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

Yosh Halberstam† Brian Knight‡

August 20, 2015

Abstract

Social media represent a new source of information for voters. Unlike mass media, information on social networking sites circulates via nonmarket interactions among individuals. In this paper, we investigate the role of homophily—a tendency to interact with similar individuals–in the diffusion of political information in social networks. We develop a model predicting disproportionate exposure to like-minded information and that larger groups have more connections and are exposed to more information. To test these hypotheses, we use data on 90 million links and over 500,000 communications between 2.2 million politically-engaged Twitter users during the 2012 elections. We find that users affiliated with majority political groups have more connections and are exposed to more information than users from minority groups. Likewise, we find that users are disproportionately exposed to like-minded information and that information reaches like-minded users more quickly.

*We are deeply indebted to Zack Hayat for getting this project off the ground and providing continual advice. We thank seminar participants at UC-Berkeley, CU-Boulder, Hebrew U, Michigan State, Stanford, Toronto, the National University of Rosario, the Central Bank of Colombia, the 2014 Media and Communications Conference at Chicago-Booth and the 2015 American Politics Summer Conference at Yale. 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.

†University of Toronto, yosh.halberstam@utoronto.ca

‡Brown University, Harvard University, and NBER, Brian_Knight@brown.edu

1 Introduction

A long literature in political economy, including Black (1958), Downs (1957), Becker (1958), Putnam et al. (1994), Besley and Prat (2006), and Besley (2007), has highlighted the importance of providing voters with information. Given uncertainty over candidate quality and ideology, information helps voters to select candidates that are both of high quality and moderate ideology, reducing adverse selection and increasing overall voter welfare. Moreover, information on the activities and accomplishments of incumbents is useful in terms of monitoring the behavior of politicians once in office and thus minimizing moral hazard.

Traditional media outlets, such as radio, newspapers, and television, represent important sources of information for voters. Given this, a significant domain for government policy involves the regulation of these media outlets. Policy objectives often involve providing voters with a large number of independent outlets. That is, fixing the degree of independence across outlets, an increase in the number of media outlets is associated with voters receiving a larger volume of information. Moreover, fixing the number of outlets, a greater degree of variety across outlets (i.e., pluralism) is associated with more independence in information across sources. These two goals can be supported, for example, by simultaneously reducing entry costs and limiting cross-ownership of media outlets.

The growing importance of social media platforms, such as Facebook and Twitter, in providing information to voters provides a new challenge to policymakers.[1] While traditional media outlets tend to be hierarchical in nature, with a small number of suppliers providing information to a large number of consumers, typical users of social networking sites can both consume and produce information on these platforms. Moreover, users of the same social media platform may be exposed to significantly different content depending upon the set of accounts that they follow. On the other hand, readers of newspapers and viewers of a television station are exposed to the same information, at least to a first approximation. Given these differences between traditional media outlets and social media platforms, practices in the regulation of traditional media outlets may not translate well to social media platforms.

To better understand voter exposure to political information on social media platforms, we conceptualize social media as a social network. If different types of individuals have different

[1]The use of social media has grown dramatically during the past decade, with 60 percent of American adults currently using social networking sites (Rainie et al., 2012). 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. 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 voted in the 2012 elections.

beliefs and tend to develop relationships with like-minded individuals, a phenomenon known as homophily, then users may be disproportionately exposed to like-minded political information. As noted above, a lack of independence across sources of information may reduce the quality of information, threatening to increase adverse selection and moral hazard in government. Moreover, due to homophily, minority groups may be exposed to less information than majority groups, potentially undermining the efficacy of democratic institutions via a reduction in electoral competition.

We begin by formalizing these ideas in a simple model of political information diffusion in a social network characterized by homophily and two ideological groups, majority and minority. The model predicts that, with homophily, members of the majority group have more network connections and are exposed to more information than members of the minority group. We also use the model to show that, with homophily and a tendency of users to produce like-minded information, individuals are disproportionately exposed to like-minded information. Finally, the model predicts that information reaches like-minded individuals more quickly than it reaches individuals of opposing ideology.

The primary contribution of this paper involves an empirical investigation of these hypotheses using data from Twitter. As the theoretical model highlights, measuring exposure to information in social networks is challenging in the sense that it requires data on both network structure and communications within the network. We overcome these challenges by constructing a political network of Twitter users and then examining the flow of information through this network. More specifically, we begin by selecting politically-engaged users, defined as those 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 network links (i.e., one user following another user). Using these links, we construct a single national network and 50 state subnetworks comprising only of users who follow candidates from the same state. We consider users to be conservative if they follow more accounts associated with Republican candidates than accounts associated with Democrats and liberal if they follow more Democrats than Republicans. To measure communications, 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 communications, 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 flows through the network.

Based upon these data, we find that the degree of homophily in the political network is similar to that documented in other social networks, such as the offline high-school friendship networks analyzed by Currarini et al. (2009). 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. Next, we examine the content of tweets, showing that the results regarding exposure and speed are stronger for political tweets by candidates than for non-political tweets. We also show that exposure to positive mentions of candidates, when compared to negative mentions of candidates, is more like-minded in nature. Taken together, our results suggest that homophily in social networks limits voter exposure to information. In particular, homophily generates a built-in advantage in knowledge for voters belonging to the majority group and increases the correlation across information sources for all groups of voters, potentially undermining the speed and likelihood of convergence to the truth. This may make it difficult for voters to select the best candidates and to monitor the behavior of politicians once in office.

The paper proceeds as follows: in Section 2, we discuss our contribution to the literature. Next, 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 exposure to information, 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 Related Literature

Research in political economy documents a causal effect of the media on voter knowledge, measures of political behavior, such as voter turnout and candidate choice, as well as other political outcomes. These include DellaVigna and Kaplan (2007) and Martin and Yurukoglu (2014), who both investigate the effect of the introduction of Fox News on turnout and support for Republican candidates. Likewise, Enikolopov et al. (2011) show that access to a partisan television station in Russia increased support for the party affiliated with the station, and Chiang and Knight (2011) document that surprising newspaper endorsements are more influential than unsurprising endorsements. Gentzkow et al. (2011) show that the entry of partisan newspapers in the United States increased voter turnout but had no impact on vote shares, and Strömberg (2004b) found that radio increased voter turnout and the federal government spent more during the Great Depression in areas with a higher concentration of radio listeners. Our work suggests that social media may influence voter behavior because homophily produces an asymmetry in exposure to information in social networks, and given the findings in the studies above, beliefs may affect political behavior. Moreover, there is growing evidence that social media affect political and civic behavior. For

example, communications on social networking sites have been shown to precipitate protests during the Arab Spring (Acemoglu et al., 2014) and reduce corruption in Russia (Enikolopov et al., 2014). Our work presents a mechanism through which to understand how behavior and aggregate outcomes induced by these new platforms are influenced by homophily.

