Open Source Social Media: @Russia-@Ukraine #Conflict

Tyson J. Quink
July 27, 2016

ABSTRACT

Security and privacy seem to be increasingly mutually exclusive concepts that are hotly debated in the era of ever growing digital transformations. In the wake of top secret documents released by Edward Snowden in 2013, which exposed the National Security Agency's (NSA) metadata monitoring program, politicians and citizens alike grew concerned about their digital privacy being monitored especially via the collection of social media data. I investigated whether historic social media data alone could identify the movements of large Russian military formations in 2014 towards Ukraine by examining volume changes in areas of known military garrisons. I connected geotagged social media posts between January 1-31, 2016, in Eastern Ukraine and Western Russia (~$550{,}000\ km^2$), from Twitter, Instagram, and VKontakte (VK), and used this data to detect events by examining the changes in volume of spatial clusters of data. The analysis produced 4,817 tangible areas of interest at a known moment in time out of a possible 122,016 possibilities, which reduces the need for possible investigation by 96%. As a result, time to discovery and faster analysis is possible with spatial clustering and temporal analysis of geotagged social media.

INTRODUCTION

Social networks allow users to share information, usually under the presumption that their content is only seen by the people they choose. Digital social networks are places where people from around the world can connect to share and receive content and engage in dialogues with likeminded people. The lineage of digital social networks can be traced back to the first email being sent in 1971, connecting two people through content, which has grown into website communities, blogs, AOL instant messenger, Myspace, and then exploding with Facebook in 2004 (O'Dell, 2011). Activity based analysis of social media content allows businesses, law enforcement, intelligence agencies, and individual users to obtain precise information about habits, interests, and locations of users without user knowledge. A YouTuber named Jack Vale (2013) conducted an experiment by using Instagram to find posts that were near his current location. He used the post to look through the individual's timeline to figure out personal information about them that would be impossible for his to otherwise know. He would search out the individual on the streets pretending to be a psychic. The people were usually confused, uncomfortable, or angry that someone they didn't recognize knew so much information about them. One of the individuals that Jack confronted said, "Thanks for invading our privacy. I'll call the police if you do that again" (Ibid.).

As smart phone usage surpasses traditional desktop computer usage, content shared on social media is becoming more real-time and robust, producing insight to information that would otherwise be unavailable. The robustness of social media content helps consumers make quicker and better informed decisions, especially with regard to location based analysis. We have the capabilities to ascertain the location of a message, image, or video posted to a social network via coordinates explicitly enabled by the user through location services, by means of inferred geolocation through text-mining, or analysis of location-based social networking (LBSN) (Musolesi & Rossi, 2014). This type of information shows where the user lives, works, and what places they frequent. Social media consumers use this location-derived data to improve business intelligence by acting quickly to customer feedback, increasing marketing

opportunities in cities where products may not be selling well, alerting consumers when there is a deal near their location, breaking up criminal activity, and helping governments exploit terrorist networks. Students from Lahore University of Management Sciences in Pakistan and Cadets from the United States Military Academy (USMA) took first and second place respectively in a "Peer to Peer" competition sponsor by the US State Department, Department of Homeland Security, and EdVentrue Partners, that used social media networks to target extremist and non-affiliated vulnerable individuals to see Islam as a moderate religion and not a means of Jihad and to speak out against terrorism. USMA's project reached over 900,000 Facebook users in two months from 25 countries (Markoe, 2016).

Humans by nature tend to follow patterns, such as waking up at a certain time, working at one location for a given number of hours, returning home at a certain time, and on weekends frequenting the same place over and over. Individuals do not travel very far on a day to day basis from where they live due to many factors like work, day care, family, etc. More than half of all Americans live within eighteen miles of their mothers, which ranged from eight miles to forty-four miles on average by region (Bui and Miller, 2015). With most people being relatively static, a single area will likely encompass the same people repeatedly posting near the same places over and over again.

Most social media does not contain explicit latitude and longitude coordinates. With over 310 million active monthly Twitter users (303 million tweets per day/2 percent geotagged) and 400 million active monthly Instagram users (80 million photos uploaded per day/5 percent geotagged), there are over 10.06 million geotagged event locations per day (3.67 billion per year) worldwide from just these two social media sites alone (Press Page, 2016; Twitter, 2016; Heine, 2014; Cairns, 2013). Many users have a lack of understanding of the privacy concerns associated with enabling the location services for phone apps or when posting to a social media site, while others are willing to assume the risk it presents in exchange for more tailored content. Much research on LBSN analysis has looked at the aggregation of data at a city level. The research uses a large amount of data, usually from a single source, to infer a user's resident city, based on previous location check-ins, geographically specific vernacular, and even the social media posts by others in one's network (Pontes et al., 2012; Thome et al., n.d.; Musolesi & Rossi, 2014). These ideas assume that people's activity is determined by a semi-fixed distance to their home and that people in like areas often share like traits and connections, while neither a good or bad thing, it does make it easier for the human experience to be modeled.

