Congressional Use of Twitter

Final Report

By Jessica Su and Sohini Khan

**I. Introduction**

**Research question**

Twitter is a microblogging site with 500 million users. [1] Users may post “tweets,” which are messages of 140 characters or less. They may post “retweets,” which are copies of others’ tweets. They may also “follow” users, which causes their tweets to show up in their newsfeed.

Since Twitter has such a large userbase, some members of Twitter have amassed a large number of followers. Among these are politicians such as Mitt Romney (1.5 million followers) and Barack Obama (30 million followers). [2] Each follower is part of their audience; they read their tweets, retweet their messages, and take appropriate actions in real life. For this reason, having large numbers of followers gives them political power. On March 19, Barack Obama posted the tweet “RT if you stand with same-sex couples around the country fighting for the right to marry who they love.” This tweet got 39,000 retweets, which means 39,000 people spread his message to all of their followers. [2]

It is obvious that politicians use Twitter to communicate with the electorate and achieve their political goals. However, we would like to formalize this by quantifying the ways that politicians use Twitter. Specifically, we would like to answer the following:

* How does congressional Twitter use correlate with personal characteristics of the congressperson, such as age, title, gender, or political party?
* How does it correlate with the voter demographics of their state?
* How does it differ between congresspeople in rural and urban states?
* Do people holding leadership positions in Congress exhibit markedly different Twitter behavior from normal congresspeople?
* What can we say about the content of congressional tweets? How do congresspeople from different parties react differently after major world events?

**Previous research**

There has been previous work done that looks at trends in the use of social media both by voters and congresspeople. In fact, social media is a very useful tool as by October 2012, 98% of Congress has been shown to use Facebook, Twitter, or YouTube as a platform of communication. [18]

A study done in 2009 by Golbeck et al. [19] analyzed Twitter use by ~70 congresspeople where they collected approximately 4500 tweets in one month. They were trying to analyze the content of tweets and found that 54.7% were informational (like “mini press releases”) and 72% contained links pointing to longer posts. They also noted how the congressional calendar affected the content of the tweets. This type of study is important because it analyzes the evolution of the content in tweets and shows how this sort of analysis can be significant in order to understand which sorts of issues are important for congresspeople to show public support for and see what they think the public cares about. The Golbeck et al. study also found that 44.8% of tweets contained links. Overall, this study helps conclude that Twitter is an accessible and useful database to gather information about congresspeople especially since they have a political presence in social media and not just a social one.

A study in 2012 by Greenberg et al. [18] followed 408 members of Congress on Twitter and studied 30,765 tweets for 3 months. They studied both the content of tweets as well as general demographics of congresspeople to see if they could find any trends. For example, they saw that 40.5% of tweets were position-taking while 25.6% were about district/state issues. They also noted how 71% of tweets were made while Congress was in session. They also saw how uneven Twitter usage is amongst all the congresspeople. In other words, the top 20% of most frequently tweeting congresspeople were responsible for 53.6% of tweets that they counted. Additionally, they found that the highest frequency user group was the Senate Republicans, that there weren’t any significantly strong geographical trends, how 80% of posts by the Senate were in the first-person while 90% of posts by the House were in the first person, and more. The Greenberg et al. study was also able to conclude that gender did not affect Twitter usage significantly but that younger members on average used Twitter more than older users.

As this study was done during an election year, it is important to have a more objective analysis of Twitter usage during a normal year as we intend to do in our study to see how congresspeople remain active even when political issues might not be as popular or trending as during an election year.

**II. Data**

We retrieved the following data: (1) Personal information for each congressperson on Twitter, including their name, age, party, title, gender, state, and Twitter handle. (2) Twitter data for each congressperson, including their join date, number of tweets, number of followers, and number of users they are following. (3) Demographics for each state, including the number of people in each age bracket and the fraction of people of each race. (4) The total population of each state.

We took the personal information from the Sunlight Foundation. [3] The table was available for direct download, and we wrote a Python script to remove information that was not pertinent to our analysis. The state demographics came from Wikipedia [4] and the 2010 US Census [5]. The total population of each state also came from the US Census. [8] In each case the data was available for direct download.

We gathered the Twitter data using the Twitter API [6] and python-Twitter [7]. The Twitter API is a set of tools Twitter built to allow developers to scrape information from the site, and python-Twitter is a tool that makes it easy to use the Twitter API from Python scripts. To use these, we first created a Twitter account, “jsuPS120,” and generated an authentication key. Then we found a “user object” for every congressperson in our list. Each object contained (among other things) the join date, number of tweets, number of followers, and number following. We used standard Python tools to export this information to a file. (Instead of writing the join date and number of tweets separately, we combined them to get the number of tweets posted per month.)

