Big Data and Social Media: Calculating Network Agenda Setting using Twitter and the 2012 U.S. Presidential Election

By Chris J. Vargo & Lei Guo

Abstract:

In the first section of this chapter, we outline the major benefits that a big data study of social media can yield in NAS. The second section outlines in detail the methodological steps needed to complete such a study. For this chapter, we rely on major studies of the 2012 election that captured millions of Tweets and performed big data analytics. The chapter closes with a discussion of future big data, social media NAS studies.

Section 1: Why Big Data & Social Media?

Big Data is a term that is thrown around loosely. It is used to represent the size of the analysis, not necessarily the scope. So what does big data really bring to the table when it comes to the third-level of agenda-setting analysis? Here we argue power and detail.

Power & Detail

Big data analysis allows for greater power in effect sizes. Basic statistics tells us that when the *n* grows, the size of the effect that is needed to be significant shrinks. The statistical value of power can often result in the difference between "approaching significance" and actually reaching it. With this power comes a need for responsibility of the researcher, which we discuss at the end of this chapter. Indeed, the power of big data studies can enable us to control for as many variables as we can muster in regression, and still have significant effect sizes.

Moreover, big data sets such as Vargo et al. (2014) allow for great flexibility, imagination and specificity of the researcher. If a researcher wanted to only measure agendas of Twitter users in southern, conservative voting districts, such an analysis is possible. Similarly, if a measurement to detect people of certain beliefs (i.e. pro-choice or pro-Obamacare) as long as the measurement can be reliably created, it can be tested against other groups. Researchers no longer have to rely on pre-structured survey questions to develop hypotheses. When a dataset has the behaviors of hundreds of thousands of people, an infinite number of subgroups of people inherently exist. Groupings of people on Twitter include: political affiliation, sets of people connected by interactions with each other, groups of people that "follow" or listen to certain users, groups of users that follow or listen to certain types of media, groups of people that mention certain issues or attributes, groups of people that talk with similar linguistic traits, groups of people living in certain geographic areas, or even groups of users that exhibit similar levels of incivility. This list scratches the surface. If a difference could be derived in a panel or survey, it may similarly be detected in a big data study using a dataset like Tweets on Twitter.

Similarly, big data studies allow us to detect a large number of issues and attributes. Unlike a survey or a panel, the data can always be retroactively summed and scored. In a panel or survey, if a participant was not asked about an issue or attribute, the response is forever lost. With big data studies, if people talked about an issue or attribute with enough salience, and the data was captured, it can be later analyzed. This allows researchers to be inductive rather than deductive. There is no need to enter a big data third-level agenda-setting study with preconceptions of what the major issues or attributes are. Instead, those features can be derived using methodologies such as term-frequency lists. In this way, researchers are less selective and more perceptive of what truly were the most important and salient issues and attributes.

Any number of studies can be done on a good corpus of big data. However, whether a study will be successful is the direct result of methodological rigor. The rest of this chapter outlines how the NAS model can be measured and tested using the ever-popular microblogging platform Twitter. The following section outlines each methodological step needed to perform a big data NAS study. The examples show how this can be done using Twitter, but similar social media platforms can be accessed in similar ways. The following is a deep dive into the data and studies on Twitter during the 2012 U.S. presidential election by Vargo et. al, (2014). Tweets from many different groups, including citizens, news media and political campaigns were retrieved and stored in a 22-gigabyte corpus. Network analysis was then performed to investigate and compare how these groups were related to each other.

Section 2: A Review of Existing Methodology

Data Collection

At the crux of every big data project is the data. While no hard and fast rule defines what size is awarded the status of "big data" – here we assume millions of units. The idea here is that the data size is so large that traditional methods such as sampling a proportion with a manual content analysis is not feasible (e.g. as in Riffe, Lacy & Fico, 2005). Studies have shown that small samples of big data sets are often not enough (Morstatter, Pfeffer, Liu, & Carley, 2013). Instead, all of the data needs to be retrieved. As such, semi-automated computer methods must be performed – *with academic rigor and understanding*.

