# ABSTRACT

Title of Document: FREEDOM TO TWEET OR TWEET TO FREEDOM: THE RELATIONSHIP BETWEEN FREEDOM STATUS AND TWEETS DURING ELECTIONS

Zahra Ashktorab, Master of Science 2013

Directed By: Jen Golbeck, Assistant Professor, College of Information Studies

In this thesis, I conduct an exploratory study of the relationship between a country's freedom and the twitter activity during elections. While there have been many studies of Twitter and elections, there has been no previous research conducted to explore the relationship between a countries' freedom and how Twitter influences elections in that given country. My goal is to identify hypotheses for future work in this area, introduce research designs and to shed light on areas of research where there seems to be little indication of relationships. I explore this space with automated analysis of the tweets' text, election outcomes, freedom ratings for the countries, and sentiment analysis. My results show that there seems to be a weak relationship between the outcome of an election and the sentiment expressed towards a candidate in tweets and that there is no relationship between the freedom in a given country and the sentiment expressed towards the incumbent. I found promising initial results regarding the relationship among content removed from links during an election and freedom status of a country, as well as the correlation between how frequently a candidate is mentioned and the election outcome. In the discussion, I present research questions in areas that are promising for future work.

# FREEDOM TO TWEET OR TWEET TO FREEDOM: THE RELATIONSHIP BETWEEN FREEDOM STATUS AND TWEETS DURING ELECTIONS

By

Zahra Ashktorab

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Masters of Science
2013

Advisory Committee:
Professor Golbeck, Chair
Jon Froehlich
Jimmy Lin

# Acknowledgements

I would like to express my deep appreciation and gratitude towards my advisor, Dr. Jen Golbeck for her support, guidance and patient mentorship. I would also like to thank the members of my committee Jon Froehlich and Jimmy Lin for their thought-provoking advice and friendly guidance.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Purpose and Significance of the Study

In the past several years, various social media platforms have become widely used on a global scale. One particular microblogging platform, Twitter, has emerged as a global Internet venue for expressing opinions. The topics discussed on Twitter vary, but almost no genre is left unaddressed. Millions of new posts are made daily about entertainment, sports news and political news.

In today’s wired world, politicians have realized the value of utilizing Twitter and the cost of ignoring it. In the United States, the 100% of the Senate and 90% of the House of Representatives have verified Twitter accounts (Sharp, 2013). Twitter is utilized in the campaigning process, to harness support for legislation and to spread information.

Certain political organizations and campaigns even find ways to exploit Twitter’s growth in prominence by political astroturphing. These are political campaigns disguised as grassroots behavior attempting to spread information (Ratkiewicz et. al, 2011). This use of social media shows that campaigns understand the importance of the role of social media in influencing opinions and ultimately influencing the outcome of an election. In the election process, candidates use social media as a means to express their views and harness support. Social media’s success in accumulating support is delineated by Barack Obama’s success in harnessing grassroots support using social media in his 2008 presidential campaign (Rasmussen & Schoen, 2010).

While the prominence of social media, particularly Twitter, has become evident in United States elections, Twitter also plays a role in elections on a global scale. Even in less democratic systems, where the general public may believe the election to be fixed, social media is used by citizens and people to anonymously express opinions about incumbents without experiencing the repercussions common to dictatorial regimes for expressing opinions freely. Globally, Twitter has played an instrumental role in elections and has even contributed to accelerating the pace of revolutions that contribute to entire regime changes (Chebib & Sohail, 2011). Egypt's revolution was influenced by social media (Eltantawy & Wiest, 2011). Western media dubbed the uprisings in Iran following the announcement of the 2009 election results, "The Twitter Revolution" (Grossman, 2009; Schleifer, 2009). In countries like Brazil, Venezuela, France, Mexico, South Korea, France and Colombia, presidential candidates have "verified" Twitter accounts and use their Twitter pages to express their views and campaign. In Russia, political hashtags affect public sentiment towards various topics (Alexanyan et. al, 2012). Thus, it is observed that social media plays a role in politics in countries all over the world with varying political climates.

The Freedom House, a non-governmental organization that administers research and promotes democracy, political freedom and human rights globally conducts a study annually that results in the assigned "Freedom Status" of all of the countries in the world. The Freedom House assigns a Freedom Status (Free, "not free" or "partly free") to all the countries in the world and assigns 'political rights' and 'civil liberties' scores  (between 1-6, 1 being most free, 6 being least). From this, I conduct an exploratory of study nine different countries with varying "freedom statuses", their tweets about presidential candidates, and the outcome of the election.

With social media burgeoning with new information streams and new users, researchers can utilize these information streams to discover trends in politics. In this thesis, I conduct an exploratory study of the relationship between a country's freedom status and the twitter activity during elections. While there have been many studies exploring Twitter’s role in elections (Gayo-Avello, 2012; Little, 2012; Livne et. al, 2011; O’Connor et. al, 2011), there has been no previous study of the relationship between a country’s freedom status and the Twitter activity during that election. I utilize global Twitter use to compare the relationship between tweets, freedom status of a country, and election outcome. Using a combination of network analysis, text analysis and metrics from the Freedom House, I explore the trends that emerge among nine different countries with different freedom statuses during their respective elections.

# Chapter 2: Review of the Literature

Much previous research has been conducted on the use of social media during elections and the after math of elections. Particularly, much research has been conducted on social media use in the countries focused in this study. Below, I discuss previous research regarding Twitter use and a few of the countries in this study. Furthermore, I outline previous work that has revealed evidence of abuse such as astroturphing, spamming and "message dilution" by political entities. I discuss previous work conducted on election prediction with Twitter in the United States, as well as criticisms of election prediction using Twitter.

## *Social Media Use During Mass Protests*

The last four years have seen an increase of social media use during mass protests globally. Egypt's revolution, one of the uprisings involved in the Arab Spring movement, was very closely linked with the widespread use of social media, particularly Twitter use. Internet use in the Middle East varies. Based on the Internal Telecommunications Union, 24% of Egyptians use the Internet. The Mubarak regime cut access to the Internet following the January 25, 2011 protests when widespread Twitter use posed a threat to the regime. Despite limited Internet access, Egyptians were able to make use of Twitter, as demonstrated by the tweet: "RT @Dima_Khatib: Mobiles around Tahrir Square are not working any more. Blocked too. Like Internet #egypt #jan25 #cairo". Tweets like the one above demonstrate that Egyptians view Twitter as an important tool in furthering their cause to democracy and freedom (Kavanaugh et. al, 2012).

Similarly, Twitter played an important role in the 2009 Iranian election. Kavanaugh et. al (2012) discusses the reasons why Twitter was important in Iran's street protests following the 2009 Iranian elections.  Protestors required immediate information in order to avoid clashes with the authorities. The government blocked access to Twitter, so the Twitter service was only available either through proxy or text message on a mobile phone.

### *Different Forms of Censorship Within "Not-Free" Countries*

Previous research outlines the various methods dictatorial regimes employ to prevent the access to information. In countries like Egypt and Iran, the government outright blocks access to sites like Facebook and Twitter in times of protest. For example, Internet traffic dropped abruptly from and to Egypt across 80 Internet Service Providers on January 25, 2011.  As a result of this government intervention, approximately 97% of Egyptian Internet traffic was lost during this time (Kavanaugh et. al, 2012).

Though Russia shares the same Freedom House classification as Iran and Egypt according the 2012 Freedom House Annual Report (2012), authorities in Russia employ other methods to stifle Twitter dissent, for instance, through the use of hashtags. Hashtags in Twitter are important during any crisis or event. Hashtags have become a mechanism for organizing conversations around topics. The government in Russia takes advantage of this feature on Twitter to employ a new type of censorship. Thomas, Greer and Paxson (2012) define "message dilution", a process that involves automated accounts posting conflicting, irrelevant and incomprehensible content with hashtags that are used by legitimate users in an attempt to "hijack" the conversation.

During the most recent 2011 Russian parliamentary elections, 25,860 fraudulent Twitter accounts "injected" 440,793 tweets into legitimate conversations about the election, in an attempt to distract the conversation from the original topic. According to the geolocations of the "injected" tweets, the majority of the spam bots used to inject tweets were not located in Russia. 39% of the IP addresses from which the spam bots tweeted belong to IP blacklists. This study relied on Twitter's internal spam detection algorithm to detect the spam (Thomas et. al, 2012). This work shows that even governments in "not free" countries such as Russia understand the importance of Twitter as a political tool.

### *Astroturphing*

While the previous section discusses "message dilution", a method employed by Russia, a "not free" country, to stifle political dissent, political campaign groups in "free" countries also attempt to use Twitter to influence public opinion. In the United States, political organizations and campaigns exploit Twitter's growth in prominence by political astroturphing. These are political campaigns disguised as grassroots behavior attempting to spread information. Rakiewtz et. al (2012) introduce a new system architecture, Truthy, to detect atroturphing and ultimately succeed in automatically detecting political memes, a term coined by Stephen Colbert to described something some claims to know that is known based on feelings, rather than facts (Ratkiewicz et. al, 2012).

In the most recent Mexican elections, the Institutional Revolutionary Party has reportedly also resorted to spam tactics. The Institutional Revolutionary Party (PRI) used about 10,000 bots to tweets words or phrases in an attempt to generate trending topics, popular topics discussed on Twitter. These spam bots take advantage of the existence of hashtags to make certain words or run-on phrases searchable. While all three dominant political parties utilized spamming in the Mexican presidential campaign, a Mexican web developer has created a site listing all of the spam accounts used by the PRI. The list of spam bots can be found at this url: http://santiesteban.org/adiosbots/en.html.

Detecting spam, astroturphing and sybil accounts involved in "message dilution" all present the same challenges. Different strategies have been deployed for the detection of such abuse on Twitter, such as analyzing the profiles and networks of spam accounts, looking at statistical properties of accounts, and detecting spam URLs (Thomas et. al, 2012).

### <u>*Election Prediction*</u>

Many studies have attempted to predict election outcomes looking at Twitter data. A study conducted by O'Connor, Balasubramanyan, Routledge, and Smith (2010) compares the results of traditional polls with sentiment provided by the text in Twitter. Looking at the text in Twitter, this study aimed to retrieve relevant information and decide whether a Tweet expressed a positive or a negative opinion. They employed a deterministic approach and used linguistic knowledge to decide whether a tweet was positive or negative. Instances of positive-sentiment words and negative sentiment words were counted. In this study, a formula is presented to represent the day's sentiment. This formula is the ratio of positive-sentiment words over negative-sentiment words. The study found that there was a strong correlation between sentiment on Twitter and what was reflected in the polls. Though this study utilizes interesting methods to measure public opinion through Twitter, it does not look at the relationship between those sentiments expressed and election outcome. Furthermore, it is specifically focused on the United States election.  I conducted this study on a global scale and look at the outcome of the election, not just public sentiment. In countries where the election is fixed and there is a likelihood of fraud, public sentiment may be inclined against the incumbent. I look at nine different countries to find trends and relationships between sentiment on Twitter and the election outcome.

Twitter prediction election has its criticisms as well. Gayo-Avello (2012) has a pessimistic view of election prediction using Twitter. He mentions several flaws of using Twitter as a means to make election predictions. He states that incumbency plays a large role in elections and that "chance is not a valid baseline", that there is no robust way to count votes on Twitter, and variations of sentiment analysis do not yield a valid result (Gayo-Avella, 2012).

While incumbency plays a large role in elections, some of the countries to which I look at for this study do not have incumbents (Brazil, Colombia, and Egypt). Additionally, this study can show the degree to which incumbency influences the outcome of the election in different types of countries (countries classified as Free, "partly free", and "not free").

In this study, I take the criticisms outlined above under consideration and do not aim to make predictions. However, I do aim to use previously utilized methods to conduct an exploratory study of global tweets and the role they play during elections in nine different countries. Though there is no robust and completely accurate way to “count” votes on twitter, number of mentions of a candidate, or hashtags associated with a candidate can express the sentiment of the tweeters and can ultimately aid in understanding the nature election in a particular country.

Using sentiment analysis, automated analysis of tweets, analysis of networks of tweets, as well as keeping in mind abuses that occur in tweets (atroturphing, message dilution, and spamming), I aim to explore the relationship of the sentiment reflected in tweets and the election outcome as well as detect insightful trends from within the Twitter data. In this exploratory study, I aim to introduce new research designs, data collections methods and selection of subjects given the preliminary results of the various analyses conducted.

# Chapter 3: Methodology

The methodology utilized in this study is multifaceted and consists of various methods of analysis of data. The results yielded from the methods below, which were part of my exploratory study, and allowed me to introduce new hypotheses and research designs for new areas of research.

I collected tweets occurring before the general election of each of the nine countries outlined in Table 2 for a week prior to the respective election date. Analyzing the sentiment of the tweets toward the candidates as well as the graphical structure of the networks resulting from these tweets, I also looked at the frequency of links, hashtags and mentions in these tweets. A "hashtag" is defined as a tag embedded into a tweet on Twitter prefixed with a hash sign, "#". Hashtags are used to organize topics around tweets. In this study, I explore two different types of mentions: (1) a mention is a when a Twitter handler user name embedded in a tweet is prefixed with the ampersand symbol, "@", or (2) a mention is any mention of a candidate's last name. For the second type of mention described above, for certain countries in which candidate last names were potentially ambiguous, both first and last names were queried.

Textual analysis and graphical analysis as well as the metrics from Freedom House were used to compare the tweets generated for all of the elections. To compile a range of countries with differing freedom levels, I used the classifications presented by Freedom House (2012). I reviewed the tweets from the countries in Table 2 during the mentioned election cycles. For presidential systems in which there are two election cycles, an initial and secondary round, I look at the election cycle that determines the winner. All of the government systems that are reviewed in this study are presidential.

