Cooperative, Dynamic Twitter Parsing and Visualization for Dark Network Analysis

Patrick M. Dudas
School of Information Science
University of Pittsburgh
pmd18@pitt.edu

Abstract

Developing a network based on Twitter data for social network analysis (SNA) is a common task in most academic domains. The need for real-time analysis is not as prevalent due to the fact that researchers are interested in the analysis of Twitter information after a major event or for an overall statistical or sociological study of general Twitter users. Dark network analysis is a specific field that focuses on criminal, terroristic, or people of interest networks in which evaluating information quickly and making decisions from this information is crucial. We propose a platform and visualization called Dynamic Twitter Network Analysis (DTNA) that incorporates real-time information from Twitter, its subsequent network topology, geographical placement of geotagged tweets on a Google Map, and storage for long-term analysis. The platform provides a SNA visualization that allows the user to interpret and change the search criteria quickly based on visual aesthetic properties built from key dark network utilities with a user interface that can be dynamic, up-to-date for time critical decisions and geographic specific.

Keywords: dark networks, visualization, social network analysis, user-design

Introduction

When looking at patterns in large datasets, a popular choice is the utilization of Twitter. It provides user-generated content, or micro-blogging, in a real-time environment as a social media outlet that is propagated from one individual to another in terms of text, links, pictures, or videos. Projects for this type of networked data is usually data-mined after the fact, because either researchers are interested in a certain past event (Lotan et al., 2011; Yardi & Boyd, 2010) or topics of interest (Tumasjan, Sprenger, Sandner, & Welpe, 2010). There may also be a need for a very large dataset and backtracking provides the best means of collecting ample amounts of information versus real-time information (Bruns, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; Kwak, Lee, Park, & Moon, 2010).

Dark networks (Raab & Milward, 2003) or terrorist informatics (Cheong & Lee, 2011) provide their own unique problem set because immediate reaction is critical to the success of authorities looking for terroristic or criminal activity. We propose an architecture and visualization to enhance user decision making using various saliencies (Lurie & Mason, 2007) from key utilities designed for dark networks (Roberts & Everton, 2011). We also connect this information processed in Twitter and map geotagged tweets to a Google Map, combining both network topology and spatial information. We built a Twitter feed that allows the user to click links embedded in the tweet directly so the user can see if a particular website, social feed, or video is causing the propagation of information or other sources they should consider viewing. An additional visualization allows the user the ability to see temporal pattern changes in the network. All of this tailored to be simplistic and individualized allowing the user to input a location and either a username, hashtag, or keyword to pull information and begin visualizing.

The novelty of this approach is that we built an interface that can process, visualize, and store real-time Twitter data in network visualization and incorporate various salient features to capture the most prudent information. This platform is also novel in that as a website, the user can log in and work on the

Dudas, P. M. (2013). Cooperative, dynamic Twitter parsing and visualization for dark network analysis. *iConference 2013 Proceedings* (pp. 623-632). doi:10.9776/13295

same dataset as another user by selecting that project. Multiple people can view multiple locations and multiple terms, but for a common goal in capturing and analyzing all information obtained in that project. We provide two mechanisms of visualizing: 1) Real-time network topology, and 2) Long term network analysis.

The project was built and designed for use-cases surrounding the Syria uprising, a topic being analyzed at the CORE Labs at the Naval Postgraduate School (Monterey, CA). Since its development, it has been showcased to a variety of different users and groups in or associated with various government agencies. Provided in the paper is also a high-level view of the data from locations of interests.

Background

More recently, a variety of different types of real-time architectures have emerged in handling and interpreting Twitter data. Murthy suggested a means of employing real-time data into a data repository/data gathering application in terms of research related to cancer (Murthy, Gross, & Oliveira, 2011). This novel approach provided a framework, but lacked the ability to visualize or create a user interface for sense-making when it came to harvesting data.

A real-time data approach to terrorism informatics was developed by Cheong (Cheong & Lee, 2011), who provided a detailed look at parsing and filtering the data from Twitter. They provided some markups on how they would visualize this information for users, including: 1) WEKA timeline analysis (Hall et al., 2009); 2) a Google Map with geotagged tweets; and 3) a self-organizing map for unsupervised clustering (Kohonen, 1984). The timeline analysis is a great mechanism for visualizing terrorism or dark networks, but lacks the ability to network analyze in real-time. Also, the Google Map approach was strictly a markup. We provide this functionality along with the ability to select the geotagged position to view the tweet involved with this geospatial information.

