverse set of sources, a separate literature has documented a tendency towards homophily, a preference for associating with like-minded individuals. While it has previously been established that groups tend to inbreed and that members of large groups have more network connections, we investigate the role of these patterns in the context of exposure to information.

In particular, we begin by developing a model in which larger groups are exposed to more information and all groups are disproportionately exposed to like-minded information. To test these hypotheses, we use data from a large network of politically engaged Twitter users. Using information on links between voters within this network, we find strong evidence of inbreeding homophily, within group associations that are disproportionate to group size. We also find that members of larger groups have more connections on a per-capita basis. Taking the network structure as given, we then examine the flow of information through the network. Consistent with larger groups having more network connections, we find that larger groups are exposed to more information. Also, consistent with inbreeding homophily, we find that voters of all groups are disproportionately exposed to like-minded information. Finally, we present evidence suggesting that, conditional on exposure, information reaches like-minded users more quickly. Taken together, these results suggest that social networks in general, and social media in particular, may be a force for increasing differences in exposure to information between majority and minority groups and may also increase exposure to like-minded information for all groups.

References

Acemoglu, D., T. A. Hassan, and A. Tahoun (2014). The power of the street: Evidence from egypt's arab spring. *Working paper*.

Barberá, P. (2013). Birds of the same feather tweet together. bayesian ideal point estimation using twitter data. *Proceedings of the Social Media and Political Participation, Florence, Italy*, 10–11.

Becker, G. S. (1958). Competition and democracy. *Journal of Law & Economics 1*, 105.

Black, D. (1958). *The theory of committees and elections*. Cambridge: Cambridge University Press.

Campante, F. R. and D. A. Hojman (2013). Media and polarization: Evidence from the introduction of broadcast tv in the united states. *Journal of Public Economics*.

Chiang, C.-F. and B. Knight (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies 78*(3), 795–820.

Currarini, S., M. O. Jackson, and P. Pin (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica 77*(4), 1003–1045.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor (1999). The rise and decline of the american ghetto. *Journal of Political Economy 107*(3), 455–506.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

DellaVigna, S. and E. Kaplan (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics 122*(3), 1187–1234.

Downs, A. (1957). An economic theory of democracy.

Durante, R. and B. Knight (2012). Partisan control, media bias, and viewer responses: Evidence from berlusconi's italy. *Journal of the European Economic Association 10*(3), 451–481.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and political persuasion: Evidence from russia. *The American Economic Review 101*(7), 3253–3285.

Flaxman, S., S. Goel, and J. M. Rao (2013). Ideological segregation and the effects of social media on news consumption. *Available at SSRN*.

Gentzkow, M. and J. M. Shapiro (2008). Competition and truth in the market for news. *The Journal of Economic Perspectives 22*(2), 133–154.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from us daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

George, L. and J. Waldfogel (2003). Who affects whom in daily newspaper markets? *Journal of Political Economy 111*(4), 765–784.

Golub, B. and M. O. Jackson (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics 127*(3), 1287–1338.

Groseclose, T. and J. Milyo (2005). A measure of media bias. *The Quarterly Journal of Economics 120*(4), 1191–1237.

Jackson, M. O. and D. Lopez-Pintado (2013). Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science 1*(01), 49–67.

Jackson, M. O. and L. Yariv (2010). Diffusion, strategic interaction, and social structure. *Handbook of Social Economics, edited by J. Benhabib, A. Bisin and M. Jackson*.

Lerman, K. and R. Ghosh (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM 10*, 90–97.

Marsden, P. V. (1987). Core discussion networks of americans. *American sociological review*, 122–131.

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

Oberholzer-Gee, F. and J. Waldfogel (2005). Strength in numbers: Group size and political mobilization*. *Journal of Law and Economics 48*(1), 73–91.

Oberholzer-Gee, F. and J. Waldfogel (2009). Media markets and localism: Does local news en español boost hispanic voter turnout? *The American Economic Review*, 2120–2128.

Putnam, R. D., R. Leonardi, and R. Y. Nanetti (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton, NJ: Princeton university press.

Rainie, L., A. Smith, K. L. Schlozman, H. Brady, and S. Verba (2012, October 19). Social media and political engagement. *Pew Internet & American Life Project*.

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

White, M. J. (1986). Segregation and diversity measures in population distribution. *Population index*, 198–221.

