Are Social Media more Social than Media? Measuring Ideological Homophily and Segregation on Twitter*

Yosh Halberstam† Brian Knight‡

May 23, 2014

Abstract

Social media represent a rapidly growing source of information for citizens around the world. In this paper, we measure the degree of ideological homophily and segregation on social media. In particular, we construct a network based upon links between politically-engaged users and find that the network exhibits previously documented homophily patterns in offline social networks. Likewise, the degree of ideological segregation in the network is similar to that in networks based upon face-to-face interactions, such as friends and co-workers, and is higher than that associated with mass media, such as television and newspapers.

*Preliminary. Please do not cite or circulate without permission of the authors. We are particularly indebted to Zack Hayat for getting this project off the ground and providing continual advice. We thank seminar participants at Berkeley, CU-Boulder, Michigan State, Stanford and Toronto for their input. Ashwin Balamohan, Max Fowler, Kristopher Kivutha and Somang Nam jointly created the infrastructure to obtain the Twitter data used in this paper. 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.

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

‡Department of Economics, Brown University and NBER, Brian_Knight@brown.edu

1 Introduction

Nations around the world grapple with the appropriate regulation of information. On the one hand, non-democracies are often accused of restricting both access to information and freedom of expression. On the other hand, there is often a presumption that democracies should provide citizens with as much information as possible and allow for freedom of expression by citizens. In particular, a long tradition of scholarship, starting with Black (1958), Downs (1957), and Becker (1958), highlights the key role of voter information in terms of selecting high quality candidates and monitoring the behavior of politicians once in office. Further, it is often argued that multiple sources of information are most valuable when they come from diverse viewpoints (Gentzkow and Shapiro, 2008). Multiple sources of information from like-minded viewpoints, by contrast, may be positively correlated and thus provide little information to voters on the margin. Taken together, these two points emphasize the importance of providing voters an ideologically diverse set of high-quality information for a well-functioning democracy (Putnam et al., 1994).

The tremendous rise in social media during the past decade, with 60 percent of American adults and over 20 percent of worldwide population currently using social networking sites (Rainie et al., 2012), has the potential for changing the type of information to which voters are exposed.[1] Indeed, this phenomenal growth in social media engagement in the U.S. and around the world has transformed the nature of political discourse. 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 U.S. voters used social media to let others know how they voted in the 2012 elections. Given this increased role of social media for voters, we are motivated to examine the degree of ideological homophily and segregation in social media. We examine these issues using data from Twitter, one of the leading social media websites in the world, with over 140 million users in the U.S. and a half a billion users worldwide.

Three key features of social media distinguish it from other forms of media and social in-

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

teractions. First, social media allow users to not only consume information but also to produce information. In traditional media, such as newspapers and television, the supply of information is concentrated in the hands of a small number of media outlets. On social media, by contrast, all users can produce new information, for example, via tweets to their followers. Likewise, users can also forward information produced by other users, via retweets. Second, the information to which users are exposed depends upon self-chosen links among users. That is, users primarily observe content produced or transmitted by the set of individuals with whom they have chosen to connect. Given this, at any given time, users may be exposed to significantly different information depending upon the set of individuals with whom they are connected, the content created and transmitted by this set of individuals, and, more generally, the structure of the network to which they belong. 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. Third, information on social media travels more rapidly and broadly than in other forms of social interactions. For example, a tweet from a user on Twitter is simultaneously transmitted to all of his followers. Further, each time one of these followers retweets this tweet, another set of followers is exposed to the information. This process leads to a substantially broader reach and more rapid spread of information than other forms of social interactions.

Given these three distinguishing features, the rapid growth of social media has the potential to effect a structural change in the way individuals engage with one another and the degree to which such communications are segregated along ideological lines. On the one hand, social media may expose individuals to a more diverse set of viewpoints by allowing users to reach beyond their traditional geographic domains. On the other hand, political content on social media may be highly polarized due to homophily, a tendency of users to develop links with like-minded users. Given this, we are motivated to compare the degree of homophily to that in offline social networks, such as friendship networks analyzed by Currarini et al. (2009). It is natural to conjecture that social media has displaced, at least partially, other sources of information for voters, most prominently traditional media, such as newspapers and television, and face-to-face social interactions,

such as friendship networks and neighborhoods. It is also well-established that face-to-face social interactions are characterized by higher ideological segregation when compared to traditional media settings (Gentzkow and Shapiro, 2011). Given all of this, we are also motivated to measure the degree of ideological segregation on social media and to compare this measure to ideological segregation on both traditional media outlets and in face-to-face social interactions.

In order to examine these issues, we construct a network of links between politically-engaged Twitter users. For this purpose, we selected Twitter users who followed at least one Twitter account associated with a candidate for the U.S. House during the 2012 election period. Among this population of over 2.2 million users, we identify roughly 90 million links, which form the network. To infer the ideology of users, we use the political party of the candidates they follow.

Based upon our preliminary analysis, we have two key findings. First, we find that the network we constructed shares important features with face-to-face interactions. Most importantly, both settings tend to exhibit a significant degree of homophily, with links more likely to develop between individuals with similar ideological preferences. Second, when computing the degree of ideological segregation and comparing it to ideological segregation in other settings, we find that Twitter is much more segregated than traditional media, such as television and radio, and is more in-line with ideological segregation in face-to-face interactions, such as among friends and co-workers. Taken together, our results suggest that social media may be a force for increasing isolation and ideological segregation in society.

The paper proceeds as follows. After reviewing the relevant literature, we provide background information on social media in general and Twitter in particular. Section 4 describes the data, Section 5 presents the empirical framework, and Section 6 provides the empirical results. Section 7 concludes and discusses the implications of our findings.

2 Literature Review

This paper contributes to a long-standing literature on network analysis, much of which is reviewed in Jackson (2010). As a starting point, this literature develops measures for describing network characteristics, such as degree distributions, average path length, and clustering. Economists have contributed to this literature in several ways. First, economists have studied the endogenous formation of networks from both a theoretical and an empirical perspective. For example, Currarini et al. (2009) develop a theoretical model of network formation and examine its empirical implications in the context of friendship networks. In particular, we use this paper to guide our analysis on ideological homophily in online social networks and employ the same homophily indices that they do to examine racial homophily in high school friendships. Second, taking the network as given, studies have examined the impact of network structure on economic outcomes. For example, Golub and Jackson (2012) examine how network structure, and homophily in particular, impacts the speed of learning.

