Detailed Description

Objectives

Our research proposal aims to shed light on the contribution of social media to the exchange of and exposure to diverse viewpoints. In Project I, drawing from the literature on network analysis, we construct a measure of ideological segregation in social media, compare it to other media platforms (e.g., radio, national newspapers, etc.) and explore the extent to which ideological segregation in social networks varies with the election cycle. Specifically, we ask whether social networks are more segregated in the weeks leading up to Election Day. Further, we explore whether patterns of information transmitted and the content of such communications may reflect voting intentions of social media users. In Project II, we examine how such content transmitted in different types of social networks, as defined by their degree of ideological segregation, maps to elections outcomes. These analyses speak to how social media influences political engagement in both static and dynamic settings, and gauge the influence electoral institutions may have on ideological segregation in social media as well as what type of content and social networks best reflect subsequent political outcomes.

Context

Social media transcends existing forms of online and offline media because it fosters social interaction and empowers individuals to influence as well be influenced by the flow of information. Social network sites, such as Twitter and Facebook, are playing increasingly important roles in elections in the United States and around the world. Given this, an important issue involves the degree of ideological segregation in this emerging media environment. On the one hand, it is possible that social media lessens ideological segregation. In particular, as documented by Putnam (1995, 2001) and others, geographically-defined communities in the U.S. are increasingly segregated by ideology, and the emergence of social media may present citizens with the opportunity to engage politically with fellow constituents beyond these traditional boundaries. On the other hand, to the extent that social media functions as an echo chamber, with conservatives linked to other conservatives and liberals linked to other liberals, social media may increase ideological segregation (Sunstein, 2001).

Whether social media increases or decreases ideological segregation has important implications for politics and policy. Putnam (1994) argues that representative government is most effective when political engagement of citizens is bipartisan. Likewise, seminal work by Downs (1957) and Becker (1958) emphasizes the importance to democracy of voters' exposure to diverse viewpoints. Along these lines, guaranteeing that individuals encounter information that may challenge their pre-existing beliefs has been a goal of media policy in the United States and around the world (Gentzkow and Shapiro, 2008).

The most closely related study to our proposed work is Gentzkow and Shapiro (2011), which shows that the Internet is unsegregated along ideological lines, especially when compared to other media platforms, such as national newspapers, and other more general social contexts, such as geographically-defined neighborhoods. Their study, however, focused on visits by Internet users to news websites, such as *www.cnn.com* and *www.nytimes.com*, and did not examine social networks on the Internet. Unlike more traditional websites, the exposure to content by users of social networks depends upon self-chosen links between individuals. That is, two individuals visiting the Twitter website may experience dramatically different political content depending upon the ideology of their

respective networks, defined as the accounts of other Twitter users that they have chosen to follow. Given this diversity of experiences, we hypothesize that the results of Gentzkow and Shapiro (2011) may not extend from the exposure to ideology on news websites to the exposure to ideology on social network sites. A recent survey by the Pew Research Center suggests that the most active and engaged political participants on social media sit at opposite ends of the ideological spectrum, yet their experiences around political material on social media are quite similar.[1] Further, a study by Conover et al. (2011) suggests that information transmission (i.e., retweets) on Twitter is significantly more partisan than information creation (i.e., mentions). Our project seeks to determine whether these initial findings survive more rigorous scrutiny.

More broadly, our proposed research contributes to a literature, including our own work, on how voters acquire and use information during elections. Chiang and Knight (2011) show that biased, or unsurprising, newspaper endorsements have less influence than unbiased, or surprising, endorsements. Durante and Knight (2012) show that television viewers in Italy respond to a change in the bias of public television by re-sorting across stations according to their ideology. Knight and Schiff (2010) examine whether late voters learn from early voters in sequential elections, and Halberstam and Montagnes (2012) provide evidence that voters learn about the ideology of candidates in congressional races from the ideology of candidates in the contemporaneous presidential race.

Project I: Ideological Segregation on Twitter

Objective

In this study, we plan to analyze the degree of segregation, or ideological isolation, that exists in a "political network" that is created by individuals using social media. We focus on the social network site Twitter during the pre-election environment (the weeks leading up to the 2012 electoral campaigns) and during the post-election, or governing, period (the time period after winning candidates take office in January 2013).

