Project Report
Who Are Trump Supporters?

CS 7930 Social Media Mining

Spring 2016, May 6, 2016

Gopal Menon
Computer Science Department
Utah State University
Logan, UT
gopal.menon@aggiemail.usu.edu

ABSTRACT

The current front runners in the 2016 Presidential Election are Donald Trump from the Republican party and Hillary Clinton from the Democratic party. The rise of Trump has taken many people by surprise and he was initially written off by supposed experts who expected his campaign to self destruct. As of the writing of this report, Trump is the only candidate left in the Republican field and is the presumptive nominee. This study aims to find out the reasons behind the rise of Trump by finding the Twitter community of his followers and then identifying the thought leaders in this community and their thoughts and ideas. By doing this I hope to identify the issues that matter most to Trump supporters.

1. INTRODUCTION

The Republican party base has been steadily moving away from traditional candidates as they reportedly feel that they do not have a say in matters that are of importance to them. The rise of the Tea Party and the influence of Talk Radio, has reportedly made the Republican base disillusioned with Congress. In the last few elections, traditionally safe Republican candidates have been unseated in the primaries by Tea Party candidates.There has been growing discontent with President Obama's perceived Liberal agenda and people in the Republican base seem to want to start afresh with new candidates as they seem to have lost trust in career politicians.

Trump is not a career politician and has no experience in government. He started out as an outside candidate who was not taken seriously. However he has succeeded in becoming the presumptive nominee by appealing to the base of the Republican party. He has been accused by many of being racist, fear mongering and wrong on things that he has put forward as facts. He is accused of being neither religious nor conservative and has still emerged as the presumptive nominee from his party. He is reportedly disliked by members of his own party, but despite this, has emerged victorious

The Democratic party traditionally has its strong base in the East and West coasts, Liberals and minority communities. Unlike the Republican party, the Democrats have not been going through any internal turmoils. However, their candidate has had her share of problems. Clinton has been associated with scandals from the past like her husband's infidelities and the Whitewater investigations. During her term as Secretary of State, she was held responsible for the death of the American ambassador along with three other persons, during the attack on the consulate in Benghazi. After her term, she was faced with the email scandal. Outside of her party she is reportedly disliked and is accused of being corrupt and is a very divisive figure.

The Democrats, unlike the Republicans, started out with just four candidates and were left with only candidates - Hillary Clinton and Bernie Sanders. At the time of writing this proposal, Clinton seems to be the likely nominee from the party.

2. WHY PROFILE TRUMP SUPPORTERS

2.1 Why is it interesting

I find this interesting because of the success of the Trump campaign. He had been written off initially by news analysts and they did not expect him to last beyond the initial phase. His continued success has taken many people by surprise. By finding the issues that matter to his supporters, I can find out what are the key factors that drive the apparent demand for change among the base. Since the career politicians have not been successful, I can find out how the Republican party is out of touch with the issues that matter most to its base. The reasons could be perceived loss of employment opportunities due to globalization and illegal immigration, changes in the demographics of the voter population, perceived Liberal policies of President Obama, opposition to gay marriage and the Affordable Care Act (also known as Obamacare), fear of terrorism or it could be something that I have not thought of. Whatever the reasons are, they will be interesting as the emergence of Trump is an unforeseen phenomenon that maybe even he cannot explain.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Figure 1: Donald Trump's Twitter Profile

2.2 Who will be interested

Social scientists, people who follow politics and key members of the two parties will potentially be interested in knowing the results of the investigation.

The Republican party will be interested as they apparently need to change their focus since all their traditional presidential candidates have been rejected by the primary voters. After the election of President Obama in 2008, many news analysts said that the Republican party needs to change its attitude towards that Latinos, who consist of the largest minority, in order to be viable in the future. They said the party may need to change in order to survive.

The Democrats on the other hand, have not had any issues with their traditional base. They would be interested in knowing what issues are important to their opponents and would want to use this information in order to to better compete.

3. PREVIOUS RESEARCH

Previous research has been done by Kloumann and Kleinberg [1] on identifying members of a community starting from a small seed set. The seed set was expanded using PageRank and other algorithms. The PageRank algorithm reportedly had good results in identifying key members of a community given the seed set.

4. METHODOLOGY

The Twitter ids of Trump (see figure 1), his family members along with those of people from the media, politicians and other famous personalities were used as the seed set[3]. The seed set details are given in figure 4 in section 8.1.