Another literature focuses on the role of media in political polarization. Some studies have shown low political polarization associated with the introduction of new media platforms: Campante and Hojman (2013) examine the introduction of television in the United States, finding a corresponding decline in political polarization. Likewise, Gentzkow and Shapiro (2011) document that media consumption on the internet (e.g., nytimes.com) is relatively unsegregated politically and comparable to traditional mass media. More recently, Flaxman et al. (2013) find that individuals are more ideologically segregated when they read opinion articles on social media than when they read descriptive news on online media. Related to this issue, we find relatively low levels of political segregation in media consumption on Twitter (e.g., @nytimes) but higher levels when measuring network segregation in terms of connections among Twitter users from different ideological groups.[2]

A body of empirical research examines the impact of network effects on decision-making, such as welfare participation (Bertrand et al., 2000; Gee and Giuntella, 2011; Furtado and Theodoropoulos, 2013), enrollment in publicly-funded prenatal care (Aizer and Currie, 2004), health-care utilization (Deri, 2005), and bankruptcy filings (Miller, 2015). These studies tend to employ similar identification strategies, typically leveraging variation in the size of ethnic groups across geographic areas and variation in knowledge across ethnic groups. Simply put, the strategy involves investigating whether being surrounded by more members of the same group changes decision-making more when those in the group have more knowledge. These studies implicitly assume the existence of homophily and its role in leading larger groups to have more network connections and receive more information. In this paper, we seek to get inside this 'black box' and provide evidence on the homophily mechanism, documenting that larger groups do indeed have more network connections and receive more information. In the political domain, research on group size has shown that political mobilization is weaker (Oberholzer-Gee and Waldfogel, 2005) and less effective (Ananat and Washington, 2009) among minority groups. Our work offers a potential explanation for this phenomenon: political mobilization of minority groups is weaker than that of majority groups because minority groups, due to homophily, are exposed to less information than majority groups.[3]

[2]We report these results in the Appendix.

[3]Further, the knowledge gap between minority and majority groups is likely exacerbated since media outlets are less likely to cover issues for which demand among its consumer base is low (Strömberg, 2004a). Moreover, politicians are more likely to target larger groups because the infrastructure for transmitting information to them has already been created by the media (Oberholzer-Gee and Waldfogel, 2009).

Empirical research on homophily in social networks has shown that this phenomenon applies to many different individual characteristics, including racial identity, gender, age, religion, and education (McPherson et al., 2001). Focusing on group size and homophily, Currarini et al. (2009) develop a theoretical model of network formation in which homophily can arise from both biases in preferences or opportunities for meetings. They use data on high-school friendship networks and, consistent with their model, find that larger racial groups have a larger fraction of same-race friendships and more per-capita friendships overall. Likewise, Marsden (1987) investigates advice networks and finds that members of larger groups tend to have more connections. We build upon these studies by examining the role of group size and homophily in the diffusion of information and by providing microfoundations for how ideological homophily in a social network can influence voter beliefs.

Most research examining homophily on the internet and, more specifically, social media measures network structure but not communications (e.g., Colleoni et al. (2014)) or communications but not network structure (e.g., Adamic and Glance (2005)). As emphasized by our model, measuring exposure to information in social networks requires measures of both network structure and communications. We are aware of three studies in computer science, all produced independently of our work, that do combine measures of network structure and communications. Constructing network clusters, Conover et al. (2011) show that users are disproportionately exposed to like-minded information in the re-tweet network but not in the mentions network. Likewise, Weber et al. (2013) examine Twitter communications in Egypt and show that communications are religiously polarized and especially so during periods of violence. And third, Bakshy et al. (2012) use Facebook data to show that users are disproportionately exposed to like-minded information. While we also study exposure to like-minded information, our work is unique in our investigation of the role of group size in social networks and the influence of homophily on the speed of information transmission.

Our paper is also related to a theoretical literature on the role of homophily in communications in social networks. 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. Consistent with average-based updating, we show that the speed of exposure to information further diverges between political groups when the tweets have political content. Turning to exposure, in Jackson and Lopez-Pintado (2013) the authors explore how homophily influences the spread of an idea throughout an entire network, and our results shed light on the extent and speed of this information diffusion.

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 ($t \in \{C, L\}$). Total population is normalized to one, and group sizes are given by w_t such that $w_C + w_L = 1$. 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 population. Specifically, if conservatives dominate the population and links are formed at random, then conservatives would appear to be homophilous and liberals 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, *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.

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^τ is the hazard rate, or the probability of group t exposure at time τ, conditional on not being exposed at time $\tau - 1$:

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

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 are no differences 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 the 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:

$$F_t^1 = qw_t\pi_s\varepsilon + q(1-w_t)\pi_d\varepsilon - q^2w_t(1-w_t)\pi_s\pi_d\varepsilon^2$$

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 $(1-w_C)\pi_d$, each of whom transmits the information with probability q. 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 liberal and conservative 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 $\tau = 0$.[4] That is, conservatives produce conservative information with probability ε_s and

[4]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

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

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

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 $\varepsilon_s = \varepsilon_d$, then 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 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 information at any given time period, it then follows that average time to exposure to 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, a higher degree of 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.

---

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 $(1\text{-}w_t) q_d \pi_d$.

4 Data

To test these hypotheses, our study uses data from Twitter, an internet platform through which users connect and communicate with each other. In order to measure ideology and given our focus on political communications, we focus our analysis on a sample of Twitter users who follow accounts of politicians. Given this selection rule, our sample may disproportionately include Twitter users with strong preferences for linking to like-minded users.[5] To shed further light on this issue, we report in the Appendix isolation indices for our political network and compare these to isolation indices in different settings analyzed in Gentzkow and Shapiro (2011). One notable finding is that isolation in our Twitter political network is similar to isolation in offline social networks of individuals who at least sometimes discuss politics with each other (i.e., "political discussants").[6] These two settings are arguably comparable since both involve political communications. We further discuss these issues below and in the Appendix.

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 from the two major parties for the House of Representatives in the 2012 general election, 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

[5]Moreover, social media users may not be not representative of voters at large along several dimensions, including age and race (Pew, 2013). However, the political network we construct is well-suited to study politically-relevant diffusion of information.

[6]In addition, using data on which media accounts are followed by which Twitter users, we show that our results, especially for the sample of Twitter users who follow politicians from both parties, are comparable to the measures of isolation in online media consumption reported in Gentzkow and Shapiro (2011).

[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.

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 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 reflects the geographic 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 politicians that they follow, and this process is depicted in Figure 4. In particular, voters who follow more Democratic than Republican politicians are coded as liberal, and voters that follow more Republican than Democratic politicians 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 politicians from the two parties.

To shed light on the validity of these measures of voter ideology, we again use user-supplied locations and correlate our measures with survey responses from the 2012 Gallup State of the States political survey. In particular, in Figure 5a we compare our estimate of the share of liberals in each state according to our 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

[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.