The first question most social science researchers have when dealing with exponentially large datasets is retrieval. What is the best method to retrieve and store the data? In the early years of social media agenda-setting research, third party services were used to retrieve and store the data (e.g. Vargo, 2011). While this can save the researcher time, several problems exist. Most data aggregation services don't offer clear methodology as to what data is being retrieved. The researcher rarely knows what sample of the data they are getting. The best bet in these cases is to get the data straight from the source. This means connecting to the social media service in question directly and downloading the data using an Application Programming Interface (API). APIs are backends that allow for the downloading of data in plain text formats. Twitter's API for instance, can be queried like a search engine. The API replies back to queries with JSON formatted data. This data can be manipulated in many different ways easy because it is structured and uniform. For instance, many programs exist to convert JSON to tabular (i.e. Excel, STATA and SPSS) formats.

In the paper at hand, version 1.0 of Twitter's API was used to download relevant Tweets during the election period. The streaming API call was used to download public messages from Twitter. If a term contained "Obama" or "Romney" it was captured. To use an analogy, the streaming API is like a radio. It is only works well for social science research when it is tuned to one radio station (e.g. Morstatter, Pfeffer, Liu, & Carley, 2013). Unaltered, Twitter's streaming API listens to all of Twitter. But due to the extremely high volume of Twitter, when instructed to do so, it only retrieves approximately 1% of all of Twitter. Studies have clearly shown that this sample is not representative enough for research (Morstatter, Pfeffer, Liu, & Carley, 2013). However, when the streaming API is used, it can also be configured to only retrieve Tweets that mention certain keywords. When used in this way, samples become very representative, provided the salience of the keywords does not approach 1% of all Twitter traffic (Guo & Vargo, under review).

Various modules exist in popular computer programming languages that work directly with Twitter's API. Dr. Deen Freelon maintains an exhaustive list of tools and modules that exist to extract data from social media platforms can be found online (Freelon, 2014).

For the study at hand, the collection started on August 1st, 2012 and ended on the Election Day, November 6th. In all, 70 million Tweets were collected and analyzed. The number of messages collected here is not as important as the time period is. While 70 million certainly makes this study “big” in nature, it also verifies that the effects we are seeing from an agenda-setting point of view are continuous and not sporadic. Vargo, Basilaia and Shaw (2015) show that agenda-setting effects vary by nature of whether they are a breaking news story or an ongoing event. The study at hand chose exclusively the latter. As such, a long period of time was needed to capture the long, debate-like nature of those ongoing issues.

Who is talking, anyway?

After the data is collected, another important question surfaces when performing an agenda-setting analysis of a social media platform such as Twitter. In traditional surveys and panels, identification of subjects by demographics and political affiliations is easy. Respondents are asked these questions directly. On social media such as Twitter or Facebook, these questions are often inferred. Twitter for instance, does not have profiles where political affiliation can be self-identified by the user. Facebook and other platforms may provide this option, but the response is voluntary, and as a result many users fail to do so.

An agenda-setting analysis must make an effort to understand whose messages are being analyzed. Twitter is a platform where almost every person, business, media and organization exists. Tweets contain the media. Tweets contain the government. Tweets contain people. Everyone is posting tweets. All API calls download messages, but provide no mechanism to differentiate between these different types of accounts. At the time of this chapter, this problem does not apply to Facebook, but does to other services such as YikYak. Vargo et al. (2014) parses through the data to show what is being measured.

There are many possible ways to infer different groups of people on Twitter. Here we will cover three possible methods: (1) through unstructured profile data (2) through the messages that users create and (3) through network characteristics.

Group Inference By Profile Data

From the 77 million Tweets in the dataset, 5.46 million unique users were found. One way to see who users are is to download their corresponding profiles. Profiles are short descriptions of who a user is on Twitter. At the time of writing this chapter, Twitter profiles were subjected to the normal API rate limit. This means that only 720 per hour could be downloaded without special (i.e. very rare or expensive) access from Twitter. This limitation seems minor, but it would have taken 12 computers 26 days to download all the profiles. Guo and Vargo (under review) downloaded profiles of users with at least 4 Tweets in the dataset. 2.97 million profiles were queried using the REST API call. This call is different from the streaming call in that it retrieves one exact piece of information. The REST "Get Users" call was able to download 69% of the profiles, or 2.05 million. The remaining users' profiles were inaccessible.

The next step was to somehow sort through these users to find those who self-identify themselves in a way that is suitable for the analysis. There are no apparent limits as to what this criterion could be. Guo & Vargo (under review) chose to look at liberals and conservatives. To do this, several rounds of manual content analysis and intercoder reliability were performed. 500 random user profiles were pulled from the user profile dataset. Two human coders labeled profile descriptions as liberal, conservative or neither. The hope here was that users that tweeted about Romney and Obama would offer political affiliations in their profiles. While most profiles offered no affiliation, those that did offered straightforward affiliations were coded. The intercoder reliability was 99%. Previous studies (e.g., Lombard, Snyder-Duch, & Bracken, 2002), consider 90% percent agreement a reasonable cutoff for robust intercoder reliability.