The seed set was expanded using a Twitter crawler that started with each seed in the seed set and found the Twitter followers of the seeds. This process was repeated in order to identify the Twitter community of Trump followers.

The PageRank algorithm was used to find the important members in the Twitter community

4.1 High Level Design

The Twitter followers crawler started crawling from the seed set. It uses a FIFO Queue (see figure 2) for implementing a Bread First Search (BFS). It starts out by loading the Queue with the seed set. Then does the following till the crawl level reaches a preset level: take a Twitter id out from the front of the Queue. Find out its followers using Twitters APIs and put the followers in the back of the Queue.

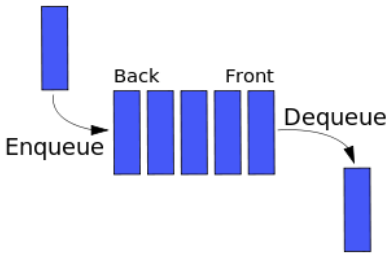

Figure 2: First in First out Queue[4]

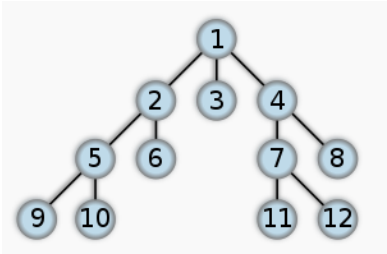

Figure 3: BFS[5] - Order in which nodes are expanded

Twitter puts in a rate limit on the number of API requests that can be made. If the rate limit is exceeded, the API fails and an exception is thrown. To overcome this problem, the crawler keep track of the number of API requests that are allowed before the rate limit is exceeded. If the rate limit is reached or if any other error is detected, the crawler writes the Queue contents to a text file and sleeps for the amount of time that the requests are blocked. If the crawler is restarted after a crash, it has the ability to start from the the last saved point where the Queue contents were written to the text file. This restart ability was built into the crawler as the crawling process was expected to take a long time and it was essential that the Twitter community retrieved that far was not lost.

The Page Rank algorithm[2] was used to identify the important members of the Trump followers Twitter community. A sparse matrix was used for the transition probability matrix in order to save space. The teleportation probability was stored separately in order to keep the matrix sparse. In order to find the Page Ranks the probability vector was multiplied with the transition matrix in a loop till the probability vector reached a steady state that represented the Page Ranks. The multiplication loop was terminated when the cosine similarity between subsequent vectors reached 0.9999 or after 75 iterations, whichever came first. The following formula was used to find the cosine similarity between two vectors A and B[6]:

$$Similarity = cos(\theta) = \frac{A \cdot B}{|A|\,|B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

The Page Ranks were sorted in decreasing order to identify the community though leaders. I plan to expand the seed set using the PageRank algorithm. Each member of

the seed set will be a Twitter account. The other accounts followed by the seed account will constitute the equivalent of the forward links on a web page. In addition to the forward links, the list of accounts that are followers of an account will be the incoming links. This is where my approach differs from PageRank. The equivalent of the PageRank case where the graph walker is transported to a random web page, will be replaced by random transport to a member of the seed set. I plan to evaluate the output both with and without the incoming links. Use of PageRank will expand the seed set and give me a big community that is linked together.

When the community is sorted by decreasing order of page rank, I will obtain the thought leaders in the community. These accounts with higher page rank will be thought leaders as they will be the ones that are most important due to the fact that more people follow them, and because more important people follow them.

Once the community though leaders have been obtained, I plan to use topic modeling on their Tweets in order to find out what issues and topics interest them and their followers.

At the end of the project, I will have obtained the thought leaders in the Twitter community for each of the two presidential candidates and the topics and issues that matter to them.

4.2 Tools and Techniques

I plan to use Twitter APIs through Python for retrieval of Tweets and for Twitter account crawling. There are some repositories on GitHub that claim to be able to crawl Twitter. I would need to investigate these. The ability to recover from errors and restart along with the capability to crawl the Twitter graph in parallel will be an important criteria to consider.

The Twitter graph of users along with followers will need to be stored in a graph database. For this purpose, I would need to evaluate and use a graph database.

I would need to code or use a function that would compute a left eigen vector for computing the PageRank as described in [2].

The R package has topic mining libraries that I plan to use for finding the major topics in the profiles and Tweets of the thought leaders of the community for each of the candidates.