SocialAction (Perer & Shneiderman, 2008) extends this model by interweaving the statistical properties in the visualization itself, imploring a higher rate of success of exploratory SNA. They even extended this expert knowledge to analysis of terrorist connection. The restriction with this approach is that the only information passed to the user is network topology and the salience changes based on the embedded statistics of that network. It lacked real-time data which is crucial to making decisions quickly and efficiently.

Visualization

Real-time

We developed a web application that allows an analyst the ability to: 1) collect information on Twitter by hashtag, username, or keyword; 2) visualize this information or network in real-time; and 3) change initial search criteria based on the visualization and applied social network analysis properties. The process starts by having the user log into the website and selecting either an existing dataset or creating a new one. Once a project is selected or created, the user then inputs city, state, country or zip code and a search term. For the search term, s/he can either select hashtag, username, or keyword to search. The web-application then utilizes Google Maps API (Google, 2005) to geocode the address given. Lastly, the user can select the type of network s/he is interested in collecting and viewing, which includes: 1) username -> username 2) username -> hashtag 3) hashtag -> hashtag. A mock example is given below.

Example:

James tweets: “Just saw @Dannie at the #Pittsburgh #Library”

Username -> Username:	James -> Dannie
Username -> Hashtag:	James -> Pittsburgh and James -> Library
Hashtag -> Hashtag:	Pittsburgh -> Library

With the geocode, search term and network type specified, the application creates a global search query, which is communicated to Twitter’s Search API while the page is active. The entire architecture is listed in Figure 1.

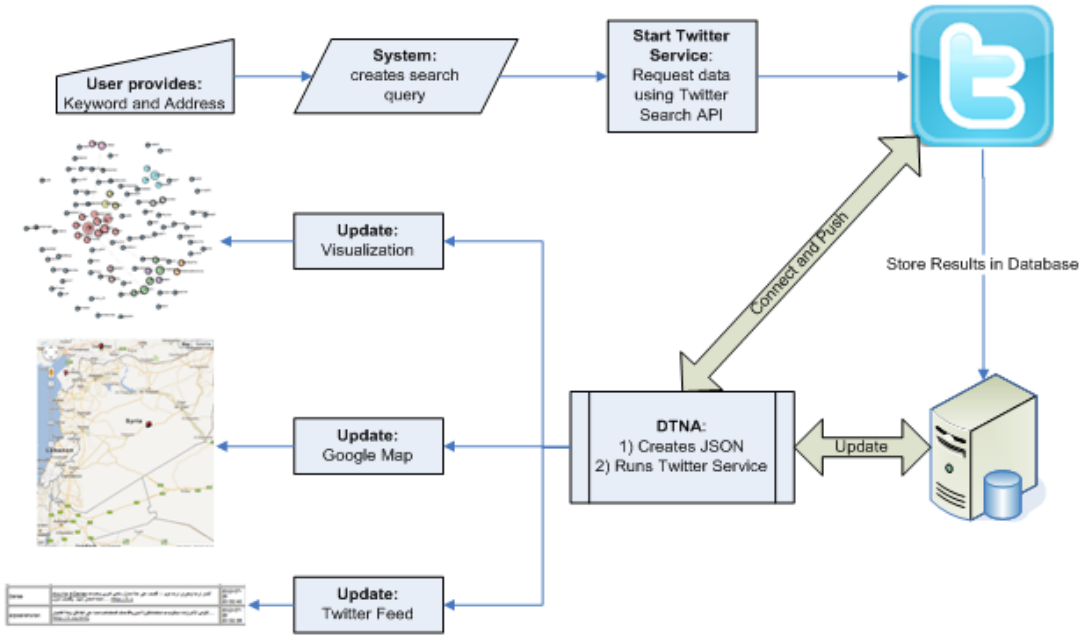

Figure 1. The Architecture of DTNA

At this point the user is provided the D3.js (Bostock, Ogievetsky, & Heer, 2011) force-directed network visualization, Google Map, and Twitter feed, all of which continually updates as long as the website remains active. The Twitter data is saved to a MySQL database, including the timestamps of tweets. The network itself has built-in aesthetic saliencies that suggests to other users, hashtags, or keywords that may provide better insight into the original search topic.