Keyword lists were then populated based on popular words found in the manual content analysis. This process was intuitive and generally contained words that would be logical to such associations such as "liberal" and "conservative." Each word was then queried, and the profile results were inspected for each word. Only words that correlated very highly with positive matches were used. Interestingly, the term "Obama" could not be used, due to an extremely high number of negative mentions (i.e. Hi I am Peter, I hate Obama). Using regular expressions and Python, if a keyword was found in a profile for one, but not both of the lists, the user was identified as liberal or conservative.

To verify the lists of keywords, another round of manual content analysis was performed to compare the results of the human annotations to those from machine coding. 400 profiles were randomly sampled and then annotated by two human coders. The two coders achieved 100% agreement and agreed on the computer annotations 97% of the time for Conservatives and 98% for Liberals. The straightforwardness of political affiliations in Twitter bios accounted for such robust reliability. In all 19,509 Liberals and 26,494 Conservatives were identified. The 46,003 users created 2.69 million messages on Twitter during the time sampled. We note at this point that the "big" dataset is beginning to get smaller. In general, the further we put restrictions on the data the smaller the dataset gets.

Group Inference by Message Characteristics

One limitation of the first method is that Twitter users rarely have much detail in their profiles. In general the descriptions are terse, and often contain something clever that may not be so straightforward to for a computer to identify via keywords. As such, Twitter users rarely identify themselves as Republican or Democrat in profiles. One way to get around this limitation of Twitter user profiles is not to use self-proclaimed political alignment, but instead to identify users by the contents of the messages that those users broadcast publicly. Vargo et. al (2014) identified two groups of users: Obama supporters and Romney supporters. They conceptualized these users as positively vocal for that given candidate. As such the authors are careful to refrain from calling them Democrats and Republicans, although affiliations likely coincide.

First, the dataset was divided into tweets that mentioned Obama and not Romney or other Republican primary candidates, and, conversely, Romney and not Obama or other Republican primary candidates. Then sentiment analysis was used to detect how the user felt about that candidate in each tweet. All of a given user's tweets about a candidate were measured and an average sentiment score for that user was calculated. The sentiment was scored using a lexicon-based sentiment analysis tool.

Sentistrength was chosen due to its development for short texts (Thelwall et al., 2010; Thelwall, Buckley & Paltoglou, 2012). The tool has "human-level accuracy for short social web texts in English," and has been widely used in a wide range of research projects (Thelwall et al., 2010). The validity of SentiStrength was again tested using similar validity checks to Guo and Vargo (under review).

Next, Microsoft Excel was used in conjunction with Power Pivot, a free add-on provided by Microsoft for PC versions of Excel. Pivot tables were created for each remaining user in the two datasets. This tool easily created summary statistics for each user. Each userid was given a row in Excel. The columns contained variables such as average sentiment score. This averaged sentiment score was created for each candidate. It was then subjected to a one-way directional t-test. The degrees of freedom was one minus the number of tweets that user had created about the given candidate. A probability of .10 was used as the cut-off. While no formal rule for the cutoff is presented, the lower the probability used, the more likely the users are to be avid, vocal supporters. The researchers identified 2,875 and 2,457 candidates as Obama and Romney, respectively.

This method ultimately identified fewer users, but did so in a way that looked at the behaviors of the users, instead of the self-identifications of the users. It also did not require a lengthy and involved profile data retrieval process. It instead used the data already retrieved to infer affiliations.

Group Inference by Network Characteristics

Other methods would likely result in identification of groups of users. For instance, a user could be considered a republican based on who they "followed" on Twitter. If a user followed Mitt Romney, and FoxNews but not Barack Obama or MSNBC, that user might be a republican. Public Twitter accounts can be queried to reveal what sources that user is following. As such, groups of users can be connected through commonalities in those sources. The downside to this method is again the rate-limiting nature of Twitter's API. If a user follows 100,000 users, it will take a significant amount of time to retrieve the data. Once that data is retrieved however, surmising that a user is likely a republican or democrat could be as simple as writing a series of IF THEN statements in a computer language, like Python (i.e. IF a user follows Obama AND not Romney THEN the user is a democrat). Like the other two methods discussed, any type of method that identifies users should be subjected to a manual content analysis using human coders to evaluate the results of the process.