The most closely related studies below use network analysis to examine political issues and polarization on Twitter. This literature, mostly authored by computer scientists, falls into two research areas: papers that explore the connection between social networks on Twitter and polarization in communications among users and papers that focus on homophily in links among users. An example of the first type of papers is Conover et al. (2011) who find that a network based on Twitter users that mention each other is less ideologically polarized than one that is based on users who retweet each other's tweets, suggesting a distinction between information creation and information dissemination on social media. Related to this, Himelboim et al. (2013) find that polarization in communication in a Twitter network increases with the proportion of partisan users. Likewise, in a dynamic setting, Yardi and Boyd (2010) pair tweets on abortion with subsequent responses to those tweets and find that a significant proportion of these pairs consist of users with opposing standpoints. Turning to papers on homophily, De Choudhury (2011) collects a large number of tweets and user characteristics. Using data on connections among these users at two points in time, the author finds that homophily in topical interests trumps homophily in most other attributes, in-

cluding gender and ethnicity, in predicting a new connection between Twitter users. Finally, Wu et al. (2011) find that elite users, those contributing a large amount of content on Twitter, are more likely to retweet tweets from each other than tweets by a non-elite user, suggesting heterogeneous effects of communications within a network based on user centrality. Relative to these studies, our research is the first to construct and analyze a follower-followee network of Twitter users that includes an ideology measure for each user. Moreover, we are the first to produce an ideological isolation index for the Twitter political network that is comparable to other settings, such as other forms of media and social interactions.

Closely related to our study is a literature that investigates ideological segregation on the internet. In particular, some have argued that the Internet may be detrimental to democracy because it allows citizens to isolate themselves within "echo chambers", groups that share similar views and experiences (Sunstein, 2001). In a challenge to this view, Gentzkow and Shapiro (2011) document that users of online media (e.g., *www.nytimes.com*) are as segregated as users of traditional (i.e., offline) media and less segregated than face-to-face (offline) social networks. This research, however, did not examine social media, which, as noted above, represent a growing source of online information.[2] Given this motivation, our research aims to compare the degree of ideological segregation on social media to ideological segregation in other settings, including the internet, mass media, and face-to-face social interactions.

There is also a related literature in economics on media bias and voter exposure to partisan information in the media. This literature has examined both the causes and consequences of media bias. Possible causes of media bias include consumer preferences (Gentzkow and Shapiro, 2010) and media ownership (Durante and Knight, 2012). Studies on the consequences of media bias have tended to focus on voting outcomes. DellaVigna and Kaplan (2007) document that the introduction of Fox News increased support for Republican candidates. George and Waldfogel (2003) document the impact of the entry of the New York Times on local political outcomes. Chiang and Knight

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

(2011) document that surprising newspaper endorsements (e.g., those for Republican candidates from left-leaning papers) are more influential than unsurprising endorsements. Enikolopov et al. (2011) show that access to a partisan television station in Russia increased support for the party affiliated with the station. Finally, our work is also related to recent research on the role of media in polarization. For example, Campante and Hojman (2013) provide evidence that the introduction of television in the United States led to a decline in political polarization.

3 Background

As noted above, social media has emerged in the past decade as a new source of information for citizens around the world. Prior to this, the two main sources of information for citizens were mass media and face-to-face social networks, with social media sharing some aspects with both. As illustrated in Figure 1, mass media include offline media, such as television, radio and newspapers as well as online media (e.g., *www.cnn.com*) and news aggregators (e.g., Google News). By contrast, until the arrival of social media, social networks were primarily formed via a wide array of face-to-face interactions, such as among family, friends and co-workers.

3.1 Social Media, Politics and Ideology

Social media has changed the media landscape in the past decade. The social media website, Facebook, launched in February 2004 and, by October 2012, claimed over 1 billion active users (Facebook, 2013). Twitter, the second most popular social networking site and ninth most popular website overall, has experienced similarly rapid growth since its inception in March 2006 (Picard, 2011); by May 2014, the site purported to have 255 million users active monthly. In February 2012, a Pew poll found that 15 percent of American adults used Twitter, a substantial increase from 8 percent in November 2010 when Pew first began polling on the subject (Smith and Brenner, 2012).

The growth of social media sites like Facebook and Twitter has influenced the manner in which Americans consume news. In 2008, only about 2 percent of all American adults reported regularly

using social media as a source of news (Kohut et al., 012b). By 2012, this figure had grown to 19 percent, with 3 percent citing Twitter specifically as a regular source of news. Paralleling its growth as a source of news generally, the use of social media as a source of political news has been rapidly expanding. During the 2012 election campaign, 20 percent of American adults received at least some campaign information through Facebook and 5 percent did so through Twitter (Kohut et al., 012a). This represents a significant increase compared to the 2010 campaign when only 6 percent received information about the campaign through any social media and less than 1 percent reported doing so through Twitter (Smith, 2011).

The remarkable growth of social media as a source of campaign news over this period has come as other sources of campaign news have begun to stagnate or even decline in importance (Kohut et al., 012a). Television has declined as a source of campaign news, from 38 percent in 2008 to 36 percent for cable network news, from 40 percent to 32 percent for local news, and from 32 percent to 26 percent for national network news. The influence of local papers saw the largest decline, with 31 percent of Americans reporting them as a regular source of campaign news in 2008 and only 20 percent in 2012. The Internet, which saw rapid growth as a source of campaign news earlier in the decade, has also stagnated in recent years, with 25 percent of American adults reporting it as a regular source of campaign news in 2012, little changed from a figure of 24 percent in 2008.

Americans are increasingly using social media for a wide variety of other political tasks as well. In 2008, 3 percent of American adults reported having used social media to post political news, a figure which reached 17 percent by 2012 (Smith, 2013). Also in 2012, 12 percent of American adults said they had used social media to "friend" a political candidate's account and 12 percent said they had started or joined a group on a social media website dedicated to a particular political or social cause, having both risen from an earlier level of just 3 percent in 2008.

To summarize, the role of social media in the American political discourse has been rapidly expanding. Its growing use as a source of political news has come as the importance of some traditional sources has begun to fall. Given this, the degree of ideological segregation and homophily on social media, relative to other sources of political information for voters, may have significant

implications for degree to which voters are exposed to an ideologically diverse set of high-quality information.

3.2 Twitter Interface and Experience

Twitter is an internet platform through which users connect with each other and gather information from a variety of sources. A Twitter user "follows" other users in order to receive updates on their posts. 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. Also, users are not updated on the activities of other users unless they actively choose to follow these accounts.

Twitter provides many avenues for users to find people to follow. When users first register for accounts, they select areas of interest and Twitter provides them with "initial suggestions" for whom to follow. Twitter frequently updates its initial suggestions based on which suggestions new users commonly follow and which suggestions they ignore. Twitter also generates suggestions based on a user's account history. Twitter will suggest users based on whom a user follows and whom those users follow. It also gathers information about which websites a user has visited in the Twitter Ecosystem (websites that have integrated Twitter widgets) and will suggest users with similar Twitter Ecosystem visitation patterns. If a user provides email account information, Twitter will suggest email contacts that have Twitter accounts. A user can also email contacts without Twitter accounts and suggest that they join Twitter.

