Filomat 32:5 (2018), 1607–1617
https://doi.org/10.2298/FIL1805607P

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Repeatable Measurement of Twitter User Impact

NASA and the Great American Eclipse of 2017

Doug Pickering[a], Mykel Shumay[a], Gautam Srivastava[a]

[a] *Brandon University, Brandon, MB, CANADA*

Abstract. **NASA** is viewed as part of the frontier of human knowledge by several generations, and is relied upon to educate the public on astronomical matters. For decades **NASA** has provided not only North America but the entire world with information and events pertaining to our Universe and beyond. With the Great American Eclipse of 2017, **NASA**'s production was crucial to the general public's awareness and understanding of the event. To date, it may have been **NASA**'s largest production of an event spanning many social media platforms and hundreds of Media outlets. With the eruption of data mining avenues and techniques available, being able to study and quantify such major events from a "reach" perspective has become of utmost importance for many of the groups involved. Our goal with this paper is to understand how the public perceived the social media coverage that **NASA** had provided, specifically in the world of **Twitter**, a free social networking microblogging service that allows registered members to broadcast short posts called *tweets*. We accomplish this through sentiment analysis and the spotting of trends within Twitter data. Furthermore, we follow a framework of study that allows simple and cost-effective analysis of discrete events of arbitrary nature.

1. Introduction

On Monday, August 21, 2017, all of North America was treated to an eclipse of the sun. Anyone within the path of totality saw one of natures most awe-inspiring sights - a total solar eclipse. The path, where the moon will completely cover the sun and the sun's tenuous atmosphere - the corona - can be seen, stretched from Lincoln Beach, Oregon to Charleston, South Carolina. Observers outside the path still saw a partial solar eclipse where the moon covers part of the sun's disk.

Total solar eclipses are not entirely uncommon, with the Earth experiencing one every 18 months on average, although these are usually viewable by only half a percent of Earth's surface. [1] The last total solar eclipse to cross coast-to-coast was in 1918, although the most recent total solar eclipse with a path of totality within the continental United States was in 1979.

With a total solar eclipse stretching from Oregon, U.S.A. to South Carolina, U.S.A., it makes for quite the compelling event. Such a strange phenomenon passing through the United States certainly would have

2010 *Mathematics Subject Classification*. 91D30.
Keywords. Data Mining; Sentiment Analysis; Twitter; NASA.
Received: 13 September 2017; Accepted: 30 January 2018
Communicated by Senlin Yan
Research supported by NASA
Email addresses: pickering@brandonu.ca (Doug Pickering), shumaymj55@brandonu.ca (Mykel Shumay), srivastavag@brandonu.ca (Gautam Srivastava)

been awe-inspiring, as most of the population had likely never before experienced a total solar eclipse, nor perhaps anything of such magnitude.

The eclipse's umbra - the shadow cast when the sun is totally obscured by the moon - is roughly 70 miles wide; this totality lasts for less than 3 minutes at any point along the path of totality. Between the first landfall of the umbra in the state of Oregon, and the last of the path over land in South Carolina, the path of totality had run its land-based course in about 90 minutes. Meanwhile, the penumbra - the shadow cast by partial obstruction - was viewable by the majority of North America, giving quite a mass of people a spectacle to behold. In comparison to the audience of other total solar eclipses, this is a huge percentage of Earth's population. For many, this was the most fascinating, or perhaps only, astronomical phenomenon that they have experienced. It gives anyone along the path of totality the opportunity to see the Sun's magnificent corona and chromosphere, both of which are normally overpowered by the surface of the Sun.

	Eclipse Begins	Totality Begins	Totality Ends	Eclipse Ends	
Madras, OR	09:06:43	10:19:36	10:21:38	11:41:06	PDT
Idaho Falls, ID	10:15:10	11:33:04	11:34:48	12:58:05	MDT
Casper, WY	10:22:21	11:42:44	11:45:09	01:09:30	MDT
Lincoln, NE	11:37:16	01:02:40	01:03:48	02:29:46	CDT
Jefferson City, MO	11:46:07	01:13:07	01:15:38	02:41:05	CDT
Carbondale, IL	11:52:25	01:20:06	01:22:41	02:47:25	CDT
Paducah, KY	11:54:03	01:22:16	01:24:38	02:49:32	CDT
Nashville, TN	11:58:31	01:27:25	01:29:23	02:54:02	CDT
Clayton, GA	01:06:59	02:35:49	02:38:23	04:01:27	EDT
Columbia, SC	01:13:08	02:41:51	02:44:21	04:06:21	EDT