A Appendix (Not for Publication)

Proof of Proposition 2: We have shown in the text that $F_C^1 > F_L^1$, and we now show that $F_C^{\tau-1} > F_L^{\tau-1}$ implies that $F_C^\tau > F_L^\tau$. Note first that:

$$F_C^\tau - F_L^\tau = F_C^{\tau-1} - F_L^{\tau-1} + (1 - F_C^{\tau-1}) f_C^\tau - (1 - F_L^{\tau-1}) f_L^\tau$$

which can be re-written as:

$$F_C^\tau - F_L^\tau = F_C^{\tau-1} - F_L^{\tau-1} + [(1 - F_L^{\tau-1}) - (F_C^{\tau-1} - F_L^{\tau-1})][f_L^\tau + (f_C^\tau - f_L^\tau)] - (1 - F_L^{\tau-1}) f_L^\tau$$

expanding the terms in brackets and re-arranging, we have that:

$$F_C^\tau - F_L^\tau = (F_C^{\tau-1} - F_L^{\tau-1})(1 - f_L^\tau) + [1 - F_C^{\tau-1}](f_C^\tau - f_L^\tau)$$

Thus, $F_C^\tau - F_L^\tau$ is positive if $f_C^\tau - f_L^\tau$ is positive. This latter difference can be written as:

$$\begin{aligned} f_C^\tau - f_L^\tau &= q w_C \pi_s F_C^{\tau-1} + q(1 - w_C)\pi_d F_L^{\tau-1} - q(1 - w_C)\pi_s F_L^{\tau-1} - q w_C \pi_d F_C^{\tau-1} \\ &= q F_C^{\tau-1} w_C(\pi_s - \pi_d) + q F_L^{\tau-1}(1 - w_C)(\pi_d - \pi_s) \\ &= q(\pi_s - \pi_d)[F_C^{\tau-1} w_C - F_L^{\tau-1}(1 - w_C)] \end{aligned}$$

This is positive under the maintained assumptions that $F_C^{\tau-1} > F_L^{\tau-1}$, $w_C > 0.5$, and $\pi_s > \pi_d$.

Proof of Proposition 3: Due to the symmetry of the model, it is the case that $C_C^\tau = L_L^\tau$ and that $L_C^\tau = C_L^\tau$ for all τ. Given this, we focus on exposure to conservative information, and, in particular, show that $C_C^\tau > C_L^\tau$ for all τ. Note first that

$$C_C^\tau - C_L^\tau = C_C^{\tau-1} - C_L^{\tau-1} + (1 - C_C^{\tau-1}) c_C^\tau - (1 - C_L^{\tau-1}) c_L^\tau$$

which can be re-written as:

$$C_C^\tau - C_L^\tau = C_C^{\tau-1} - C_L^{\tau-1} + [1 - C_L^{\tau-1} - (C_C^{\tau-1} - C_L^{\tau-1})][c_L^\tau + (c_C^\tau - c_L^\tau)] - (1 - C_L^{\tau-1}) c_L^\tau$$

re-arranging, we have that:

$$C_C^\tau - C_L^\tau = (C_C^{\tau-1} - C_L^{\tau-1})(1 - c_C^\tau) + [1 - C_L^{\tau-1}](c_C^\tau - c_L^\tau)$$

Thus, the sign of $C_C^\tau - C_L^\tau$ involves a comparison of c_C^τ and c_L^τ, which can be written as:

$$\begin{aligned} c_C^\tau - c_L^\tau &= q0.5\pi_s C_C^{\tau-1} + q0.5\pi_d C_L^{\tau-1} - q0.5\pi_s C_L^{\tau-1} - q0.5\pi_d C_C^{\tau-1} \\ &= q0.5(\pi_s - \pi_d)(C_C^{\tau-1} - C_L^{\tau-1}) \end{aligned}$$

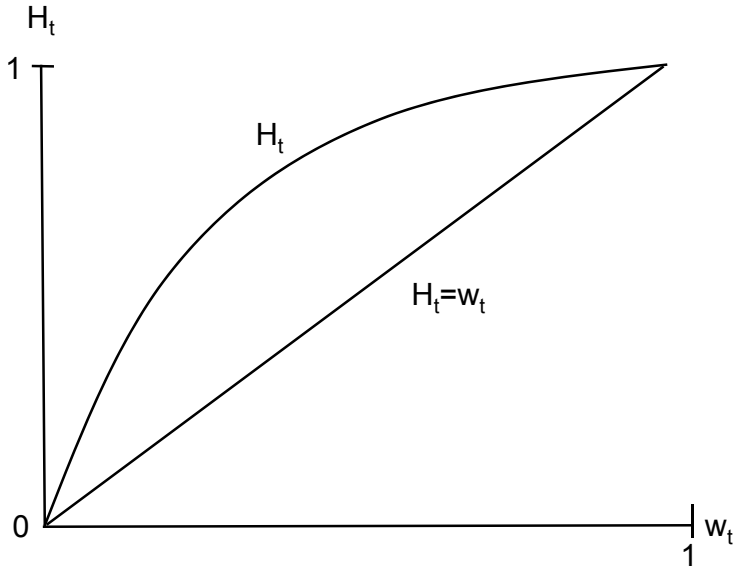
This is positive under the maintained assumption that $C_C^{\tau-1} > C_L^{\tau-1}$, and thus $C_C^\tau - C_L^\tau$ is also positive. Finally, we show that $C_C^1 > C_L^1$, which is implied by:

$$\begin{aligned} C_C^1 &= q0.5\pi_s\varepsilon_s + q0.5\pi_d\varepsilon_d \\ C_L^1 &= q0.5\pi_d\varepsilon_s + q0.5\pi_s\varepsilon_d \end{aligned}$$

