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 (see figure 1) and the influence of Talk Radio (see figure 2), 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.

Figure 1: Tea Party logo

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

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

Figure 2: Talk Radio hosts Sean Hannity and Rush Limbaugh

Figure 3: Obama Care (The Affordable Care Act)

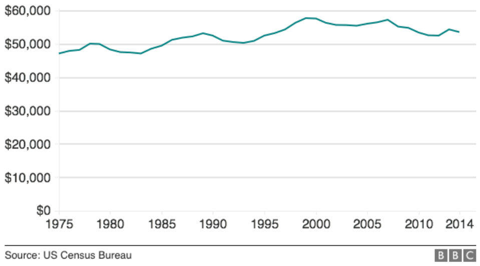


Figure 4: US median household income (adjusted for inflation)[1]

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.

There is a lot of anger among the voters. The reasons could be

- There has not been any real progress to middle-class and working-class Americans over the past 15 years[1]. Although the country may have recovered from the recession - economic output has rebounded and unemployment rates have fallen from 10% in 2009 to 5% in 2015 - Americans are still feeling the pinch in their wallets. Household incomes have, generally speaking, been stagnant for 15 years (see figure 4). In 2014, the median household income was $53,657, according to the US Census Bureau - compared with $57,357 in 2007 and $57,843 in 1999 (adjusted for inflation).

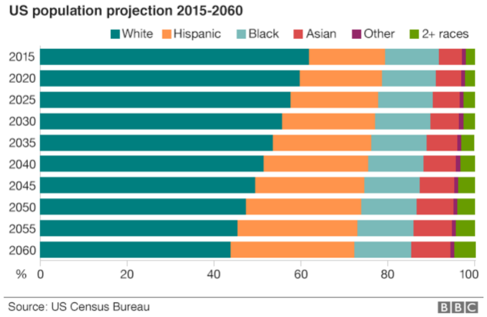


Figure 5: Changing Demographics of United States

- There is also a sense that many jobs are of lower quality and opportunity is dwindling[1]. For the left the culprits are the billionaires, the banks, and Wall Street. For the right it is immigrants, other countries taking advantage of us and the international economy.

- America's demographics are changing[1] - nearly 59 million immigrants have arrived in the US since 1965, not all of whom entered the country legally. Forty years ago, 84% of the American population was made up of non-Hispanic white people - by last year the figure was 62%, according to Pew Research. It projects this trend will continue, and by 2055 non-Hispanic white people will make up less than half the population. Pew expects them to account for only 46% of the population by 2065. By 2055, more Asians than any other ethnic group are expected to move to US. Some older, whiter voters do not recognize the country they grew up in. See figure 5.

- When asked if they trust the government[1], 89% of Republicans and 72% of Democrats say ”only sometimes” or ”never”, according to Pew Research. Six out of 10 Americans think the government has too much power, a survey by Gallup suggests, while the government has been named as the top problem in the US for two years in a row - above issues such as the economy, jobs and immigration, according to the organization. The gridlock on Capitol Hill and the perceived impotence of elected officials has led to resentment among 20 to 30% of voters. People see politicians fighting and things not getting done - plus the responsibilities of Congress have grown significantly since the 1970s and there is simply more to criticize.

- America is used to being seen as a superpower but the number of Americans that think the US ”stands above all other countries in the world” went from 38% in 2012 to 28% in 2014[1], Pew Research suggests. Seventy percent of Americans also think the US is losing respect internationally, according to a 2013 poll by the centre. For a country that is used to being on top of the world, the last 15 years haven't been great in terms of foreign policy. There's a feeling of having been at war since 9/11 that's never really gone away, a sense America doesn't know what it wants and that things aren't going our way. The rise of China, the failure to defeat the Taliban and the slow progress in the fight against

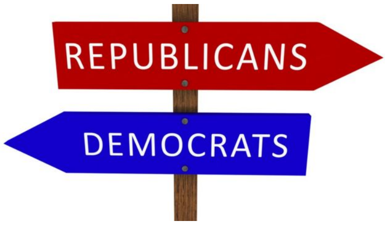


Figure 6: A nation divided

the so-called Islamic State group has contributed to the anxiety.