The Twitter data was slightly challenging to get because the Twitter API is rate limited. That means you can only get data for 180 users every 15 minutes. Since there are only 471 congresspeople on Twitter, this was not prohibitive to our analysis. However, we did have to write the code in a way that respected the rate limits. Originally, we tried using the “sleep” command to force the program to wait for 15 minutes every time it processed 170 users. But we kept getting strange bugs, so we gave up and split the congresspeople into three lists. While this was fine for our current purposes, our approach would not scale to larger datasets.

After getting all this data, we used a MySQL database to correlate the tables. We needed to do this because the congressional data, the Twitter data, and the state demographics were all in different tables, and integrating the tables by hand would have been impossible. We created four MySQL tables: “Twitter” (containing Twitter data), “statepop” (containing state populations), “race” (containing race demographics information), and “age” (containing age demographics information). Then we correlated these tables using queries of the form

SELECT age, (num\_following / population) AS followingPC FROM

Twitter INNER JOIN statepop

ON Twitter.state = statepop.state

INTO OUTFILE 'ageVSfollowingPC.csv'

FIELDS TERMINATED BY ','

ENCLOSED BY '"'

LINES TERMINATED BY '\n';

Although the queries were standard SQL, executing these queries was slightly challenging because (1) we originally didn’t have permission to write data to any files (2) when we did get permission, MySQL wrote the data to a directory I didn’t know about (3) the directory was inaccessible to non-root users. We also had to set up a MySQL server on my computer, which was a fairly nontrivial undertaking.

**Shape of the data**

Using MySQL we gathered the following information about the congresspeople on Twitter:

**Tweets**

* The most prolific congressperson has 13773 tweets, and the least prolific has 0. Only one congressperson had 0 tweets.
* The average congressperson has posted 1084 tweets (with a standard deviation of 1378). However, only 145 out of 471 congresspeople have posted more than that number.
* The congressperson with the most tweets per month posts 275 tweets per month. However, only 22% of congresspeople post over 50 tweets per month.
* The average congressperson posts 36 tweets per month, but the median congressperson posts 29.

This indicates that the distribution of tweets posted is heavily skewed to the right and any statistics may be dominated by prolific outliers.

**Followers and following**

The average congressperson has 14,749 followers. The median congressperson has 4834. The least followed congressperson has 103 followers and the most followed has 1,798,403. This is another distribution where our statistics were dominated by outliers. I plotted log(followers) here instead of followers because the data spans several orders of magnitude.

The average congressperson is following 1240 people and the median follows 371, indicating another right-skewed distribution. Although we haven’t run the correlations, we conjecture that these phenomena are related. In particular, congresspeople with a large “Twitter presence” should have more tweet activity, more followers, and be following more people.

**Followers and following per capita**

There’s also the issue that politicians from large states should be more well-known and thus have an unfair advantage in followers. So we used state populations to calculate the number of followers (and following) per capita. Unfortunately, congresspeople tend to have very few followers compared to the population of their state. Not everyone in their state uses Twitter, and most of the ones who do aren’t politically involved enough to follow their local congressperson.

Since we only reported numbers to four decimal places, we were often left with only one significant digit, and some numbers were converted to “0.0000.” When we realized this, we stopped using the data in our regressions. However, in future work we should recalculate these numbers and multiply all of them by 1000 to avoid losing information. Multiplying by 1000 would also allow us to perform log-linear regression (we tried it before, but it didn’t work because we were taking the log of 0.000).

**Congressional demographics**

* The average House member on Twitter is 57 years old. The average senator on Twitter is 62. This is similar to real life, where the average senator is 60 and the average representative is 55. [15]
* 219 out of 254 Democrats are on Twitter. So are 250 out of 277 Republicans (and 2 out of 2 independents). [16] These fractions are pretty similar (86% vs. 90%).
* 86 out of 100 senators are on Twitter, and so are 385 out of 433 representatives. Those fractions are also about the same.

Overall there are no interesting correlations here, probably because the vast majority of congresspeople use Twitter.

**III. Congressional demographics**

**Hypothesis**

The first question we asked was how Twitter behavior depends on personal characteristics of the congressperson. We suspected that older congresspeople would use Twitter less often. This is because Twitter is more popular among younger audiences, specifically people aged 18-44. [9]

We also suspected Democrats would use Twitter more often than Republicans. We guessed this because Barack Obama is very active on Twitter, with 32 million followers and nine thousand tweets. By comparison, Mitt Romney has 1.5 million followers and one thousand tweets, and John McCain has 1.8 million followers and five thousand tweets. Furthermore, Twitter users are more likely to lean liberal (57% of Twitter users are Democrats, as compared to 46% of the general public.) [17]

We didn’t believe the gender of a congressperson would affect Twitter usage. However, if we had to guess, female congresspeople would have more Twitter activity than male congresspeople, since Twitter is more popular among women. [9]

We thought senators might be more active on Twitter than representatives, since there are fewer of them and therefore they might be more prominent. Prominent politicians are more likely to be well-funded and have large congressional staffs that update their social media accounts for them. Therefore, they should post more tweets per month. They are also likely to have more followers, since they are more well-known.