### *The Freedom House*

The freedom classifications in this study are based on the *The Freedom in the World* survey, an annual survey evaluates the status of global freedom. The classifications are according to two categories: civil liberties and political rights. The survey includes analytical reports and numerical ratings of 195 countries and 14 select territories. The report also includes a summary for each country of the last years major developments. The ratings are based on checklist of 10 political rights questions and 15 civil liberties questions. The questions were rated by 59 analysts and 20 senior-level academic advisors using a variety of information sources: academic analyses, foreign and domestic news reports, think tanks, nongovernmental organizations, individual professional contacts, and visits to the region. Based on these sources of information, each country is assigned a civil liberties and political rights score. These scores are averaged for each country to determine whether the country is “free”, “partly free” or “not free”. A country receiving an average rating between 1.0 -2.5 is considered “free”, an average score of 2.0 – 5.0 “partly free” and average score of 5.5-7.0, “not free” (The Freedom House).

### *Freedom Classifications*

According to the Freedom House, a country can be classified as “Free”, “partly free”, or “Not free”. These three classifications are defined by the Freedom

House in its “Freedom in the World 2012” annual report. In this report, a “free” country is defined as a country “where there is open political competition, a climate of respect for civil liberties, significant independent civic life, and independent media.” A “partly free” country is defined as one “in which there is limited respect for political rights and civil liberties.” The freedom house concludes that “partly free” countries “suffer from an environment of corruption, weak rule of law, ethnic and religious strife, and a political landscape in which a single party enjoys dominance despite a certain degree of pluralism.” Finally, a “not free” country is defined as “one where basic political rights are absent, and basic civil liberties are widely and systematically denied.”

The resulting metrics from the 2012 Annual Report were used as the primary Freedom House Classifications for this study. Additionally, the Freedom House offers a second set of classifications that scope the results of the research conducted in this study. Freedom House offers “internet freedom” scores. However, the Freedom House metric scores were most relevant to the research in this paper because go beyond just Internet freedom and encompass political climate and civil liberties, themes and topics reflected in the tweets in this study.

| **Country** | **Press Freedom Score** | **Internet Freedom Score** |
|---|---|---|
| The United States | Free | Free |
| Brazil | Partly Free | Free |
| France | Free | N/A |
| Colombia | Partly Free | N/A |
| Venezuela | Not Free | Partly Free |
| Mexico | Not Free | Partly Free |
| Egypt | Partly Free | Partly Free |
| Iran | Not Free | Not Free |
| Russia | Not Free | Partly Free |

**Table 1: 2013 Internet freedom scores by the Freedom House for each country in this study.**

These particular metrics were taken into consideration as the study was conducted. Iran is the only country in this study with an Internet freedom status of "not-free." This freedom status was reflected in the challenges I faced when collecting tweets that occurred during the Iranian election, as well as in the amount of content removed from links tweeted during the Iranian election.

### *Determining which Countries to Analyze*

In order to ensure that countries were chosen that had potentially relevant tweets, I investigated election cycles that occurred after 2008. Barack Obama's success in harnessing grassroots support through social media in the 2008 United States election is a milestone that marks the beginning of the utilization of social media by political campaigns globally. I stipulate that elections after this date are relevant (Rasmussen & Schoen, 2010). The two countries with the most Twitter users are the United States and Brazil (Evans, 2010). For this reason, the United States and Brazil are obvious choices for being representative of "free" countries to analyze in this study. The third country I selected to analyze, France, was chosen because it held a presidential election in the past year, and is additionally rated as one of top twenty countries worldwide with the most Twitter users (Evans, 2010). From the nine countries explored in this study, France's 2012 election was also the only example of an election in which the incumbent lost the election.

In order to select three "partly free" countries to explore for this study, I considered two criteria: twitter usage and type of political system. I selected Colombia, Mexico and Venezuela as the "partly free" Countries in this study because all three of these countries are rated as one of top twenty countries worldwide with the most Twitter users. Some of the Twitter usage in these “partly free” countries rates even higher than twitter usage in some “free” countries (Evans, 2010).

The most challenging part of the selection process was choosing countries classified as “not free” by the Freedom House. Countries that are classified as “not free” are classified as such because citizens do not have as many civil liberties or human rights as their “free” country counterparts. Lack of human rights often translates to limited access to information resources, like social media. Countries that had high Twitter usage (with respect to other “not free” countries) and had held presidential elections since 2008 were selected as “not free” countries for this study. As a result, Russia, Iran, and Egypt were selected. While some countries like Iran simply block content on social media, other countries like Russia have attempted to manipulate social media to their advantage by message dilution, discussed in the Chapter 2. The attempt by Russia to “hijack” hashtags shows that Twitter plays an influential role in Russian politics (Thomas et. al, 2012). For this reason, I chose Russia as a “not free” country to explore as a part of this study. I chose Egypt as another representative sample of a “not free” country because Twitter played an instrumental role in the 2011 Revolution and the “Arab Spring” as well as the most recent election in Egypt (Kavanaugh et. al, 2012).

In Iran, the government blocks access to Twitter (Kavanaugh et. al 2012). Given this censorship, the candidates of such countries often avoid using social media mainly because the general public may not be able to access it. For example, I found no Twitter usage by Iranian candidates. However, the significant role of Twitter use in Iran is demonstrated by the usage of Twitter to spread news about protests in June 2009 during the “Green Revolution”. Iran’s significant Twitter usage is also demonstrated by the fact that it is listed in the top twenty countries with the most Twitter users (Evans, 2010). Though the candidates in Iran don’t use Twitter, the people are very much involved, despite accessibility issues due to censorship (Burns, 2009). For this reason, I chose to analyze Iran as one of the “not free” countries in this study.

### *Data Collection*

Candidate tweets were collected for six months prior to the election date. General tweets in which the content mentioned candidate names were also collected from one week prior to the election. In order to accurately collect candidate tweets, I ensured that Twitter accounts that claimed to represent the candidate were “verified” by Twitter. Not all of the candidates in this study had “verified” twitter accounts. For example, none of the Iranian or Russian candidates had “verified” twitter accounts. For countries in which I do not have candidate twitter names, I focused my analysis on tweets six months prior to election by the general public. I have chosen countries and election cycles in which Twitter played a role so that in the event of insufficient candidate twitter data, tweets by the general public will be available for analysis.

From the candidate twitter accounts, three lack the “verified” tag. These Twitter accounts belong to: Gabriel Quadric de la Torre of Mexico, Sergey Maroon of Russia, and Virgil Goode of the United States. All of these accounts contain information that imply that they belong to the their owners. Torre’s twitter account has 216,513 followers and links to his official site http://nuevaalianza.mx/. Maroon’s site has 81, 656 followers and also points to his official site http://mironov.ru. Goode's account has 957 followers and links to his campaign website for the 2012 presidential election http://goodeforpresident2012.com. Without the "verified" tag, however, it is difficult to be completely certain that these twitter accounts belong to whom they claim to belong. I kept in mind this uncertainty as I conducted the analysis. For all candidates who have a Twitter account, including the two un-verified twitter accounts, the Twitter Search API was used to collect tweets for six months prior to the election date.

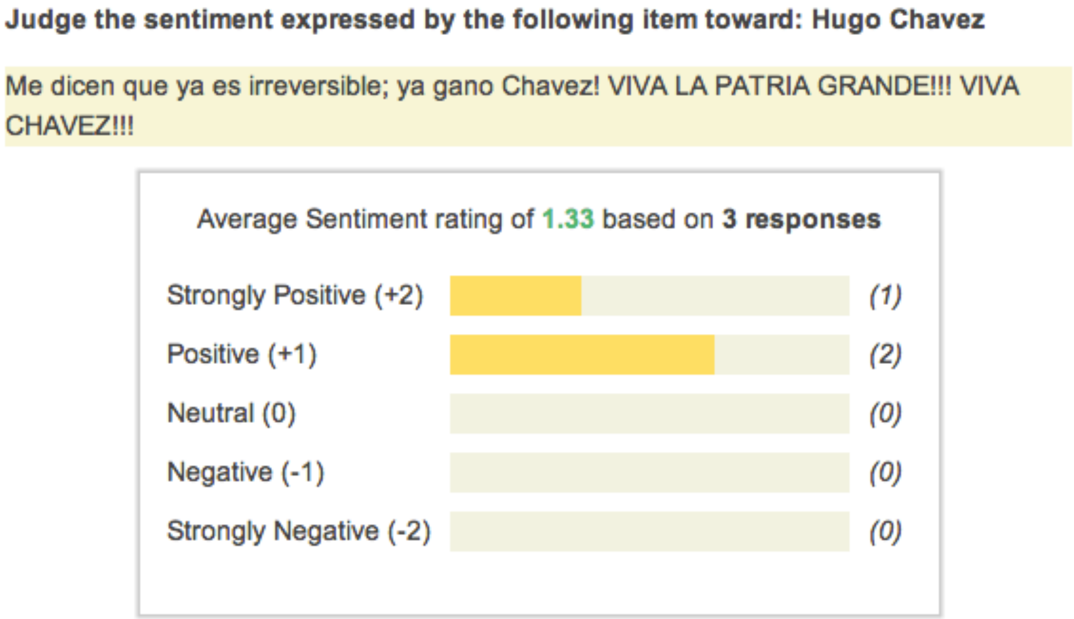
In order to analyze tweets made about candidates during the elections of each of these countries, I collected tweets from one week prior to elections that contained the names of any of the candidates from http://www.topsy.com. Topsy, a service that has access to Twitter's stream of information, allows users to search tweets that occurred during the window of time indicated by the query. The smallest window of time allowed by Topsy is one hour. In order to maximize the number of tweets collected, queries were constructed for tweets that contained candidates' names for every hour within one week prior to the election date. Candidate names were queried both in English as well as the native language of the country in which the election occurred.

Twitter's current API does not allow searching for old tweets. Ideally, tweets collected directly from Twitter would have yielded a more representative distribution of actual tweets during these election cycles. However, Topsy yields a representative sample of tweets adequate for this study.

*Sentiment Analysis*

To measure sentiment towards candidates, sentiment analysis was conducted on candidates that received more than 10% of the popular vote for each country. For the sentiment analysis of tweets, I opted for the Naïve Bayes Algorithm through the Natural Language Toolkit (NLTK).  Naïve Bayes is an efficient and effective tool in language learning (Tumasjan et. al, 2012). To use this method, I trained the classifier to classify a tweet as either positive or negative towards each candidate. To accomplish this, users on Amazon's Mechanical Turk rated 700-800 tweets and determined whether a tweet was positive or negative towards a candidate for the USA, France, Brazil, Colombia, Venezuela, and Mexico. The training set for Egypt, Iran and Russia were rated by Arabic, Farsi and Russian speakers respectively using the identical rating method employed by Amazon Mechanical Turk users. Each tweet on Mechanical Turk was rated three times as either very negative, negative, neutral, positive, and very positive. These ratings are associated with the scores -2, -1, 0, 1, and 2 respectively. The sentiment score of a tweets is the average of the three scores by Mechanical Turk users. Tweets in the training set with a score less than -.25 were classified as negative, tweets with a score between -.25 and .25 were classified as neutral, and tweets with a score greater than .25 was classified as positive.

I then provided this training set of pre-rated tweets to the classifier for each candidate in order to train the system. Using the classifier, I derived a sentiment classification of negative, neutral, or positive for each candidate. A total sentiment score was calculated towards each candidate by subtracting 1 from the score if the tweet was positive, adding a 1 to the score if the tweet is negative and doing nothing if a tweet was classified as neutral. This number was then divided by the total number of tweets to yield a sentiment score towards a particular candidate.



**Judge the sentiment expressed by the following item toward: Hugo Chavez**

Me dicen que ya es irreversible; ya gano Chavez! VIVA LA PATRIA GRANDE!!! VIVA CHAVEZ!!!

Average Sentiment rating of 1.33 based on **3 responses**

| | |
|---|---|
| Strongly Positive (+2) | *(1)* |
| Positive (+1) | *(2)* |
| Neutral (0) | *(0)* |
| Negative (-1) | *(0)* |
| Strongly Negative (-2) | *(0)* |

**Figure 1: Example of Interface for rating sentiment towards candidate on Mechanical Turk.**

*Network Analysis*

I analyzed the network of tweets for the election cycle of each country. From the entire set of tweets collected, I derived a sample of 5000 tweets for each country in order to generate the network. I opted for Random Node Sampling because the Twitter API rate severely limits my options. Though Random Node Sampling does not retain power-law degree distribution, it is currently the best option for sampling for the data collected (Leskovec, 2006). Current Topsy data contains only a screen name, tweet content, and an influence score. Three additional Twitter API calls need to be made for each user to obtain (1) their user id, (2) followers, and (3) friends. Due to time constraints, this would be an unreasonable amount of API calls for a network of 50,000 twitter profiles. Optimally, Forest Fire Sampling and Snowball Sampling would yield better samples of the data, however, the number of API calls to sample 10% of the total tweets by applying these approaches to any of the data I have collected would take too long to accomplish with the current Twitter API limitations.

Additionally, I considered non-uniform Node Sampling as well as Edge Sampling. However, both Node Sampling and Edge Sampling require additional calls to the Twitter API and exhaust the number of calls that I have available per hour. I also considered looking for nodes with the highest edge degrees in a particular network. However, such a method would require traversing the entire data set and making calls to the Twitter API for all of the data. For this reason, I have used random node sampling for my data.

### *Clustering*

In order to detect communities, I clustered the sample networks using the Clauset-Newman-Moore algorithm. I analyzed Twitter profiles in each cluster to accurately classify communities in each network. I then classified clusters in each of the networks and calculated the overall influence score of certain clusters, particularly in countries in which other clusters' tweets did not originate in the country being studied.

Using network analysis, I attempted to identify key players in the network based on in-degree of nodes and betweenness centrality of nodes. The average degree, network diameter, average path length, number of shortest paths, density, modularity, number of weakly connected components, number of strongly connected components, average clustering coefficient, and number of communities were calculated for each of the networks.