[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).

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 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. There is also support at the individual level for the validity of our ideology measure. Using information on voter registration with parties, Barberá (2013) matches a sample of voters from Ohio to their Twitter accounts and finds a strong correlation between party registration and the parties these voters follow on Twitter. In summary, these results suggest that our measures of voter ideology are reasonable and do capture some underlying measure of political preferences. It is still possible that higher segregation on Twitter may be observed due to public exposure and pressure to conform with one's ideological reference group (Perez-Truglia and Cruces, 2015).

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: 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.

[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 Empirical Framework

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

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 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 Exposure to Information

We next develop measures of exposure to like-minded 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 3 and the measures developed in Section 4, we next present our empirical results on network structure. We begin by describing our results on homophily at the national level before turning to results in state political networks.

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. Turning to the homophily measures, we have that liberals have 40 liberal followees on average and 59 total followees. This implies a homophily index for liberals equal to 69 percent. 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 at the national level 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. Taken together, Table 1 suggests a significant degree of homophily in this national Twitter political network.

6.2 State Political Networks

We next investigate the degree of homophily in state-level subnetworks. Relative to the national level networks, focusing on state-level networks provides us variation in group size, allowing for further investigation of the predictions of Proposition 1 regarding group size. In particular, we investigate whether: (a) larger groups form a larger share of their friendships with users of their own type, (b) groups inbreed and (c) larger groups form more links 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,

conservative or liberal, 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 relates followees per capita for each group to the group's share in the population, where the linear fit is presented to demonstrate the general trend. As shown, an increase from 0 to 1 in the share of the population increases the number of followees per capita from about 40 to 60, a 50 percent increase. Thus, our data are also consistent with the prediction that larger groups have more followees per capita.

To summarize, the results from analyses of state-level subnetworks support the key predictions of Proposition 1 regarding group size. In particular, all groups tend to inbreed, and larger groups exhibit a greater degree of homophily and have more network connections per capita. Given these results, we next examine the implications of this homophily in network structure for network communications.

7 Results on Network Communications

Having documented evidence of network structure consistent with the theoretical model and the existing literature on homophily, we next examine how information flows through this political network. That is, as a result of 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 users more quickly? We also examine how these measures vary depending upon the content of the information, distinguishing between political and non-political information and positive and negative sentiment.

7.1 Production and Transmission 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. While we examine the production of information via mentions of candidates by voters, we also examine how voters transmit information in the network via retweets of tweets initially produced by candidates.

As shown in Table 2, there is a strong correlation between voter ideology and candidate party in the transmission of information via retweets. In particular, 91 percent of retweets of tweets by Democratic candidates are transmitted by liberal voters, and almost 99 percent of retweets of

tweets by Republicans candidates are transmitted by conservative voters. While this may reflect a preference for producing like-minded information, it may also reflect the exposure 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 transmission of first retweets, with 86 percent of retweets of tweets by Democratic candidates transmitted by liberal voters, and almost 98 percent of retweets of tweets by Republicans candidates transmitted by conservative voters.

Next, we examine the production of mentions, and, as shown in the final two columns of Table 2, 66 percent of mentions of Democratic candidates are produced by liberal voters. Likewise, 77 percent of mentions of Republicans candidates are produced by conservative voters. Thus, using data on candidate mentions, we find that voters disproportionately produce like-minded information.

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 in section 6.5.

7.2 Communications in the National Political Network

Having established homophily in network structure and the production of like-minded information, we next test the predictions of the model regarding exposure to like-minded information. In particular, we 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. That is, we develop analogs to our homophily measures based upon the exposure to tweets and retweets from, along with mentions of, like-minded candidates (i.e conservative voters and Republican candidates and liberal voters and Democratic candidates).

As shown in Table 3, 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 candidate accounts. Likewise, exposure to like-minded information for conservative voters is also 90 percent, with 63 out of 70 tweets originating from Republican candidate accounts. Since 48 percent of tweets were produced by Democratic candidates, liberal voters exposed randomly to

tweets 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, documenting that communication serves to amplify an already significant degree of homophily in this Twitter political network.[13]

We next turn to exposure to like-minded information based upon retweets, which account for multiple exposures to the same candidate tweet. That is, if a candidate tweet is retweeted by multiple followees of a voter, the tweet-based exposure index, as described above, would count this as one exposure, whereas the retweet-based exposure index would count this as multiple exposures. 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 3, exposure to like-minded information is even higher (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 mentions of candidates by voters are provided at the bottom of Table 3. 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. We investigate this issue further below in the content analysis in Section 6.5.

To summarize, consistent with the predictions of Proposition 3, we find that voters in the Twitter political network are disproportionately exposed to like-minded information. This holds true when measured by exposure to candidate tweets, exposure to candidate tweets via retweets, and exposure to mentions of candidates by voters.

[13]Recall that our measure of voter ideology is based upon the set of candidates followed by each user. Given this, when measuring exposure to candidate tweets, we ignore exposure to the initial tweet produced by the candidate since this would bias our results towards finding disproportionate exposure to like-minded information. That is, we only measure exposure to candidate tweets via re-tweets from other voters in the network. Thus, there is no issue of circularity in terms how we measure voter ideology and voter exposure to political information.

7.3 Communications in 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, consistent with Proposition 2, larger groups receive more information on a per-capita basis than smaller groups in the presence of homophily. 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 group size exposure to retweets of candidate tweets on a per-capita basis, and, in panel b), the corresponding relationship using data on mentions of candidates. Consistent with the model, exposure to information, in terms of both retweets and mentions, clearly 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 a 19 percent increase in exposure to mentions. This result is consistent with Proposition 2, which predicted that majority groups are exposed to more information on a per-capita basis than minority groups.

Turning to exposure to like-minded information, we next examine communications within state networks using data on tweets produced 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. That is, in the absence of homophily, voters are exposed to ideological content in proportion to that produced in the state-level network.

We illustrate this connection between exposure to like-minded information and this baseline measure of exposure in Figure 8. In panel a), we show this relationship using retweets. As shown, in all states, and for both conservative and liberal groups, 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. This relationship is analogous to the positive relationship between H and w, as documented in Figure 6a, 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 like-minded in nature. 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 than 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 exhibits a similar downward slope but is less pronounced when compared to the trend 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 is exposed to more information in general and to like-minded information in particular.

7.4 Speed Analysis

We next consider measures of speed, or time to exposure, in the flow of information through the network at the national level. In particular, Proposition 4 states that, with homophily and the production of like-minded information, individuals are exposed to like-minded information 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 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 candidate 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). By including tweet fixed effects, differences in speed are identified via the time difference in exposure for a given tweet between like-minded users (i.e., liberal voters and Democratic candidates and conservative voters and Republican candidates) and those with a mismatched ideology.

As shown in Table 4, 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.[14] To provide results in percentage terms, we next run a similar regression but with the natural log of minutes as the dependent variable.[15] 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.