- Democrats and Republicans have become more ideologically polarized than ever[1]. The typical (median), Republican is now more conservative in his or her core social, economic and political views than 94% of Democrats, compared with 70% in 1994, according to Pew Research. The median Democrat, meanwhile, is more liberal than 92% of Republicans, up from 64%. The study also found that the share of Americans with a highly negative view of the opposing party has doubled, and that the animosity is so deep, many would be unhappy if a close relative married someone of a different political persuasion. This polarization makes reaching common ground on big issues such as immigration, healthcare and gun control more complicated. The deadlock is, in turn, angering another part of the electorate. See figure 6.

Whatever the reasons are, they will be interesting as the emergence of Trump is an unforeseen phenomenon that maybe even he cannot explain.

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 [2] 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.

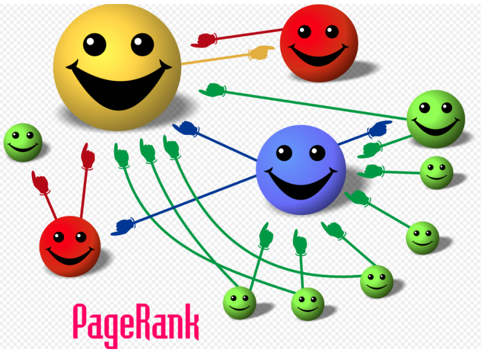
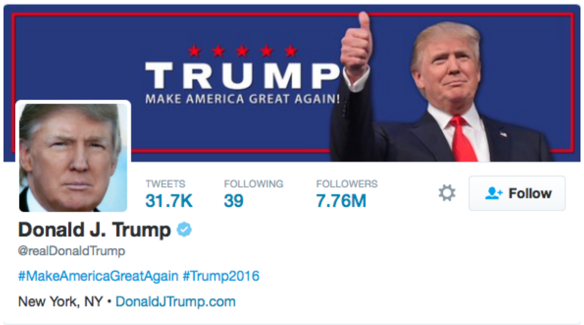


Figure 7: Cartoon illustrating the basic principle of PageRank[8]. The size of each face is proportional to the total size of the other faces which are pointing to it.



TWEETS 31.7K FOLLOWING 39 FOLLOWERS 7.76M

Donald J. Trump
@realDonaldTrump
#MakeAmericaGreatAgain #Trump2016
New York, NY · DonaldJTrump.com

Figure 8: Donald Trump's Twitter Profile

PageRank is an algorithm for quantifying the importance of a website on the world wide web by the number of important websites that link to it. It is a recursive definition. The Page Rank of a website may be thought of as the probability that a random walker on the world wide web graph of websites and their links will land on that web page.

4. METHODOLOGY

The Twitter ids of Trump (see figure 8), his family members along with those of people from the media, politicians and other famous personalities were used as the seed set[4]. The seed set details are given in figure 12 in section 9.1.

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 9) for implementing a Bread First Search (BFS). See figure 10 for breadth first search. 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

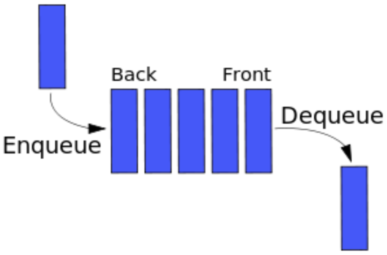


Figure 9: First in First out Queue[5]

- find out its followers using Twitters APIs and put the followers in the back of the Queue.

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 keeps 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[3] 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[7]:

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

The Page Ranks were sorted in decreasing order to identify the top ten community though leaders.

4.2 Tools and Techniques

A Twitter followers crawler was implemented in Java using the Twitter4J API. As described above, the crawler uses BFS for building the Twitter follower graph. The crawler stores the follower list of each profile it crawls, as a text file with the name of the text file being the Twitter id being followed. The PageRank process runs on the folder containing the text files and builds the transition probability matrix and finds the Page Ranks by multiplying the probability vector in a loop with the transition probability matrix.