*Links, Mentions and Hashtags*

The set of tweets for each candidate, compiled over six months prior to the election date, were analyzed to extract a list of the most frequent links, hashtags and mentions for all of the tweets during a particular election. I aimed to see which hashtags were popular and which candidates were mentioned more frequently.

| **Country** | **Freedom Status** | **Political Rights** | **Civil Liberties** | **Candidates** (Incumbents are **bold**) | **Twitte Handle** (non-verified are red) | **Election Date** |
|---|---|---|---|---|---|---|
| United States | Free | 1 | 1 | **Barack Obama** | BarackObama | 11/6/12 |
| | | | | Mitt Romney | MittRomney | |
| | | | | Jill Stein | jillstein2012 | |
| | | | | Gary Johnson | GovGaryJohnson | |
| | | | | Virgil Goode | VirgilGoode | |
| Brazil | Free | 2 | 2 | Dilma Rousseff | Dilmabr | 10/31/10 |
| | | | | Jose Serra | jossers_ | |
| France | Free | 1 | 1 | Francoise Hollande | Hollande | 4/22/12 |
| | | | | **Nicolas Sarkozy** | NicolasSarkozy | |
| Venezuela | Partly Free | 5 | 5 | **Hugo Chavez** | Chavezcandanga | 10/7/12 |
| | | | | Henrique Caprioles Radonski | Hcapriles | |
| Colombia | Partly Free | 3 | 4 | Juan Manuel Santos | JuanManSantos | 6/20/10 |
| | | | | Antenas Mockus | AntenasMockus | |
| Mexico | Partly Free | 3 | 3 | Enrique Pena Nieto | EPN | 7/1/12 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  |  | Andres Manuel Lopez Abrader | lopezobrador_ |  |
|  |  |  |  | Josefina Vazquez Mota | JosefinaVM |  |
|  |  |  |  | Gabriel Quadric de la Torre | g_quadri |  |
| Russia | Not Free | 6 | 5 | Vladimir Puttin | N/A | 3/4/12 |
|  |  |  |  | Gennady Zyuganov | N/A |  |
|  |  |  |  | Mikhail Prokhorov | N/A |  |
|  |  |  |  | Vladimir Zhirinovsky | N/A |  |
|  |  |  |  | Sergey Maroon | mironov_ru |  |
| Iran | Not Free | 6 | 6 | Mahmoud Ahmadinejad | N/A | 6/12/09 |
|  |  |  |  | Mir-Hossein Mousavi | N/A |  |
|  |  |  |  | Mehdi Karroubi | N/A |  |
|  |  |  |  | Mohsen Rezaee | N/A |  |
| Egypt | Not Free | 6 | 5 | Mohamed Morse | MuhammadMorsi | 6/16/12-6/17/12 |
|  |  |  |  | Ahmed Shafik | Ahmedshafikeg |  |

**Table 2: Countries, freedom status, political rights score, civil liberties score, candidates, candidate Twitter handles, and election cycles in this study.**

# Chapter 4: Findings

In this chapter, I present my sentiment analysis results and classify communities within each network of the election tweets for all nine countries in this study. In order to conduct this exploratory research to determine the research design for hypotheses I conducted sentiment analysis of tweets, clustered the network of tweets to find communities, and extracted the top links, hashtags and mentions for each election cycle.  My methodology involves analyses that allow me to draw conclusions about the nature of each network. I chose sentiment analysis in order to understand how the tweeters for each of the elections felt towards each candidate. I clustered the network to detect communities within the network and to understand the structure of the network. And lastly, I looked at the top links, mentions and hashtags to gain an additional understanding to the sentiment analysis of the content of the tweets. These three methods yielded initial promising results that allow me to introduce hypotheses and research designs further discussed in Chapter 5.

## *Sentiment Analysis*

Following the analysis of my results, I discovered that there is no relationship between the sentiment of a country towards the incumbent and its freedom status. Additionally I discovered that, regardless of a country's freedom status, there is no correlation between the sentiment presented towards a candidate and the outcome of an election.

For each of the countries in this study, the overall sentiment towards a candidate in every tweet during an election cycle was rated as either negative, neutral, or positive towards a particular candidate by Mechanical Turk users. These ratings were used as a training set to find the overall sentiment score towards a particular candidate. A score system was derived to reflect sentiment towards the top two contenders in every election cycle. Overall sentiment scores towards each candidate of the entire tweet set during the election cycle were calculated by adding one point to the score if the sentiment score of a tweet was positive, subtracting one point if the sentiment of a tweet was negative, and doing nothing (adding zero) if the sentiment of a tweet was neutral or irrelevant.

**Sentiment in "Free" Countries**

The sentiment expressed towards candidates in the "Free" countries in this study reflected that the winner of the election has a lower sentiment score than the competing candidate. The tweets in the United States show a negative sentiment score for both candidates, with Barack Obama scoring a -.41, and Mitt Romney scoring a -.26. Even though Barack Obama won the election, his sentiment score was lower than his competitor Mitt Romney. The top hashtags in the United States election provide a clue as to why Barack Obama's score was lower than that of Mitt Romney (Table 13). The hashtag "#tcot", standing for "Top Conservatives on Twitter" is the most used hashtag. In contrast, "#tlot" (Top Liberals on Twitter) appears lowest on the list of top hashtags.

In France, Nicolas Sarkozy, the incumbent at the time of the election, received a higher sentiment score than François Hollande, the winner of the election. While both candidates received positive scores, Nicolas Sarkozy received a score of .15, while François Hollande received a score of .00. In Brazil, Dilma Roussef, the winner of the election, scored a -.25, while Jose Serra, scored a .31. Even though the general sentiment towards Dilma Rousseff was negative, the election outcome did not reflect this score. Among the "free" countries in this study, all three countries (United States, Brazil, and France) reflected that the winner of the election has a lower sentiment score than the loser of the election.

### Sentiment in "Partly Free" Countries

While the sentiment scores in the "partly free" countries in this study did not consistently reflect the winner of the election, two of the countries, Colombia and Venezuela, demonstrated that the winner of the election had a higher sentiment score than the defeated candidate. In Venezuela, Hugo Chavez scored a sentiment score of 0.62, the highest candidate sentiment score in this study, while Henrique Capriles Radonski scored a 0.28. In Colombia, Juan Manuel Santos scored a -0.01, higher than the losing candidate, Antanas Mockus, who scored a -0.06.

While the scores in these two countries might indicate that the sentiment in "partly free" countries accurately reflects the outcome of the elections, Mexico's results reflected otherwise when Enrique Peña Nieto, the winner of the election, received the lowest sentiment score in all of Mexico's presidential candidates. Nieto scored a -.12, compared to competing candidates Josefina Vazquez Mota and Andres Manuel Lopez Obrador, who scored a 0.09 and -.08 respectively.

### Sentiment in "Not Free" Countries

Similar to the pattern observed in the "free" countries in this study, the sentiment score of the candidate who lost the election in "not free" countries was higher than that of the candidate who won the election. The sentiment analysis of tweets during the election cycles in Russia, Egypt and Iran reflect that the candidate with the lower sentiment wins the election. In Egypt, the winner of the election, Mohammad Morsi, scored a -0.04 while his competitor, Ahmed Shafik, scored a 0.04. In Iran, Mahmoud Ahmadinejad, the incumbent as well as the winner of the election, scored -.51. His competitor, Mir Hussein Mousavi, scored a 0.07. In Russia, Vladimir Putin, the winner of the election, scored -0.07 against Gennady Zyuganov, who scored a 0.08. In all three "not free" countries, the winner of the election received a lower sentiment score than that of the loser. A similar trend was observed for "free" countries. Based on these scores, it is impossible to correlate freedom and sentiment.

### The relationship between Incumbency and Sentiment

The sentiment scores reflected that there is no relationship between incumbency, freedom status and sentiment. Out of the nine countries in this study, four held an incumbent at the time of the election: The United States, France, Venezuela, and Iran. Of the tweets belonging to these four countries, tweets from the United States and Iran reflected a negative sentiment towards the incumbents Obama and Ahmadinejad respectively. Venezuela and France's tweets reflected a positive sentiment towards the incumbents Chavez and Sarkozy respectively. The United States and Iran, as well as Venezuela and France, have different freedom statuses. Based on these scores, it is observed that there is no relationship between sentiment, incumbency and freedom status.

### Predicting Election Outcome based on Sentiment

If election predictions were made solely on the sentiments conducted in this study, only Colombia and Venezuela would yield the correct election outcome. The sentiment scores from the other election cycles reflected a more negative score for the winner of the election. The sentiment analysis of tweets in our study reflects simply that sentiment analysis of tweets cannot be used to predict election outcome.

| **Candidate** | **Sentiment Score** | **Country** | **Freedom Status** |
|---|---|---|---|
| **Barack Obama** | **-0.41** | **United States** | **Free** |
| Mitt Romney | -0.26 | United States | Free |
| **Dilma Roussef** | **-0.25** | **Brazil** | **Free** |
| Jose Serra | 0.31 | Brazil | Free |
| Nicolas Sarkozy | 0.15 | France | Free |
| **François Hollande** | **0.00** | **France** | **Free** |
| **Hugo Chavez** | **0.62** | **Venezuela** | **Partly Free** |
| Henrique Capriles Radonski | 0.28 | Venezuela | Partly Free |
| **Juan Manuel Santos** | **-0.01** | **Colombia** | **Partly Free** |
| Antanas Mockus | -0.06 | Colombia | Partly Free |
| Josefina Vazquez Mota | 0.09 | Mexico | Partly Free |
| **Enrique Peña Nieto** | **-0.12** | **Mexico** | **Partly Free** |
| Andrés Manuel López Obrador | -0.08 | Mexico | Partly Free |
| **Mahmoud Ahmadinejad** | **-0.51** | **Iran** | **Not Free** |
| Mir Hussein Mousavi | 0.07 | Iran | Not Free |
| **Mohammad Morsi** | **-0.04** | **Egypt** | **Not Free** |
| Ahmed Shafik | 0.04 | Egypt | Not Free |
| **Vladimir Putin** | **-0.06** | **Russia** | **Not Free** |
| Gennady Zyuganov | 0.08 | Russia | Not Free |

Table 3: Sentiment scores for candidates

The results of the sentiment analysis conducted in this study did not yield a correlation between sentiment of overall tweets, freedom status of state, and election outcome. One reason can be attributed to a lack of sufficient tweets to calculate an accurate sentiment score for tweets. The Topsy service was used in this study to collect tweets. Topsy only yields “influential” tweets in its queries. Such queries result in tweets whose writers are influential, meaning that that they have many followers. A large percentage of tweets collected belonged to established news outlets or organizations whose tweets do not reflect the opinion of a single individual. While tweets from organizations, especially political parties, are relevant for this study, it is difficult to calculate the sentiment score of a candidate when the tweets of organizations and news agencies are included in the overall scores.

### *Network Characteristics*

In this section, I classify communities within a random sample of tweets for each of the election cycles from each of the nine countries in this study. I used the Clauset Newman Moore algorithm to cluster tweets to view visible communities within each of these sample networks (Clauset, 2004). In the networks shown below, communities were formed based on a variety of attributes. The communities in each of these networks reveal that there are classifiable groups within each country. Communities formed according the languages used to tweet, the tweet’s country of origin, or based on similar interests such as entertainment. Clustering tweets into communities aids in understanding the network and communities that form it.

#### **Brazil**

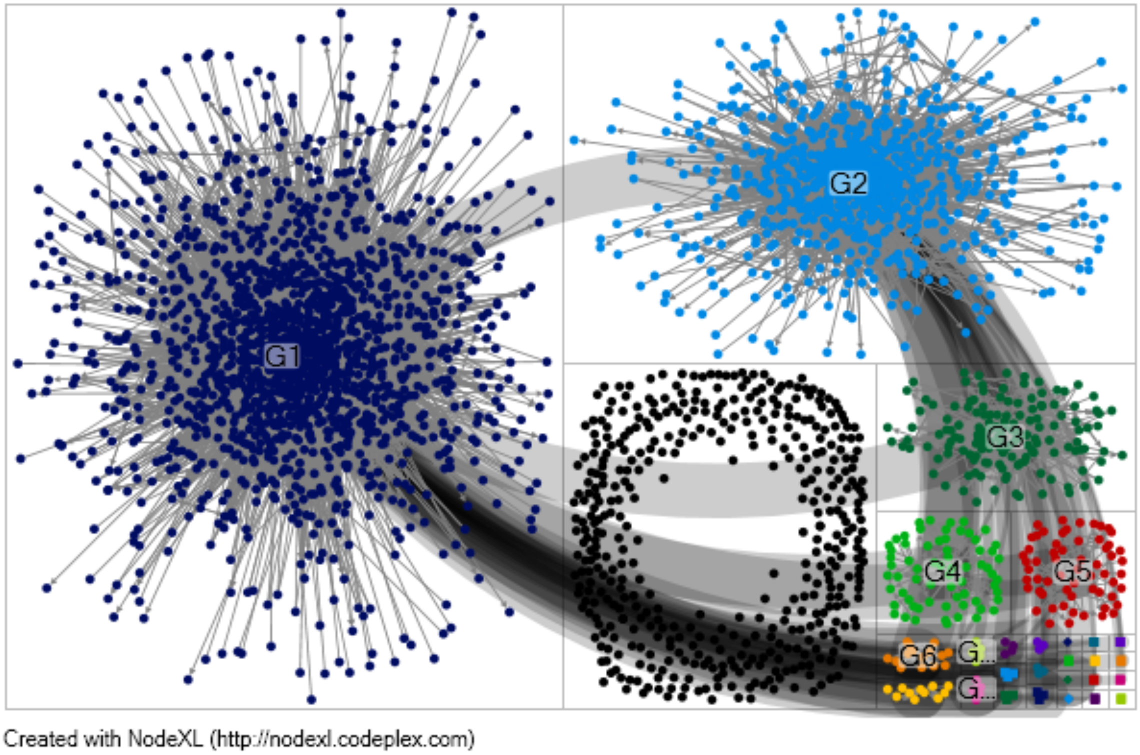


**Figure 2: Network of communities tweeting the week before Brazil's election**

Five groups emerged from the tweets resulting from the Brazilian election. The two largest communities ("G1" and "G2") consisted of journalists, bloggers and news agencies located in Brazil or specifically Sao Paulo. These communities tweeted primarily in Portuguese. A third group, "G3", consists of Latin American news agencies, bloggers, or popular Twitter individuals tweeting in Spanish. The tweets in "G3" were primarily in Spanish and originated in various locations around South America including Mexico and Venezuela. "G4" consisted of Brazilian entertainers and individuals including comedians, adult entertainers, and students. These tweets were primarily in Portuguese. "G5", also consisted of tweets in Portuguese, reported exit poll information. 40% of the tweets produced by profiles in "G5" reported exit poll information.