Methodology

In terms of the pre-election period, we define a Twitter user as a "voter" in this political network if they "follow," defined as having chosen to view content from, at least one candidate running for federal office during the 2012 election. We can further characterize voters as Democrats or Republicans based upon the party affiliation of the candidates that they choose to follow. We collected data on followers of candidates on a daily basis during the two months leading up to Election Day and also in the two months following the election.[2] To construct the political network, we plan to harness information on links between individual voters (using information on which voters follow one another), and we are currently working to collect this information from Twitter. In terms of the post-election environment, we are currently working to update and find any new Twitter

[1] For more information on the survey, see: http://pewInternet.org/~/media//Files/Reports/2012/PIP_SNS_and_politics.pdf

[2] In particular, we have assembled and have been tracking over 1,200 Twitter accounts daily since early September 2011. These include approximately fifty of the U.S.'s top media outlets, all Twitter accounts of Democratic and Republican congressional and presidential candidates (some have multiple accounts) as well as incumbents not running for reelection and senators not up for reelections (we plan on using these as control groups in future analyses). The basic daily information we collect for each account is: the number of tweets, followings, followers and mentions. We also collect the content of the tweets, mentions of each candidate, and other information, such as the exact time of a tweet, the number of retweets of the tweet, the unique identification of the user who tweets, etc. The unique identity of the user who tweets can then be used to retrieve any information from that user's Twitter account.

accounts for all candidates who won their respective elections in November and currently hold office as of January 2013.

Based upon these data, we will compute the degree of ideological segregation among Twitter users, and investigate several other measures. First, we can compute the isolation index, defined as the difference between the probability that two linked voters are from the same political party and the probability that two linked voters are from different political parties:

$$\textit{isolation} = \Pr(\textit{two linked voters from same party}) - \Pr(\textit{two linked voters from different parties}) \quad (1)$$

If Democrats are only linked to Democrats and Republicans are only linked to Republicans, ideological segregation is maximized, and this measure will equal one. If Democrats are equally likely to be linked to Republicans and Democrats, ideological segregation is minimized, and this measure will equal zero. This measure, which varies between 0 and 1, is analogous to the isolation index formulated by Gentzkow and Shapiro (2011) and will allow us to compare our results to their findings on the isolation indices of traditional media platforms and other social contexts. This comparison sheds light on whether social media are a force for increasing or decreasing ideological segregation.

This isolation index, as defined above, does not account for indirect links between voters. That is, if A is directly linked to B but not C and B is directly linked to C, then A is indirectly linked to C.[3] To account for these indirect links, we plan to use the tools developed in a burgeoning literature in economics and related fields on the analysis of networks (Jackson, 2008). This field has developed summary statistics, such as the degree of clustering, that are designed to measure segregation in a network. Similarly, we can calculate the number of links required, on average, to travel in the network from a Democratic voter to another Democratic voter, relative to the number of links required, on average, to travel from a Democratic voter to a Republican voter. Finally, we also plan to implement a new measure of social segregation within networks developed by Echenique and Fryer (2007).

We will compute these measures of ideological segregation both during the pre-election and the post-election periods. To the extent that followers of the losing candidate rally around the winning candidate and choose to follow this candidate after the election is resolved, then the post-election network may be less segregated along ideological lines than the pre-election network. This result would suggest that election periods might be more polarized than non-election periods, during which policies are implemented to cater to broader post-election constituencies.

Finally, in addition to measuring the network of voters, in terms of who follows whom on Twitter, we also plan to investigate the content of political communication within the network. This analysis will introduce measures both of the intensity of communications, based upon mentions and re-tweets in Twitter, as well as the tone of such communications. Regarding the latter, we plan to use sentiment measures developed by Topsy Pro Analytics. In particular, Topsy has developed a sentiment measure that distinguishes between positive, neutral, and negative mentions depending upon the content of the tweet.

[3] If indirect links are formed through non-candidate followers, we plan on using media outlets or other user followings to determine the political affiliation of a user. See Footnote 3 for more information on using media outlets for political inference.