Detecting the News Media

Agenda-setting research must also measure the media's agenda. This step appears to vary in agenda-setting research that involves social media. Vargo (2011) used a manual content analysis of traditional newspaper articles. However Twitter is not just people. The data itself contains messages from news organizations of all shapes and sizes. In this way, news media can be extracted from Twitter's streaming API. Usernames of news organizations can be gathered by researchers. This makes the extraction process rather simple. Vargo et al. (2014) took Tweets from the official accounts of top 25 U.S. newspapers by circulation and major broadcast news networks. Extraction was as simple as telling a computer programming language to only retrieve messages that came from these accounts. The researchers used Python, but Excel formulas or MySQL could be easily used.

Detecting Political Campaign Messages

Just as news media were retrieved using the usernames tied to the official accounts, Guo and Vargo (under review) used the REST "GET User Tweets" call to retrieve all the messages from the Obama and Romney official Twitter accounts. This method was more exhaustive than the streaming API in that it ensured that all campaign messages on Twitter were included.

Issue Selection and Coding

To this point we've outlined how to separate users into groups. Now that the data is separated, the analysis is ready to be performed. To do an agenda-setting analysis, issues and/or attributes need to be selected and then analyzed.

Vargo et. al (2014) used term-frequency lists of single word and word pairs to determine the most popular issues in the 2012 election. This process was manual and inductive. Words were sorted into constructs of keywords to indicate each issue. Again, rounds of reliability tests enabled the researchers to refine the keywords lists. Guo and Vargo (in review) used a more direct method. A stratified sample of 600 Tweets was pulled by each Tweet category (news media, citizens and campaign messages). Two coders were asked to assign each Tweet by the issue it mentioned. Agreement was calculated for each group to ensure reliable results. Words that correlated positively with the coders' annotations were placed into corresponding issue lists for computer-assisted analysis. Again, manual content analysis was performed to compare the human and machine coding results.

After several rounds of reviewing the samples and adjusting the issue lists, acceptable agreement was reached, with no issue scoring below 92%. In all 16 issues were identified. When operationalized using Python and regular expressions, two sets of words were established for identifying issues. Exact matching and non-exact matching was used to reduce false positives (i.e. so tweets containing “jobs” would only return valid results, not results that mentioned “Steve Jobs”).

Once keyword lists are derived a computer program must annotate each message as either containing or missing (1 or 0) an issue. This process needs repeated for each issue. This can be done by simply appending a column at the end of the dataset for each issue and filling the columns with the corresponding 1’s and 0’s.

Analyzing The Data Using Network Analysis

With the data prepared, any level of agenda setting (ie. 1st, 2nd or 3rd) or agendamelding analysis can be preformed. Since this chapter focuses on the 3rd, we will cover the two statistical methods that have been recognized for NAS.

In Guo and Vargo (under review) a candidate’s issue ownership network is represented as an ego network in which the candidate acts as the “ego” at the center of the network. The ego is always at the center in this NAS analysis. It is connected with other nodes (i.e. the other 16 issues). If a Tweet mentions a candidate and one of the nodes, a link was identified. The issue nodes can have ties with each other. When two nodes were mentioned in the same Tweet, the connection was recorded in a co-occurrence matrix.

Again, computer languages such as Python can generate and populate co-occurrence matrices. The standalone program Ucinet also is capable to generating such matrices from tabular formatted data (Borgatti et al., 2002). Matrices were used to represent the “associative” issue networks regarding the two candidates in news coverage, campaign messages and public discussions on Twitter. To create “competent” issue ownership networks, sentiment analysis was performed to detect positive association between a candidate and the set of issues using the same method outlined earlier in the chapter (e.g. Thelwall et al., 2012). Using sentiment analysis scores, positive ties were tallied between the ego and issues and among different issues to construct “competent” issue ownership networks of the two political candidates in the form of matrices.

The statistical method of assessment used most commonly for NAS studies is the quadratic assignment procedure (QAP). This regression analysis was used to assess the relationship between different issue ownership networks. For each candidate, the tests regressed the public opinion data during a later period on, the news media and political campaign data from an earlier period. Using the network analysis software Ucinet, QAP regression tests were performed.