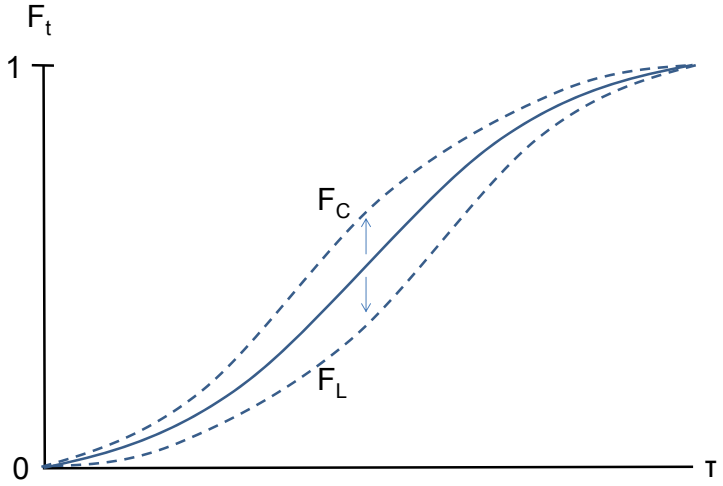
Taking the difference, we have that:

$$C_C^1 - C_L^1 = q0.5(\pi_s - \pi_d)(\varepsilon_s - \varepsilon_d) > 0$$

Proof of Proposition 4: Focusing again on conservative information (without loss of generality), let expected time to exposure for conservatives and for liberals be given, respectively, by $T^C = \sum_\tau \tau(C_C^\tau - C_C^{\tau-1})$ and for liberals $T^L = \sum_\tau \tau(C_L^\tau - C_L^{\tau-1})$. Using summation by parts, the difference in expected time to exposure can be written as: $T^C - T^C = \sum_\tau (C_C^\tau - C_L^\tau)$, which, as shown in Proposition 3, is negative.

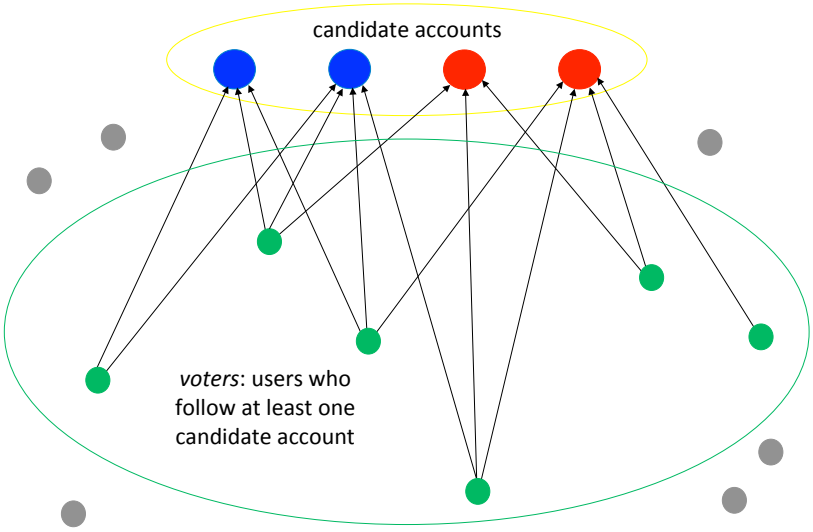


(a) Homophily and Group Size

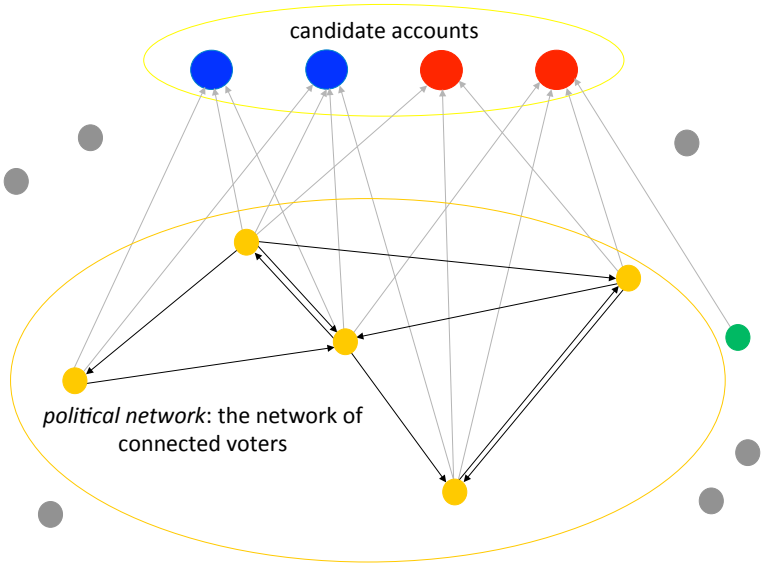


(b) Group Size and the Diffusion of Information

Figure 1: Theoretical Figures



(a) Selecting sample of users (*voters*)