## Colombia

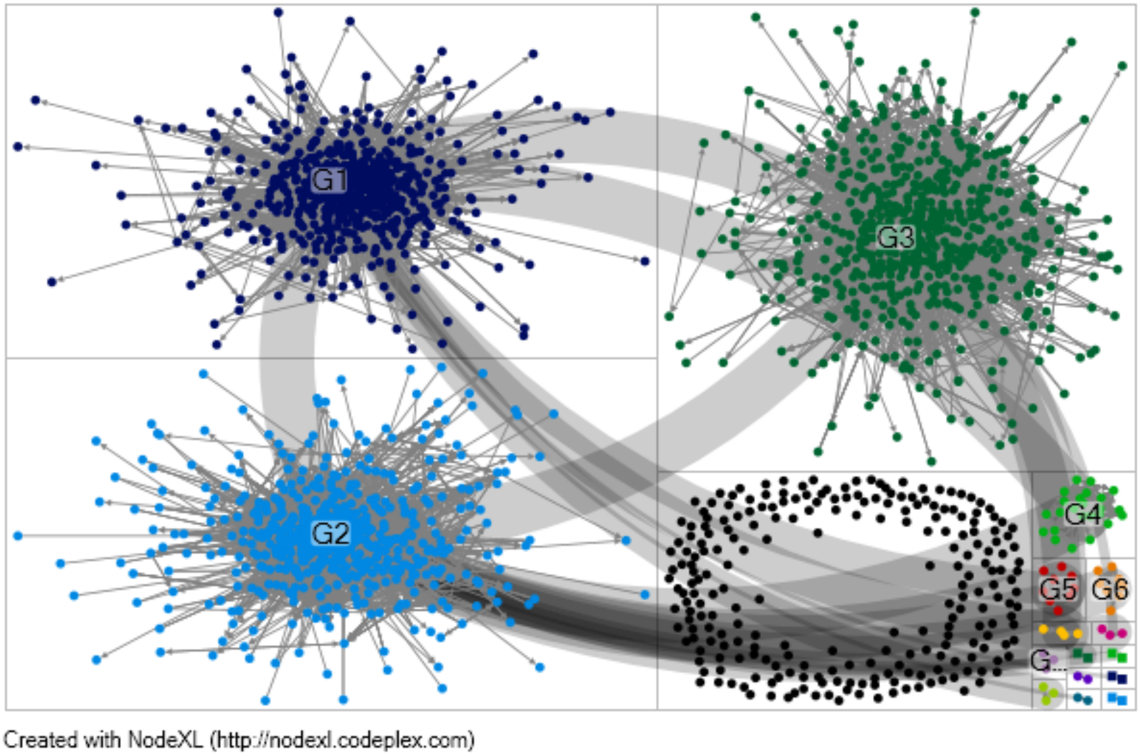


Figure 3: Network of communities tweeting the week before Colombia's election

Colombia's network can be grouped into two main communities. The first largest group, "G1", consists of both candidates, Antanas Mockus and Juan Manuel Santos, media outlets, organizations, journalists, and the Official Green Party. The second group, "G2", includes columnists, journalists and news agencies from other Latin American countries including Peru and Venezuela.

## Egypt

Clustering the network of tweets from the Egyptian election yielded three significant groups. The largest group, "G1", consisted of Arab media outlets and political organizations that tweet primarily in Arabic. The second largest community, "G2", also consisted primarily of media outlets, political activists, and political organizations, however the tweets were primarily in English and locations self-reported within the profile were not located in Egypt. The third community of tweeters that emerged during the Egyptian Election cycle was a community of Egyptian individuals primarily tweeting in Arabic from within Egypt. The individuals in this group self reported that they were tweeting from locations within Egypt.

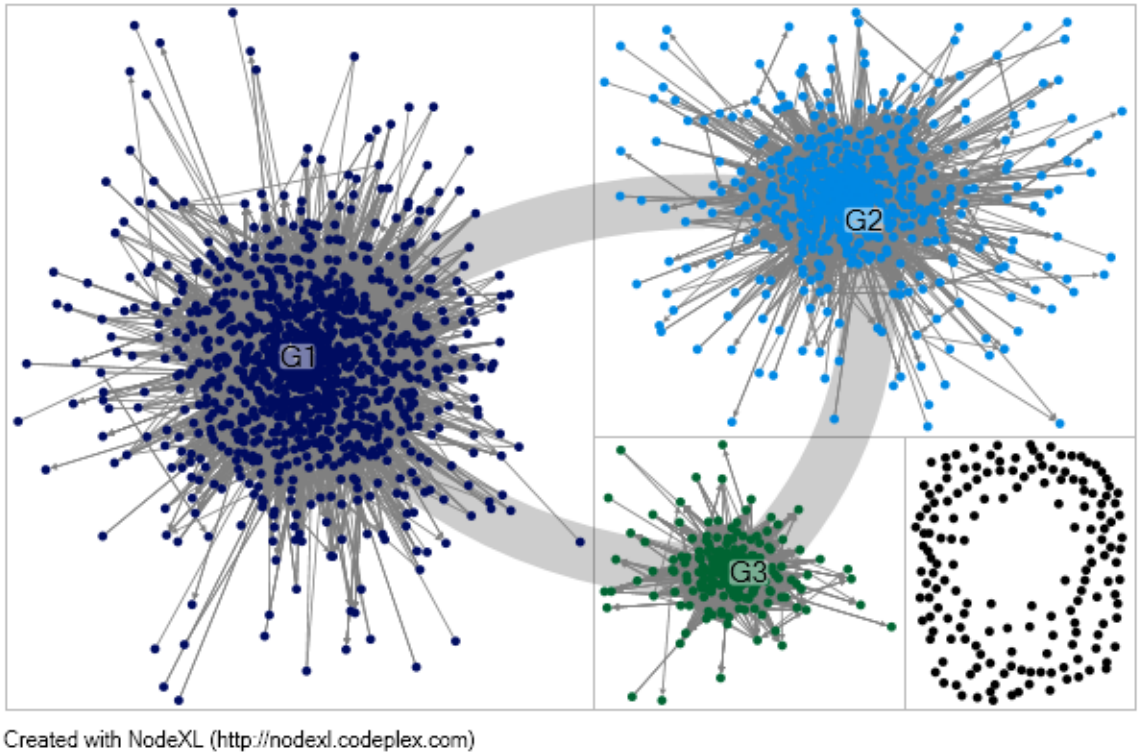


**Figure 4: Network of communities tweeting the week before Egypt's election**

**France**

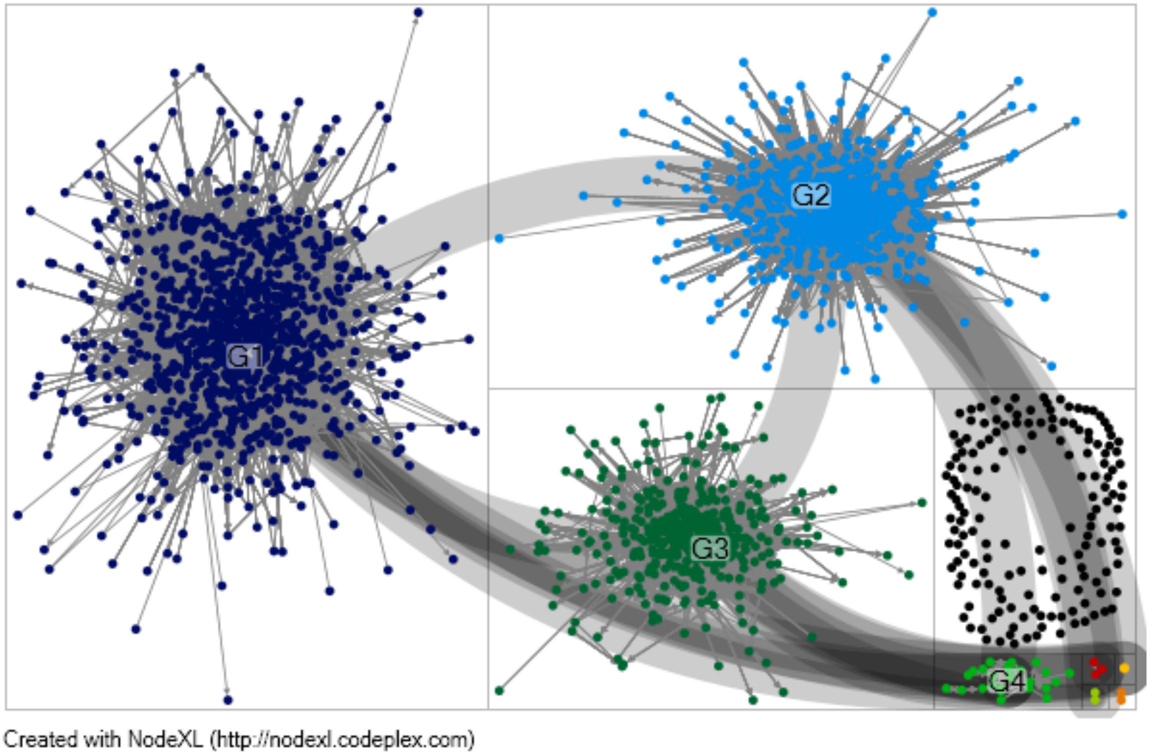


Figure 5: Network of communities tweeting the week before France's election

France's network consists of four main communities. The largest community, "G1" consists of Spanish media outlets reporting on the French election. Tweets from "G1" were primarily in Spanish. The second largest community, "G2" consists of French news agencies and journalists. "G3" consists of personal accounts tweeting from France. The tweets in this group were primarily in French. The majority of accounts in "G4" were personal accounts or journalist accounts. These individuals self-reported that they were located in France and the majority of their tweets were in French.

**Iran**

Iran’s network of tweets yielded nine groups from the Clauset Newman Moore algorithm (including a group of tweeters that were not connected to others in the network). This group, placed on the upper left hand corner in the visualization is the largest group in the community. “G1”, the second largest group in this network, consists of media outlets and news figures such as: The Guardian, Ann Curry, and Anderson Cooper. For this study, these tweets were not significant because they did not represent the sentiment of Iranian citizens. One community, “G4” stood out as representing tweets from individuals in Iran. The profiles belonging to members of “G4” claimed that they live within Iran and tweeted primarily in Farsi. One individual within this group described himself as a “cyber citizen”. A sentiment score was calculated for this group specifically towards both the candidates, Ahmadinejad and Mousavi. Because a majority of the tweets from the Iran election did not originate in Iran, a new sentiment score was calculated only based on the tweets in “G4”. A score of -0.16 was calculated towards Ahmadinejad while a score of 0.24 was calculated towards Mousavi. These sentiment scores more accurately reflect the sentiment of Iranian tweeters than the sentiment represented in Table 3, because the sentiment score above took into consideration tweets from all of the other communities in the network, which do not originate in Iran.

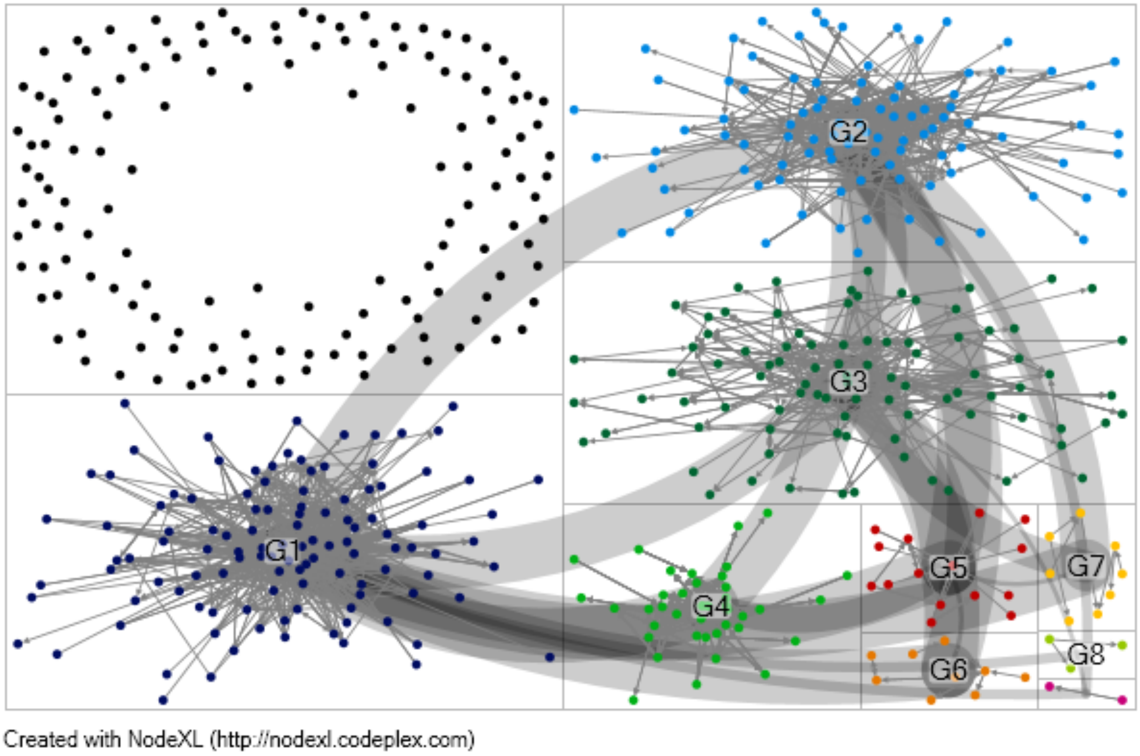


Figure 6: Network of communities tweeting the week before Iran's election

### Mexico

Three main groups emerged following clustering of the Mexican election tweets. The first group primarily consisted of news agencies, organizations, and journalists in Mexico. These tweets were primarily in Spanish. "G2" consisted of personal accounts and journalist accounts. These tweets were primarily in Spanish. "G3" consisted of personal accounts. "G4" consisted of personal accounts belonging to young adults and teenagers in Mexico whose handle names referenced popular music.