Social media lets anyone with access to the internet post whatever, whenever, and often without understanding the bigger picture. Most large companies have an individual or team of people whose job it is to interact with the public, because companies understand information that is made public can be harmful. On occasion these types of posts have proven to expose a lot of information about the actions of a larger group. On July 5, 2014, Russian soldier Alexander Sotkin posted a picture of himself on Instagram inside an armored personnel carrier. The picture alone does not provide much content other than an image of his face, but the metadata exposed his location, which was well within the Ukraine boarder (Szoldra, 2014). This helped confirm the pro-Russian military support to the rebel fighters in the contested areas of eastern Ukraine, which is countered information stated by Vladimir Putin's administration. Another such incident of a social media accident came when an ISIL fighter posted a social media post revealing the location of a command and control center at an unspecified location likely within either Iraq or Syria. That information helped the U.S. Defense Department find and reduce the site with three Joint Direct Attack Munitions (JDAM) (Castillo, 2015).

METHODS

Using the concept that near things are more related than distant things (Tobler, 1970), it is necessary to look at how to properly aggregate the month-long (January 2016) dataset for the project spanning an area approximately 775 by 715 kilometers with over 131 thousand points without a loss of individual incidences. Although only about two percent of posts are geotagged, if this two percent is consistent with their geotagging, we can use the sample to represent the entire population.

I obtained data from DigitalGlobe's social media platform. The data was cleaned and preprocessed by DigitalGlobe to work in Esri's ArcGIS Desktop products. Through a local add-in in ArcMap, I queried my area of interest to collect all the data for January 2016, which resulted in a dataset of 131,401 individual posts. In the dataset there are 9,902 unique screen names, which range from posting as few as 1 post during the specified time period to other users posting over 2,000 times. 34% of all users in the dataset posted only one time during January 2016. The mean number of posts per user is 13.3, and the median is 3 posts.

Second, I used a basic fishnet binning over the data set in (Figure 1) to show how volume changes within the fishnet when the cell size changes. The major problem with this spatial binning methodology is that each bin represents a discrete value that is independent of the neighboring bins. The spatial bins should account for the relationship of each point compared to its neighboring points. This can be accomplished with spatial point clustering. Irregularly shaped clusters provide a more accurate aggregation for the temporal analysis and event detection compared to that of square bins, because they get their shape from known locations of human interaction. I used spatial clustering of the data in order to determine irregularly shaped polygons that were spatially consistent with the density of the dataset. The machine-learning python module scikit learn, was used to cluster the dataset using Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996). DBSCAN "finds core samples of high density and expands clusters from them," which allows for the detection of irregularly shaped clusters consistent with abnormal shapes in spatial patterns (Scikit Learn, n.d.). The DBSCAN algorithm has a number of configurable parameters such as the epsilon value (ε) and the minimum samples value. The epsilon value represents "the maximum distance between two samples for them to be considered as in the same neighborhood" (Scikit Learn, n.d.). The epsilon value is key in determining how compact or expansive the resulting clusters will be. The other configurable parameter I used was the minimum samples value, which sets "the number of samples (or total weight) in a neighborhood for a point to be considered as a core point. This includes the point itself." DBSCAN uses these parameters to determine what are the core points, points that meet the minimum sample value, and what are the border points, points that are not core points, but are reachable from the core point by the epsilon value, and finally if it is neither a core point or a border point, it is then is considered as noise (Ester et al., 1996).

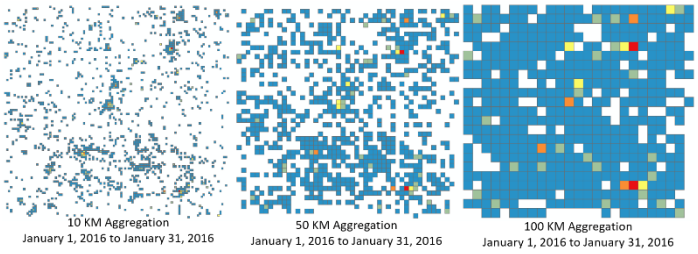

Figure 1. Varying square bins of the dataset showing how aggregation has an effect on our understanding of information.

The DBSCAN clustering algorithm used with an epsilon value of 0.08 and a minimum sample size of 10, returned 164 unique clusters and 528 points considered as noise. The values 0.08 and 10 were chosen because the results appeared visually consistent with the original data. The initial run of the data had only four clusters, requiring re-calibration. Running DBSCAN on an un-projected version of the same data resulted in five more clusters than the projected data using the Europe Equidistant Conic.