(b) Connecting selected users (*political network*)

Figure 2: Constructing the Network of Politically-Engaged Twitter Users

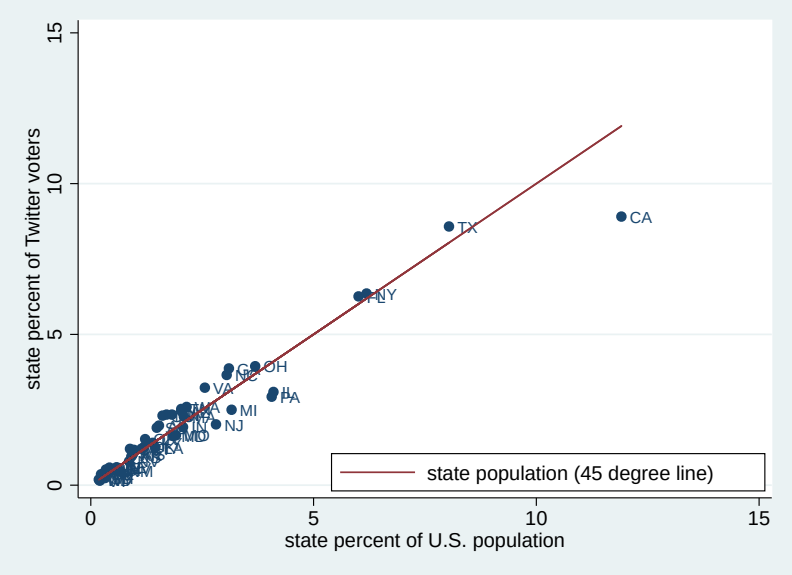


Figure 3: Spatial Representation of Twitter Voters

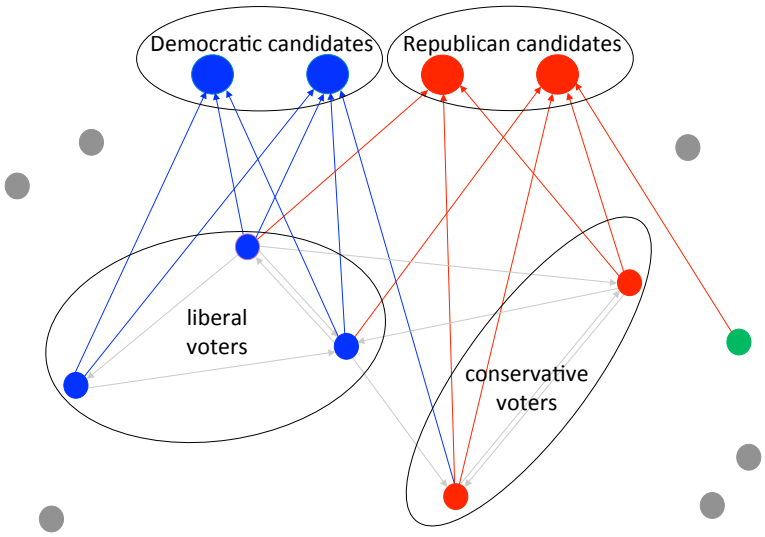
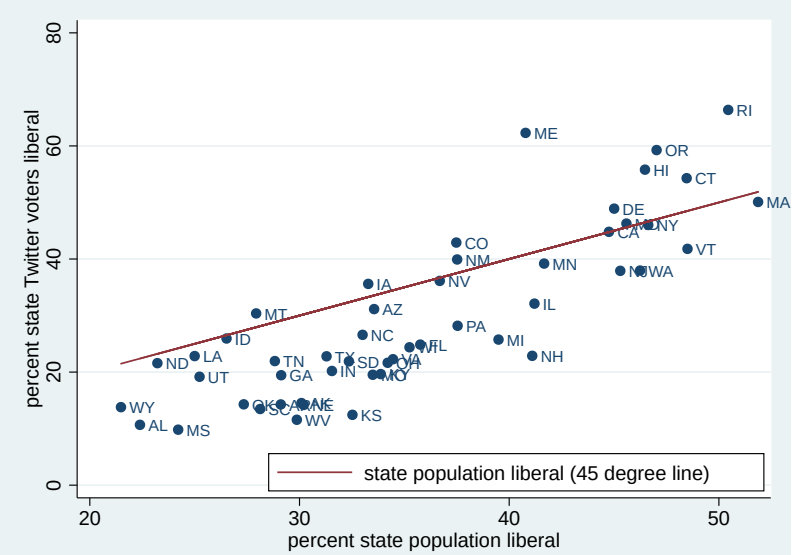
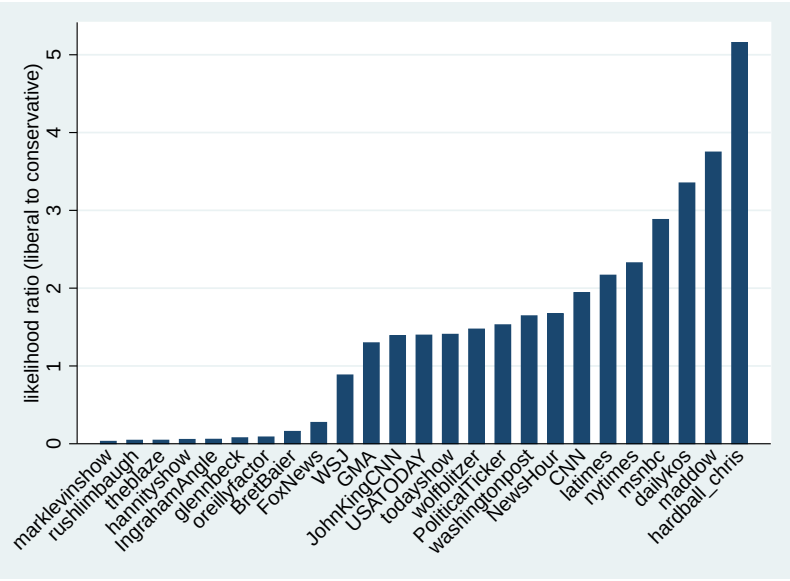