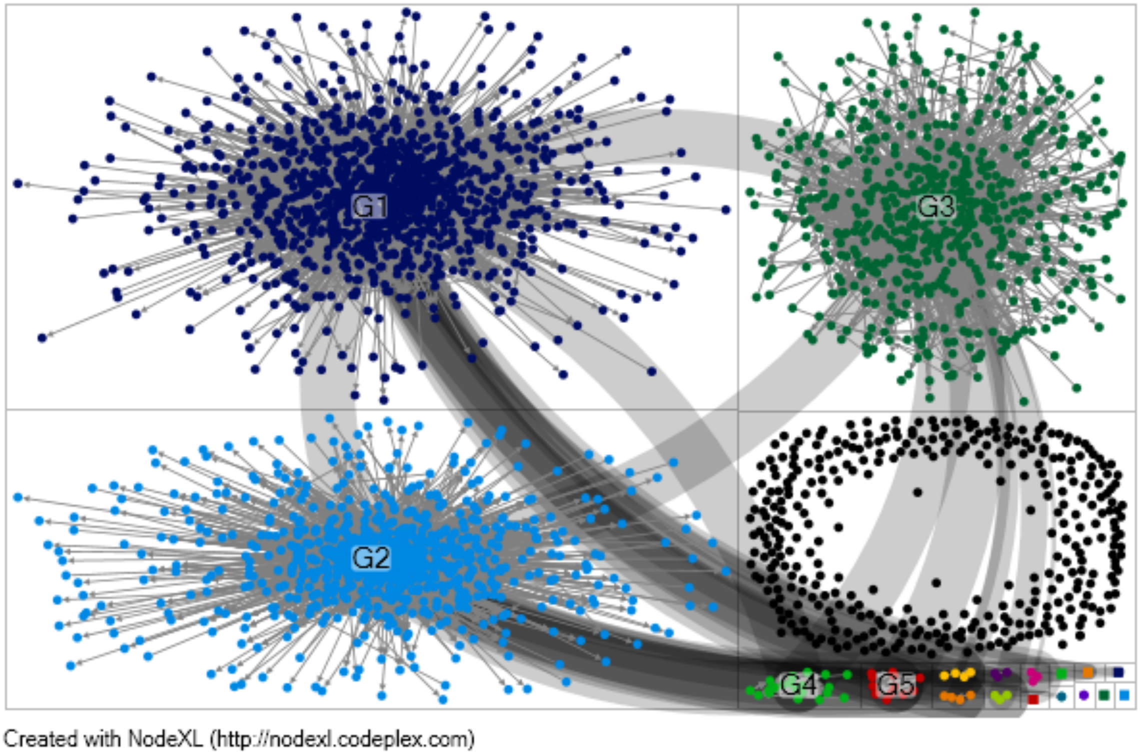


**Figure 7: Network of communities tweeting the week before Mexico's election**

**Russia**

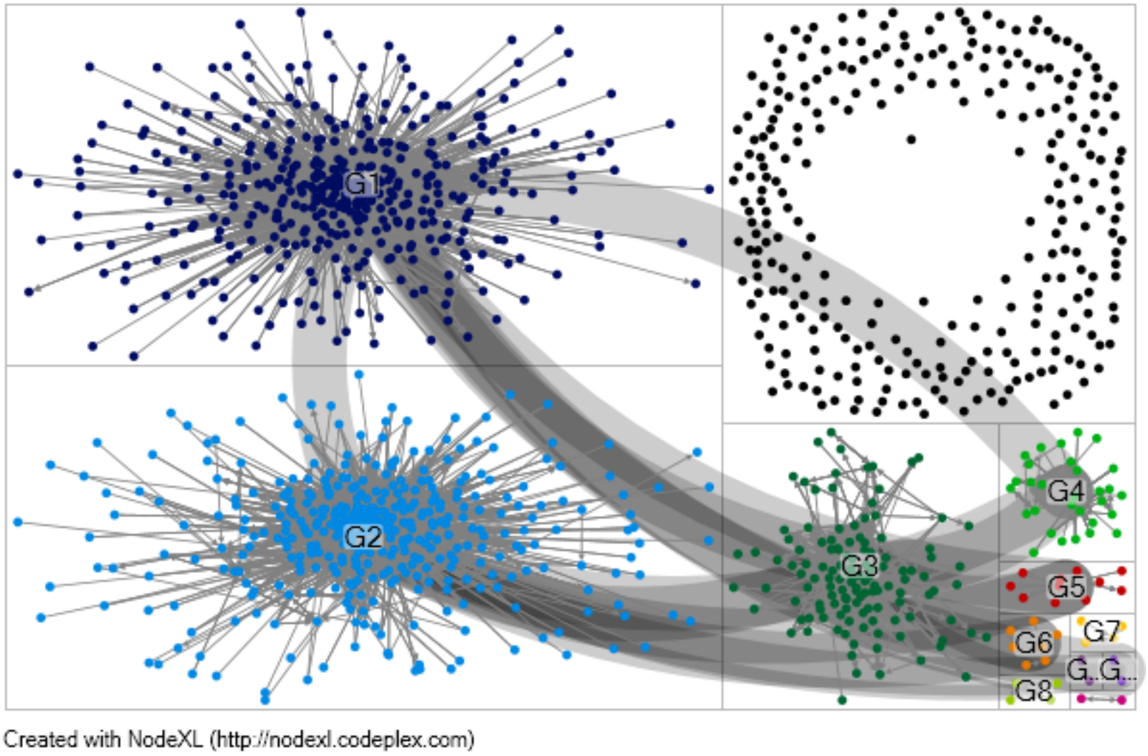


Figure 8: Network of communities tweeting the week before Russia's election

Russia's sample network was clustered into seven classifiable communities. The largest group, "G1" consisted of Russian news agencies and popular individuals. The tweets in "G1" were primarily in Russian. The second group, "G2", consisted of Russian news agencies like "Moscow Times" that tweeted primarily in English. The "G3" group tweeted both in English and Russian. This group's profiles belong to individuals who self-report their locations as within Russia. A fourth group, "G4", primarily consisted of individuals self reporting that they are tweeting from within Russia. The majority of tweets in "G4" were in Russian. "G5" contains tweeters that primarily tweet about exit poll information. These accounts are personal and tweets are primarily in Russian. The "G6" group is a very small cluster of tweets in English.

### The United States of America

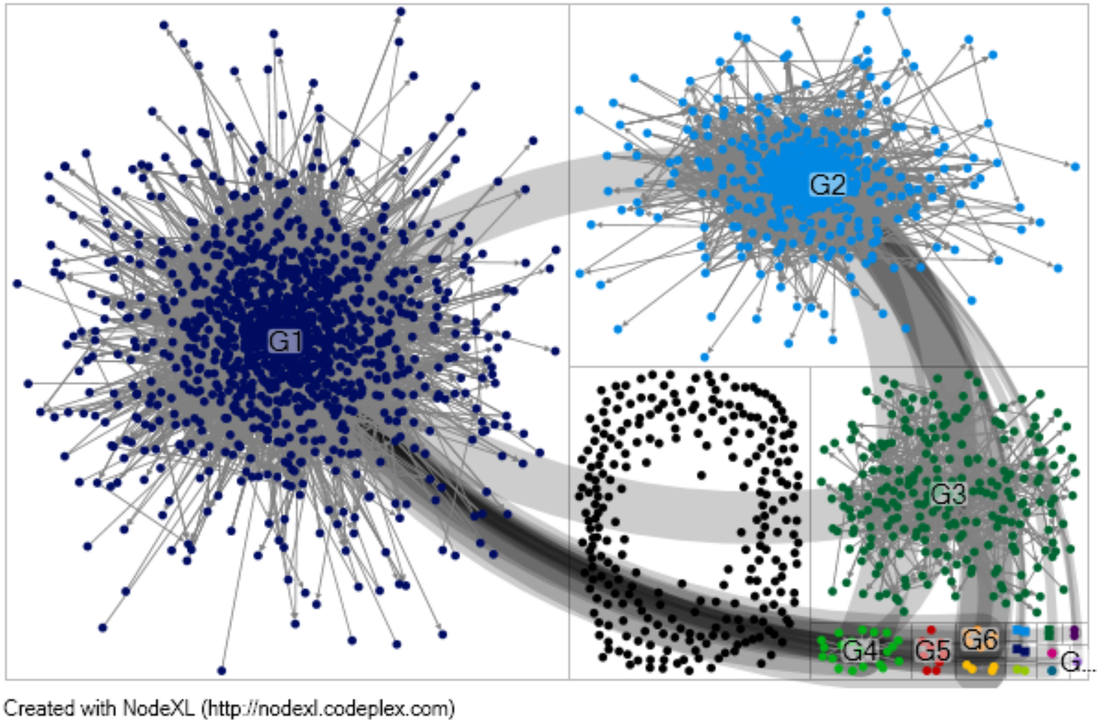


Figure 9: Network of communities tweeting the week before the United States' election

The United States network in this study can be divided into four main groups. The largest group, "G1" consists of verified individuals, journalists, and media outlets. Individuals in this group include: Barack Obama, The While House, ABC World News, The New Yorker, ABC, USA Today, Huffington Post, and BBC World. The second group, "G2" consists of more conservative individuals such as Governor Mike Huckabee, Glenn Beck, and Fox News.  The "G3" group consists of primarily young tweeters who tweeted in support of Obama or against Romney. The fourth community of tweets primarily consisted of individuals tweeting in support of presidential candidate Gary Johnson.

**Venezuela**

Five main communities emerged from clustering the Venezuela network tweets. The majority of tweets in groups “G1”, “G2”, “G3” and “G4” are in Spanish. Tweets in the group “G5” are in English. The groups “G1” and “G2” both consisted of journalists, individuals, and news media outlets that were located within Venezuela. The “G3” group consists of individuals and news agencies located in other Latin American Countries as well as other Spanish news agencies. The group “G4” consists of entertainers and younger individuals from Venezuela.

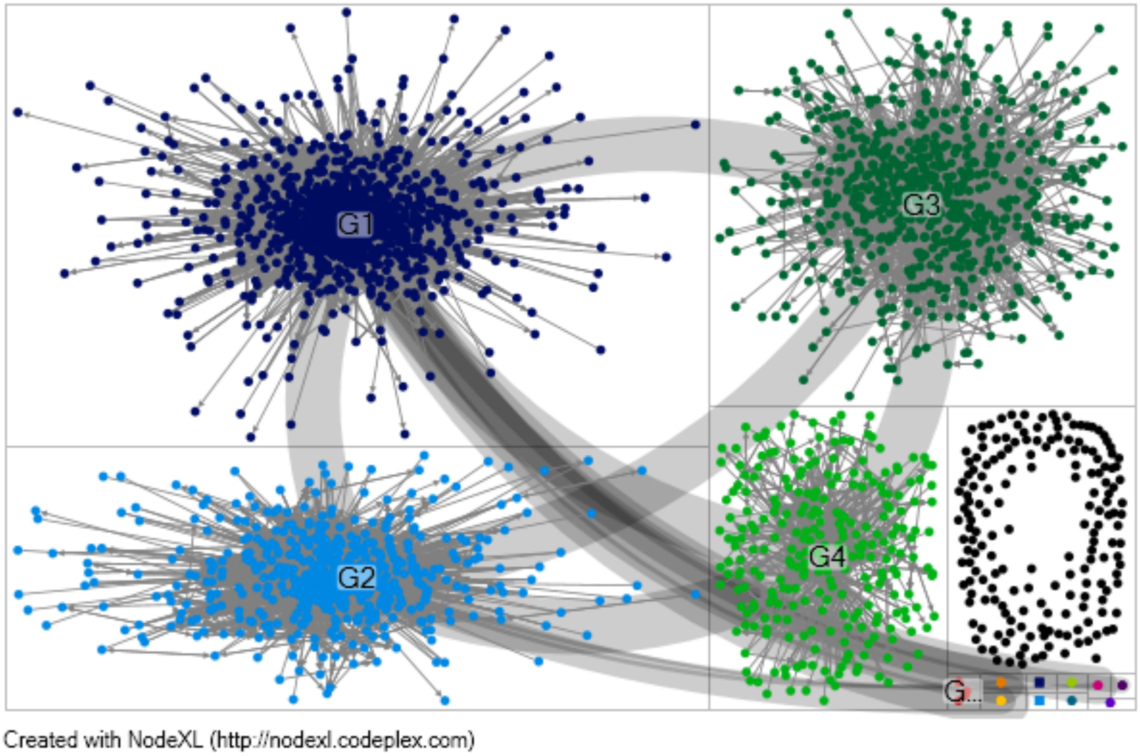


Figure 10: Network of communities tweeting the week before Venezuela’s election

*Network Characteristics*

In this section I present the network characteristics of the nine countries in this study. For each network, I calculated the average degree, network diameter, shortest number of paths, and number of communities. These calculations aid in understanding the characteristics of each of the networks during the election cycles in the nine countries in this study.  I clustered each network using the Markov Clustering algorithm (MCL), which detects communities within a network based on simulation of stochastic flow (Markov, 2009). Table 4 shows the features of the networks including: average degree in the network, average weighted degree, network diameter, average path length, number of shortest paths, density of the network, modularity, number of communities (as detected by the MCL algorithm), number of weakly connected components, number of strongly connected components, and average clustering coefficient.

### Number of Communities

The number of communities within each network varied. The country with the most amounts of communities, Brazil, is classified as a "free" country, and the country with the least amount of communities detected, Iran, is a not-free country. However, it is important to note that there is no clear correlation between the numbers of communities detected in the sample networks and freedom status of a country.

### Modularity

A network with a high modularity indicates that the connections between Twitter users in a particular community are dense, but communities are not connected to one another. The three countries in this study with the highest modularity are the USA, Colombia, and Iran, which each hold different freedom statuses. The three countries in our study with the lowest modularity are France, Russia, and Egypt, which also hold different freedom statuses. There is no relationship between the modularity of a network and a country's freedom status.

### Connected Components

I detected no relationship between connected components (either weakly or strongly connected) and freedom statuses of countries. Brazil and Mexico, a “free” country and “partly free” country respectively, had the most strongly connected components, while Iran and France, a “not free” country and a “free” country respectively, had the least amount of strongly connected components.