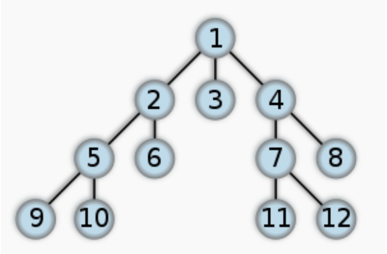


Figure 10: BFS[6] - Order in which nodes are expanded

4.3 Risks

- As was seen in assignment 1 for retrieving Tweets, the process took a long time and did not return many results. As a result, the process of retrieving the Twitter graph of users and followers was expected to take a lot of time.
- Twitter has limits on the retrieval rate that would limit data collection for the project.
- The Twitter API calls fail intermittently even when the rate limit was not exceeded.
- 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

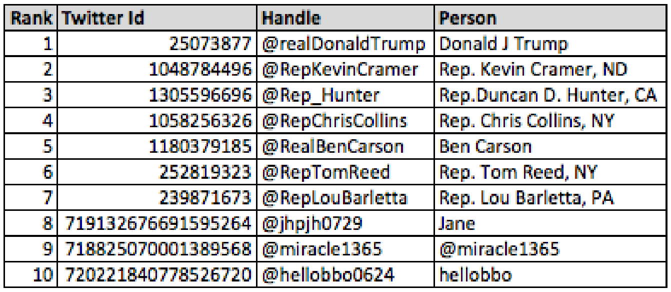
- In order to mitigate the risk of not getting enough data due to slow retrieval, I started the process early.
- I planned to create multiple Twitter accounts in order to get around the limits on the retrieval rate. Retrieval was to be interspersed with delays so that maximum limits are not reached. In order to further speed up the retrieval process, I would need to do parallel retrieval. The Tweetf0rm GitHub repository (link shown below) seemed capable of handling these requirements.

  https://github.com/bianjiang/tweetf0rm.git
- The retrieval process would need to be able to recover from failure and restart automatically.
- I decided to use the Trump family, supporters and endorsers as the seed set.

5. ROADBLOCKS

The Tweetform software showed promise after initial testing. It could use three Twitter profiles and proxy servers to get around the per profile rate limit and the limit on requests from one IP address. It could send asynchronous requests using multiple Twitter profile authorities. However Tweetf0rm did not work on the Trump seed set. Possibly because of the large number of Trump Twitter followers.

Since Tweetf0rm could not be used, a Twitter Crawler needed to be written. The Twitter4J API was chosen for



Rank	Twitter Id	Handle	Person
1	25073877	@realDonaldTrump	Donald J Trump
2	1048784496	@RepKevinCramer	Rep. Kevin Cramer, ND
3	1305596696	@Rep_Hunter	Rep.Duncan D. Hunter, CA
4	1058256326	@RepChrisCollins	Rep. Chris Collins, NY
5	1180379185	@RealBenCarson	Ben Carson
6	252819323	@RepTomReed	Rep. Tom Reed, NY
7	239871673	@RepLouBarletta	Rep. Lou Barletta, PA
8	719132676691595264	@jhpjh0729	Jane
9	718825070001389568	@miracle1365	@miracle1365
10	720221840778526720	@hellobbo0624	hellobbo

Figure 11: Top Ten Page Ranks for Twitter follower graph

this purpose and the crawler was written in Java.

However this resulted in a delay in the execution of the project due to

- Time was lost in developing a new Twitter crawler as this was not planned for.
- The Twitter crawler that was developed, does not have the ability to do Twitter followers retrievals in parallel. This resulted in the follower graph not being as deep as would have been possible with Tweetf0rm, had it worked.

6. RESULTS

- A Twitter follower graph with 470,914 Twitter id nodes was downloaded in three weeks.
- With the seed set as level 0, the Twitter crawler was able to crawl through only partly through level 2. That is, the crawler started at the seed set and was able to reach the followers of the seed set followers.
- The PageRank process needed 7 iterations in order to reach a steady state where the cosine similarity between subsequent probability vectors crossed a threshold of 0.9999. The PageRank process took 54 hours to complete.
- The Page Ranks were sorted in descending order and the top ten results can be seen in figure 11.