Once connected, users communicate with one another via a variety of means. In terms of producing information, the starting point is a "tweet", defined as a message or update that is posted by a user and cannot exceed 140 characters.[3] Each user has a Profile that displays all of his or her published tweets. Tweets are public by default and can be seen by anyone, including people

[3]Within a tweet, users can mention other users by including an "@" symbol proceeded by another username. When a user is mentioned, it will only appear on the profile of the user who authored the tweet, and the mentioned user's profile will not display the tweet. If a user begins a tweet with an @username, it will only show up on the timelines of users who follow both the author and the user mentioned. Twitter assumes they are the only users who are interested in the direct interaction between two people.

without Twitter accounts. If an account is private, users manually approve every user who wishes to follow them and see their tweets. By 2012, 340 million tweets were being posted daily. Users can also transmit information produced by their friends via re-tweets, a republishing of another user's tweet. If a user clicks on the retweet button on a tweet, it will republish the tweet from his or her account with a small retweet icon. Retweets display the user that originally tweeted and the user that retweeted it.[4]

In terms of the consumption of information, the default interface is labeled Home.[5] This interface displays the user's timeline or feed, where users are exposed to tweets from their entire set of followees. These tweets are listed with the most recent on top, and the display of tweets is updated in real time.

4 Data

4.1 Construction of the Political Network

Our goal is to construct a social network of social media users who are politically engaged. Given this and lacking a direct measure of the ideology of Twitter users, we focus on Twitter users who follow politicians, defined here as candidates for the House of Representatives in 2012, and we use the party affiliation of these politicians to infer the ideology of the Twitter user.[6] 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]

[4]Instead of pressing the retweet button, some users will manually copy a tweet and label it "RT". Users often manually retweet if they want to add a comment to the tweet and do not want to retweet the entire tweet verbatim.

[5]Other interfaces includes Discover and Connect. The Discover tab displays tweets that Twitter believes a user will find interesting. These tweets are not necessarily from accounts that the user follows. Twitter's weekly emails are meant to resemble a shorter version of the Discover tab, which displays tweets that Twitter thinks the user will enjoy. Twitter carefully selects which tweets to display based a user's personal interests and search history. The user does not have to follow another user to see their tweets in his or her email or Discover tab. The @Connect tab displays the user's interactions with the world of Twitter. It includes new followers and tweets in which the user was mentioned, favorited, or retweeted.

[6]In some specifications below we also use information on accounts for Senators and candidates for the U.S. Senate.

[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.

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

To construct the network, we use information on links among voters, and this process is depicted in Figure 2. In particular, we downloaded the list of followers of each of the 2.2 million voters.[8]Out of our sample of roughly 2.2 million voters, we found over 1.5 million voters who are either following or are followed by other voters. These links among voters form the raw data necessary to construct our Twitter political network.

4.2 Voter Geography

In some specifications we employ information on voter location. This is useful for understanding the extent to which ideological segregation in social media is explained by geographical segregation. For example, if voters tend to disproportionately link to other votes within their state, and given that some states tend to be left-leaning and others right-leaning, then, due to geography, conservatives will tend to be linked to other conservatives and likewise for liberals. In the extreme, segregation in social media may be fully determined by geographic segregation.

In particular, we develop two separate measures of the location of voters, which is not automatically revealed. Our primary measure simply assumes that the location of the voter is the same as the location of the candidate. This allows us to measure both the state and the Congressional district for each user. In some cases, we also define state-level and district-level sub-networks, and, in these cases, we allow voters who follow multiple candidates to be assigned to multiple states and districts.

Second, roughly one-quarter of voters supply their location voluntarily. Voter-supplied location entries vary in specificity and format. We have used a simple procedure for inferring a user's state

[8]These data were particularly challenging to obtain since we needed to initially download the full set of followers, whether they were voters or not, for each voter. This amounted to 88 million Twitter accounts.

from the information he or she supplies, with a focus on two letter postal codes or full state names. To provide some sense of the accuracy of these user-supplied locations, Figure 3 plots the percent of Twitter voters from a given user-supplied state against the state's percent of U.S. 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 U.S. population.[9] This finding suggests that our set of Twitter voters closely reflect the distribution of actual voters in the United States.

4.3 Voter Ideology

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

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

As further evidence on our proxies for ideology, we have also downloaded information on Twitter accounts associated with significant media outlets and computed the fraction of liberal voters following each media outlet.[10] Using this information, Figure 5b plots, for the 25 outlets

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

[10] 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),

with the most followers in our sample of voters, the likelihood that a liberal voter follows a given outlet, relative to the likelihood that a conservative voter follows the same outlet. As shown, media outlets and programs traditionally considered to be right leaning, such as Rush Limbaugh, The Hannity Show, and Fox News, have a very low likelihood ratio. 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 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. In summary, these results suggest that our measures of voter ideology are reasonable and do capture some underlying measure of political preferences.

5 Empirical Framework

Based upon these Twitter data, we aim to describe the network of voters. In particular, the literature on social and economic networks has focused on two attributes of networks formed by heterogeneous agents: the degree to which types are homophilous and segregated from each other. Homophily is the tendency of voters of different types (liberal or conservative) to be linked to voters with the same ideology, relative to their tendency to be linked to voters with different ideologies. Segregation captures the degree to which voters of different types have distinct patterns of exposure, in terms of the set of voters they follow. The more these sets of followers are distinct from one another, the more isolated are the voter types from each other. Comparing the two measures, homophily captures within-group attributes, while segregation measures captures cross-group attributes.

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).

5.1 Measures of Homophily in Social Networks

For measures of homophily, we follow Currarini et al. (2009). In our setting, there are two different voter types, liberals and conservatives, denoted by $t \in \{l, c\}$. Let I be the total number of voters who follow at least one voter and I_t be the total number of type t voter followers, such that $I = I_l + I_c$. 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 followees of 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 $d_t = \frac{1}{I_t}\sum_{i \in I_t} v_{i-t}$ denote the average number of non-type t voters followed by type t voters (different). With these in hand, we define several related measures of voter type homophily.

DEFINITION 1: 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 friendships that are with voters of the same type t. Note that this basic index does not account for the distribution of types in the populations. Specifically, if a) liberals dominate the population and b) friendships are formed at random, then liberals would appear to be homophilous while conservatives would appear heterophilous. To address this issue, the literature has also focused on relative homophily. In particular, if the degree of homophily increases in group size in a given network, then the network is said to satisfy relative homophily.

DEFINITION 2: The profile (s_l, d_l, s_c, d_c) satisfies *relative homophily* if $w_t > w_{t'}$ implies $H_t > H_{t'}$.

Alternatively, in the case where friendships are assigned at random, the share of same type friends would be the same as their share in the population. Such a case is defined as baseline homophily.