7.5 Content Analysis

In this section we investigate heterogeneity in our data according to the content of the communications. We distinguish between whether the candidate tweets include political or non-political information. In addition, we investigate differences between positive and negative mentions of candidates by voters.

Starting with the distinction between political and non-political communications, we investigate whether the patterns of exposure to like-minded information differ between political and non-political candidate tweets.[16] In terms of the production of information, we find some evidence in panel a) of Table 5 that the production of political information tends to be more like-minded in nature, when compared to 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

[14] These results effectively assume that users are continuously monitoring their Twitter feed. While this assumption is unlikely to hold in practice, reducing the time in which a voter could be potentially exposed to this information will reduce expected time to exposure even if users are not continuously monitoring their Twitter feed. For example, suppose that a followee of a user posts some information at either $t = 1$ (probability p) or at time $t = 2$ (probability $1 - p$). Then, we would measure time to exposure as $p + 2(1 - p) = 2 - p$, which is clearly decreasing in p. Further, suppose that this user logs in at $t = 1$ with probability q and logs in at $t = 2$ with probability one. Then, expected time to first exposure equals $pq + 2(1 - pq) = 2 - pq$, which is also decreasing in p so long as $q > 0$.

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

[16] To classify candidate tweets, we designed a survey 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 particular, 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.

small, with conservative voters accounting for almost 99 percent of political information and 98 percent of non-political information.

In Table 6, we next investigate whether these differences in production translate into differences in exposure. As shown in panel a), we do find that exposure to political tweets is more like-minded in nature when compared to exposure to non-political tweets. In terms of magnitudes, however, the differences between political and non-political retweets are relatively small. Liberal voters have an exposure index of 92 percent for political information and 91 percent for non-political information, and conservative voters have an exposure index of 94 percent for political information and 89 percent for non-political information. The fact that the differences in exposure between political and non-political information are small is not surprising given, as noted above, that differences between the like-minded production of political and non-political retweets are relatively small.

In Table 7, 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.

Turning to mentions, we next investigate whether exposure to positive mentions tends to be more like-minded in nature, when compared to exposure to negative mentions of candidates. For example, a conservative voter may mention Republican candidates using positive sentiment and may mention Democratic candidates using negative sentiment. To do so, and given the large sample of mentions, we distinguish between positive and negative sentiment using a 10% random sample of mentions of candidates by voters.[17]

As shown in panel b) of Table 5, we do find significant differences in the production of like-minded information 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. Turning to exposure, as shown in panel b) of Table 6, we also find significant differences in exposure between mentions with positive and negative sentiment, with high exposure to like-minded information for positive

[17] In the survey on sentiment for mentions we asked "What is the sentiment expressed in this tweet?".

sentiment mentions and low exposure for negative sentiment mentions. Taken together, due to homophily and differences in the production of information, we find that voters are disproportionately exposed to positive mentions of affiliated candidates (e.g., conservative voters and Republican candidates) and are disproportionately exposed to negative mentions of candidates from the other party (e.g., conservative voters and Democratic candidates).

To summarize, the content analysis documents that production of and exposure to political information is more like-minded in nature, when compared to non-political information, although some of the differences are small in magnitude. Political information also reaches like-minded users more quickly than non-political information. Finally, we find that the sentiment of communications matters, with voters disproportionately exposed to positive mentions of affiliated candidates.

8 Conclusion

This paper begins by developing a model that predicts that larger groups are exposed to more information and all groups are disproportionately exposed to like-minded information. To test these hypotheses, we use data on network connections and political communications for over two million Twitter users who follow political candidates during the 2012 US elections. We split the sample of users into conservatives and liberals based upon the political party of candidates most followed by the user. Using information on links between voters within this network, we find strong evidence of homophily, with conservatives more likely to link to conservatives and liberals more likely to link to liberals. To investigate the role of group size, we then define state subgroups comprising users who follow candidates in a given state, providing cross-state variation in group size. Using this, we 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 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 media may be a force for further exacerbating the majority-minority gap 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*.

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43. ACM.

Aizer, A. and J. Currie (2004). Networks or neighborhoods? correlations in the use of publicly-funded maternity care in california. *Journal of Public Economics 88*(12), 2573–2585.

Ananat, E. O. and E. Washington (2009). Segregation and black political efficacy. *Journal of Public Economics 93*(5), 807–822.

Bakshy, E., I. Rosenn, C. Marlow, and L. Adamic (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528. ACM.

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.

Bertrand, M., E. F. Luttmer, and S. Mullainathan (2000). Network effects and welfare cultures*. *The Quarterly journal of economics 115*(3), 1019–1055.

Besley, T. (2007). Principled agents?: The political economy of good government. *OUP Catalogue*.

Besley, T. and A. Prat (2006). Handcuffs for the grabbing hand? media capture and government accountability. *The American Economic Review*, 720–736.

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.

Colleoni, E., A. Rozza, and A. Arvidsson (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication 64*(2), 317–332.

Conover, M., J. Ratkiewicz, M. Francisco, B. Gonalves, F. Menczer, and A. Flammini (2011). Political polarization on twitter. In *ICWSM*.

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.

Deri, C. (2005). Social networks and health service utilization. *Journal of Health Economics 24*(6), 1076–1107.

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

Enikolopov, R., M. Petrova, and K. Sonin (2014). Social media and corruption. *Available at SSRN 2153378*.

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*.

Furtado, D. and N. Theodoropoulos (2013). Ssi for disabled immigrants: Why do ethnic networks matter? *The American Economic Review 103*(3), 462–466.

Gee, E. R. and G. O. Giuntella (2011). Medicaid and ethnic networks. *The BE Journal of Economic Analysis & Policy 11*(1).

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

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011). The effect of newspaper entry and exit on electoral politics. *American Economic Review 101*(7), 2980–3018.

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.

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

Martin, G. J. and A. Yurukoglu (2014). Bias in cable news: Real effects and polarization. Technical report, National Bureau of Economic Research.

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

Miller, M. M. (2015). Social networks and personal bankruptcy. *Journal of Empirical Legal Studies 12*(2), 289–310.

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.

Perez-Truglia, R. and G. Cruces (2015). Partisan interactions: Evidence from a field experiment in the united states. *Available at SSRN 2427148*.