Similarly, there was no relationship detected between weakly connected components and freedom status of countries. Iran, Egypt, and France were countries with the least weakly connected components. While Iran and Egypt are both countries classified as “not free”, France is classified as a “free” country. Brazil, Mexico and Russia, which have freedom statuses of Free, “partly free”, and “not free” respectively, have the most amount of weakly connected components.

| | **France** | **USA** | **Brazil** | **Colombia** | **Mexico** | **Venezuela** | **Russia** | **Egypt** | **Iran** |
|---|---|---|---|---|---|---|---|---|---|
| **Average Degree** | 16.39 | 36.739 | 17.805 | 19.491 | 17.532 | 31.742 | 9.278 | 37.051 | 4.315 |
| **Average Weighted Degree** | 144.346 | 47.864 | 21.438 | 36.748 | 23.698 | 68.012 | 33.75 | 111.891 | 9.217 |
| **Network Diameter** | 10 | 11 | 13 | 13 | 10 | 10 | 9 | 8 | 10 |
| **Average Path Length** | 3.404035806 | 3.346320186 | 3.435923973 | 3.82412542 | 3.192610821 | 3.320053737 | 3.409486177 | 2.941047672 | 3.901283649 |
| **Number of Shortest Paths** | 3921645 | 3346718 | 4884393 | 2658275 | 3779554 | 5301416 | 482871 | 2288578 | 71437 |
| **Density** | 0.011 | 0.015 | 0.006 | 0.009 | 0.006 | 0.011 | 0.007 | 0.018 | 0.008 |
| **Modularity** | 0.302 | 0.435 | 0.382 | 0.539 | 0.292 | 0.391 | 0.274 | 0.255 | 0.472 |
| **Number of Communities:** | 165 | 265 | 442 | 245 | 365 | 191 | 290 | 168 | 150 |
| **Number of Weakly Connected Components:** | 161 | 260 | 436 | 240 | 361 | 185 | 282 | 164 | 143 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Number of Strongly Connected Components:** | 637 | 994 | 1405 | 703 | 1349 | 837 | 783 | 851 | 332 |
| **Average Clustering Coefficient:** | 0.193 | 0.154 | 0.168 | 0.19 | 0.136 | 0.21 | 0.176 | 0.247 | 0.121 |
| **Freedom Status** | Free | Free | Free | Partly Free | Partly Free | Partly Free | Not Free | Not Free | Not Free |

**Table 4: Network characteristics of the nine countries in this study.**

## *Twitter Mentions, Links, and Hashtags*

In order to further understand the type of information tweeters were sharing, I extracted several key pieces of information from within tweets. I counted the number of times a candidate was mentioned by name (without the formal ampersand symbol "@" allowed by Twitter). Furthermore, I extracted the most frequent links, hashtags and "@" mentions by Twitter users. Hashtags reveal the topics discussed during the elections and mentions reveal to whom tweets are being addressed. Links also reveal the kind of topics discussed.

### Mentions of Candidate Names

Tweets collected from Topsy were queried for all variations of candidate last names, including multiple ways of spelling as well as spelling with non-roman alphabet letters. Searching for all variations of a candidate name was particularly relevant for querying mentions of candidates in Egypt, Iran and Russia.

Table 5 below shows the frequency of mentions for each candidate from the tweets that were collected. In all of the countries in this study, regardless of freedom status, the most frequently mentioned candidates won the election. The candidate names marked with an asterisk (*) below indicate that there are multiple spelling variations and language variations for this candidate name.

| Candidate Name(s) Queried | Total mentions | Country |
|---|---|---|
| **Dilma Rousseff** | **4,006** | **Brazil** |
| Josè Serra | 53,177 | Brazil |
| Antanas Mockus | 2,167 | Colombia |
| **Juan Manuel Santos** | **3,535** | **Colombia** |
| Ahmed Shafik* | 17,028 | Egypt |
| **Muhammad Morsi*** | **17,249** | **Egypt** |
| Nicolas Sarkozy | 26,575 | France |
| **Françoise Hollande** | **27,915** | **France** |
| Mohsen Rezaee* | 5 | Iran |
| Mehdi Karroubi* | 47 | Iran |
| Mir Hussein Mousavi* | 222 | Iran |
| **Mahmoud Ahmadinejad*** | **600** | **Iran** |
| Josefina Vazquez Mota | 7,013 | Mexico |
| **Enrique Peña Nieto** | **13,203** | **Mexico** |
| Andrés Manuel López Obrador | 12,153 | Mexico |
| Gabrielle Quadri de la Torre | 8,483 | Mexico |
| Gennady Zyuganov* | 3,833 | Russia |
| **Vladimir Putin*** | 10,923 | Russia |
| Vladimir Zhirinovsky* | 4,066 | Russia |
| Mikhail Prokhorov* | 5,900 | Russia |
| Sergay Mironov* | 3,001 | Russia |
| Gary Johnson | 2,054 | USA |
| Virgil Goode | 392 | USA |
| Jill Stein | 2,842 | USA |
| Mitt Romney | 39,759 | USA |
| **Barack Obama** | **42,444** | **USA** |
| Henrique Capriles Radonski | 20,450 | Venezuela |
| **Hugo Chávez*** | **33,306** | **Venezuela** |

Table 5: Number of mentions for each candidate.

Twitter's 140 character microblogging statuses allow for tagging other users using a "@" and using hashtags "#" to organize the topic of a particular tweet. For all of the tweets collected during the aforementioned election cycles, I searched for the most frequently mentioned profiles in tweets (using the "@" sign), the most frequent hashtags (marked with a "#" sign), and the most frequently shared links. Above, I reveal a table of results for each of the countries in this study. The analysis and implications of the results yielded from this study are discussed in Chapter 5.

### Brazil

Neither candidate Jose Serra nor Dilma Rousseff appears in the top mentions for Brazil. However, hashtags reflect support for the candidates. The top three hashtags “#voude13”, “13neles” and “#soumaisdilma” are pro-Dilma hashtags, the first two relating to the Worker’s party, the party to which Dilma Roussef belongs. Hashtags supporting Jose Serra also appear in the top ten most frequently tweeted hashtags: “#serra45” and “#serra”. The top tweeted link is a live blogging site for Dilma and Serra’s debate. Another link is for live tweeting about the candidates to TV personality, Bemvindo Sequeira. The significant role Twitter plays in Brazilian politics is reflected by the fact that there are multiple links encouraging live-tweeting in the top ten most shared links. These two links also reflect the degree of accessibility that Brazilians have to the Internet – so much so that they are able to live tweet about events as they happen.  This accessibility is a reflection of Brazil’s freedom status.

| Rank | Link | Mentions | Hashtags |
|---|---|---|---|
| 1 | http://migre.me/1RABx | @el_pais | #voude13 |
| 2 | http://bit.ly/br45il | @ptnacional | #13neles |
| 3 | http://twitcam.com/2jrt2 | @ConversaAfiada | #soumaisdilma |
| 4 | http://twitcast.me/_PAvg | @sensacionalista | #virada45 |
| 5 | http://bit.ly/bRim3R | @Le_Figaro | #euquero45 |
| 6 | http://t.co/y2UCgcX | @g1eleicoes | #serra45 |
| 7 | http://bit.ly/b5bdqE i | @KeshaSuja | #serra |
| 8 | http://pud.im/eop | @exilado | #vaidarvirada |
| 9 | http://twb.ly/aFncAT | @BlogdoNoblat | #DebateGlobo |
| 10 | http://bit.ly/jsclipe | @luisnassif | #Br45il |

Table 6: Top ten mentions, links and hashtags for Brazil's election.

### Colombia

While Juan Manuel Santos won the Colombian election, his name appears below Antanas Mockus' name in the most frequently tagged profiles. Additionally, Santos' name appears in top hashtags only below Mockus' name. In the previous section I observed that Juan Manuel Santos was mentioned more than Antanas Mockus (without the "@" symbol.). The content shared on the top links varies. The content includes links to articles by various media outlets, a link to a Twitter picture, as well as a link to a Blog expressing joy about Santos' victory over Antanas.

| Rank | Link | Mentions | Hashtags |
|---|---|---|---|
| 1 | http://bit.ly/a6euzV | @globovision | #Mockus |
| 2 | http://bit.ly/9tTfcx | @partidoverdecol | #olaverde |
| 3 | http://www.globovision.com/news.php?nid=152564 | @DanielSamperO | #elecciones2010 |
| 4 | http://bit.ly/9oysf6 | @vladdo | #Santos |
| 5 | http://twitpic.com/1ykocl | @semanadigital | #JMSantos |
| 6 | http://www.globovision.com/news.php?nid=148743 | @AntanasMockus | #alianzaciudadana |
| 7 | http://url.ie/6khy | @JornalOGlobo | #SoySemana |
| 8 | http://bit.ly/aWXpCb | @JuanManSantos | #elecciones2010rcn |
| 9 | http://tinyurl.com/2u8rtwk | @ElUniversal | #votebien |
| 10 | http://tinyurl.com/35ddpn2 | @caracolradio | #WorldCup |

Table 7: Top ten links, mentions, and hashtags for Columbia's election.

### Egypt

The top mentions in Egypt's election cycle reveals that Youtube was a primary source of information during the elections. Among the "free" and "partly free" countries, all of the candidates are tagged directly in the top ten mentions. However, neither Mohammad Morsi nor Ahmed Shafik appear in the top ten tagged profiles, though "ikhwanweb" the official Twitter account for the Muslim Brotherhood, the political party to which Mohammad Morsi belongs to, is in the list of most frequently tagged profiles. The hashtag “#Shafik” is used more often than the hashtag for “#Morsi”.

Four out of the five of the Twitter pictures shared in the top ten links are no longer accessible. Content that is no longer accessible represents a suspicious trend, one that is observed in the top links for Iran. The implications of this trend are further discussed in Chapter 5. Other top links include articles about the candidates from American news agencies and Arab news agencies.

| Rank | Link | Mentions | Hashtags |
|---|---|---|---|
| 1 | http://t.co/ZgCvJaaK | @YouTube | #Egypt |
| 2 | http://t.co/KRoEPQgW | @ahramonline | #Shafik |
| 3 | http://t.co/2wMc1L9K | @guardian | #Morsi |
| 4 | http://t.co/IguCjaTs | @Shorouk_News | #Shafiq |
| 5 | http://t.co/JuGvhuK1 | @egyindependent | #EgyElections |
| 6 | http://t.co/h0a7loO9 | @AlMasryAlYoum_A | #tahrir |
| 7 | http://t.co/KSTzfOMK | @M_ibr | #EgyPresElex |
| 8 | http://t.co/Pa7FIFq1 | @shadihamid | #SCAF |
| 9 | http://t.co/hALKAIg6 | @ikhwanweb | #jan25 |
| 10 | http://t.co/bDLi3soa | @FRANCE24 | #ikhwan |

Table 8: Top ten links, mentions, and hashtags for Egypt's election.

**France**

For France’s top mentions, I observed that Francoise Hollande, the winner of the election was tagged more than the incumbent, Nicolas Sarkozy, who lost the election.  However, Sarkozy’s name appears in the top hashtags above Hollande’s name. Top links shared in France consist of news articles from French, US and Spanish media outlets as well as a twitter picture meme circulating regarding the incumbent Nicolas Sarkozy.

| Rank | Links | Mention | Hashtag |
|---|---|---|---|

| Rank | Links | Mention | Hashtag |
|---|---|---|---|
| 1 | http://t.co/XEe7C439 | @YouTube | #Sarkozy |
| 2 | http://t.co/sUqWLBtA | @el_pais | #Hollande |
| 3 | http://t.co/7PWL2eYD | @lemondefr | #LePen |
| 4 | http://t.co/fFk38k6m | @fhollande | #France2012 |
| 5 | http://t.co/L7u8rxoO | @lesoir | #Mélenchon |
| 6 | http://t.co/nFrHzvCI | @NicolasSarkozy | #elysee2012 |
| 7 | http://t.co/aeVM5HLc | @guardian | #Francia |
| 8 | http://t.co/FCEaencW | @LeHuffPost | #AvecHollande |
| 9 | http://t.co/clsR9rla | @Reuters | #Résultats |
| 10 | http://t.co/AjFVkOcA | @2012resultats | #FH2012 |

Table 9: Top ten links, mentions and hashtags for France's election.

### Iran

Out of the top ten most frequently tweeted links in the Iranian election cycle, three of them had been removed. The top ten mentions reveal that there is indeed twitter activity by one of the Iranian candidates or supporters of one of the Iranian candidates. The account "@mousavi1388" does not claim to represent Mir Hussein Mousavi. The bio reads, "MirHossein Mousavi is standing for election in the upcoming Iranian presidential election 2009. With Khatami, Vote Mousavi." The last tweet from this profile was on February 20, 2011. While the candidates in this election did not have official Twitter pages, the top two names of candidates appear in the top ten hashtags (#Ahmadinejad and #Mousavi).

| Rank | Link | Mention | Hashtags |
|---|---|---|---|
| 1 | http://tinyurl.com/mko2q4 | @add this | #Iran |
| 2 | http://www.alisanaei.com | @can | #fib |
| 3 | http://friendfeed.com/vahid9 | @RIA_Novosti | #cnn |
| 4 | http://iranvote.wordpress.com | @Drudge_Report | #tcot |
| 5 | http://bit.ly/h9l3s | @MelissaTweets | #Ahmadinejad |
| 6 | http://www.rfi.fr | @maddow | #iranelection |
| 7 | http://tinyurl.com/n56yeh | @TIME | #elections |
| 8 | http://tinyurl.com/oe8egw | @bbcworld | #election |
| 9 | http://bit.ly/wao6z | @andersoncooper | #US |
| 10 | http://tr.im/oj9J | @mousavi1388 | #Mousavi |

Table 10: Top ten links, mentions and hashtags in Iran's election.