DEFINITION 3: The profile (s_l, d_l, s_c, d_c) satisfies:

1. *baseline homophily* if for all type *t,*

$$H_t = w_t.$$

2. *inbreeding homophily* for type *t* if

$$H_t > w_t.$$

3. *heterophily* for type *t* if

$$H_t < w_t.$$

5.2 Measuring Ideological Segregation

For comparison with existing measures of ideological isolation, we compute the isolation index following Gentzkow and Shapiro (2011). For each followee $j \in J$, 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 account *j* as the fraction of his or her followers who are conservative:

$$share\ conservative_j = \frac{v_{jc}}{v_{jl} + v_{jc}}.$$

Defining $F_{ij} \in \{0,1\}$ as an indicator for voter *i* following account *j,* we can then define conservative exposure for each voter *i* as follows:

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

Taking averages across voters within groups, we then have conservative exposure for conservatives and conservative exposure among liberals. With these in hand, the isolation index is then given by:

$$isolation = conservative\ exposure_c - conservative\ exposure_l.$$

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 each group's exposure to content supplied by a distinct distribution of followed voters.[11]

Note that this measure does not include any information on the ideology of the followee. As an alternative measure that uses this information, we also compute a measure of conservative exposure by first defining the share of the followees of voter i who are conservative:

$$share\ conservative_i = \frac{v_{ic}}{v_{il+}v_{ic}}.$$

Then the conservative exposure of group t equals the average share of conservative voters that type t voters follow:

$$conservative\ exposure'_t = \frac{1}{I_t}\sum_{i \in I_t} share\ conservative_i.$$

As above, isolation follows by comparing conservative exposure for conservatives to conservative exposure for liberals. We present results using this alternative specification below as a robustness check to our baseline specification.

6 Results

Using the data described in Section 4 and the measures developed in Section 5, we next present our results. We begin by describing our homophily results followed by our isolation measures. Finally, we discuss and investigate possible selection issues involving the construction of our network.

[11] To illustrate the measure, consider the following example with two liberal voters (1 and 2) and two conservative voters (3 and 4). In terms of within-group links, assume that 1 and 2 follow each other and that 3 and 4 follow each other. In terms of cross-group links, assume that 2 follows 3 and that 4 follows 1. There are no other links across groups. Then, the share conservative for individuals 1, 2, 3, and 4 is given by 0.5, 0, 0.5, and 1, respectively. Further, the conservative exposure of individuals 1, 2, 3, and 4 is given by 0, 0.5, 1, and 0.5, respectively. Averaging within groups, conservative exposure for liberals equals 0.25, conservative exposure for conservatives equals 0.75, and isolation equals 0.50.

6.1 Ideological Homophily in Social Media

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

Using these measures, Table 2 provides estimates of homophily at the national level. The share of each group in the population (w) is identical to those in Table 1. The column reporting the average number of same-type friends (s) indicates that a liberal is likely to have 40 liberal followees on average out of a total of 59 followees across both ideological groups. Similarly, conservatives have 58 same-type followees out of total 68 followees. As shown, relative homophily holds for both of these groups since homophily is highest for the largest 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.

Table 3 provides alternative measures of homophily. To investigate the role of geography, we begin by comparing a) homophily in links between voters with the same user-supplied state of residence to b) homophily in links between voters with a different user-supplied state of residence.[12] That is, the former focuses only on within-state connections, whereas the latter focuses on cross-state connections. Presumably, if geography plays a role in facilitating connections among voters with similar ideologies, then our measures of homophily should vary across these voter networks. Overall, except for conservative voters from different states, where homophily falls from the baseline of 0.844 to 0.779, the homophily measures are broadly similar to each other and to those we obtained for the baseline network. This suggests that voter geography neither facilitates nor impedes one's tendency to connect with ideologically similar voters.

We next expand the network by taking into account followers of Senators and candidates for Senate. We find that homophily does not vary substantively when considering a larger sample of voters who follow either House or Senate candidates; however, when we look at the network of

[12]Note that this analysis does not incorporate links in which at least one of the two voters does not provide a state.

Senate voters alone we make two observations. First, liberals now account for the majority (0.511) of the voters. This is in contrast to the majority share of conservatives in the previous networks. Second, since homophily is greater among conservatives than liberals, the network does not exhibit relative homophily in this case.

Moving next to sub-networks, we investigate several characteristics of the political network we have constructed. In particular, using our state-level homophily estimates, we investigate whether our data support findings of Currarini et al. (2009) on friendship networks of high school students. In particular, we focus on the following findings: (a) larger groups form a larger share of their friendships with people of their own type, (b) groups inbreed and (c) larger groups form more friendships per capita.

Using variation in group size at the state level, Figure 6a plots the homophily index (H) for each type against their share in the population (w). Each point in this figure is an ideological group at the state level. As shown, almost all observations are above the 45 degree line, implying that inbreeding homophily is satisfied ($H > w$). Thus, our results support the finding (b) regarding group inbreeding. Also, the general pattern is consistent with relative homophily as well since homophily is broadly increasing with w; however, strictly speaking the definition of relative homophily is not satisfied since some groups that have a higher share in the population than other also have a lower homophily estimate. Thus, our data tend to support finding (a) that larger groups tend to form a larger share of own-type friendships. Again using state-level data, Figure 6b presents scatter plots of followees per capita for each groups against the group's share in the population. The linear fit is presented as well to show the general trend. As shown, a move from 0 to 1 in the share of the population increases the number of followees per capita within the group from about 40 to 60 followees. Thus, our data are also consistent with finding (c) that larger groups have more followees per capita.

Finally, Figure 7 presents corresponding results at the level of the Congressional district. As shown, the results again support the findings of Currarini et al. (2009) in the sense that (a) larger groups tend to be more homophilous, (b) groups inbreed, and (c) larger groups tend to have more

connections per-capita.

6.2 Segregation in Social Media

Having demonstrated that the Twitter political network exhibits patterns of homophily similar to those observed in other social networks, we next examine whether the degree to which groups are isolated in social media is similar to other face-to-face social interaction or more closely resembles isolation of groups with respect to exposure to traditional media outlets.

In Table 4, we report conservative exposure estimates for liberals and conservatives at both the national-level and at the state-level. As shown, at the national-level, conservative exposure among conservatives is 0.776, and conservative exposure among liberals is 0.372. Thus, the isolation index at the national-level is 0.403.

To put our results in context, Table 4 places this estimate of national isolation in social media to other media and face-to-face interactions from Gentzkow and Shapiro (2011). Comparing the isolation indices, we find that social media is highly segregated, with an isolation index of 0.403, which is similar to face-to-face interactions with political discussants, the second most segregated environment. We also find that isolation in social media is significantly higher than other forms of face-to-face interactions. Finally, we note that national isolation in social media is much higher than in traditional media, which vary from 0.018 for broadcast news to 0.104 for national newspapers.