4.3 Risks

- As was seen in assignment 1 for retrieving Tweets, the process may take a long time and may not return many results. As a result, the process of retrieving the seed set or the Twitter graph of users and followers may take a lot of time.
- The Tweet retrieval process would fail intermittently.
- Twitter has limits on the retrieval rate that would limit data collection for the project.
- The seed set may not be of the right size or the right mix of accounts. This would limit the ability of the crawling process to uncover the important members of the community.

4.4 Mitigation of Risks

- In order to mitigate the risk of not getting enough data due to slow retrieval, I would need to start the process early.
- The retrieval process would need to be able to recover from failure and restart automatically.
- I would need to create multiple Twitter accounts in order to get around the limits on the retrieval rate. Retrieval would need to be interspersed with delays so that maximum limits are not reached. In order to further speed up the retrieval process, I would need to do parallel retrieval using the multiple CPU cores available.
- I would need to do further research in order to find out the optimal seed set composition for the project.

5. MILESTONES

- High Level Design complete - mid March
- Seed set identification complete - 3^{rd} week of March
- Twitter account crawling complete - 1^{st} week of April
- Topic Modeling complete - 3^{rd} week of April
- Project Report complete - end of April

6. EVALUATION CRITERIA FOR RESULTS

I plan to compare the set of thought leaders obtained by following only forward links versus following both outgoing and incoming links. I expect the seed set expansion to be more complete in the latter case.

I expect to get a list of Twitter community thought leaders for each of the two candidates and their main topics and issues.

There is no way to predict what results I am going to get, although I am sure that they will be interesting. The interesting for me will be the answer to the question *who would want to support Donald Trump.*

7. REFERENCES

[1] Kloumann, I.M. and Kleinberg, J.M., 2014, August. Community membership identification from small seed sets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1366-1375). ACM.

[2] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval/Christopher D.

[3] "List of Donald Trump Presidential Campaign Endorsements, 2016." Wikipedia. Wikimedia Foundation, n.d. Web. 05 May 2016.

[4] "Queue (abstract Data Type)." Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

[5] "Breadth-first Search." Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

[6] "Cosine Similarity." Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

8. SUPPLEMENTARY MATERIALS

8.1 Seed Set

The seed set used is shown in figure 4. The actual seeds used were the Twitter Ids corresponding to the names and handles shown in the figure.

Who	Handle	Twitter Id
Donald J Trump	@realDonaldTrump	25073877
Donald Trump Jr.	@DonaldJTrumpJr	39344374
Melania Trump	@MELANIATRUMP	108471631
Ivanka Trump	@IvankaTrump	52544275
Eric Trump	@EricTrump	39349894
Sarah Palin	@SarahPalinUSA	65493023
Sen. Jeff Sessions, AL	@SenatorSessions	47975734
Sen. Scott Brown, MA	@USSenScottBrown	117537998
Rep. Lou Barletta, PA	@RepLouBarletta	239871673
Rep. Chris Collins, NY	@RepChrisCollins	1058256326
Rep. Kevin Cramer, ND	@RepKevinCramer	1048784496
Rep. Scott DesJarlais, TN	@DesJarlaisTN04	235312723
Rep. Renee Ellmers, NC	@RepReneeEllmers	213634439
Rep.Duncan D. Hunter, CA	@Rep_Hunter	1305596696
Rep. Tom Marino, PA	@RepTomMarino	240363117
Rep. Tom Reed, NY	@RepTomReed	252819323
Saba Ahmed	@SabaRMC	206113016
Fmr. Rep. Doug Ose, CA	@DougOse	1465982610
Patrick J. Buchanan	@PatrickBuchanan	19599446
Jeff Lord	@JeffJlpa1	397545273
Joseph E. Schmitz	@josepheschmitz	258343236
Chris Christie	@GovChristie	90484508
Paul LePage	@Governor_LePage	637143497
Rick Scott	@FLGovScott	131546062
Jan Brewer	@GovBrewer	40923070
Joe Arpaio	@RealSheriffJoe	44951059
Ben Carson	@RealBenCarson	1180379185
Ann Coulter	@AnnCoulter	196168350
Michael Savage	@ASavageNation	66019768
Alex Jones	@RealAlexJones	109065990
Ted Nugent	@TedNugent	17879692

Figure 4: Seed Set