Saliency and Network. The graph for the tweets is created every time a query is requested from Twitter. The network, when visualized, is limited to only 200 nodes to limit the amount of attention needed by the user. The nodes themselves are selectable and will augment the original search query to include this hashtag, username, or keyword. This is a measure to help limit the amount of noise in a given graph. A clustering algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) is applied to the network and groupings are assigned and represented by color assignment. Based on log-based degree centrality, the size of the node reflects the number of connection each username or hashtag accompanies. Considering Twitter is a directed network, edges reflect the direction of the message. Figure 2 highlights both the groupings and centrality, with a bias towards increased inner-cluster strength versus out of cluster nodes. The importance of these saliency changes is to highlight nodes of interest or key players, which is important in dark networks (Everton, 2012).

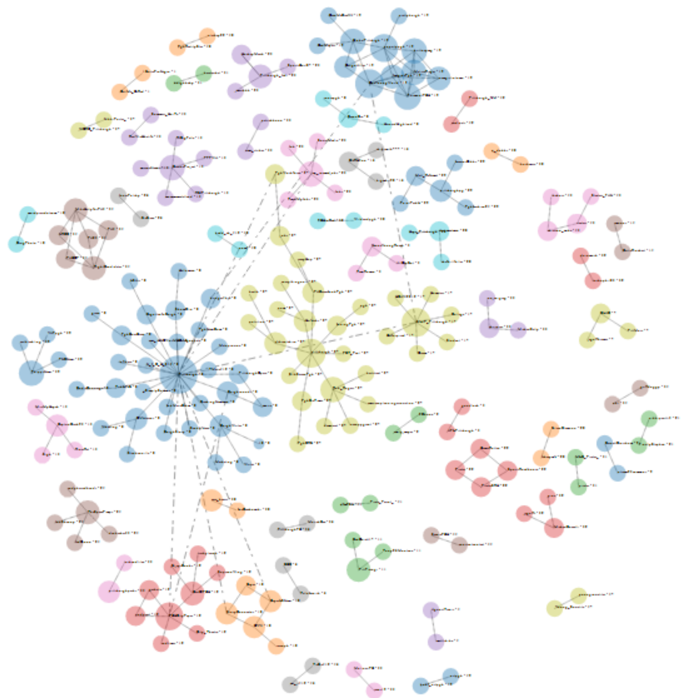

Figure 2. Network Visualization

Geographical Display and Twitter Wall. We appended a Google Map that shows geotagged tweets with the tweets themselves (Figure 3). This allows the user the ability to view the tweet and from where it originated. We created a text GUI that displays the tweets (Figure 4) along with the URLs, hashtags, and users. The URLs are selectable and will show the website in a new window. This will allows the user to see if a news article, YouTube video, blog, social networking post, etc. is causing an increased number of tweets.

Figure 3. Google Map with tweet

Daraa	#syria #Daraa [illegible] ... http://t.c	2012-07-29 20:02:40
aljzeerahoran	[illegible]... http://t.co/ANC	2012-07-29 20:02:39

Figure 4. The Twitter Feed (Including: username, the tweet, and timestamp)

Long Term

We provide a mechanism to group various datasets together for a single visualization. This can then be outputted as a .gexf, .net, or .dl network type. To visualize this closer to real-time, we developed a stream in a social network analysis application called Gephi (Bastian, Heymann, & Jacomy, 2009) and uses Gephi Streaming Network add-on to view the network as it evolves. Figure 5 below contains only 5 minutes of data collection and using the username -> hashtag connection. It was a grouping of both Syria and FSA search terms in Syria.