The top seven Twitter Ids are from the seed set and the last three seem to be unrelated to politics in any way. The reason for this results seems to be that the Twitter follower graph is not deep enough and as a result there was not enough depth for the graph to have Twitter followers following people across levels, thereby creating a realistic graph of followers.

7. FUTURE IMPROVEMENTS

Given more time the following features could be incorporated in order to get better results

- A proxy server could be used in order to get around Twitter's rate limit on requests from a single IP address. Twitter4J allows the use of a proxy server, but does not have the ability to use multiple proxy servers in order to send Twitter API requests in parallel.
- Nowadays most computers have multi-core processors that are capable of running processes in parallel. This capability could be utilized in order to run four processes in a computer having a dual core hyper threaded CPU. Each CPU thread could use a different Twitter profile and send requests through its own proxy server.
- Given sufficient time and storage a Twitter follower graph could be built with more levels. I had planned for 25 levels, but in three weeks, I could only reach three levels.
- The PageRank process does multiplication operations on a vector and a matrix. Instead of doing naive multiplication, a more efficient algorithm like Strassen Algorithm[9] can be used for better performance.
- Matrix operations can also be done in parallel to a large extent. e.g. operations on one row could be made independent of operations on another row. Intel's Threading Building Blocks[10], a template library can be used to do operations in parallel on multi-core processors.
- The Hadoop Map-Reduce framework can be used to speed up PageRank computation if suitable hardware is available. In fact if the Twitter follower graph is built to 25 levels, like it was planned, the Map-Reduce framework may be the only setup capable of executing such large computations.

8. REFERENCES

[1] Barford, Vanessa. ”Why Are Americans so Angry?” BBC News. N.p., n.d. Web. 06 May 2016.

[2] 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.

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

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

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

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

[7] ”Cosine Similarity.” Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

[8] ”PageRank.” Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

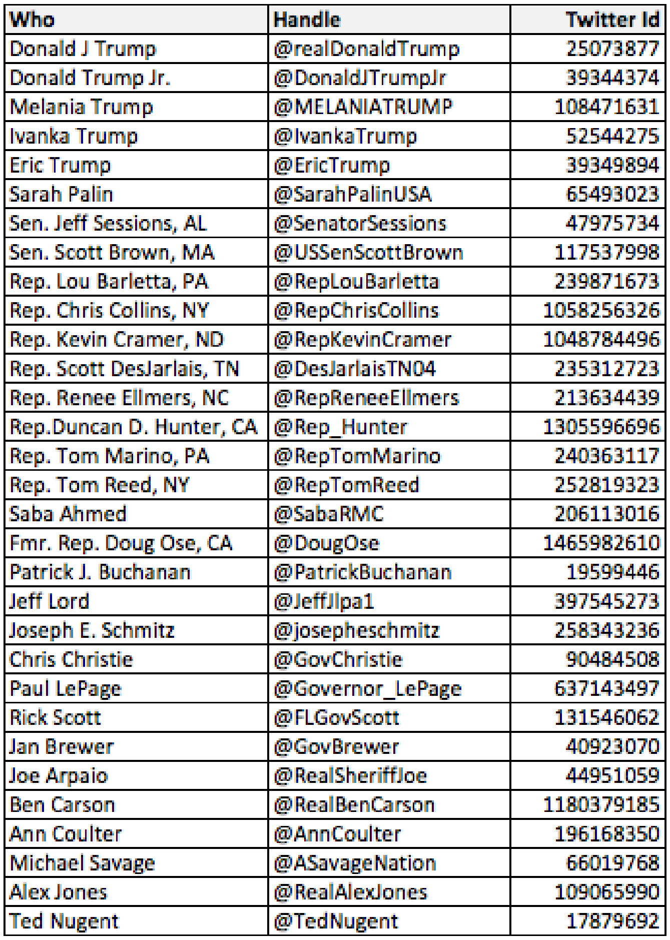
[9] ”Strassen Algorithm.” Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

[10] ”Threading Building Blocks.” Wikipedia. Wikimedia Foundation, n.d. Web. 06 May 2016.

9. SUPPLEMENTARY MATERIALS

9.1 Seed Set

The seed set used is shown in figure 12. 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 12: Seed Set