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

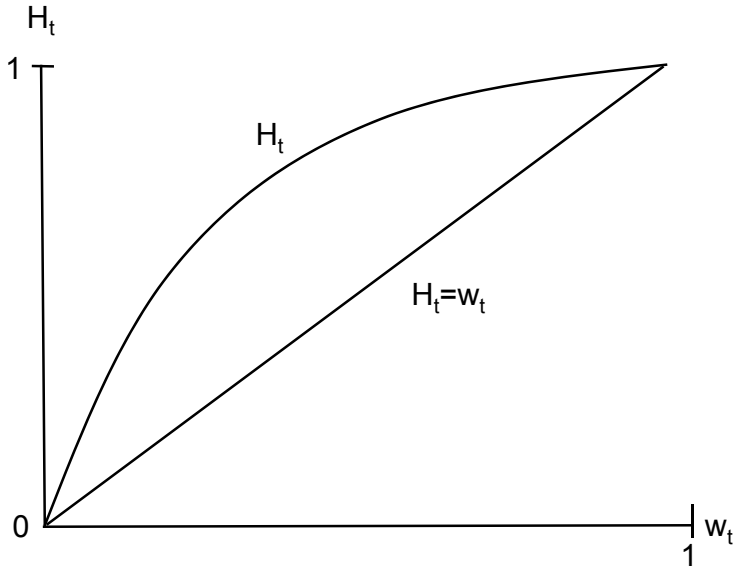
Rainie, L., A. Smith, K. L. Schlozman, H. Brady, and S. Verba (2012, October 19). Social media and political engagement. *Pew Internet & American Life Project*.

Strömberg, D. (2004a). Mass media competition, political competition, and public policy. *The Review of Economic Studies 71*(1), 265–284.

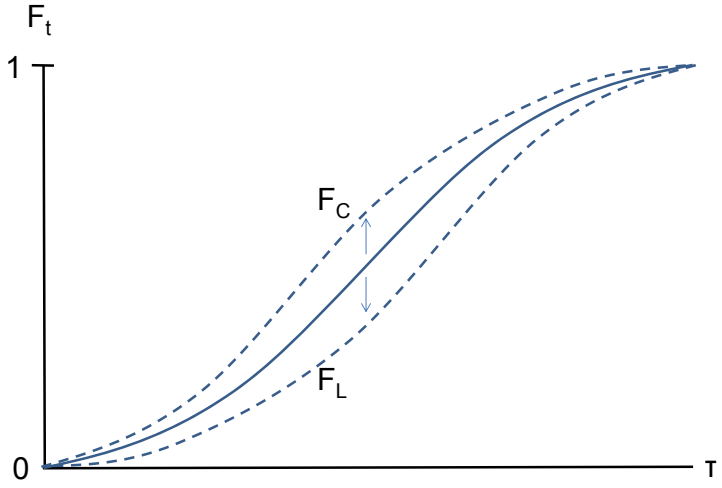
Strömberg, D. (2004b). Radio's impact on public spending. *The Quarterly Journal of Economics*, 189–221.

Weber, I., V. R. K. Garimella, and A. Batayneh (2013). Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 290–297. ACM.