**Method**

We organized the MySQL data in Microsoft Excel and created scatter plots. Then, we added linear trendlines along with the R-squared and equations on the graphs. For t-tests, we used Microsoft Excel’s t.test( ) function to run two-tailed t-tests assuming different standard deviations for both sets of each pair of data.

**Analysis**

First we looked at congressional age, aggregating the Twitter data to look for significant trends. Linear univariate regressions were done for congressional age vs. (1) number of followers, (2) age vs. number of followers per capita, (3) age vs. number of users followed, (4) followed per capita, and (5) tweets per month. The R squared or coefficient of determination for the linear regressions are stated below in Table 1.

From this data, it could be concluded that the strongest linear correlation was between congressional age and the average number of tweets made per month while the weakest one was between congressional age and number of users following per capita. However, it should also be noted that none of the linear relationships were strong in the first place and that is probably because of the very large number of congresspeople who have made so few tweets, or have so few followers or number following that it skews the linear relationship. In fact, most of the data looks like random scatter plots as visible in the graphs presented in Figure 2.

But the result that congressional age correlates linearly with the tweets per month makes sense since they are probably the group of people who are most comfortable with using social media as evident in Figure 1. The correlation between congressional age and their tweets per month is clearer if we exclude a few outliers as pointed out in orange as the R-squared values rises to 0.1328 making the correlation stronger.

When we match the outlying data with the actual congressperson’s name to see whether there is a reason for their increased Twitter usage, we see the following results as summarized in Table 2. These congressmen were outliers because as younger than 40, they made less than 35 tweets per month and as older than 60 made more than 50 tweets per month. The only exceptions were the 53 year-old who made 209 tweets per month (Dennis Ross) and the 59 year-old who made 275.46 tweets per month (Darrell Issa). John Boehner, 63 years old making 151.17 tweets per month, is an obvious outlier as he is the current Speaker of the House and so his prominence as a public and legal figure probably explain his increased social media presence on Twitter. Another very relevant outlier would be John McCain who at the age of 76 has a record of 94.3725 tweets per month probably because of his candidacy for President in the 2008 election against President Obama. Interestingly, two outliers, the 39 year old who made 0.10 tweets per month (Devin Nunes) and the 31 year old who made 29.33 tweets per month (Tulsi Gabbard) assumed office in January 2013. As they have not been in office for long, their lack of Twitter usage may be just because they have only been congresspeople for the past 6 months. It would be interesting to analyze the data and see how Twitter usage correlates with time since assuming office in Congress. Nonetheless, without the outliers, a general downward trend is quite apparent for tweets per month vs. age.

The data for tweets per month was also correlated based on three types of tweeters: congresspeople who make a) < 30 tweets per month, b) 30-100 tweets per month, c) > 100 tweets per month. These three categories were then analyzed for correlations as shown in Figure 3. These results show no apparent correlation between tweets per month vs. age even when considering light, medium, and heavy Twitter users separately. There is a downward trend in all three cases, but the strongest linear correlation is for the category of 30-100 tweets made per month and age. This indicates, therefore, that no one age group of congressperson uses Twitter significantly more or less than other age groups, but is most predictable in the group of congressmen who make 30-100 tweets per month.

It is also interesting to see that the correlation with congressional age and the number of followers per capita is weaker than the other correlations; this means that regardless of age, people must follow congresspeople in equal numbers as seen in Figure 2. But, the number of followers may not be indicative of active use of Twitter. Also, there seems to be no correlation whatsoever between congresspeople’s age and the number of users they follow which probably indicates that following users is not a strategy that congresspeople use in order to further their political agenda on Twitter.

Then, the data was further analyzed in Figure 4 by separating congresspeople into two groups – those with < 100,000 followers and those with > 100,000 followers. There was an upward trend for both groups of congresspeople, but this was much stronger for those with > 100,000 followers with an R2 value of 0.2754. By looking at Table 3, it is quite evident that John McCain’s over one million followers is helping to keep the positive linear trend data in the > 100,000 folllowers category. Nonetheless, with increasing age, there seems to be an linear increase in the number of followers congresspeople have both in the < 100,000 followers and at the > 100,000 followers categories.

The data of the number of people congresspeople are following as depicted in Figure 2 was also analyzed by separating congresspeople into two groups – those following < 10,000 people and those following > 10,000 people in Figure 5. From this analysis, once again there is no significant correlation within the two groups of congresspeople between the number of people congresspeople are following and their age.