Project II: Using Twitter to Predict Election Outcomes

Objective

The idea behind the "wisdom of the crowd" is that, while individual predictions may tend to be inaccurate, aggregate predictions tend to improve accuracy as errors in individual predictions cancel out one another. Applying this idea to the social media context, this project explores whether or not online content generated by a large number of users can be used to predict real-world phenomena. In particular, we propose to gauge the extent to which Twitter activity in the days leading up to the 2012 election can be used to predict political outcomes.

Methodology

To this end, we have begun to collect data on aggregate Twitter activity, as provided by Topsy Pro Analytics. Two aspects of the data are central to addressing the question of how social media aggregate predictions. First, they have developed a method for geocoding Twitter users based upon a variety of factors, including content, and provide location information for 90 percent of all tweets. Second, as mentioned above, they have developed a measure of "sentiment" that distinguishes between positive, neutral, and negative mentions. This allows us to measure whether a user mentioning Obama, for example, is doing so in order to express his or her support or opposition to the candidate.

As preliminary evidence on whether Twitter content can be used to predict election outcomes, we have collected data on Twitter activity by state and sentiment during the seven days leading up to the election (October 29 to November 5). Using these data, we calculate a "sentiment index" for Romney, relative to Obama, on a state-by-state basis as follows:

$$\ln(\textit{Romney positive}) - \ln(\textit{Romney negative}) - \ln(\textit{Obama positive}) + \ln(\textit{Obama negative}) \quad (2)$$

Then, using actual state-level voting returns for the 2012 election, we calculated an analogous "voting index" as follows:

$$\ln(\textit{Romney votes}) - \ln(\textit{Obama votes}) \quad (3)$$

Comparing these two indices, we document a correlation of 0.69 that is statistically significant. Further, to illustrate the importance of the sentiment of the tweets in this analysis, we computed the correlation between the voting index and a measure of Twitter activity that is independent of sentiment (the "overall-mentions index"):

$$\ln(\textit{Romney mentions}) - \ln(\textit{Obama mentions}) \quad (4)$$

Surprisingly, the correlation between these two indices is negative, suggesting that individuals are primarily using Twitter to express a negative sentiment about a candidate, perhaps in order to encourage others to vote against this candidate. This highlights the importance of distinguishing between a tweet that has positive versus negative content. Finally, we run a specification in which the voting index is regressed on both the sentiment index and the overall-mentions index. Consistent with the hypothesis that Twitter is largely used to denigrate a candidate, only the coefficient on the

sentiment index is statistically significant. After controlling for sentiment, the coefficient on overall mentions is statistically insignificant.

Building on this preliminary evidence, we plan to expand our research in three directions. First, we hope to conduct this analysis of presidential voting returns at finer geographic levels, such as U.S. counties. Second, we plan to extend the analysis to other federal races, such as those for the U.S. House and the U.S. Senate. Compared to the presidential election, polls in congressional races are far less common, and thus Twitter activity may represent a valuable opportunity to predict outcomes in these elections. And third, we plan to identify the degree to which different types of user or group sentiments can predict election outcomes. For example, using information on media outlets and political candidates followed by each Twitter user, we can associate each tweet with the ideological score of its user.[4] This ideological score data allows us to examine whether, for example, the sentiment of moderate users better predicts election outcomes than the sentiment of partisans. Furthermore, we can examine how this relationship might depend on the centrality of the user who supplies the tweet (e.g., by the number of followers or number of mentions) and the political diversity of his followers.

In addition to these directions, we plan to expand the analysis in this project to social networks on Twitter by exploring the implications of indirect connections among Twitter users. Using the methods we developed to measure the isolation of social networks in Project I, we will examine how the sentiment index of different types of networks, as characterized by the degree to which they are isolated, relates to election outcomes.

[4] There are several existing methods used to compute ideological scores for media outlets. Recent examples include papers by Milyo and Groseclose (2005) and Gentzkow and Shapiro (2010) who calculate scores for U.S. newspapers; An et al (2012) and Golbeck and Hansen (2011) use information on media outlet Twitter followers to infer media ideology.