**Mexico**

The winner of the election, Enrique Peña Nieto (@EPN) is tagged most after Youtube. However the top two hashtags support Andres Manel Lopez Obrador (#panistasconAMLO and #HoyVotoPorAMLO). The third hashtag, "#yosoy132", is part of a Mexican protest movement pushing for democratization against the winner of the election, Enrique Peña Nieto. Not only are the top tweeted hashtags supporting Obrador, the third most tweeted hashtag is a protest against the incumbent, Nieto. Even though Nieto's sentiment score was less than all other competing candidates, he still succeeded in winning the election.  The top most shared link is a Facebook note regarding the flaws of the Mexican presidential system and the different political parties involved. This second most shared link is an article, titled, "25 Reasons Why to Vote for Josefina Vazquez". The top links and hashtags are both anti-Nieto and despite this, Nieto still succeeded in winning the election.

| Rank | Link | Mentions | Hashtag |
|---|---|---|---|
| 1 | http://t.co/bFJdBIHk | @YouTube | #PanistasConAMLO |
| 2 | http://t.co/G1mAhKNb | @EPN | #HoyVotoPorAMLO |
| 3 | http://t.co/55SBbXxL | @lopezobrador_ | #Yosoy132 |
| 4 | http://t.co/hDMqSMqm | @sharethis | #ConfíoEnAMLO |
| 5 | http://t.co/GBpKaKSw | @sdpnoticias | #MiVoto2012 |
| 6 | http://t.co/Sjcm2dOm | @AMLO_si | #LoLograsteJosefina |
| 7 | http://t.co/ydztgIin | @aristeguionline | #AMLOGanaráPorque |
| 8 | http://t.co/AvS7eCwV | @LAURAZAPATAM | #TodoMexicoEnElZocaloConAMLO |
| 9 | http://t.co/IAD6rNJc | @JosefinaVM | #MañanaVotoPorElla |
| 10 | http://t.co/gO2mPm4C | @bernimarin | #PreguntasExistenciales |

Table 11: Top ten links, mentions, and hashtags in Mexico's election.

**Russia**

Among the most frequently tweeted links in Russia’s election, the content on the top two tweeted links listed below are no longer accessible. This is discussed in further detail in the Chapter 5. Furthermore, Youtube is a primary source of information appearing to be the top mentioned profile. The top hashtags include tweets in both Russian and English terms, the most frequently used hashtag being, “#4марта” or “March 4”, the date of the Russian election. The top mentions include media outlets and journalists.

| Rank | Links | Mention | Hashtags |
|---|---|---|---|
| 1 | http://t.co/J3oq6hkD | @Youtube | #4марта |
| 2 | http://t.co/VNkYRNyl | @rianru | #Russia |
| 3 | http://t.co/LXrmx8Mp | @navalny | #Putin |
| 4 | http://t.co/1FIksyy5 | @varlamov | #novosti |
| 5 | http://t.co/V8TFhZRx | @YouTube | #RT |
| 6 | http://t.co/G9vWayOo | @VRSoloviev | #twisident |
| 7 | http://ads.adfox.ru/173362/goDefaultLink?p1=beeky&p2=emux | @mdp2012 | #anekdot |
| 8 | http://t.co/jHf4rmMh | @naumovnk | #FreelandFile |
| 9 | http://t.co/Mw0Jnpdx | @Dobrokhotov | #Prokhorov |
| 10 | http://t.co/5Tt64Mmj | @Reuters | #выборы |

Table 12: Top ten links, mentions and hashtags for Russia's election.

**The United States**

By exploring the top mentions that emerged during the election cycle in the United States, it becomes apparent that many users share content about the election using Youtube, the popular video sharing website. Furthermore, I observed that Barack Obama was tagged or mentioned more than twice as much as Mitt Romney was tagged. Top links included articles from The New York Times, Politico, Think Progress, and an Internet meme.

| Rank | Link | Mention | Hashtag |
|---|---|---|---|
| 1 | http://t.co/AVlagOjq | @Youtube | #tcot |
| 2 | http://t.co/5xuHlwy4 | @BarackObama | #P2 |
| 3 | http://t.co/5yAMcolL | @mittromney | #OBAMA |
| 4 | http://t.co/A5bH8d8V | @cspanwj | #Romney |
| 5 | http://t.co/YMjYl84m | @FiveThirtyEight | #teaparty |
| 6 | http://t.co/IkPBNvDy | @SpikeLee | #sandy |
| 7 | http://t.co/Hw8LVFVB | @Europe1 | #GOP |
| 8 | http://t.co/SfJlc1kw | @GovGaryJohnson | #tlot |
| 9 | http://t.co/tvrI58B4 | @AP | #Benghazi |
| 10 | http://t.co/cCHf29x3 | @JillStein2012 | #election2012 |

Table 13: Top ten links, mentions and hashtags in United States' election.

**Venezuela**

In Venezuela's top hashtags, "#Capriles" appears at the top of the most frequently used hashtags. Following "#Capriles", "#HoyGanaChavez", a pro-Chavez hashtag meaning "Chavez wins today" is the top hashtag. The most frequently shared link is an article reporting pictures from a pro-Chavez rally and is titled, "Vea Las Fotos De Las 7 Avenidas que Dejaron abierta a Capriles" which translates to, "See Pictures of the 7 Avenues that left Capriles' mouth open", referring to the magnitude of pro-Chavez protestors. The majority of the other top links shared are news articles from news agencies like the Wall Street Journal, Globovision, and CNN Español. One link frequently shared is a campaign twitter picture encouraging a global twitter "tuitazo" for Hugo Chavez. This particular link reveals that Venezuela's government and Hugo Chavez's campaign understand the potential influence Social Media and Twitter have on influencing public opinion.

| Rank | Link | Mentions | Hashtags |
|---|---|---|---|
| 1 | http://t.co/9UDtYCjj | @chavezcandanga | #Venezuela |
| 2 | http://t.co/9UDtYCjj | @hcapriles | #Capriles |
| 3 | http://t.co/2FnaipOV | @FOROCANDANGA | #HoyGanaChávez |

| Rank | Link | Mentions | Hashtags |
|---|---|---|---|
| 4 | http://t.co/ZEqdO0uw | @abc_es | #HayUnCamino |
| 5 | http://t.co/vbsPqzd4 | @TRIBUNA_PCV | #eleccionesVenezuela |
| 6 | http://t.co/zHBy9Ucw | @Jan_Herzog | #7O |
| 7 | http://t.co/gVvBBhG7 | @noticiaaldia | #TuVoto |
| 8 | http://t.co/7BQnFaxt | @LucioQuincioC | #Elecciones2012 |
| 9 | http://t.co/kqsQC2cq | @tongorocho | #Venezueladecide |
| 10 | http://t.co/a7VBJSJJ | @danielscioli | #SeVeSeSabeMañanaSeVaChávez |

Table 14: Top ten links, mentions and hashtags for Venezuela's election.

# Chapter 5: Discussion

In this study, I conducted an exploratory study of the relationship between a country's freedom and the Twitter activity during elections. I studied tweets occurring during elections in nine different countries: Brazil, Colombia, Egypt, France, Iran, Mexico, Russia, The United States, and Venezuela. I explored this space with automated analysis of the tweets' text, election outcomes, freedom ratings for the countries, and sentiment analysis. I found promising initial results regarding the relationship between the removal of links shared on Twitter during elections and the freedom status of a country, the relationship between the number of disconnected profiles in a network and the freedom status of a country, and the relationship between number of mentions of a candidate and election outcome. In this section, I have identified hypotheses for future work and research designs based on the initial results from this study. For each of the hypotheses presented, I establish a research design, method of data collections and method of selection of subjects. Furthermore, I present areas of research where there seems to be little indication of relationships.

## *Sentiment Analysis and Election Prediction*

The sentiment analysis conducted in this study reveals that sentiment does not reflect the outcome of the election. If election outcome were based purely on the sentiment score derived in this study, only Colombia and Venezuela would yield correct election outcomes. The majority of the sentiment scores in this study do not accurately correlate with the outcome of the respective elections. In the beginning of the paper, I discussed the concerns raised by Gayo-Avella regarding predicting election outcome with tweets (2012). He articulates that there is no commonly accepted way of counting votes on Twitter simply because not all tweets are trustworthy and Twitter is not representative of the entire demographic population. This trend was observed within the countries in this study. The tweets I collected from Brazil, Colombia, Egypt, France, Iran, Mexico, Russia, The United States, and Venezuela are simply not representative of the entire demographic of any of the nine countries with which we conducted sentiment analysis. Also, politically active individuals tend to tweet more, so self-selection bias is ignored. In this study, the top ten tweeters for each of the countries contributed to a significant portion of the total tweets for each of the elections. For example, tweets from the top ten tweeters in Colombia made 15.9% of the total tweets. The top ten tweeters in each of the countries in this study respectively make up less than 1% of the total profiles for each of these countries and yet they contribute significantly to the overall sentiment score. Thus the sentiment score is biased towards those who tweet more often. Election prediction using sentiment analysis prediction is not feasible, regardless of the freedom status of a country.

| **Country** | **Percentage of Total Topsy Tweets that are produced by the top 10 Tweeters** |
|---|---|
| USA | 4.08% |
| Brazil | 4.81% |
| France | 6.15% |
| Colombia | 15.9% |
| Venezuela | 6.21% |
| Mexico | 3.75% |
| Iran | 13.2% |
| Egypt | 10.19% |
| Russia | 5.5% |

**Table 15: Percentage of total Topsy tweets for each country that are produced by the top ten tweeters.**

### *Sentiment and Freedom*

I discovered that sentiment scores of candidates from almost all of the countries in this study were higher for the candidate that lost the election. Venezuela and Columbia were the only two exceptions to this pattern. I explored whether freedom and sentiment were related and aimed to see whether the freedom status of a country was related to the sentiment expressed towards the incumbent or towards a particular candidate. Only one of the three "not free" countries, Iran, had an incumbent for the election cycle in which we were studying. While Vladimir Putin was not an incumbent, he has previously held the presidential position in Russia and thus enjoys the same publicity and name recognition an incumbent would. Both the sentiment scores expressed towards Putin and Ahmadinejad were less than that of their competing candidates. However, this phenomenon cannot be conclusively attributed to Iran and Russia's "not free" freedom status. While the incumbents or candidates/previously serving as presidents studied in the "not free" countries all had lower sentiment scores than the candidates with which they were competing, the results were mixed for "free" and "partly free" countries. The sentiment scores for incumbents in Mexico and Venezuela are higher than the candidates with which they were competing. The sentiment scores among the "free" countries demonstrated mixed results in regards to the relationship between incumbency and sentiment score. In the United States, Barack Obama's sentiment score was less than that of Mitt Romney. In France, the incumbent, Nicolas Sarkozy, received a higher sentiment score than Francoise Hollande's sentiment score, even though Nicolas Sarkozy lost the election. While tweets from both Russia and Iran reflect a negative sentiment towards the incumbent, the varying results of the "free" countries, especially the United States sentiment score reflecting a negative sentiment towards Barack Obama show that there is no relation between sentiment score towards an incumbent and the freedom status of a country.

### *Hypotheses*

This research is an exploratory work that revealed several insights that could lead to future research. I found promising initial results with respect to the relationship between content removed from links during an election and freedom status of a country, the number of disconnected twitter profiles in the network structure of "not free" countries, and a strong correlation between the number of times a candidate name is mentioned and the election outcome.  Below, I present each hypothesis in detail along with a research design, providing evidence from my research.

**Hypothesis 1: Links shared on Twitter during elections of "Not Free" countries are more likely to be removed than links shared on Twitter during elections of "Free" countries.**

Because a Twitter status is limited to 140 characters, using a hyperlink in a tweet allows a user to share a large amount of information despite Twitter's constraints. The content shared via hyper links is diverse. In my data set, hyperlinks are used to share blog posts, personal websites, Youtube videos and news articles. I studied the top ten most frequently shared links for each country in this study. Links are often used to share news articles by established media outlets. While the content of these news articles are certainly relevant, the content from user generated activity such as blog posts or a Youtube video uploaded by a user more accurately reflect how individuals feel about candidates. For this reason, the percentages below report the top ten links that did not include news articles by established news outlets. From the top ten most shared links among Iranian tweets (ignoring links from Western media outlets like Sunday Times and The Guardian), 70% of the content on these links had been removed. The content of these links varied. One site, http://alisanaie.com was completely inaccessible. A Youtube video was no longer accessible and had been removed by the user. Several blog posts had been removed and several sites were met with a 404 error.

I also looked at the top ten links shared during the Egyptian election. Ignoring articles from established news outlets, the top ten shared links comprised of only five posts containing user-generated content. The user-generated content in the Egyptian election consisted of Internet memes/photos shared on Twitter. Four out of five of the top pictures shared are no longer accessible. While Egypt's freedom classification has recently been promoted to "partly free", it shared this common characteristic with the other "not free" countries in this study.

In the tweets for the Russian election, the content of the two top links has been removed. These links were collectively shared hundreds of times, but are now no longer accessible. All three "not free" countries have top links that are no longer accessible or have been removed. In contrast, with the exception of live twitter stream links that are time dependent, the majority of top links were still visible and accessible in the "partly free" and "free" countries.

I stipulate that the content from links in “not free” countries are not accessible for two possible reasons. Fear of persecution may have driven a poster to voluntarily remove content posted on the web. A second possibility is coercion or being forced to remove content following persecution. During the Iranian election, the Iranian regime cracked down on Iranian bloggers and Internet activists. Similarly in Egypt, bloggers have been arrested for the content that they post (Booth, 2012). There have been no reports of Russian arrests related to Internet activity. This might be related to the fact that Russia had less links removed than that of its “Not-Free” country counterparts, Iran and Egypt.

This exploratory study provided enough evidence to suggest the following hypothesis: links shared on Twitter during elections of “not free” countries are more likely to be removed than links shared on Twitter during elections of “free” countries. To prove this hypothesis, more tweets from countries with “not free” freedom statuses and presidential elections should be collected, studied and compared to tweets from “free” countries. Below, I outline a research design that would ultimately prove whether the hypothesis above is valid.

### *Data*

The data required to conduct this experiment would involve tweets during elections of “not free” countries. While the Topsy service could be used to collect this data, a more representative distribution of tweets would result from collecting tweets directly from Twitter. However, since these tweets must be collected in real time and thus are difficult or even impossible to access for elections that have already occurred, Topsy is the most convenient method of accessing such tweets.