In Table 5, we provide a series of alternative measures of ideological segregation using our data. First, using our measure of voter state, we distinguish between cases in which the two linked voters are from the same state and cases in which the two linked voters are from different states. As above, note that this analysis does not incorporate links in which at least one of the two voters does not provide a state. As shown, the network is somewhat less segregated when linked voters are from different states than when voters are from the same state. The differences between these two measures, however, are relatively small.

Moving to our measures of isolation in sub-networks, we first present results for state-level

networks. As shown, when averaged across states, segregation rises somewhat, from 0.403 in the baseline to 0.430 at the state level. Moving to the district level, segregation falls from its baseline of 0.403 to 0.358. These differences are small in general, and putting together these results with those for same-state and different state networks, we conclude that geography does not play a primary role in explaining ideological segregation.

Next, while our baseline analysis focuses on House politicians, we next extend our sample to include followers of Senators and candidates for the U.S. Senate. As shown in Table 5, when we combine followers of House and Senate candidates, the measure of segregation falls from the baseline value, and when we focus only on followers of Senate candidates, the measure of segregation increases.

As a robustness check, we next present results using the alternative isolation measure described above. In particular, we measure the difference between conservative and liberal voters in terms of their share of followees who are conservative. Unlike the baseline measure, this uses information on ideology of the followee. As shown in Table 5, the result of 0.469 suggests a higher degree of segregation in the network.

6.3 Sample Selection

Given the construction of our political network based upon Twitter users who follow politicians, a natural critique of our analysis is that our sample may not be representative of either a) voters at large, or b) Twitter users and social media users at large. On the first point, it is well established that, when compared to the general population, users of social media are younger, more highly educated, more likely to be non-white, and more likely to use mobile devices (Pew, 2013). While we concede that our sample is not representative of the population at large, it is important to emphasize that social media represent a significant and growing source of information for voters. That is, social media is an important sector to study in and of itself. Moreover, a focus on social media allows us to compare our results to those in other settings.

On the second point, it is quite plausible that our sample, constructed by selecting users who

follow politicians, may tend to disproportionately include individuals with strong ideological views and stronger preferences for linking to like-minded users. While we again argue that followers of politicians are an important subsample of social media users, especially when considering ideological segregation, our selection criteria may make it difficult to compare our results to other measures. To address this issue, we first compute ideological segregation for our sample of voters in terms of media consumption, using information on the followers of our sample of media outlets described above. Note that this measure does not use any information on links between voters. Instead, we treat our voters as if they are only consuming information from these media outlets on Twitter. As shown in Table 5, 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 baseline measure of isolation (0.403). The latter point is true even when focusing on the subsample of users who follow both candidates and media outlets, for whom network-based segregation equals 0.394. Thus, these same Twitter users experience lower segregation when using Twitter to consume news from traditional media outlets than when using Twitter as a social network by linking to other voters. This suggests that the differences between our baseline results and Gentzkow and Shapiro (2011) are driven, at least in part, by differences between the consumption of news from media outlets on the internet and exposure to information from self-chosen links on social media, rather than by our specific criteria for creating the social network or our focus on a single social network, Twitter.

As a second attempt to address this selection issue, we split our sample into "extremists", those who follow only one party, and "moderates", those who follow at least one candidate from each of the two parties. By focusing on moderates, those with less extreme ideological views and possibly weaker preferences for linking to like-minded users, our sample may be more representative of Twitter users and internet users at large. Consistent with this view, we find that network-based segregation is indeed much higher for extremists (0.417) than for moderates (0.217). Finally, we compute isolation in media consumption for this group of moderates. As shown in the final column, segregation in media consumption for this group equals 0.067. Thus, the measure is again lower

than the network-based measures for this group (0.217) but is now comparable to the measure of 0.075 in Gentzkow and Shapiro (2011). The similarity of these results suggests that this group of voters may more accurately reflect the sample of internet users at large.

Finally, we use this distinction between extremists and moderates in the context of homophily. As shown in the final two rows of Table 3, homophily is lower among moderate liberals (0.599) and moderate conservatives (0.783), when compared to the baseline measures for liberals (0.688) and conservatives (0.844). The opposite pattern is evident for extremist liberals (0.695) and extremist conservatives (0.849). These results are consistent with the results for ideological segregation in the sense that extremist voters may have stronger preferences to link to like-minded users than moderate voters.

7 Conclusion

There is an active debate worldwide regarding the role of social media in societies. On the one hand, authoritarian regimes may view social media as a threat to stability. On the other hand, governments may wish to harness social media for the purpose of providing information to citizens. The issue of ideological segregation is important when providing such information. Exposure to diverse viewpoints in a society helps to ensure that information is disseminated with little friction across a large number of people. When a community is polarized and is divided into factions, by contrast, information may spread unevenly and may miss intended targets. Our results suggest that social media are highly segregated along ideological lines and thus emphasize these potential problems associated with the flow of information in segregated networks.

To summarize, we investigate the degree of homophily and ideological segregation on Twitter, one of the leading social media websites. Using information on links between followers of accounts associated with candidates in the 2012 election, we find that Twitter is segregated along ideological lines. In particular, followers of Republican candidates are much more likely to be linked to other followers of Republican candidates and likewise for followers of Democratic candidates. This

Twitter political network is at least as segregated along ideological lines as face-to-face interactions and is much more segregated than traditional media outlets, such as newspapers and television. Taken together, these results suggest that social media may be a force for increasing ideological segregation.

References

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

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

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

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

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.

De Choudhury, M. (2011). Tie formation on twitter: Homophily and structure of egocentric networks. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 465–470. IEEE.

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

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

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

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

Facebook (2013). Timeline.

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

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

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

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

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

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

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

Himelboim, I., S. McCreery, and M. Smith (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*.

Jackson, M. O. (2010). *Social and economic networks*. Princeton, NJ: Princeton University Press.

Kohut, A., C. Doherty, M. Dimock, and S. Keeter (2012a, February 7). Cable leads the pack as campaign news source. *Pew Center for the People and the Press*.

Kohut, A., C. Doherty, M. Dimock, and S. Keeter (2012b, September 27). In changing news landscape, even television is vulnerable. *Pew Center for the People and the Press*.

Picard, A. (2011, March 20). The history of twitter, 140 characters at a time. *The Globe and Mail*.

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*.

Smith, A. (2011, January 27). 22politics in 2010 campaign. *Pew Internet & American Life Project*.

Smith, A. (2013, April 25). Civic engagement in the digital age. *Pew Internet & American Life Project*.

Smith, A. and J. Brenner (2012, May 31). Twitter use 2012. *Pew Internet & American Life Project*.

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

Wu, S., J. M. Hofman, W. A. Mason, and D. J. Watts (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 705–714. ACM.

Yardi, S. and D. Boyd (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society 30*(5), 316–327.