Seconds may vary depending on your location. View the interactive map for more information:
https://eclipse2017.nasa.gov/sites/default/files/interactive_map/index.html

Figure 1: Viewability of Eclipse Across North America

Public engagement within science may be defined as intentional, meaningful interactions that provide mutual learning between scientists and members of the public. The measurement of the impact of these interactions is a difficult problem in itself. In today's world, millions of people take to social media every day to voice their opinions, thoughts, feelings, and experiences, providing readily-available sources of information on their lives. Twitter posed as the most attractive choice to us, as the limiting of tweet size forces the user to make frequent, concise tweets, allowing us to track the change in sentiment of a large population in small increments of time.

It followed that we make use of Twitter to gather quantitative results for **NASA**'s social media campaigns. Over our collection timespan, from late June to the end of August, we obtained roughly 13 million English tweets from around the world.

2. Methodology

One of our main goals was to make our approach as repeatable as possible. Given the input variables of *Users* and *Keywords*, one should be able to repeat this process for a future event with ease. Twitter was chosen as the platform to focus on as it is a major player in the social media space, and carries an emphasis for keeping text short and to the point. Moreover, for the vast selection of social media platforms available, it is known that Twitter is the most closely related for most scientific communities.

To gather as much relevant data as possible at low cost, tweets were gathered in real-time from Twitter's Public Streaming API. [5] Python 3.4 and the Tweepy library were used to access this API, which required a small portion of simple code and provided adequate reliability.

For reasonable redundancy and overlap on downtime, we had two machines listening to the entire Streaming API without filtering, and another three machines listening to a filtered version; this was filtered both by specific user accounts and keywords. Each machine was under its own network in a unique physical location. This redundancy protected us from multiple outages and network overloads, and, even

if not used, storing the entire Streaming API's output allows us the ability to go back and attempt the recovery of missing data without the worry of rate limiting. Through this, only a small sample of the actual pool of tweets were obtained, as the Public Streaming API only delivers somewhere around 1%. [6] There may be slight biases in how the Streaming API functions, although this seems minimal and not relevant to our analysis. [2]

Followed Users: *@NASA*, *@NASAedu*, *@NASASun*, *@Exploratorium*

Tracked Keywords: *NASA*, *Exploratorium*, *eclipse*, *umbra*, *totality*, *astronomy*

If a keyword is only part of a word in a tweet, it is still captured. For example, searching for *umbra* will also provide us with tweets that contain only *penumbra* or *#umbra*. Along with the listed keywords, we also tracked some possible misspellings and a small collection of specific hashtags.

Although all listeners were usually operational at the same time, it was not uncommon for one to receive tweets that the others had missed out on. This may be due either to simple network issues or from some nuances of the Twitter's Streaming API. This difference seemed to crop up in each of the listeners and was usually well below 1%; however, sometimes this became 10% or greater for extended periods of time, particularly during high traffic periods.

NASA's social media campaign was officially launched June 21^{st}, 2017, with further additions on July 21^{st}, 2017. Our Twitter listeners began on June 18^{th}, 2017, however we changed our filtering criteria on June 22^{nd}, 2017. The main reason for the change in criteria was to remove more generic words such as *#sun* and *#corona*, which more often than not pointed to tweets regarding warm weather and beer respectively.

2.1. Sentiment Analysis

To acquire some measurement on the public's mindset, we made use of VADER (Valence Aware Dictionary and sEntiment Reasoner) [4], a simple rule-based model for general sentiment analysis. The inherent nature of social media content poses serious challenges to practical applications of sentiment analysis. VADER outperforms individual human raters (F1 Classification Accuracy = 0.96 and 0.84, respectively) and generalizes more favourably across contexts than any current benchmark.

On top of this, we made use of a sentiment dictionary to determine the **Happiness** level of the tweets. Most dictionaries were used from [3], but others could easily be used in this analysis to fit other needs. Each tweet was split into its constituent words, each of which was looked-up in the dictionary for its corresponding values; this sum of values was then averaged across the number of words in the tweet.

2.2. STEM Talk

A possible clue to the reach of **NASA**'s social media campaigns is the shift in vocabulary; an increase in the occurrence of STEM words (Science, Technology, Engineering, and Mathematics) within tweets related to the eclipse may indicate that the public is becoming more comfortable with these terms, and learning about the phenomenon in greater detail.

2.3. Geolocation

Although it seems that only a minority of users actually provide their geolocation information with their tweets, we were interested to see the results of plotting these, looking particularly for strong outliers and sections of sparseness within the continental United States. Studying these may give important information regarding the public's engagement surrounding the event.