Then we wanted to test whether gender (male/female), party (Democratic/Republican), or status (House/Senate) affect congressional use of Twitter in terms of the number of followers, following, or number of tweets per month. From this analysis, we calculated the p-values using two-tail t-tests assuming different standard deviations for both sets of each pair of data in Table 5. If we consider the significance level to be 0.10, then we see that very significantly, differences in Twitter usage are due to status (House/Senate) and party (Democratic/Republican). Specifically this means that the number of tweets per month, the number of users followed per capita, the number of followers per capita is statistically different for House and Senate members. And, the number of users followed per capita and the number of users followed is statistically different for Republicans and Democrats. Further analysis showed that senators on average produce more tweets per month than representatives, follow more users per capita, and have more followers per capita. Additionally, Republicans have a higher average number of users they follow per capita and more users they follow.

(We must be careful with our analysis because of the skewed nature of the data. The t-statistic depends on the means of both samples, which are strongly affected by values that are extremely high. Specifically, that one congressperson with 1.7 million followers probably changed his sample mean considerably and may have pushed one mean above the other. Furthermore, the outliers make it more difficult to get statistically significant results because they increase the standard deviation, which lowers the magnitude of the t-statistic.)

Then, we reran the t-tests by excluding what we considered extreme Twitter activity: > 30 tweets per month, having > 100,000 followers, or following > 10,000 people on Twitter and tabulated the data in Table 5 in the column “p-value excluding outliers”. From this data, we got a whole new set of results – that the number of tweets per month is not statistically different for House and Senate members, that the number of followers is statistically different for House and Senate members, that the number of tweets per month is statistically different for Republicans and Democrats. There was also confirmation that with and without outliers, there is a statistical difference between the number of people followed by Republicans and Democrats supporting that Republicans follow more users than Democrats. (When we were redoing the t-tests, we did not include the per capita data just because many of the data points had only one significant figure at the fourth decimal point.) Further analysis shows that after excluding outliers, Senators have more followers than Representatives, Republicans make more tweets per month than Democrats, and Republicans have more followers.

**IV. Voter demographics**

**Hypothesis**

We then moved on to look at voter demographics. Specifically, we wanted to see if congressional Twitter activity depended on the age and race breakdown of their constituents. We suspected that congresspeople in states with a larger fraction of Twitter users would have more Twitter activity. For example, black people are more likely to use Twitter. 37.2% of black students use Twitter, compared to 20.1% of white students. [14] So we thought that states with a higher percentage of black people would have congresspeople with more Twitter activity. This effect should be especially pronounced for the number of followers, since the number of followers a congressperson has should directly depend on the Twitter usage of his constituency.

We also thought states with a higher percentage of young people would have congresspeople with more Twitter activity, since 18-44 year olds are more likely to use Twitter. [9] Again, we thought this would have its most direct impact on the number of followers.

**Methods**

We wrote MySQL queries to get .csv files that contained the following information: (1) the percent of 18-24 year olds in a congressperson’s state vs. the number of followers he has (2) percent of 18-24 year olds vs. number of followers per capita (3-5) percent of 18-24 year olds vs. number following, following per capita, and tweets per month (6-10) percent of people 65 and over vs. Twitter data (11-15) median age vs. Twitter data (16-18) percent of whites, Hispanics, blacks, Asians, and mixed-race people vs. number of followers, following, and tweets per month.

We loaded tables (1-15) into Excel spreadsheets and used the LINEST command to run linear regressions. Since tables 16-18 were multiple regressions, we loaded them into MATLAB and used the “regress” command.

**Analysis**

We tabulated the linear correlations in Table 6. We saw that there was no linear correlation between voter age and their Twitter activity in terms of tweets per month, followers, or people following. There aren’t any outstanding outliers either that might help to improve the correlation of the data. This means that voter age does not provide much scope for further analysis.

We also found no linear correlation between race demographics and Twitter activity. The R^2 value for the followers regression was 7.45 x 10^-3, which indicates no correlation at all. The R^2 values for the other regressions were similarly dismal. Knowing there were outliers, we split the congresspeople up into groups, where one group had congresspeople with more than 100000 followers and the other had congresspeople with less than 100000 followers. However, this didn’t really help matters. Neither did splitting the congresspeople into groups that follow more than 10000 people and less than 10000 people. We concluded that race demographics and Twitter activity were not linearly related.

**V. Leadership positions**

**Hypothesis**

We conjectured that those who hold leadership positions in Congress would have more Twitter activity than other congresspeople. We expected that people as important as the Speaker of the House would have large congressional staffs tasked with updating their social media websites. They should also have more followers, since (1) as influential politicians, more people are interested in what they have to say (2) since they hold positions of national leadership, they are more interesting to people outside their state.