Figure 4: Inferring Voter Ideology

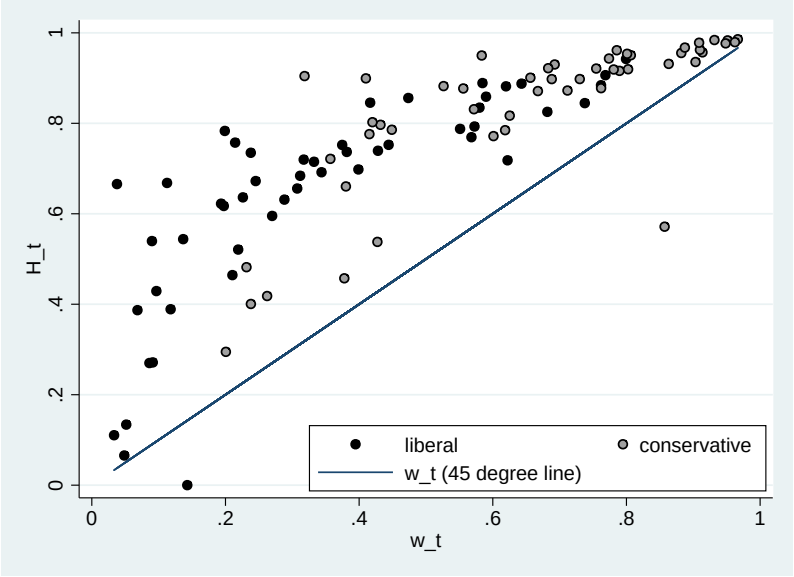


(a) Share of State Liberal Voters and Liberal Twitter Users

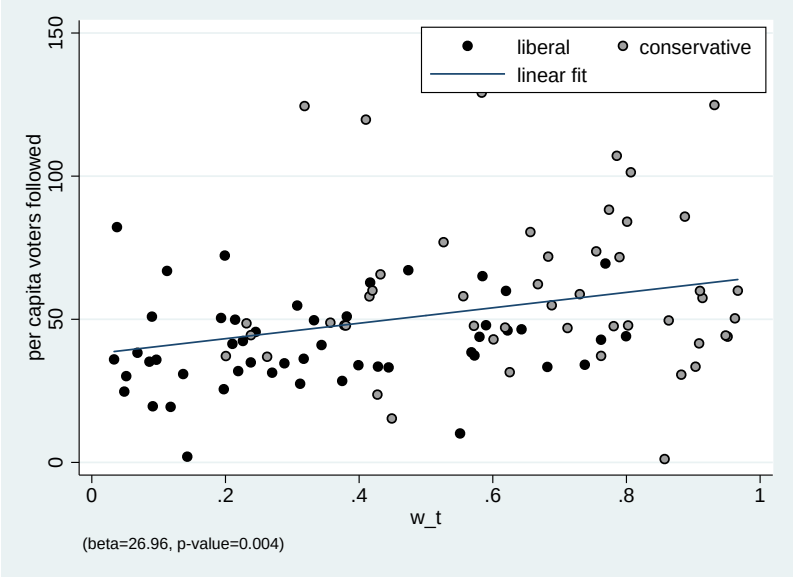


(b) Likelihood Ratio of Following Media Outlets

Figure 5: Validation of Ideology Measure for Twitter Users

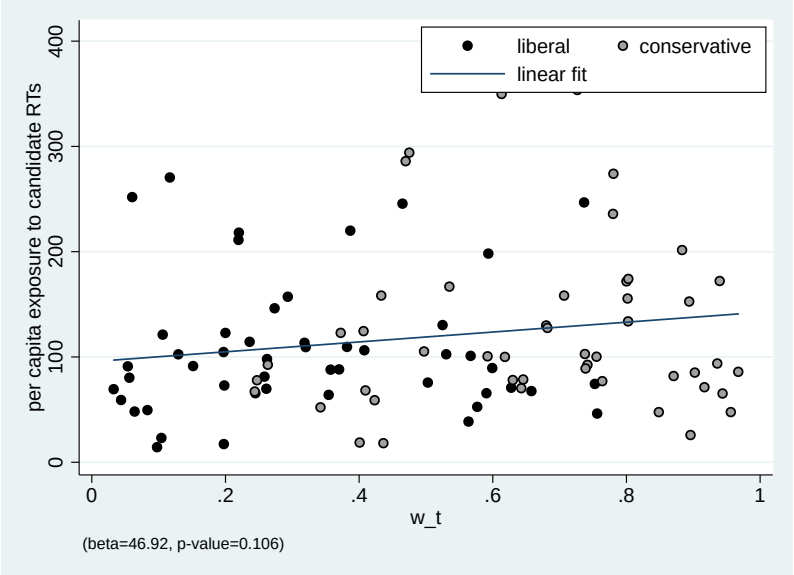


(a) Homophily in Connections

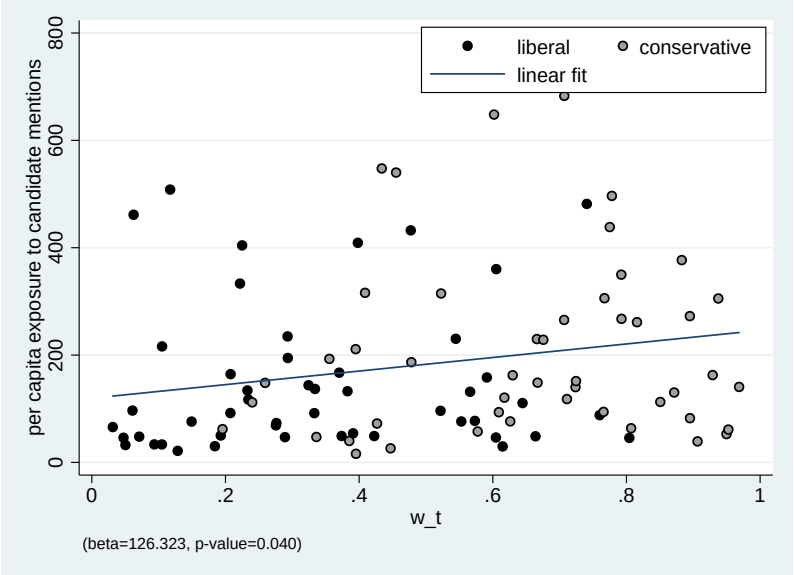


(b) Group Size and Per Capita Connections

Figure 6: Network Connections

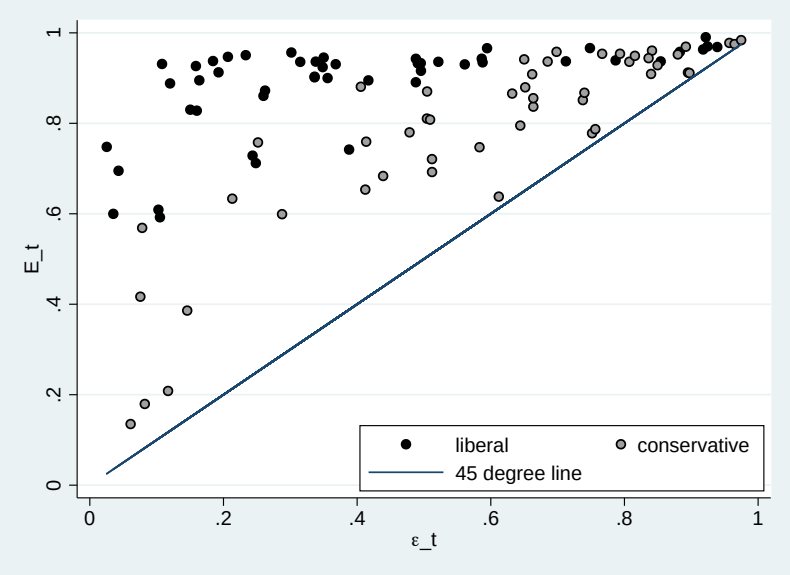


(a) Per Capita Exposure to Retweets