3. Results

In this section we present the results of our Twitter data collection analysis. Along our collection period we had collected roughly 200 GB's of unique data from the filtered Streaming API, which makes for over 35 million tweets. 24 million of these tweets were English, which we further filtered down to 16.5 million after unwanted material was removed; a large portion was related to astrology or other unrelated topics,

which we removed along with any tweets containing profanity. We did not attempt to distinguish between real users and bots.

It was found that only 120,000 tweets provided geolocation data, with 88,000 of those being English. Regardless of the small quantity, these data gave a good representation of the levels of engagement across the world. The graphs shown here illustrate the different avenues that may be taken with the data mining techniques used, along with the quality of results that may be produced.

3.1. Sentiment Analysis

To eliminate the impact of spam on our results of sentiment analysis, we further filtered the tweets by taking the inner text of every tweet and removing the duplicates; each analyzed text is unique and has the same weight as any other tweet. This also takes away the impact of retweets, leaving only new additions to the discussion.

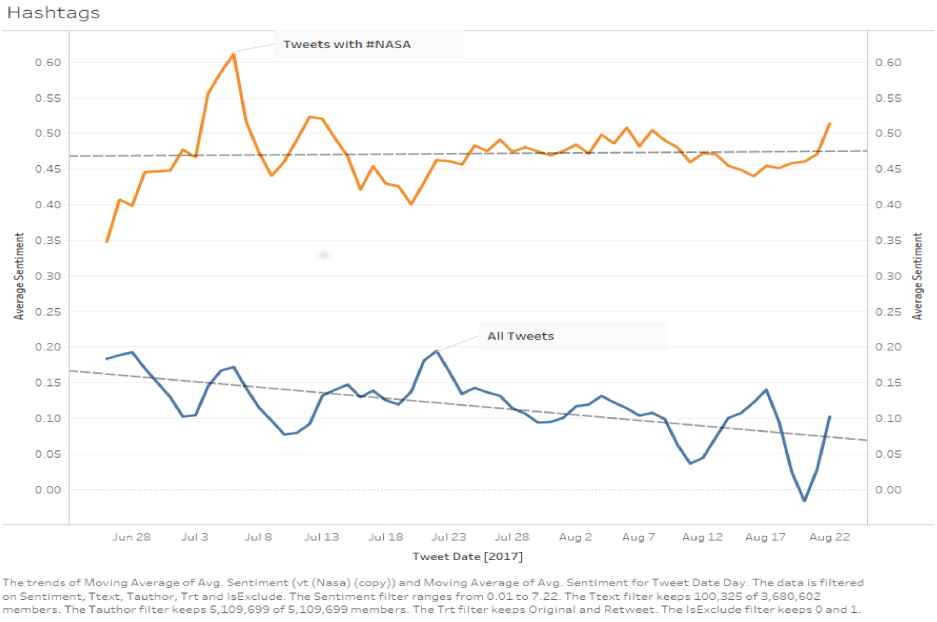

Figure 2: VADER Sentiment Comparison

In Figure 2, the average sentiment of all tweets was determined by VADER for each day from late June through to August 22nd. Interestingly, the trend line for all gathered tweets was in declining sentiment, although tweets specifically containing *#NASA* saw a slight improvement. The marketing campaign additions of July 21st seem to have possibly had a significant effect on sentiment.