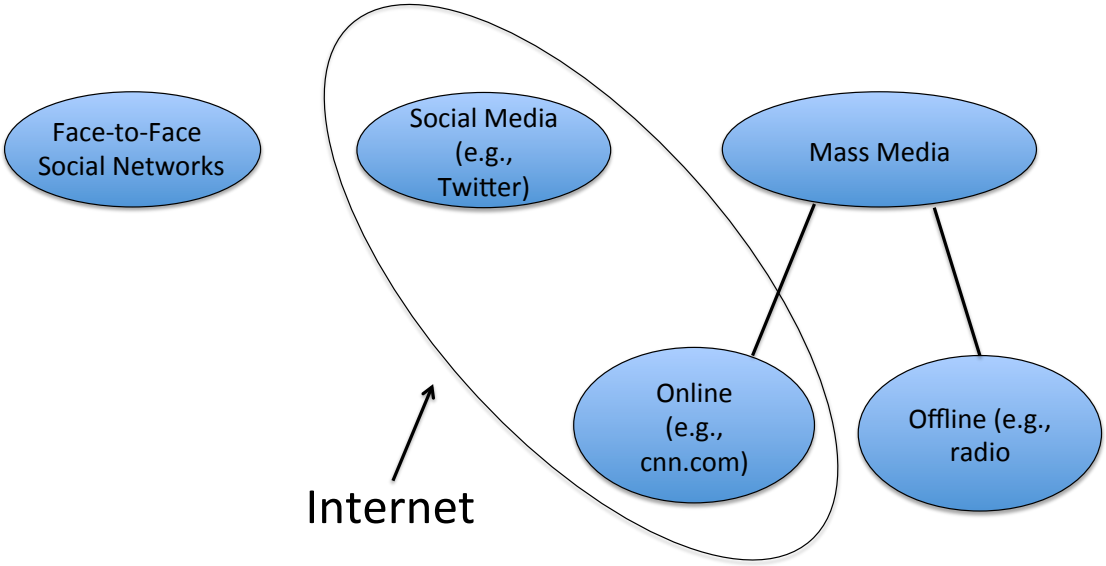
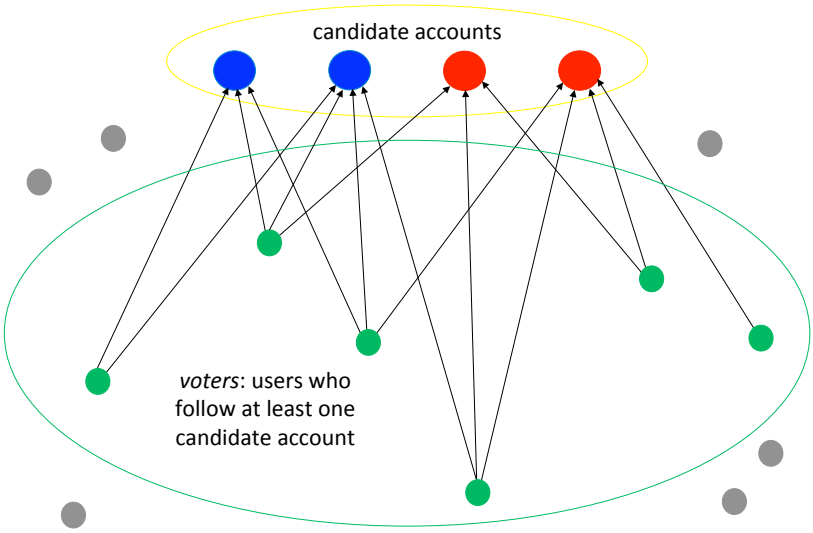
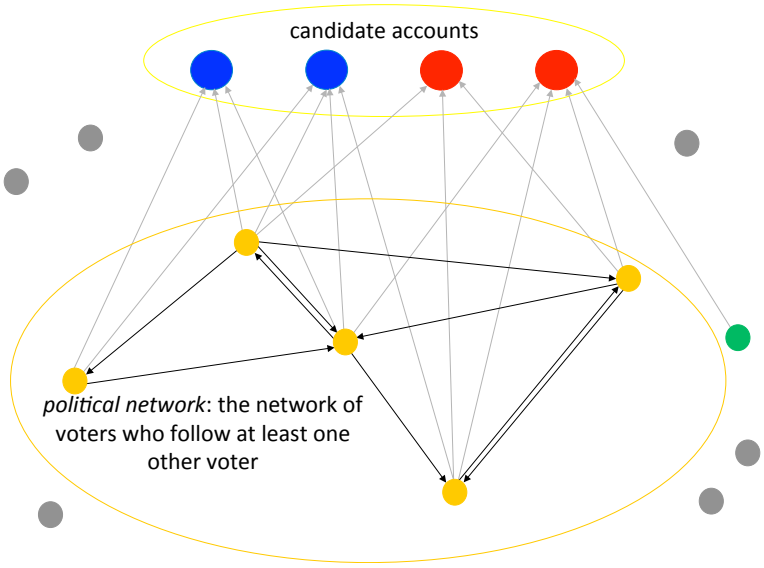