**Methods**

We found a list of congressional leaders at thecapitol.net. [11] The leaders of the House of Representatives were (1) the Speaker of the House (2) the House Majority Leader (3) the House Majority Whip (4) the House Minority Leader (5) the House Minority Whip (6) the Republican Conference Chairwoman (7, 8) the House Republican Policy Committee Chairman and Vice-Chairman (9) the Assistant Democratic Leader (10) the Democratic Caucus Chairman (11) the Democratic Caucus Vice Chairman.

The leaders of the Senate were (12) the President of the Senate (13) the President Pro-Tempore of the Senate (14, 15) the Senate Majority and Minority Leaders (16, 17) the Senate Majority and Minority Whips (18, 19) the co-chairmen of the Democratic Policy Committee (20) the Democratic Conference Secretary (20) the Republican Conference Chairman (21) the Republican Policy Committee Chairman (22) the Republican Campaign Committee Chair.

Our database did not contain a Twitter id for Joe Biden, the President of the Senate. Although Joe Biden does have a Twitter account, we left him out of the dataset for consistency. There were many people in our dataset whose Twitter accounts were missing even though they had one. Most were probably people who are low-profile or joined Twitter fairly recently (Joe Biden joined in June of last year).

**Analysis**

We ran t-tests to check if congressional leaders significantly differed from non-leaders in (1) age (2) number of Twitter followers (3) number of users they are following (4) number of tweets posted per month. Although there were only 22 leaders, we believe the sample size is still large enough that we can use the Central Limit Theorem to assume the sample mean is normally distributed.

Several of our results were significant:

|  |  |
| --- | --- |
| **t-test measuring differences between \_\_\_\_\_\_\_ for leaders and non-leaders** | **p-value** |
| Age | 0.2810 |
| Number of Twitter followers | **0.0117** |
| Number of users they are following | **1.0 x 10^-6** |
| Number of tweets posted per month | **7.4 x 10^-4** |

Although our leaders were older than our non-leaders on average (average was 60.13 years vs. 57.57), our p-value was too large for these differences to be statistically significant. If there was in fact no difference between the population means, there would be a 28.1% chance that the results occurred due to chance.

While we have no evidence that leaders differ from non-leaders in age, leaders do tend to have more followers. The average non-leader had 12458 followers. By contrast, the average leader had 61494 followers, something only twelve congresspeople can claim. Interestingly, the leaders didn’t really overlap with the most followed people; only four people with more than 61494 followers were leaders, and only nine leaders had more than 20000 followers. Presumably many leaders hold positions the public doesn’t care about, such as “Assistant Democratic Leader” and “House Republican Policy Committee Chairman.” The average follower count for leaders is probably heavily influenced by the outliers John Boehner (468004 followers), Eric Cantor (142391 followers), and Nancy Pelosi (320304 followers). However, despite the high standard deviation induced by these outliers, we were still able to get a p-value of 0.0117, indicating that if there was really no difference in the population means, there would be a 1.17% chance that the results were due to chance.

Leaders also tend to follow more people, but they still follow relatively few. The average leader follows 4134 people and the average non-leader follows 1098. While leaders are more likely to have congressional staff manning their Twitter account who will follow everyone worth knowing, leaders certainly do not follow everyone who follows them. There are 32 congresspeople who follow more than 4134 people and only 7 of them are leaders, so the following count is less influenced by outliers than the follower count. This is probably why our p-value was so low. If there was truly no difference in the population means, there would be a 10^-6 chance that the results occurred by chance.

Leaders post considerably more tweets per month. The average leader posts 58.17 tweets per month, while the average non-leader posts 35.40. This is probably not due to outliers since 10 leaders (and 67 non-leaders) post more than 58.17 tweets per month. Since the p-value on our t-tests was 7.4 x 10^-4, if there was truly no difference in the population means, there would be a miniscule chance that our results were due to chance.

Leaders have also had their accounts for a longer time on average (42 months vs. 31), although we didn’t test the significance of that result. This may have given them more opportunities to amass followers and follow more people. Perhaps in future work we will measure variables such as “followers per month” and “number followed per month.”

**VI. Rural and urban states**

**Hypothesis**

We conjectured congresspeople in densely populated states would have more followers than those in rural states. This is because we expected that (1) congresspeople would mostly be followed by their constituents, so those with more constituents would have more followers (2) people in rural areas may be less likely to use the Internet. 60% of people in rural areas have broadband Internet, a figure 10% lower than that for urban areas. [10] We also thought urban congresspeople would post more tweets and follow more people, since with a larger Twitter-using constituency they would have more incentive to be active on Twitter.

**Methods**

We defined an “urban” state as one with more than five million people and a “rural” state as one with less than five million people. This is an imperfect definition, since many states (like New York) have both urban and rural areas. We used it since we already knew which state each person was from and didn’t have data on which district they were from. In future work we should refine this analysis by finding which district each congressperson is from and gathering information about those districts.