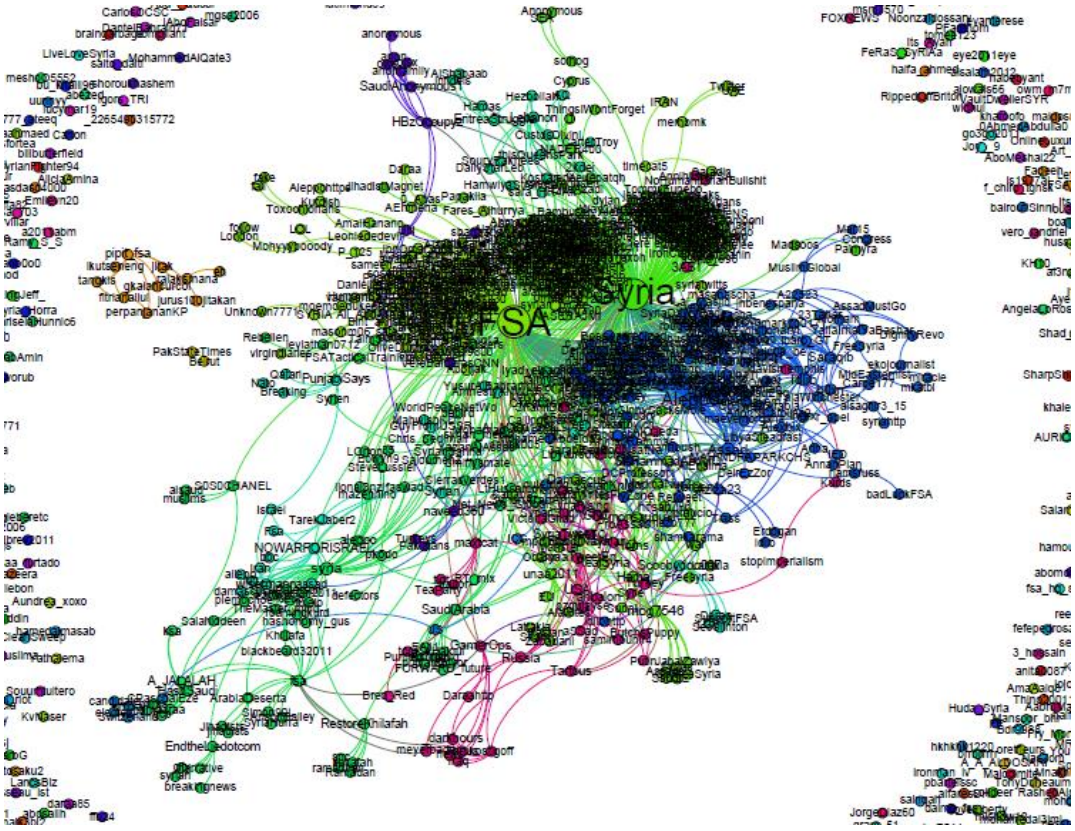

Figure 5. Gephi Visualization

We also developed our own means of visualizing entire projects called TweetViewer. This web application is to be used with DTNA and will display the completed network. Figure 6 shows the network in TweetViewer. TweetViewer for DTNA promotes the temporal data associated with the network by providing a timeline to allow a user the ability to specify at which point in time they are interested in viewing the network or the network's evolution over time. As DTNA's network visualization, edges reflect the direction of the message. Additionally, we differentiated the nodes that originated from the tweet by augmenting and animating the nodes' borders as time progresses. Figure 7 shows TweetViewer as it progresses through the timeline. Lastly, users can search based on the tweet to showcase nodes that include certain text. Figure 8 shows a network built using the hashtag "#FSA" in Homs, Syria and is highlighted by tweets that include the word "Syria."