Figure 1: Sources of Information



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



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

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

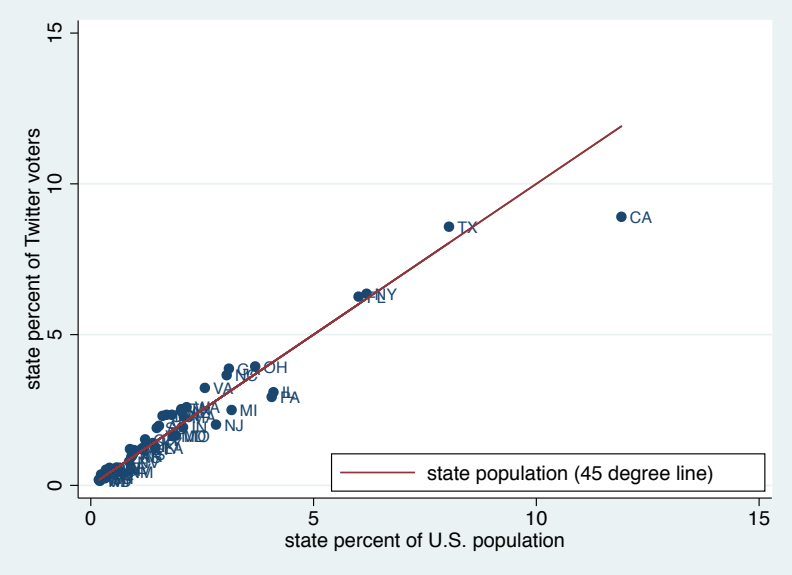


Figure 3: Spatial Representation of Twitter Voters

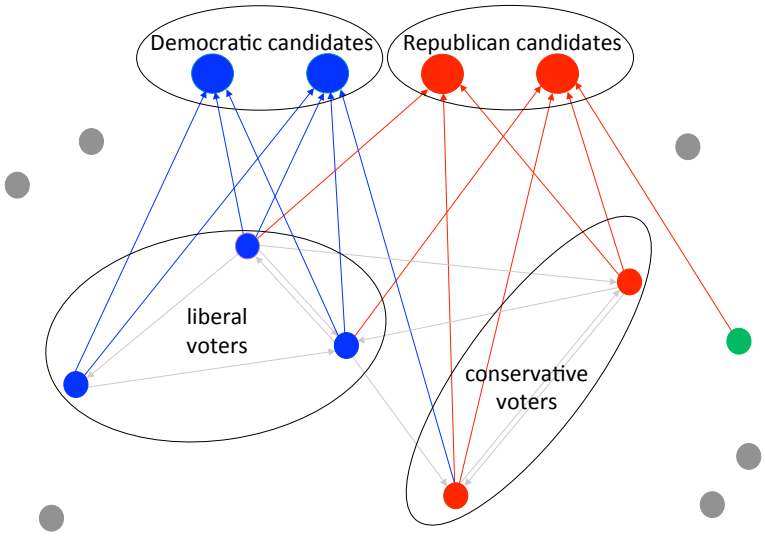
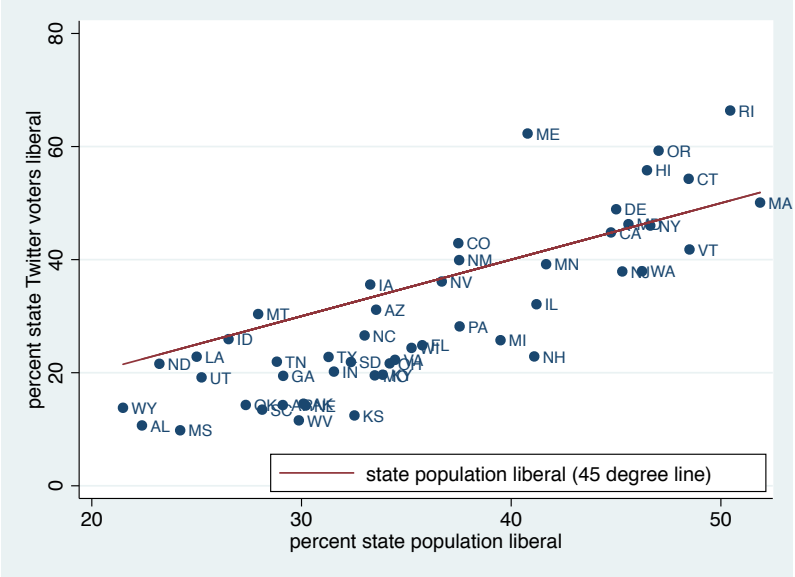
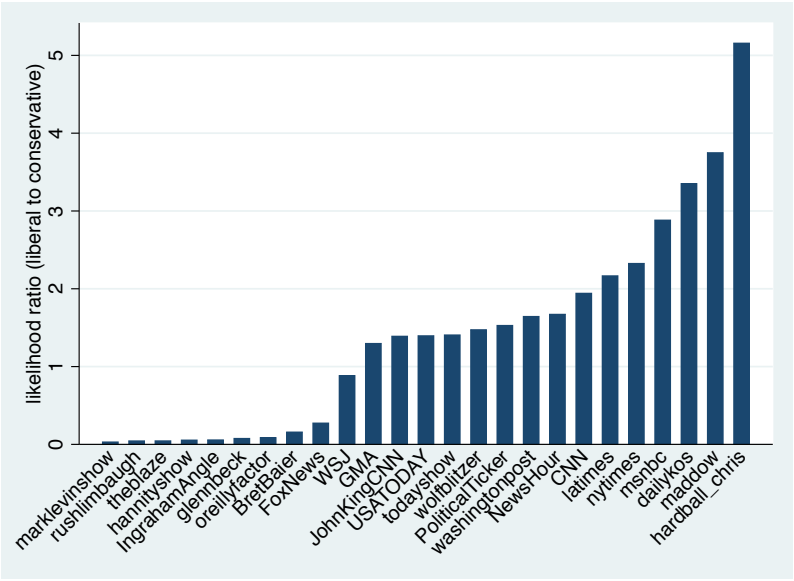