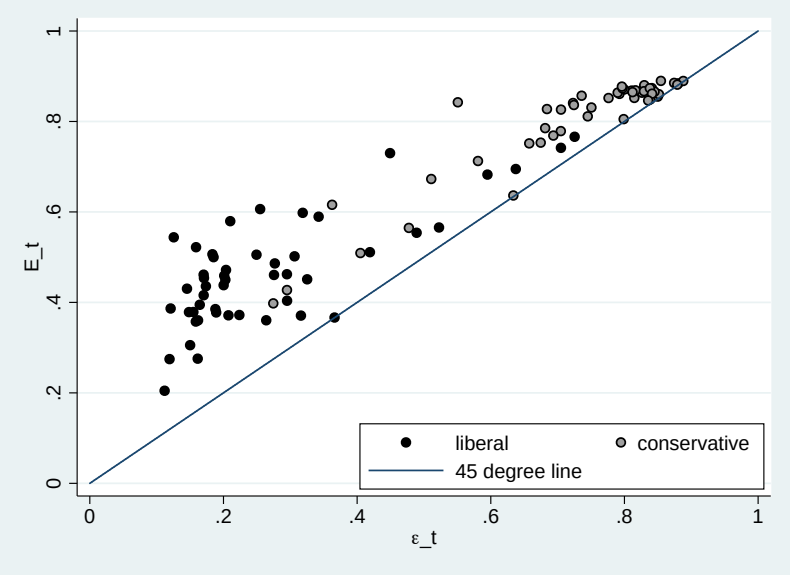


(b) Per Capita Exposure to Mentions

Figure 7: Group Size and Per Capita Exposure to Information

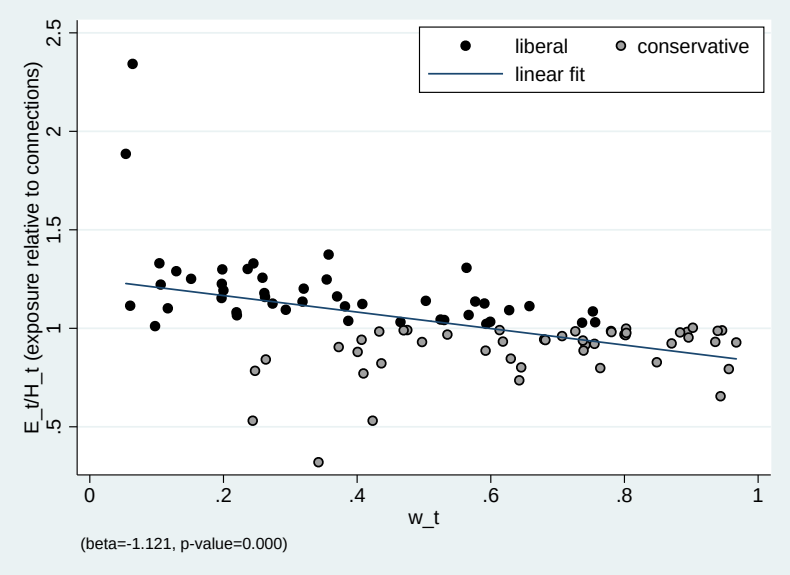


(a) Exposure to Like-Minded Retweets

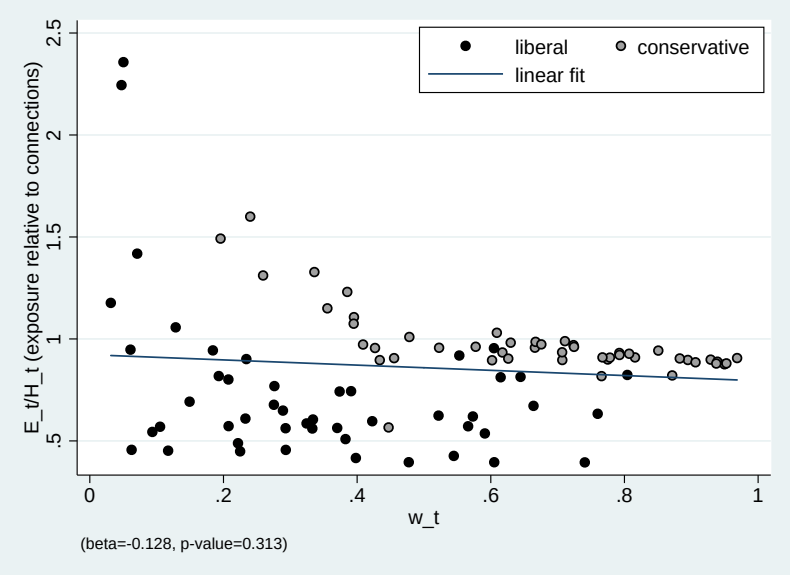


(b) Exposure to Like-Minded Mentions

Figure 8: Homophily and Exposure to Like-Minded Information



(a) RTs and Connections



(b) Mentions and Connections

Figure 9: Group Size and Relative Exposure to Like-Minded Information

Table 1: Descriptive Statistics of Political Network

		Percent Followed		Voters Followed	
	Percent	Liberal	Conservative	Average Same-Type	Per Capita
Liberal voters	36.06	67.11	32.89	40.416	58.756
Conservative voters	63.94	20.25	79.75	57.828	68.486

Table 2: Group Homophily and Ideological Segregation

	Liberals		Conservatives		Conservative Exposure of		
	Share	H Index	Share	H Index	Conservatives	Liberals	Isolation
Baseline	0.361	0.688	0.639	0.844	0.776	0.372	0.403
Followers of both parties	0.337	0.599	0.663	0.783	0.716	0.499	0.217
Followers of one party only	0.361	0.695	0.639	0.849	0.776	0.358	0.417