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

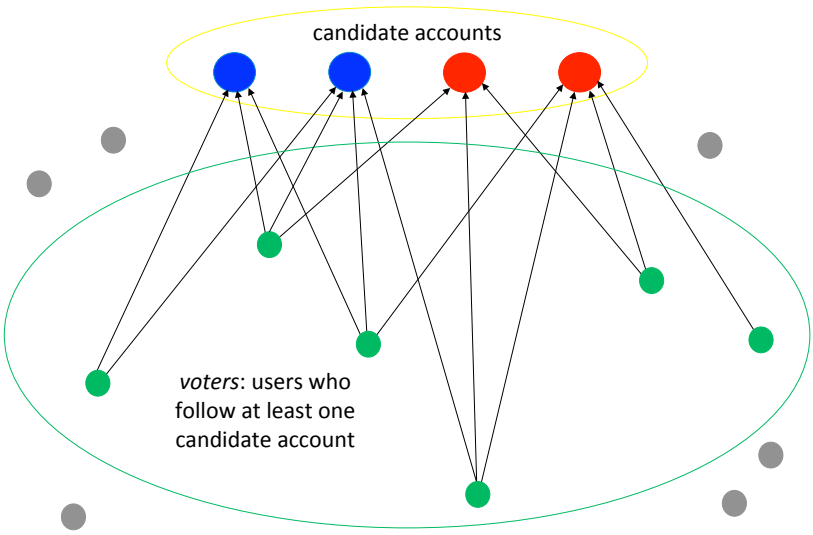


(a) Homophily and Group Size

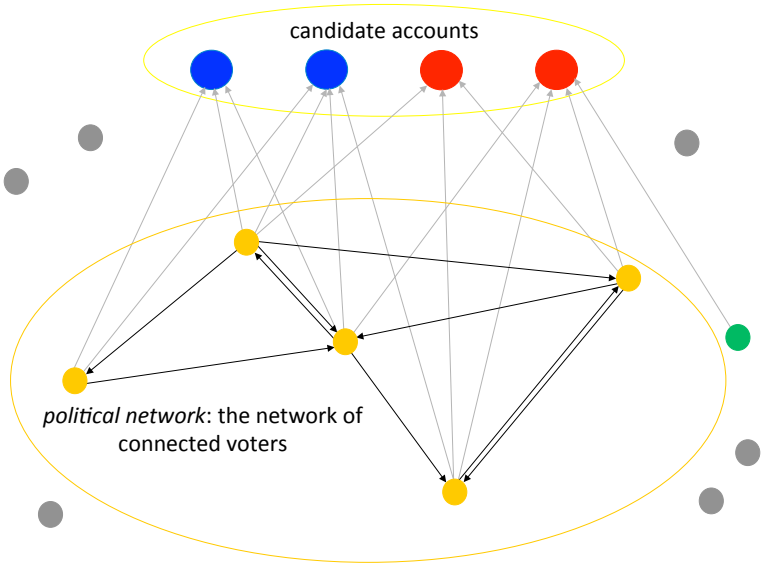


(b) Group Size and the Diffusion of Information

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

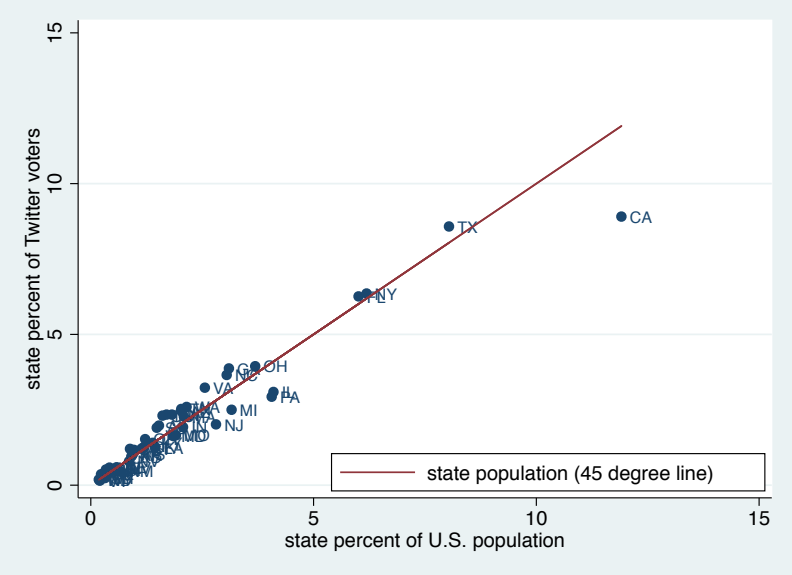


Figure 3: Spatial Representation of Twitter Voters

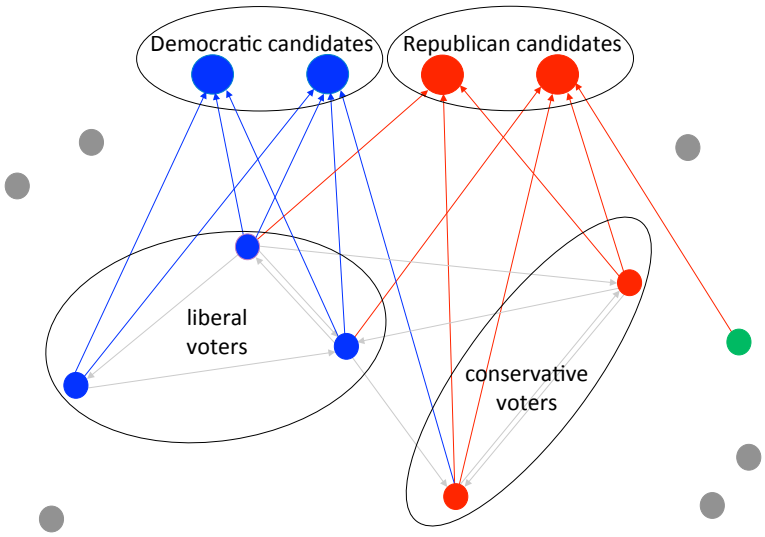
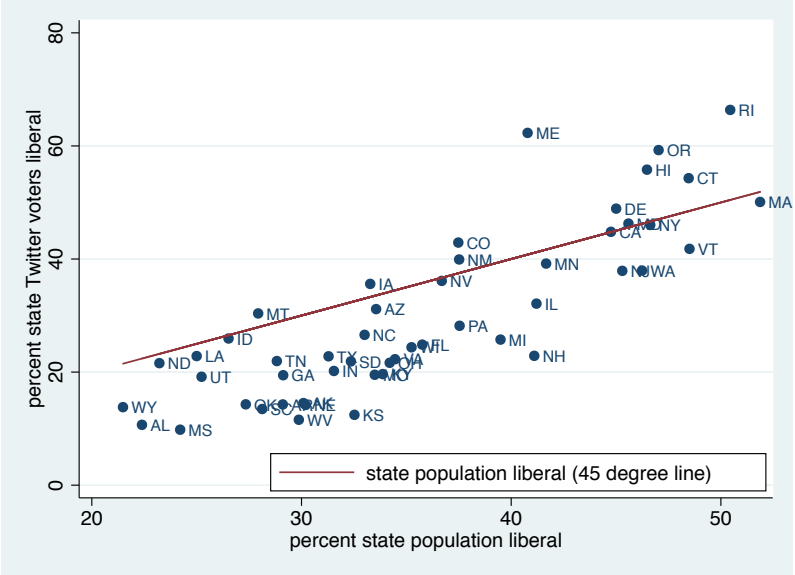
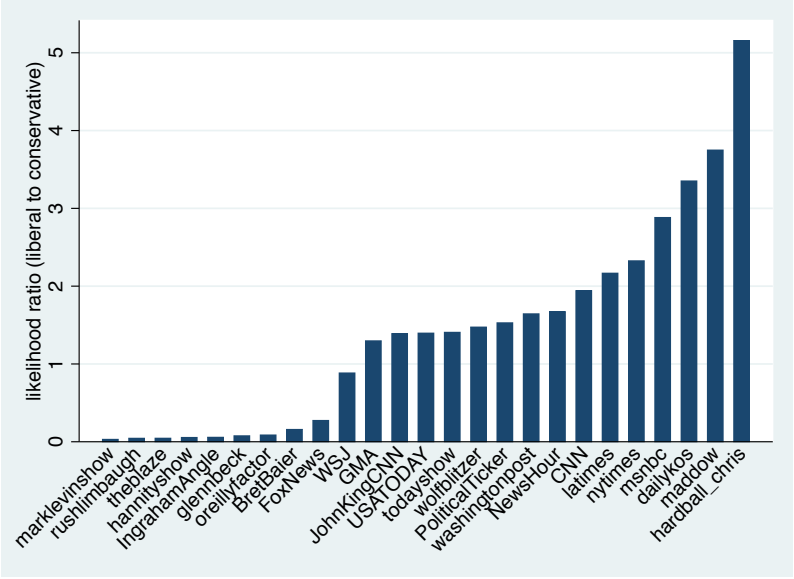