Figure 4: Inferring Voter Ideology

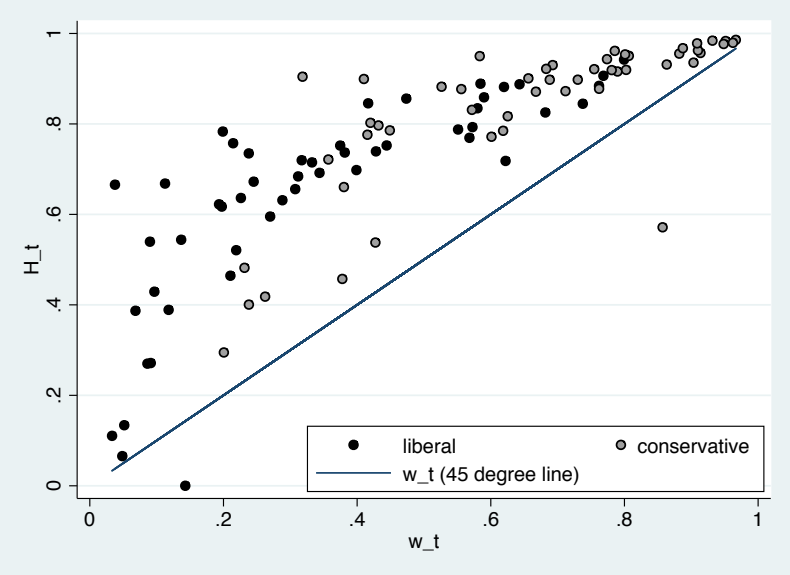


(a) Share of State Liberal Voters and Liberal Twitter Users

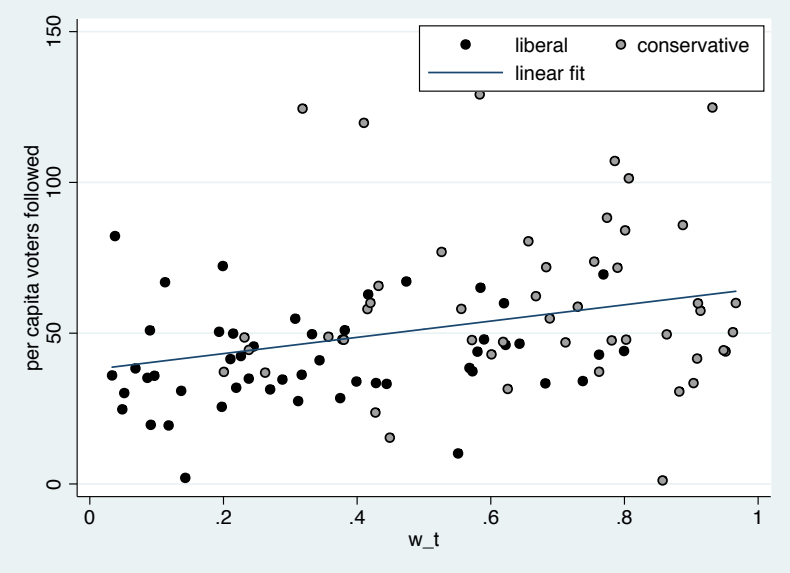


(b) Likelihood Ratio of Following Media Outlets

Figure 5: Validation of Ideology Measure for Twitter Users

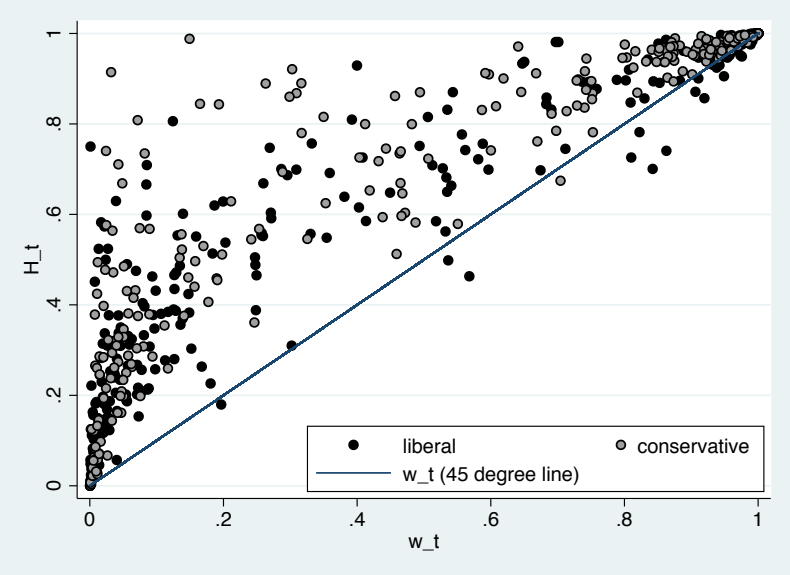


(a) Group Size and Ideological Homophily

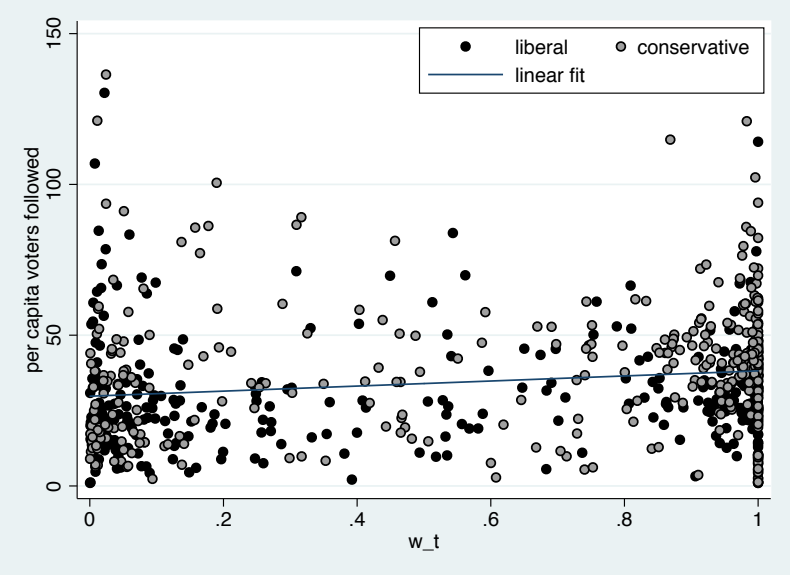


(b) Group Size and Per Capita Twitter Voters Followed

Figure 6: Group Size and State level Interactions with Voters



(a) Group Size and Ideological Homophily



(b) Group Size and Per Capita Twitter Voters Followed

Figure 7: Group Size and District level Interactions with Voters

Table 1: Tabulation of Voter Links

	Ideology of Voter Follower	
	Liberal	Conservative
Percent followed	n = 489,479	n = 868,001
by ideology	36.06	63.94
Liberal	67.11	20.25
Conservative	32.89	79.75

Table 2: National Ideological Homophily

	Population share	Same-type voters followed	Per capita voters followed	H index
Liberal	0.361	40.416	58.756	0.688
Conservative	0.639	57.828	68.486	0.844

Table 3: Comparing Homophily Estimates

Sample restrictions and measurement	Sample of candidates	Liberals		Conservatives	
		Share	H index	Share	H index
Baseline	House	0.361	0.688	0.639	0.844
Linked voters are from the same state	House	0.338	0.684	0.662	0.779
Linked voters are from different states	House	0.342	0.659	0.658	0.845
Baseline	House and Senate	0.385	0.692	0.615	0.824
Baseline	Senate	0.535	0.814	0.465	0.873
Followers of Democrats and Republicans	House	0.337	0.599	0.663	0.783
Followers of one party only	House	0.361	0.695	0.639	0.849

Table 4: Ideological Segregation

	Conservative exposure of Conservatives	Liberals	Isolation index
Social Media			
Baseline	0.776	0.372	0.403
Face-to-face interactions			
Political discussants	0.796	0.402	0.394
People you trust	0.675	0.372	0.303
Family	0.69	0.447	0.243
Neighbourhood	0.627	0.439	0.187
Work	0.596	0.428	0.168
Voluntary associations	0.625	0.48	0.145
County	0.682	0.622	0.059
ZIP code	0.637	0.543	0.094
Media			
National newspapers	0.612	0.508	0.104
Internet	0.606	0.531	0.075
Local newspapers	0.695	0.647	0.048
Magazines	0.587	0.54	0.047
Cable	0.712	0.679	0.033
Broadcast news	0.677	0.66	0.018

Note: Source for data on face-to-face interactions and media is Gentzkow and Shapiro (2011)

Table 5: Comparing Ideological Segregation Estimates

Sample restrictions and measurement	Sample of candidates	Conservative exposure of Conservatives	Conservative exposure of Liberals	Isolation index
Baseline	House	0.776	0.372	0.403
Linked voters are from the same state	House	0.772	0.374	0.399
Linked voters are from different states	House	0.780	0.414	0.366
State-level networks (average across states)	House	0.832	0.402	0.430
District-level networks (average across districts)	House	0.752	0.394	0.358
Baseline	House and Senate	0.754	0.372	0.382
Baseline	Senate	0.750	0.278	0.472
Alternative isolation specification	House	0.798	0.329	0.469
Followers of media and candidates (network segregation)	House	0.780	0.387	0.394
Followers of media and candidates (media segregation)	House	0.789	0.547	0.241
Followers of one party only	House	0.776	0.358	0.417
Followers of both Democrats and Republicans	House	0.716	0.499	0.217
Followers of media and both parties (media segregation)	House	0.723	0.656	0.067