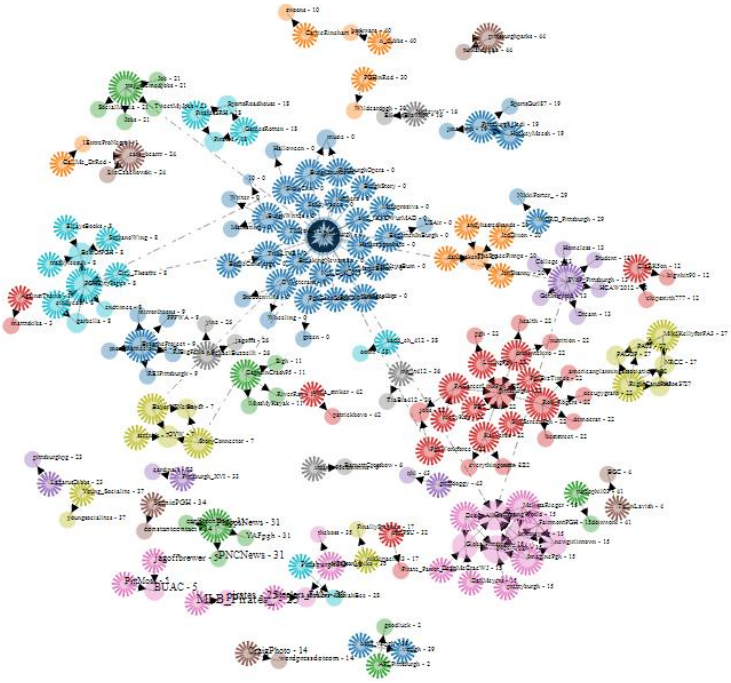

Figure 6. TweetViewer visualization

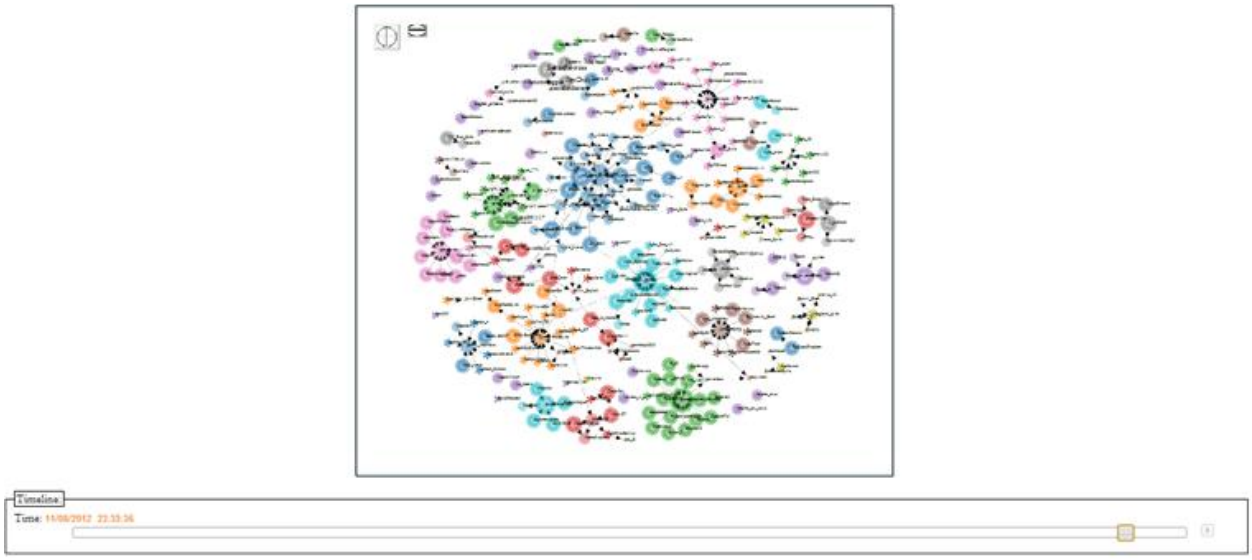

Figure 7. Network in TweetViewer using timeline

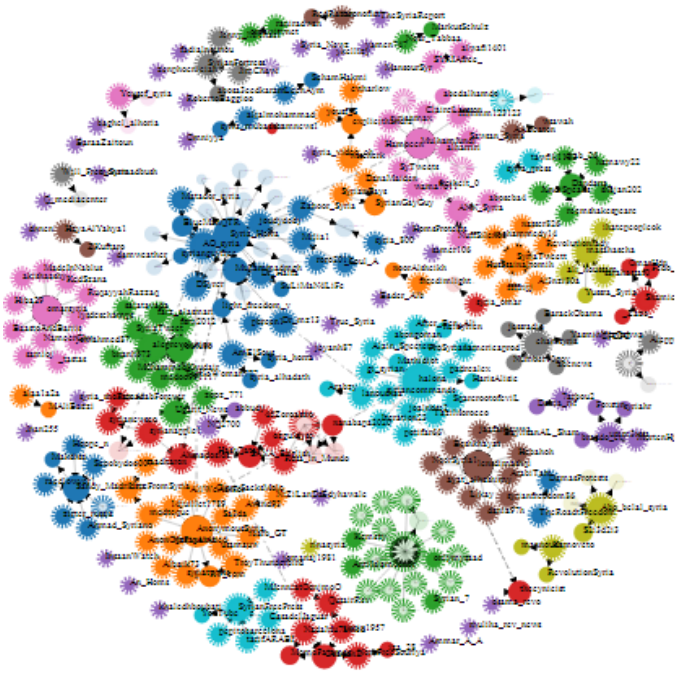

Figure 8. Searching TweetViewer for keyword “Syria”

Current Usage and Initial Findings

This project was only recently developed, but has already been employed by 30 users from a variety of military groups or organizations, ranging from civilian use to higher ranking officers including majors, first lieutenants, captains, and lieutenant colonels. This usage has accounted for 182 queries and has pulled roughly 86,000+ tweets as of publication. The majority of the tweets have centered on search areas (geographically) in Syria and the Philippines. Only 18% of the tweets pulled from these areas are geotagged. The most common search term is the hashtag “#FSA” for Free Syrian Army. There were roughly 8,000+ URLs in each tweet, which suggest a high influence on external websites driving the tweets.