Table 3: Isolation in Media and Social Networks

	Network Segregation			Media Segregation		
	Conservative Exposure of			Conservative Exposure of		
Followers of	Conservatives	Liberals	Isolation	Conservatives	Liberals	Isolation
Media and candidates	0.780	0.387	0.394	0.789	0.547	0.241
Media and both parties	0.717	0.489	0.228	0.723	0.656	0.067

Table 4: Production of Information by Voters

	Percent of RTs		Percent of First RTs		Percent of Mentions	
Candidate party	Democrat	Republican	Democrat	Republican	Democrat	Republican
Liberal voters	90.91	1.29	85.68	2.16	65.87	23.23
Conservative voters	9.09	98.71	14.32	97.84	34.13	76.77

Table 5: Group Exposure to Like-Minded Ideological Information

	Fraction of Tweets	Same-type Tweets received	Per capita Tweets	E index
Liberal	0.484	52.462	58.368	0.899
Conservative	0.516	63.449	70.351	0.902
	Fraction of RTs	Same-type RTs received	Per capita RTs	E index
Liberal	0.312	74.856	81.443	0.919
Conservative	0.688	103.280	110.949	0.931
	Fraction of mentions	Same-type mentions received	Per capita mentions	E index
Liberal	0.230	59.014	152.981	0.386
Conservative	0.770	165.746	197.344	0.840

Table 6: Diffusion of Information and Time to Exposure

	Linear Regression		Cox Survival Analysis
	minutes	ln(minutes)	
Liberal voter	1.4502***	0.0287***	-0.0147***
	(0.1000)	(0.0008)	(0.0006)
Ideology mismatch	9.9670***	0.1418***	-0.0776***
	(0.1000)	(0.0008)	(0.0006)
Tweet FE	Yes	Yes	Yes
N	48,443,770	48,443,770	48,443,770
Dependent variable mean	102.04	2.57	102.04

Notes: *** denotes significance at the 99 percent level, ** denotes significance at the 95 percent level, and * denotes significance at the 90 percent level, The dependent variable is minutes to exposure in column 1 and the natural log of minutes to exposure in column 2. Column 3 estimates a Cox survival model, using data on minutes to exposure. In all specifications, the unit of observation is an exposed voter-candidate tweet. Ideology mismatch indicates either a conservative voter and a Democratic candidate tweet or a liberal voter and a Republican candidate tweet.

Table 7: Information Production by Type of Content

(a) Percent of RTs by Type

	Political RTs		Non-Political RTs	
Candidate party	Democrat	Republican	Democrat	Republican
Liberal voters	92.45	1.20	84.55	2.04
Conservative voters	7.55	98.80	15.45	97.96

(b) Percent of Mentions by Type

	Positive Mentions		Negative Mentions	
Candidate party	Democrat	Republican	Democrat	Republican
Liberal voters	84.60	10.54	42.71	32.42
Conservative voters	15.40	89.46	57.29	67.58

Table 8: Voter Exposure to Information by Content

(a) RTs by Information Type

Content	Ideology	Fraction of RTs	Per Capita RTs	E index
Political	Liberal	0.470	58.108	0.919
	Conservative	0.530	80.178	0.937
Non-Political	Liberal	0.521	13.018	0.910
	Conservative	0.479	16.007	0.889

(b) Mentions by Information Type

Content	Ideology	Fraction of Mentions	Per Capita Mentions	E index
Positive	Liberal	0.268	7.421	0.678
	Conservative	0.732	10.780	0.896
Negative	Liberal	0.200	18.803	0.195
	Conservative	0.800	20.860	0.798
Full sample	Liberal	0.232	22.927	0.376
(10 percent)	Conservative	0.768	29.138	0.836

Table 9: Diffusion of Political versus Non-Political Information

	Linear Regression		Cox Survival Analysis
	minutes	ln(minutes)	
Liberal voter	3.4773***	0.0445***	-0.0205***
	(0.2995)	(0.0023)	(0.0016)
Ideology mismatch	7.1990***	0.1131***	-0.0672***
	(0.2995)	(0.0023)	(0.0016)
Liberal*political	-3.6740***	-0.0264***	0.0093***
	(0.3270)	(0.0025)	(0.0018)
Ideology mismatch*	3.8548***	0.0519***	-0.0217***
political	(0.3270)	(0.0025)	(0.0018)
Tweet FE	Yes	Yes	Yes
N	34,428,571	34,428,571	34,428,571
Dependent variable mean	103.71	2.61	103.71

Notes: *** denotes significance at the 99 percent level, ** denotes significance at the 95 percent level, and * denotes significance at the 90 percent level, The dependent variable is minutes to exposure in column 1 and the natural log of minutes to exposure in column 2. Column 3 estimates a Cox survival model, using data on minutes to exposure. In all specifications, the unit of observation is an exposed voter-candidate tweet. Ideology mismatch indicates either a conservative voter and a Democratic candidate tweet or a liberal voter and a Republican candidate tweet. Political indicates whether a tweet is political in nature, as opposed to non-political in nature.