We wrote MySQL queries to create the following files: (1) number of followers for each congressperson in a rural state (2) number of followers for each congressperson in an urban state (3) number of users each rural congressperson follows (4) number of users each urban congressperson follows (5) number of tweets posted by each rural congressperson per month (6) number of tweets posted by each urban congressperson per month.

**Analysis**

We wrote a short MATLAB script to run two-tailed t-tests on each of these files. The tests checked if the average number of followers, following, and tweets per month differed significantly between rural and urban states. The main assumption we needed to check was that the sample means followed a standard normal distribution. Since our sample is large (335 urban congresspeople and 133 rural congresspeople), the Central Limit Theorem applies, and we can make this assumption.

Unfortunately, none of the results were statistically significant. We failed to find evidence that there are differences in congressional tweet activity in rural and urban states. If the population means were the same, there would be a 51%, 24%, and 71% chance (respectively) that the results occurred due to chance.

|  |  |
| --- | --- |
| **t-test measuring difference between \_\_\_\_\_\_\_\_ in urban and rural states** | **p-value** |
| Number of followers | 0.5129 |
| Number of users the congressperson is following | 0.2418 |
| Number of tweets per month | 0.7094 |

**VII. Non-Twitter users**

**Hypothesis**

We wanted to examine demographic differences between congressmen who do and don’t use Twitter. Specifically, we wanted to see if the groups differed in (1) average age (2) percent of congressmen in each party (3) percent of representatives and senators. We thought there would be no gender gap, so we didn’t run a t-test for gender.

We suspected Twitter-using congressmen would be younger on average than non-Twitter users. This is because Twitter is more popular among 18-44 year olds [9] and is less popular among the older demographic. We didn’t know whether the party balance would differ, although we had a slight leaning towards Democrats because Barack Obama has a very active Twitter page and Democratic congressmen might be following in his footsteps. We had no idea whether representatives or senators would be more likely to use Twitter, but we suspected that since there are fewer senators for each state, each one would be more influential and have a greater Twitter presence.

**Methods**

We wrote a Python script against the original Sunlight Foundation data to find the information of congressmen without a Twitter id. We then wrote MySQL queries to get lists of ages, parties, and titles for Twitter and non-Twitter users. We coded Republicans as “1” and Democrats as “0” so it would be easier to measure the percentage of people in each party. (To simplify our analysis, we omitted Independents from the data set; there were only two anyway.) We also coded Senators as 1 and Representatives as “0.”

There were 64 non-Twitter users in our data set and 471 Twitter users. We strongly suspect that the data was contaminated, since many “non-Twitter users” that we googled did in fact have Twitter accounts. (The most egregious example was Joe Biden, who is President of the Senate and has had a Twitter account for a year.) Therefore, any differences we find between the two groups are likely to be weaker in our data set than in reality.

**Analysis**

We ran two-tailed t-tests to compare Twitter users with non-Twitter users. np and n(1-p) were greater than 5 in the party and title tests, where n is the size of each sample and p is the percent of people in the sample that are in a certain category. Therefore, the sample mean (which is binomially distributed) approximates a normal distribution, and we can use t-tests to analyze the data.

There was no significant difference in the average age of the two groups. Non-Twitter users were 57.20 years old on average, and Twitter users were 57.69 years old.

There was almost a significant difference in the balance of Republicans and Democrats. The Twitter users were 53.3% Republican and the non-Twitter users were 40.6% Republican. However, the p-value was 0.057, which barely exceeds the threshold for statistical significance. (If the population means were the same, there would be a 5.7% chance that our results occurred due to chance.) Either way, our results contradicted our hypothesis that the Twitter users would be more likely to be Democrats.

There was no significant difference in the balance of representatives and senators. The Twitter users were 18.2% senators and the non-Twitter users were 21.8% senators. We failed to reject the null hypothesis with a p-value of 0.4872.

**VIII. Tweet content**

**Hypothesis**

We then moved on to look at the content of some of the tweets. Specifically, we wanted to measure how Republicans and Democrats reacted differently to major news stories. We chose to analyze the Boston bombing because it was a relatively nonpartisan event that got a lot of attention. Based on our observations of non-congresspeople, we had a vague suspicion that Republicans and Democrats would react differently to the event. While both would express sorrow at the tragedy, Democrats might take the opportunity to discuss the need for gun control while Republicans might be spreading politically charged antagonism towards the Democratic party or President Obama.