I joined the cluster category assigned with each point back to the original points. The points were then converted into Thiessen polygons and dissolved by cluster category into individual spatial bins that represented each cluster. These polygons were used to standardize the data in order to identify changes in the events happening over time, as different data exists in the same place over time, and were divided into hour time bins within their respective polygons resulting in 744 time steps for each cluster. Over the same time period the top three hashtags from each cluster were extracted. The data, largely written in either Russian or Ukrainian, is easier for a non-speaker to look at smaller segments of trending words in order to provide context to the numerical results.

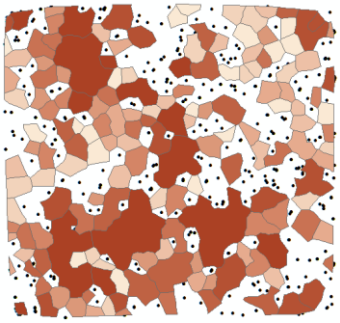

Figure 2. Irregularly shaped spatial clusters showing total volume. The black dots are the points that were considered noise.

After processing each hour time slice across each polygon, a moving average was applied to each cluster, analyzing the count value of each time period to identify outliers within each respective cluster. The moving average, looks at the average of the previous *n* number of hours to determine the current value. For example, a moving average value of 5 for 10 PM, it would sum the count values from 9 PM to 5 PM and divide that total by 5. 10 PM's value is an outlier if its count is outside of one standard deviation of all time values. A moving average of 3 was used to look for values that would be considered events. A moving average of 3 was chosen after running multiple iterations of the analysis. Anything smaller than 3 appeared to be independent of previous values and anything bigger than 3 seemed to marginalize the data.

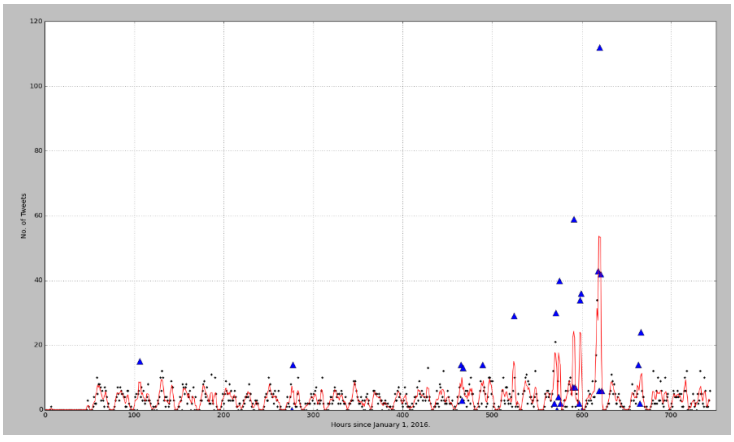

Figure 3. Cluster number 14 showing the moving average line (red), point value (black), and the outliers (blue), for the dataset from Jan 1 to Jan 31, 2016

An hourly count of cluster showing discrepancies that should be investigated as events. (Figure 3) does not have a perfect prediction rate, but produces a finite number of event detections and for faster analysis.

I highlight Cluster 14 (Figure 3) as one of the more dramatic examples of event detection. The moving average is rhythmic, since people are awake and asleep during certain time periods. Across the majority of the month, the data ranges from 0 to 10 social media posts per hour with massive spikes of activity in this clustered region 14 beginning January 24, 2016 with the largest spike of 112 posts between 8 and 9 PM on January 26, 2016.

RESULTS

Geotagged social media volumes can be used solely as a representative indicator that an event has happened. Using volume change exclusively in conducting a time series analysis, it was evident that significant changes in volume could be detected to indicate a possible event. A total of 4,817 time periods, or 3.95% of all time periods, were detected as possible events. The three densest clusters consist of 54.7% of all posts and have an average of 8.3 events detected per cluster over the month time span, compared to the 29.8 events detected per cluster for the other 161 clusters. With 62.2% of all clusters having 100 or fewer total posts over the month time frame, it is telling that having enough data is imperative to minimizing the number of excessive detections.

DISCUSSION

I found challenges obtaining data for this project. Although I sought data from sources like Twitter and Instagram, via their public APIs to scrape geotagged data, I would have only received a sample of data, and not the complete stream. Instead I reached out to Twitter's GNIP to get 568 days of historic data covering all of Russia. The 568 days of historic data was for the original scope of the project, spanning from January 2014 to the day I made the original request at the end of July 2015, to look at known Russian garrisons to identify mass troop movements. 120 million social media posts would cost $52,000. Having a full dataset, two or more years' worth of data, I believe would produce more significant results. One could compare how volume changes based on holidays, or other days of the year where volume changes would be expected.