Analyzing the Data Using Timeseries

The QAP method is robust and well accepted for regression tests of network data. However for rapidly changing agendas, it may not address fully address fluctuations for network agendas. To address this issue, Vargo et. al (2014) calculated centrality measures for of each issue and used that as the parameter for analysis. The degree centrality measurement was used for the study. It is calculated by tallying the number of connections a node has with all other nodes (Wasserman and Faust, 1994). The more ties an issue has with other issues in describing a given candidate, the higher degree centrality value the issue has, and the more centrally it is located in the resulting networks.

Once these scores were calculated for each issue across time (i.e. by week) the data was modeled using linear regressions. An OLS regression was performed with a Durbin-Watson statistic. The Durbin-Watson statistic inside of the OLS regression determined the relationship between dependent and independent variables separated from each other by a given time lag. Provided that the Durbin-Watson assessment could address the autocorrelation of the dependent and independent variables, then the autocorrelation was a violation of typical OLS assumptions.

Section 3: Future Big Data, Social Media NAS Studies

Replication

Because of the advantage of power that we mentioned in the beginning of this chapter, spurious correlations or regression coefficients may result. As such, replication is extremely important. The argument can be made that a big data study is robust with its huge *n.* However, an argument can be made that any one agenda-setting study has an $n = 1$. As with all studies, a big data finding should be replicated before it can build on theory.

We argue that as with traditional agenda-setting research, before NAS theory can be truly built, an effect should be observed 1) across issues or attributes, 2) at different time periods and 3) by different groups of people. Just because an effect is found for one issue or attribute does not make it a theory. How can we know that this effect was the rule, or the exception? This is a limitation of social media agenda-setting research: the agenda-setting effect does vary from issue to issue (Guo & Vargo, under review). You can observe this anecdotally by reading the news. Sometimes, a man on the street tweets a breaking story. Other times, someone on Twitter discovers a fugitive and where they are hiding. Taking this into account, leading scholars have adopted the view that agenda-setting theory still applies on social media. Excepting exceptions of course, which most scholarship observes to be relatively infrequent. Because of this finding alone studies in NAS should always be replicated across different time periods, or across long periods of time. Finally, true theory will account for, or explain effects of different actors. Studies such as Vargo et al. (2014) clearly show that agenda-setting research does vary by the audience in question. Republicans don't behave like democrats. This is clearly just scratching the surface.

TERGMS

Unfortunately, an OLS regression requires one dependent variable. Therefore, Obama and Romney supporter agendas in Vargo et. al (2014) were tested at the individual issue level with a model for each of the eight issues. Issues in all three media types were used as independent variables, totaling in 24 models for each candidate supporter. The researchers entered all issues as possible explanatory variables in the regression. The argument stands that the use of the variables in each model is necessary due to the interconnectedness of the degree centrality measure. Only offering one independent variable would ignore the networked agenda, and with it the network characteristics that offer explanatory power. Adjusted r^2 values were computed because of the seasonality that the 17-week cycle inevitably possesses.

OLS is the only known time series method that has been used to study NAS. However, here we note that other, more sophisticated and intricate measures methods such as Time series Exponential Random Graph Models (TERGMS) might be able to explain network agendas in similar fashions. Such methods of analysis may allow for an entire network to me analyzed at once, instead of one issue's centrality. This test may be more beneficial to fully explain networked agendas.

References

Borgatti, S. P. (2002). *Netdraw Network Visualization*. Harvard, MA: Analytic Technologies.

Freelon, D. (2014). Social media data collection tools. Accessible at: https://docs.google.com/document/d/1UaERzROI986HqcwrBDLaqGG8X_lYwctj6ek6ryqDOiQ/edit

Guo, L. & Vargo, C. (under review). The power of message networks: A big-data analysis of the Network Agenda Setting Model and issue ownership.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. Proceedings of ICWSM, Boston.

Riffe, D., Lacy, S. & Fico, F. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis in Research.* Routledge: London, U.K.

Thelwall, M. B. (2010). Sentiment detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. doi: 1.1002/asi.21416

Thelwall, M. B., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63*(1), 163-173. doi: 1.1002/asi.21662

Vargo, C. (2011, August). Twitter as public salience: An agenda-setting analysis. Paper presented at the AEJMC annual conference, St. Louis, MO.

Vargo, C., Basilaia, E. & Shaw, D. (2015). Event vs. Issue: Twitter Reflections of Major News, a Case Study. *Communication and Information Technologies Annual 2014: Politics, Participation, and Production. Emerald Studies in Media and Communication.*

Vargo, C., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on Twitter during the 2012 US presidential election. *Journal of Communication*, *64*(2), 296–316.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.