**Methods**

To gather the tweets, we still used the Twitter API, but instead of finding the user objects, we found the user timelines. These contained the 200 most recent tweets for each congressperson. Once we had the tweets, we stored them in a MySQL database with the identity of the tweeter. Using MySQL queries we produced .csv files containing (1) all tweets made by Democrats (2) all tweets made by Republicans (3) all tweets made by Democrats during the week after the Boston bombing (April 15-22, 2013) (4) all tweets made by Republicans during the week after the Boston bombing. We downloaded a word-count script from StackOverflow and used it to count the thirty most common words in each of these files (that weren’t in Wikipedia’s list of the one hundred most common English words).

Note that 89 out of 471 congresspeople had fewer than 200 tweets. The content analysis may have been biased against those users, since their tweets would show up less often than their compatriots’. One alternative approach would be to give those users’ tweets a higher weight. However, that would be unfair in a different way, since a tweet from an inactive user probably has less influence in the Twittersphere than a tweet from an active user. Another strategy would be to use a number smaller than 200, but that would have jeopardized our sample size. For these reasons, we went ahead with the naïve approach.

**Analysis**

Overall, 69.0% of the 30 most frequent words in tweets were things associated with websites like “http”, “t”, “co”, and so the word analysis was done excluding these words. (Also 61.5% of the 30 most frequent words in tweets the week after the Boston bombing was also associated with links.)

First, the overall Twitter data for the 30 most frequent words made within the 200 most recent tweets of all congresspeople was analyzed. In the top Venn diagram in Figure 7, we see the kinds of words that are used by Republicans and Democrats overall. It is interesting to see that the top 30 words Republicans use in their tweets include much more politically relevant words like “Obama”, “Obamacare”, “President”, “Benghazi”, “IRS” compared to Democrats who only really used “Congress” as a politically related word. Otherwise, Democratic congresspeople seem to be using many more emotionally appealing words like “families”, “help”, “need”, “support”. It is also interesting to see that “women” is one of the most top tweeted word by Democrats whereas Republicans have no parallel use of word that sides with either gender. These results can help conclude that perhaps Democrats’ tweets are more sentimental and directed towards creating some sort of personal connection with their followers whereas Republicans’ tweets are much more based on fact and current events. Further analysis looking at more than just 200 recent tweets, sentiment analysis, and more precise methods would of course need to be done in order to ensure that these observations are conclusive or real.

These results may also be due to the ongoing partisanship between Republicans and Democrats in the Senate and House ever since President Obama was elected president. Since then, Republicans have been trying to argue against Democratic policies while Democrats have tried to maintain their support of the President. The top 30 words or so overall seem to reflect this trend in the 200 most recent tweets posted.

Words that are used by both parties include things like “bill”, “budget”, “House”, “jobs”, and “thanks” which are applicable to negative and positive comments made about the ongoing political climate. In Figure 6 (top), it is also interesting to note some of the differences between words used by Democrats and Republicans in the 200 most recent tweets made. Significantly, Republicans have been posting more tweets including “RT” or retweet, “here”, “more”, “house”, and “budget” than Democrats. Specifically, this indicates that perhaps Republicans are retweeting other people’s messages more often than creating their own tweets. Further analysis into this would be interesting as it would help to indicate whether Democrats and Republicans are using Twitter in different ways. Nonetheless, as there are more Republican congresspeople, this might add to the increased frequency of certain words used by Republicans compared to Democrats since we are looking at the 200 most recent tweets made by all congresspeople. So further analysis needs to be done which has some sort of normalization system to ensure that these preliminary results are conclusive.

Next, we consider the top 30 words gathered from all the tweets made by congresspeople between April 15 and April 22 right after the Boston bombing. As we see in Figure 7 (bottom), Republicans had “RT” as one of their 30 top words whereas Democrats didn’t. This indicates that Republicans were retweeting more often than Democrats during this phase. Additionally, the Republican tweets included in their 30 top words things like “taxday” and “taxreform” during this time period which are clearly not related to the Boston bombing but very relevant to the same time that taxes were due. In contrast Democrats’ tweets included in their 30 top words more related to the Boston bombing like “gun”, “marathon”, “support”. Although both parties overall used words like “Boston”, “families”, “prayers”, “thoughts”, and “victims” as shown in the Venn diagram in Figure 7 (bottom), in addition to these tweets, Republicans’ tweets seemed to include more unrelated words than Democrats. It was also interesting to see that Democrats used words like “vote” and “CISPA” (the Cyber Intelligence Sharing and Protection Act) which was passed by the House during the Boston bombing. Perhaps lawmakers were trying to use the emotions of the Boston bombing in order to raise awareness of bills related to national security. But, it would be interesting to investigate why Democrats and not Republicans were tweeting as much about CISPA. It was also intriguing to note that after the Boston bombing, Republicans weren’t tweeting with as much political antagonism as they were in their 200 most recent tweets as their tweets between April 15 and April 22 didn’t include “Obama” or “president”. In contrast, Democrats used “House”, and “Senate” amongst their 30 most used words.