If this dataset was from the beginning of the Russian intervention into Ukraine in early 2014 the spike in volume in (Figure 3) could be an indication of some type of mass movement has happened that the locals are talking about, which would be a good indicator for intelligence officials to look at. However, when we look at the hashtags around the spike in data (112 posts) there is a hashtag *#PurposeTourToUKRAINE,* which is the top hashtag in 12 of the top 13 highest time periods within cluster 14. The Purpose Tour is the name of Justin Beiber's 2016 world tour, which does not have a concert scheduled for Ukraine.

CONCLUSION

Overall, using mass data aggregation and time series analysis of geotagged social media data I was able to identify events or changes from the norm. Unfortunately, I do not have a quantifiable measure of success, since this type of analysis is not linked with either a national security or business result. An event with respect to a business decision, a national security issue, or a media report is only valuable if it can be linked to an outcome. Most of the graphs produced as an end product of the analysis showed unremarkable results. Spatial clusters that were over areas of large populations seemed to give results that are more easily interpreted because of the volume of data. Clusters that didn't have a lot of volume, which was the majority of the clusters, identified a lot more possible events, because a change from zero posts to one post would identify that change as an event. This type of analysis produces tangible areas and times of interest, but would be best served as an accessory to other data to focus the results. Further tuning of the time series analysis to ignore changes in small volume shifts in clusters would likely produce much more relevant results. Social media users are going to produce data that is going to give context, time, and location that a bad actor would otherwise not disclose. When looking at this type of analysis at a global scale a distributed computing environment

and access to a firehose of data are paramount to make sure as complete of a picture as possible can be drawn as fast as possible.

REFERENCES

Bui, Q., & Miller, C. (2015, December 23). The Typical American Lives Only 18 Miles From Mom. *The New York Times*. Retrieved from http://www.nytimes.com/interactive/2015/12/24/upshot/24up-family.html?_r=0

Cairns, I. (2013, August 22). Get More Twitter Geodata From Gnip With Our New Profile Geo Enrichment [Web log post]. Retrieved from https://blog.gnip.com/twitter-geo-data-enrichment/

Castillo, W., (2015, June 5). Air Force Intel uses ISIS 'moron' post to track fighters. *CNN*. Retrieved from http://www.cnn.com/2015/06/05/politics/air-force-isis-moron-twitter/

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 2nd International Knowledge Discovery and Data Mining, Portland, 1996. Portland, OR: AAAI Press

Heine, C. (2014, October 28). 14 Instagram Data Findings That Every Marketer Needs to Know Here's a Tease: @mentions inspire 56% more engagement. *AdWeek*. Retrieved June 25, 2016 from http://www.adweek.com/news/technology/14-instagram-data-findings-every-marketer-needs-know-160969

Press Page. (2016, n.d.). Retrieved from https://www.instagram.com/press/?hl=en

Markoe, L. (2016, February 3). To fight ISIS, West Point cadets secretly build Facebook page. *Religion News Service*. Retrieved from http://religionnews.com/2016/02/03/fight-isis-west-point-cadets-secretly-build-facebook-page/

Musolesi, M. & Rossi, L. (2014). It's the Way you Check-in: Identifying Users in Location-Based Social Networks. COSN '14 Proceedings of the second ACM conference on Online social networks, Dublin, Ireland, 2014. New York, NY: ACM

O'Dell, J. (2011, January 24). The History of Social Media [Infographic]. *Mashable*. Retrieved from http://mashable.com/2011/01/24/the-history-of-social-media-infographic/#mtqc5woPYkqG

Pontes, T., Vasconceleos, M., Almeida, J., Kumaraguru, P., & Almeida, V. (2012). We Know Where You Live: Privacy Characterization of Foursquare Behavior. UbiComp '12 Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburg, PA, 2012. New York, NY: ACM

Scikit Learn. (n.d.). sklearn.cluster.DBSCAN. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Szoldra, P. (2014, July 31). A Russian Soldier's Instagram Posts May Be The Clearest Indication Of Moscow's Involvement In East Ukraine. *Business Insider*. Retrieved from http://www.businessinsider.com/russian-soldier-ukraine-2014-7

Thome, D., Bosch, H., Kruger, R., & Ertl, T. (2014). Using Large Scale Aggregated Knowledge for Social Media Location Discovery. 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, 2014. Waikoloa, HI: IEEE

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–40.

Twitter. (2016, March 31). Twitter Usage/Company Facts. Retrieved from https://about.twitter.com/company

Vale, J. [Jack Vale Films] (2013, November 18). *Social Media Experiment*. [Video File]. Retrieved from https://www.youtube.com/watch?v=5P_0s1TYpJU