Future Development

There are three avenues of research that we are currently interested in pursuing: 1) visualizing sentiment in Twitter posts, 2) determining the value of networks created using either username -> username, username -> hashtag, or hashtag -> hashtag connections, and 3) visualizing temporal data using more specific temporal dependent cluster algorithms and network centrality.

Sentiment Analysis Visualization

We are interested in not only the network that develops from this application, but also the sentiment for keywords. The way we are looking to address this is by implementing a sentiment lexicon (Nielsen, 2011), which includes word lists that are given a score from 5 to -5, based on their positive or negative connotations, respectively. This could be used to evaluate social movements or key terms and how over time these connections are influenced via sentiment lexicon over time. This does not have to be limited to strictly a global overview. It can be topic-based as well, allowing for searching of words or parts of words and connecting their positive or negative connections. The sentiment lexicon would be fixed around a circle, and the keywords would be placed in the center (Figure 9). Based on the number of occurrences of a key word or phrase in conjunction with a sentiment word would determine the nodes

polarity towards the positive or negative side of the circle. Below are two instances of this analysis at two different timestamps (Figure 10).

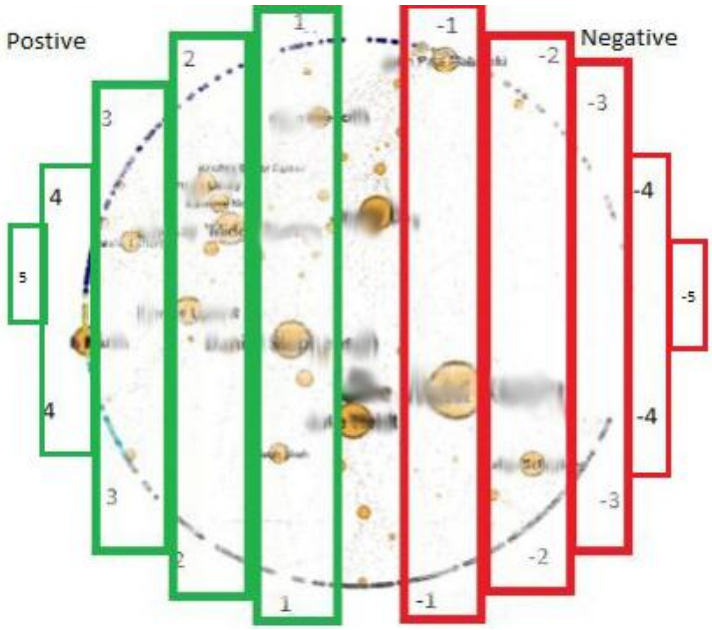

Figure 9. A markup of the sentiment visualization

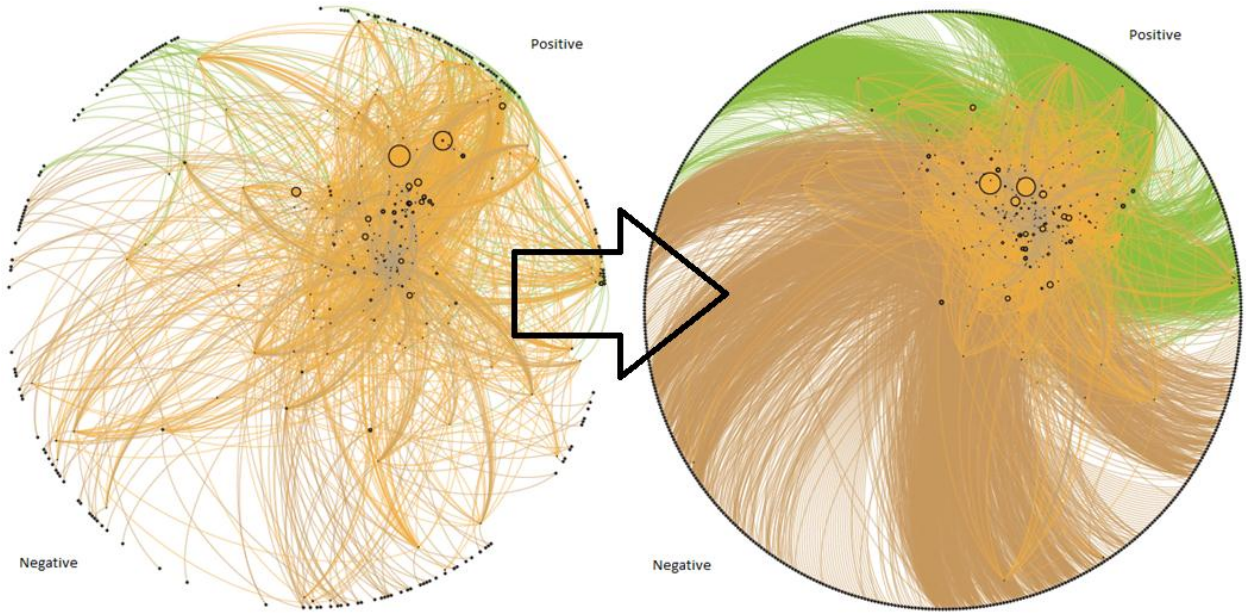

Figure 10. A markup of the sentiment visualization as the nodes shift from one side of the circle to the other (from negative to positive)

Value of Twitter Hashtag Networks

As of this of paper, there has been little to no work done on the quality or even the characteristics of a network built on a variety of different measures, such as: 1) username -> hashtag, or 2) hashtag -> hashtag. Username -> username is quite common, but username -> hashtag and hashtag -> hashtag may yield new insights into social structures. This would go beyond a social semantic ontology in that the user specifies these keywords (hashtags), which connects both a personal meaning by the person tweeting the information and by a global meaning in the concatenation of keywords (hashtags), which could be "trending" in the Twitter realm.

Visualizing Temporal Data

We are looking to implement some characteristics that are becoming popularized in evolving networks into our TweetViewer visualization. There is a need to not only show a network over time but to provide a visualization that allows for prediction. Macskassy presents multiple factions of evolution that could be applied to a visualization (Macskassy, 2012). This includes: 1) SNA metrics, such as Bonacich Centrality (Bonacich, 1987), with temporal parameters, and 2) a machine learning approach to predict and forecast. These measures take into account that communication evolves over time and should not be judged simply as a single, static, aggregated snapshot.