In Figure 6 (bottom), if we look at the common most used words by Republicans and Democrats, what we see is that Republicans had more tweets including words like “prayers”, “victims”, “thanks” than Democrats. But, since Republicans seem to have made more tweets including all the words in Figure 6 (bottom), this might just be an artifact of the fact that there are more Republican congresspeople.

Tables 7 and 8 helps to analyze the specific words that are in the top 30 list and the proportion with which they are used excluding words that refer to links like “http”, “t”, and “co”. Some of the most interesting results from this included “RT” as the 8th most used word amongst all congresspeople and Republicans, but the 16th most used word amongst Democrats amongst the 200 most recent tweets. After the Boston bombings, these tables also clearly show how Democrats and Republicans differ in their rankings of political words during this time. To better understand tweet content, it will be interesting to look at beyond just the 30 most used words and beyond just 200 most recent tweets to get a bigger picture. Additionally, more systematic analysis of the characteristics of the word than the qualitative observations will need to be done in the future.

**IX. Discussion and conclusions**

Most of our results were insignificant. In some cases this is not surprising. With congressional demographics, most of our hypotheses came from observations about the Twitter-using public. However, congresspeople do not use Twitter in the same way that regular people do. While regular people tweet about their lives, congresspeople tweet carefully polished press releases. Furthermore, most Twitter users have a relatively small following, but congresspeople are usually followed by tens of thousands of people. To politicians, Twitter is a completely different product. So even if certain types of people are more likely to use Twitter (women [9], liberals [17], younger people [9]), these correlations may not carry over into the political sphere.

It’s interesting that congressional Twitter activity doesn’t depend on the voter demographics of their state (or even its population). We thought congresspeople deliberately planned their Twitter strategy based on data about their constituency, but it’s more likely that congressional Twitter feeds are left in the hands of social media interns. The volume of their Twitter activity should depend on the disposition of their interns, whose usage patterns are essentially random. We’re not sure how to test this new hypothesis, but it would be an interesting avenue for future research.

The congressional leadership results were significant and expected. We found strong evidence that leaders had more followers and overall activity on Twitter. It makes sense for leaders to have more followers because they’re closer to the public eye and are more interesting to people outside their state. It also makes sense for them to post more tweets because they probably have a larger and better-organized campaign staff. While our results are statistically significant, we think most people would find them obvious.

Perhaps the most interesting part of this work was the content analysis. For example, it was very interesting that after the Boston bombings, Republicans mentioned “taxday” more than they mentioned “victims” or “families.” While these facts make for nice dinner conversation, it would be lovely to draw more quantitative conclusions. Another avenue for future work would be comparing the number of at-mentions by Republicans and Democrats, or the number of times they are retweeted. We could also examine linguistic differences; it would be fascinating if one party used bigger words than the other. Imagine all the conclusions the Internet could draw about their intelligence and education level!

Overall, the Twitter data presents many interesting avenues for research. While we answered some questions, there are many more left to explore.

**References**

[1] Holt, Richard. Twitter in numbers. *The Telegraph*, 21 Mar 2013.

[2] Bluff, Brian. Who uses Twitter? *Site-Seeker*, 7 Nov 2012.

[3] Sunlight Congress API @ Github. Sunlight Foundation.

[4] Wikipedia. Demographics of the United States. Retrieved 6 May 2013.

[5] U.S. Census Bureau, "Demographic Profiles: Census 2010." 2010.

[6] Twitter REST API documentation. Retrieved 6 May 2013.

[7] The Python-Twitter Developers. Python Twitter: A Python wrapper around the Twitter API. Google Code. Retrieved 6 May 2013.

[8] U.S. Census Bureau. "Table 1. Annual Estimates of the Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2012 (NST-EST2012-01)" Dec 2012.

[9] Site info for Twitter.com. Alexa. Retrieved 7 June 2013.

[10] Severson, Kim. Digital Age is Slow to Arrive in Rural America. *The New York Times*, 17 Feb 2011.

[11] Leadership and Committees of the 113th Congress, 1st Session. TheCapitol.Net. Retrieved 7 June 2013.

[13] Wikipedia. Most Common Words in English. Retrieved 7 June 2013.

[14] Chen, Adrian. Why So Many Black People Are On Twitter. Gawker.com, 17 May 2011.

[15] Colby, Jennifer. The grayest Congress. NBC News, 2013.

[16] Wikipedia. 113th United States Congress. Retrieved 6 May 2013.

[17] Beaujon, Andrew. Twitter is more liberal and more conservative than public opinion. Poynter.org. 5 Mar 2013.

[18] Greenberg et al. Congress + Social Media. University of Texas-Austin. 22 Oct 2012.

[19] Golbeck et al. Twitter use by the US Congress. Journal of the American Society for Information Science and Technology 61.8 (2010): 1612-1621.