Figure 4: Inferring Voter Ideology

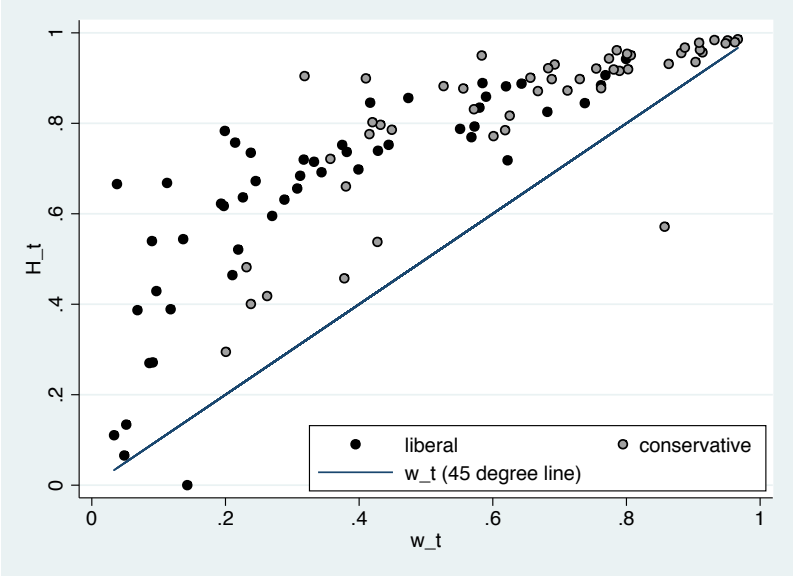


(a) Share of State Liberal Voters and Liberal Twitter Users

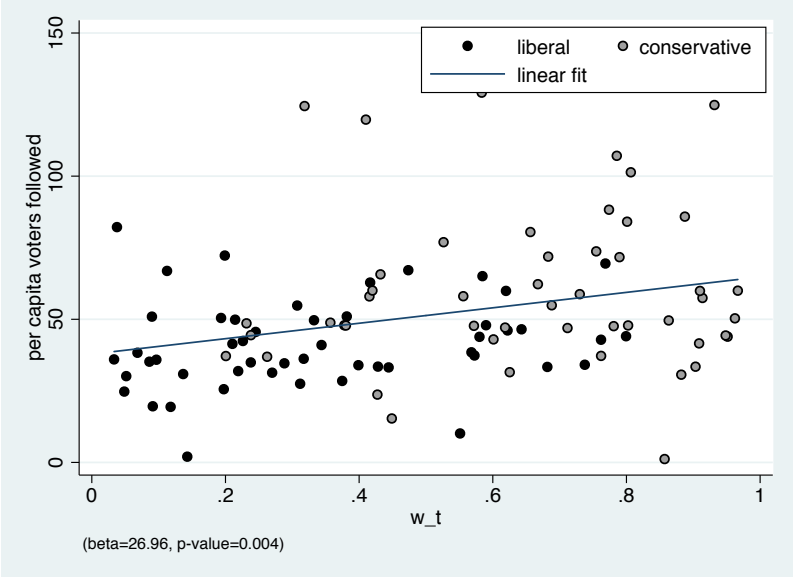


(b) Likelihood Ratio of Following Media Outlets

Figure 5: Validation of Ideology Measure for Twitter Users

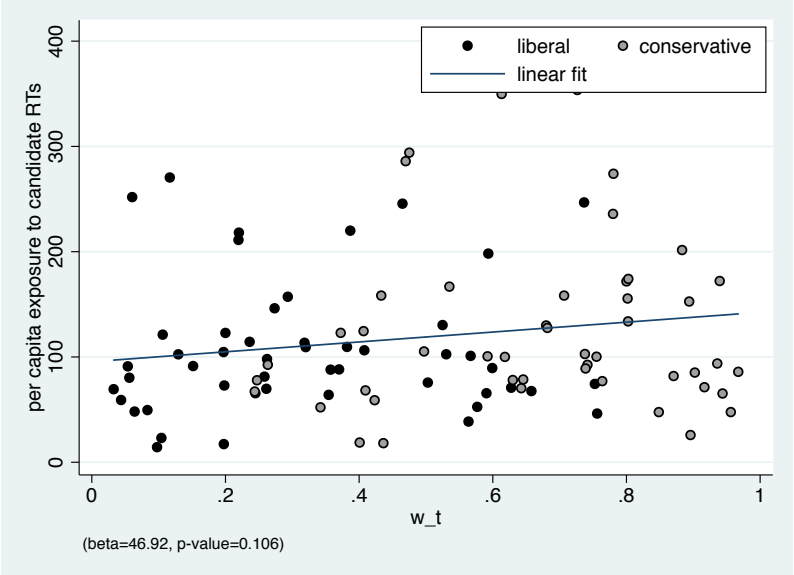


(a) Homophily in Connections

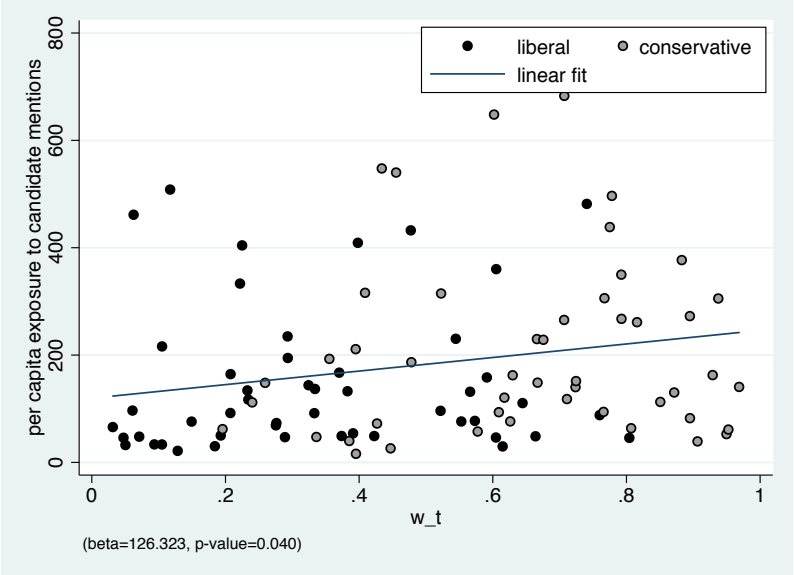


(b) Group Size and Per Capita Connections

Figure 6: Network Connections

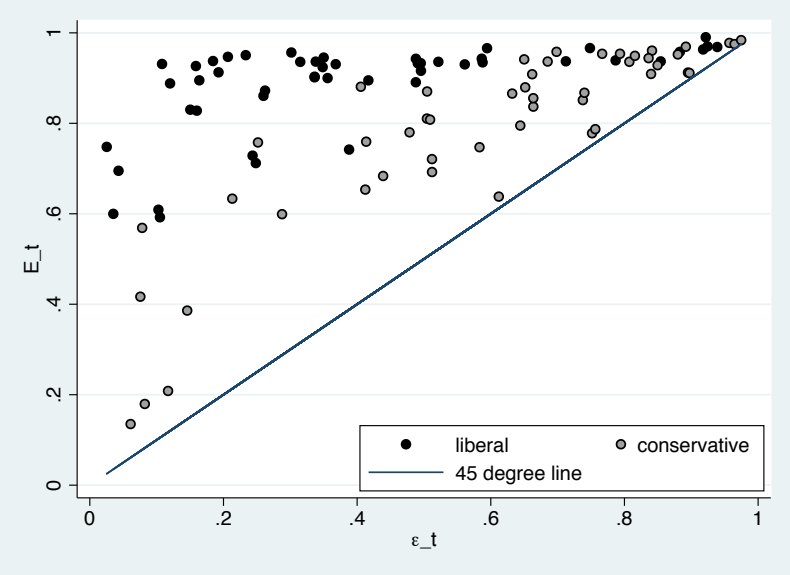


(a) Per Capita Exposure to Retweets

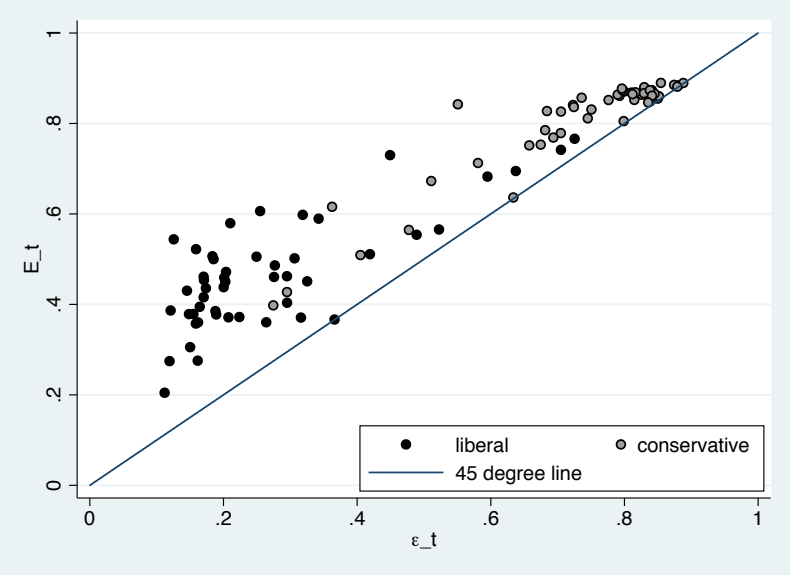


(b) Per Capita Exposure to Mentions

Figure 7: Group Size and Per Capita Exposure to Information

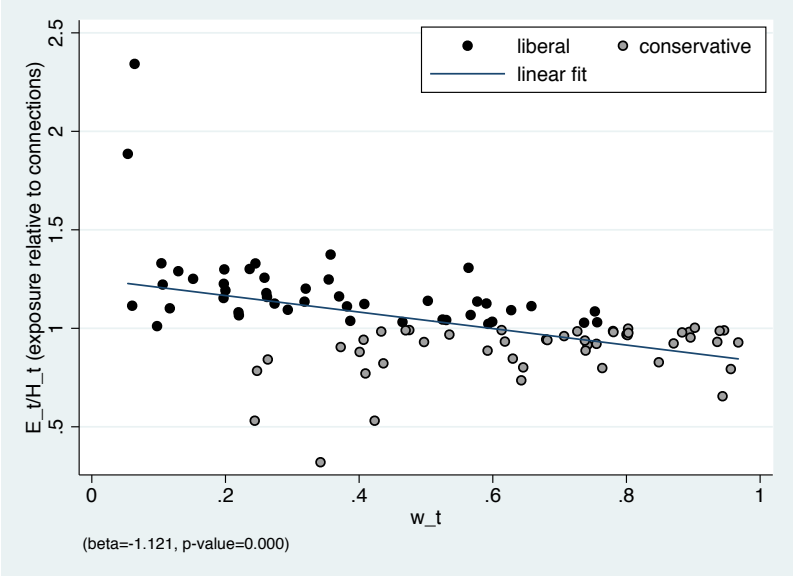


(a) Exposure to Like-Minded Retweets

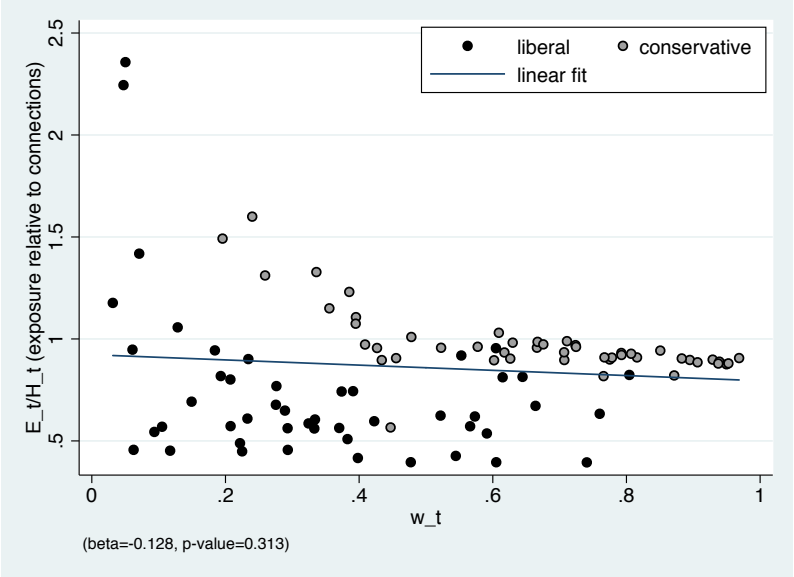


(b) Exposure to Like-Minded Mentions

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: Homophily in the Political Network

	Percent Followed			Voters Followed		
	Percent	Liberal	Conservative	Same-Type	Total	Percent same type
Liberal voters	36.06	67.11	32.89	40.416	58.576	0.688
Conservative voters	63.94	20.25	79.75	57.828	68.486	0.844

Table 2: Production of Information by Voters

	Percent of retweets		Percent of first retweets		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 3: Group Exposure to Like-Minded Ideological Information

	Fraction of tweets	Same-type tweets	Total tweets	Exposure index
Liberal	0.484	52.462	58.368	0.899
Conservative	0.516	63.449	70.351	0.902
	Fraction of retweets	Same-type retweets	Total retweets	Exposure 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	Total mentions	Exposure index
Liberal	0.230	59.014	152.981	0.386
Conservative	0.770	165.746	197.344	0.840

Table 4: 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 5: Information Production by Type of Content

(a) Percent of retweets: Political versus Non-political

	Political retweets		Non-Political retweets	
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: positive versus negative

	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 6: Voter Exposure to Information by Content

(a) retweets by Information Type

Content	Ideology	Fraction of retweets	Per-capita retweets	Exposure
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
10 percent sample	Liberal	0.232	22.927	0.376