Conclusion

In this text we outline a cooperative, dynamic, and interactive visualization tool to collect and navigate through Twitter tweets. This includes a dynamic network visualization, a tweet wall post, and map with geotagged tweets. Additionally, we create a separate but homogeneous visualization that allows for exploration using timestamps and text from the tweets themselves. This application has been implemented into military outlets and has received high praise for its ability to data mine. We believe this tool has a plethora of interesting usages in dark networks and provides real-time information to people who most need this type of data.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 1170-1182.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 Data-Driven Documents. *Visualization and Computer Graphics, IEEE Transactions on, 17*(12), 2301-2309.

Bruns, A. (2011). How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K.P. (2010). *Measuring user influence in twitter: The million follower fallacy*.

Cheong, M., & Lee, V.C.S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers, 13*(1), 45-59.

Everton, Sean F. (2012). Network Topology, Key Players, and Terrorist Network. *Connections, 32*(1), 12-19.

Google. (2005). Google Maps API. from https://developers.google.com/maps/

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Kohonen, T. (1984). Self-organization and associative memory Springer. *New York Berlin Heideberg*.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?*

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & Boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication, 5*, 1375-1405.

Lurie, N.H., & Mason, C.H. (2007). Visual representation: Implications for decision making. *Journal of Marketing, 71*(1), 160.

Macskassy, S. (2012). Evolve: Analyzing Evolving Social Networks: DTIC Document.

Murthy, D., Gross, A., & Oliveira, D. (2011). *Understanding Cancer-Based Networks in Twitter Using Social Network Analysis*.

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Arxiv preprint arXiv:1103.2903*.

Perer, A., & Shneiderman, B. (2008). *Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis*.

Raab, J., & Milward, H.B. (2003). Dark networks as problems. *Journal of Public Administration Research and Theory, 13*(4), 413-439.

Roberts, N., & Everton, S.F. (2011). Strategies for combating dark networks. *Journal of Social Structure, 12*(2).

Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2010). *Predicting elections with twitter: What 140 characters reveal about political sentiment*.

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society, 30*(5), 316-327.