To conduct this study, tweets need to be collected from countries that previously have had presidential or parliamentary elections. Countries that fit these criteria include: Russia, Iran, China, Chad, Congo, Kazakhstan, Tajikistan, Ukraine, Uzbekistan, Turkmenistan, and Algeria. Similarly, tweets from election cycles of "free" countries should be collected during their respective presidential elections. Examples of "free" countries with adequate Twitter usage include: The United States, France, South Korea, Argentina, Chile, Portugal, The United Kingdom and Germany (Evans, 2010).

*Analysis*

To conduct an analysis of the data described above, the links resulting from tweets would be crawled to see if they have indeed been removed. Removed content presents itself in various ways. A broken or dead link results in the 404 or Not Found error message, which is a standard HTTP response code that indicates the web page is not accessible. However, a removed YouTube video, a removed blog post, or a removed twitter pictures will not yield a 404 error message. The host site will simply notify the user that the content has been removed. All possibilities must be considered when automatically detecting removed content from links. Once all links tweeted in "free" and "not free" countries are crawled, a conclusion can be derived as to whether content more often is removed from tweets following the elections in "not free" countries as compared to "free" countries.

One challenge in such a study is that people in "not free" countries do not have equivalent access to social media like Twitter to people in "free" countries. This must be considered while conducting the study. The number of tweets yielded from an election in the Congo for example would certainly be less than the amount of tweets resulting from the United States election. The samples of tweets collected must account for such inequalities.

*Significance of Study*

This study is important in understanding if and why there is a correlation between a country's freedom status and amount of content removed from links following an election. Fear of persecution or self-censorship may have driven a poster or author to voluntarily remove content posted on the web. An alternative cause for removal of content can be attributed to coercion or being forced to remove content following persecution. The outlined study above can explain why this phenomenon occurs. Upon discovering which links have been removed, the sources and authors of these links can be traced. Interviews can be conducted with the authors of links, posters of videos, or bloggers to learn about the reason of removal and whether the government played a role in the removal of content. This study would aid in understanding the relationship between governments, self-censorship, and tweeters in "not free" countries.

**Hypothesis 2: The Twitter networks of "Not Free" countries have more singletons, or disconnected profiles than "Free" countries.**

Using the Clauset Newman Moore Algorithm to cluster and visualize tweets for all of the countries in this study, I found that two of the "not free" countries (Iran and Russia) had a larger community of disconnected profiles than their "partly free" and "free" counter parts. The singletons in Iran's network were part of the largest group, while the singletons in Russia were part of the second largest group. If there is indeed a correlation between the network structure of tweets during elections and the respective freedom status of a country, then Egypt should also have had a very large group of singletons. However, Egypt presents a unique case in this study. The report in which this study is based on was published in 2012, prior to the Egyptian election, but after the Egyptian revolution. The most recent freedom classifications by the Freedom house classify Egypt as “Partly free”. While Egypt also has a large community of singletons in its sample network, it is not as large as Russia and Iran’s community of singletons. I consider that the events that occurred in 2012 have now altered the resulting freedom status in Egypt. Egypt was promoted to “partly free” in 2013. However, its sample network still has a large community of singletons, but not as large with respect to the singletons in Iran and Russia. Based on these results, I hypothesize that “not free” countries have a large community of disconnected users tweeting about the election. Below, I describe data and analysis required to conduct an experiment to prove this hypothesis.

### *Data*

The data required to conduct this experiment would be much like the data described for Hypothesis 1. This data could be retrieved from either Topsy or Twitter. While Topsy could be used to collect this data, tweets directly from Twitter would result in a more representative data set. However, as described above, these tweets are difficult to access and tweets from Topsy would be sufficient for such a study.

To conduct this study, tweets need to be collected from the election cycles in “not free” countries with presidential or parliamentary elections. In addition to Russia and Iran, countries that fit these criteria include: China, Chad, Congo, Kazakhstan, Tajikistan, Ukraine, Uzbekistan, Turkmenistan, and Algeria. Similarly, tweets from election cycles of “free” countries should be collected during their respective presidential or parliamentary elections. These countries can include the countries mentioned in Hypothesis 1: The United States, France, and Brazil in addition to South Korea, Argentina, Chile, Portugal, The United Kingdom and Germany.

### *Analysis*

In order to create a network of profiles that tweet during the election, the followers and friends of the tweeters must be accessed. While finding the connections between all the tweets collected would result in a more comprehensive network, the Twitter API limits access to Twitter, and given time and API constraints, such a task would not be plausible. As done in this study, random node sampling can be used to conjure a sample network. While other sampling methods like edge sampling, forest fire sampling and snowball sampling might yield a more connected network, the aim of this study is not to study a connected network, but to compare the number of singletons in each network. For this reason, the random node sample would be the best option for this study.  Once the sample network is constructed for all “free” and “not free” countries in the study, the number of singletons, or disconnected profiles can be counted and compared for all “free” and “not free” countries.

### *Significance of Study*

This study outlined by the research design above is important because it attempts to understand if and why there are a higher number of disconnected profiles within networks for “not free” in comparison to networks for “free” countries. Upon finding a correlation, the outlined research would involve studying the disconnected profiles within the network to learn how often they tweet, if indeed they have friends and followers and to whom they are connected. Disconnected tweeters might stay disconnected in “not free” networks in order to stay anonymous to protect their safety. For example, some of the disconnected profiles in the Iran and Russia sample network did not have any tweets, despite tweeting in the past about the election. At some point, these Twitter users deleted their tweets. Such a study would prompt one to question why these tweeters opt to stay disconnected and anonymous and why have their tweets have been deleted. The number of friends and followers can be compared to the number of friends and followers of other tweeters in the network. Such a study could potentially lead to the detection of spammers who have a large amount of tweets but very few followers. The research outlined above could answer these questions and aid in understanding the networks of tweets in “not free” countries.

**Hypothesis 3: Number of mentions of a candidate’s name correlates with the election outcome.**

Predicting elections using Twitter has many flaws and criticisms, as addressed by Gayo-Avella (2012). However, Tumasjan et. al (2010) concluded in his study that looking at the number of mentions of a political party came close to traditional polls and is a plausible indication of voter shares. In this study, I looked at candidate name mentions. I did not look at political party mentions because not all of the countries in this study have official political party names. However, I looked at the number of mentions of each candidate and found that, except for Brazil, mentions of a candidate are indeed indicative of the outcome of an election regardless of a country’s freedom status. It is important to note that the winner of Brazil's election was female and candidate names were only queried for last names. In elections, female candidates are more likely to be referred to by their first names than male candidates (Reeves, 2009). The data collected from Brazil revealed that "Dilma" was mentioned more than "Rousseff", thus reflecting a degree of gender bias in referring to female candidates by their first names.

All of the countries reflected (except for Brazil) that the candidate with the most mentions wins the elections. Coupled with the sentiment analysis I conducted, these results show that when looking at tweets and the outcome of the election, it is not necessarily important what is being said about the candidate (many of the candidates had negative sentiment scores), but how many times a candidate is being mentioned. While it cannot be conclusively stated that the number of mentions of a candidate directly correlates with the outcome of an election globally, the results of this research indicate that such a hypothesis is indeed plausible. Below, I describe the data and analysis required to prove such a hypothesis on a global scale.

### *Data*

For this particular study, Topsy data would not be sufficient. The data needs to be representative of all tweets during an election and not just the "influential" tweets resulted from Topsy queries. While tweets resulting from queries from the Topsy service are sufficient for some studies, a study that looks at the raw number of mentions requires data that is representative of all the tweets and thus needs to be collected directly from the source. To prove that the number of mentions correlates with the outcome of the election, data needs to be collected for election cycles of several countries. Even though this data is difficult to access because of Twitter API limitations, a selection of 10-20 countries would be sufficient for this study.

### *Analysis*

To conduct an analysis of the data described above, the tweets need to be searched for mentions of the names of candidates and political party names. Mentions can include a reference to the candidate name or a direct tag using the ampersand symbol (@) if the candidate has a verified Twitter account. Different spelling variations of candidate names must be considered when searching for mentions of candidate names in different countries. Mentions through hashtags should be considered as well for such a study. Often, during elections hashtags in support of a particular candidate are shared. For example, in Brazil's election, the hashtags "#voude13" and "#13neles" do not mention Dilma Rousseff's name, but they are supporting her and the party to which she belongs. An automated method of detecting which hashtags support which candidates can count the number of hashtags in support of a candidate. For example, a tweet in favor of a particular candidate will most likely use hashtags in support of that candidate. Taking this into consideration would allow a researcher to draw a connection between an obvious hashtag in support of a candidate and one that is not so obvious. For example the hashtags "#soumaisdilma" is in obvious support of the candidate and includes the name of the candidate within the hashtag. If "#soumaisdilma", "#voude13" and "13neles" are all used in the same tweet, it can be deduced that "#voude13" and "13neles" are in favor of Dilma Roussef, the presidential candidate for the Brazilian election. These hashtags should potentially be considered in the count of mentions. In this research design, "mentions" should be redefined to include hashtags, direct tags (@) of political candidates as well as mentions. Upon conducting this search, the number of mentions for each candidate should be compared to voter shares to see how closely it correlates with the outcome of the election.

### *Significance of Study*

While the research design above is not attempting to predict the election outcome based on the number of mentions of a candidate, it can demonstrate whether the number and type of mentions of a candidate correlates with election outcome globally. Furthermore, if such a correlation is discovered for all of the countries studied, types of mentions and their relationship with the outcome of the election can be more clearly defined. For example, in this study it was observed that in some countries, the winner of the election was tagged directly using the ampersand symbol (@), while in other countries, the winner of the election was not tagged directly but had multiple top hashtags in support of her/him. While I found that there is no relationship between sentiment expressed towards a candidate and election outcome, the initial results showed that number of mentions correlated with election outcome regardless of freedom status. In addition to mentions by name, number of hashtags (#), and direct tags (@) can be counted and compared to see the way in which users communicate about a political party or candidates and how the method of mentioning a candidate is related to the outcome of the election.  This research would aid in understanding the relationship between different forms of referencing political parties and candidates on Twitter (tags, hashtags, and mentions) and the outcome of the election on a global scale.

# Chapter 6: Conclusion

In this study, I explored the Twitter activity during nine election cycles within countries with nine different freedom statuses. In this study, I found promising initial results showing that tweets from "not free" countries are more likely to have content removed from links that are shared from the web, networks of tweets occurring during elections of "not free" countries have more disconnected profiles and that there is a strong relationship between election outcome and number of mentions of a candidate. Based on my results, I presented three hypotheses with a research design that can be pursued in future work. My results also show that the sentiment expressed towards a candidate by tweeters during an election cycle does not indicate who will win the election and that the sentiment expressed towards an incumbent does not correlate with the freedom of a given country.

## References

Alexanyan, K., Barash, V., Etling, B., Faris, R., Gasser, U., Kelly J., Palfrey, J., and Roberts H. (2012). Exploring Russian cyberspace: Digitally-mediated collective action and the networked public sphere. *Berkman Center Research Publication,* 2. Retrieved from http://ssrn.com/abstract=2014998

Booth, W. (2012, September). Egyptian blogger Alber Saber's arrest underlines differences on freedom of speech. The Washington Post. Retrieved from http://articles.washingtonpost.com/2012-09-26/world/35497000_1_egyptian-blogger-violent-anti-american-protests-egyptian-authorities

Burns, A. and Eltham, B. (2009). Twitter free Iran: An evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 election crisis. *Communications Policy & Research Forum 2009*. Sydney: University of Technology.

Clauset A., Newman, M., Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), doi:10.1103/physreve.70.066111

Eltantawy, J., Wiest, J. (2011). Social Media in the Egyptian revolution: Reconsidering resource mobilization theory. *International Journal of Communication*, 5, 1207–1224.

Evans, M. (2010, January 22). The Top Twitter countries and cities (part 2). Retrieved from http://blog.sysomos.com/2010/01/22/the-top-twitter-countries-and-cities-part-2/

*Freedom in the World 2012.* (2012, October 17). *Freedom House*. Retrieved from http://www.freedomhouse.org/sites/default/files/FIW%202012%20Booklet_0.pdf

Gayo-Avello, D. (2012, April 28). I wanted to predict elections with Twitter and all I got was this lousy paper: A balanced survey on election prediction using Twitter data. *Cornell University Library*. Retrieved from arXiv:1204.6441v1

Kavanaugh, A., Sheetz, S., Hassan, R., Yang, S., Elmongui, H., Fox, E., . . . Shoemaker, D. (2012, April). Between a rock and a cell phone: Communication and information technology use during the 2011 Egyptian uprising. *Proceedings of the 9th International ISCRAM Conference*.

Leskovec, J., Faloutsos C. (2006). Sample from large graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 631-636, doi:10.1145/1150402.1150479

Little, A. (2012, October). Fraud and monitoring in noncompetitive elections. *New York University.* Retrieved from https://files.nyu.edu/atl250/public/little_fmne_web.pdf

Livne, A., Simmons, M., Adar, E., Adamic, A. (2011). The party is over here: Structure and content in the 2010 election. *Association for the Advancement of Artificial Intelligence*. Retrieved from http://wwwpersonal.umich.edu/~ladamic/papers/blogosphere/ICWSMLivnePoliTweet.pdf

O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the International AAAI Conference on Weblogs and Social Media.* Retrieved from http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf

Reeves, K. S. (2009). Media Gender Bias in the 1984 and 2008 Vice Presidential Elections (Master's thesis). Retrieved from: http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1031&context=honors

Sharp, A. (2013, January 18). 100 Senators and the 57th Inauguration. Retrieved from http://blog.twitter.com/2013/01/100-senators-and-57th-inauguration.html

Thomas, K. Gier C., Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In Proc. of Leet.

Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852

Vorobyev, Dmitriy. (2010, October 1). Growth of electoral fraud in non-democracies: The role of uncertainty. *CERGE-EI Working Paper Series, 420*. Retrieved from http://ssrn.com/abstract=1692347 orhttp://dx.doi.org/10.2139/ssrn.1692347

Wang, X., Wei, F. Liu, X., Zhou, M. Zhang, M. (2011). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 1031–1040. doi10.1145/2063576.2063726

Zhang, H. (2004). The optimality of naïve bayes. *FLAIRS Conference.* Retrieved from http://courses.ischool.berkeley.edu/i290dm/s11/SECURE/Optimality_of_Naive_Bayes.pdf