	Conservative	0.768	29.138	0.836

Table 7: 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.

A Appendix (For Online 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:

$$\begin{aligned} C_C^1 &= q0.5\pi_s\varepsilon_s + q0.5\pi_d\varepsilon_d - q^2 0.25\pi_s\varepsilon_s\pi_d\varepsilon_d \\ C_L^1 &= q0.5\pi_d\varepsilon_s + q0.5\pi_s\varepsilon_d - q^2 0.25\pi_s\varepsilon_s\pi_d\varepsilon_d \end{aligned}$$

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.

Measuring Ideological Segregation: Following Gentzkow and Shapiro (2011), we also compute segregation in media consumption using information on the set of media outlets followed by each voter on Twitter.[18] For comparison purposes, we also compute network isolation. 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

[18] This measure has been developed by White (1986) and Cutler et al. (1999), and widely applied to study ethnic and urban segregation.

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.

Table 8 reports the results from computing these isolation measures. As shown in the first row, conservative exposure among conservatives for the network-based measure 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), with a baseline estimate of segregation equal to 0.075. 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 focus on moderates, those who follow both parties. As shown in Table 2, and consistent with the view that these moderates have weaker preferences for linking to like-minded users, we find that segregation is lower for moderates, when compared to the entire sample. Second, as noted above, we use information on the followers of our sample of media outlets to compute segregation in media consumption on Twitter. As shown in the third row, 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. 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), but significantly lower than network segregation for this group, which equals 0.228.

Table 8: Ideological Segregation in Media and Social Networks

	Network Segregation			Segregation in Media Consumption		
	Conservative Exposure			Conservative Exposure		
Followers of	Conservative	Liberals	Isolation	Conservative	Liberal	Isolation
Baseline	0.776	0.372	0.403	n/a	n/a	n/a
Both parties	0.716	0.499	0.217	n/a	n/a	n/a
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