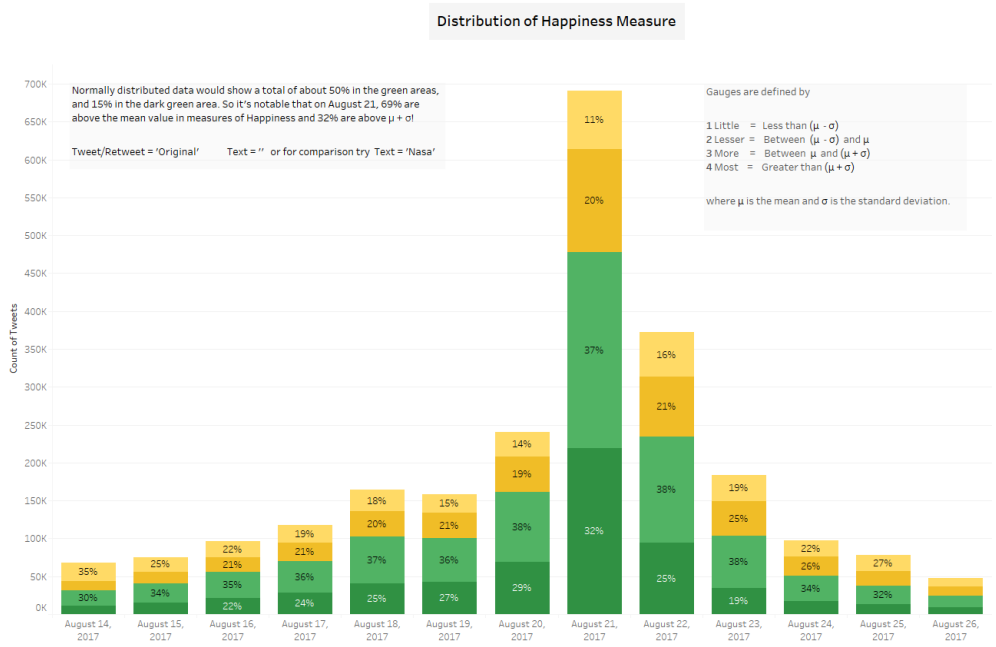

Figure 3: Daily distribution of happiness measure in late August.

Figure 3 illustrates the Happiness level distribution of each day in relation to the Happiness mean and standard deviation. For each day surrounding the eclipse, there is a clear majority in high happiness levels, indicating that the discussions surrounding the eclipse were very much positive.

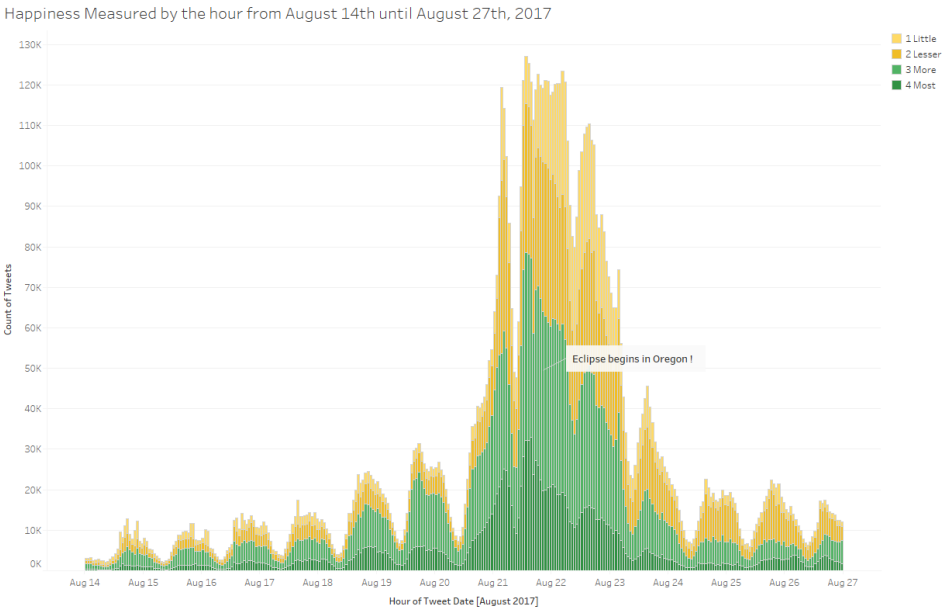

Figure 4: Hourly distribution of happiness measure in late August.

Looking at Figure 4, there are clear swells in activity during North America's daytime, with increases in magnitude approaching the eclipse date. For each day prior to the eclipse, the greater portion of tweets

were consistently of positive sentiment, with the overall count of tweets increasing at a steady rate. On August 20^{th}, the day before the eclipse, the number of tweets increased substantially; on the night of the 20^{th} and early morning of the 21^{st}, the number of tweets maintained a very large amount, showing great excitement for the upcoming event even throughout the night.

Starting in the early morning of the 21^{st}, the number of tweets spiked substantially as the start of the eclipse drew near, with a clear majority being in positive sentiment. While the volume of tweets was maintained throughout the day, the sentiment somewhat shifted negatively. Possible factors resulting in the shift in sentiment could have been reactions to bad weather and traffic; more detailed analysis would have to be undergone to verify this. After the 21^{st}, the percentage of negative sentiment became more dominant over the positive, and by August 24^{th} the volume returned to levels near that of the days leading up to the eclipse.

3.2. STEM Talk

In educating the public about the eclipse, **NASA** would hope that other users would become more comfortable in the new vocabulary that **NASA** would be using; the greater use of these STEM words would point to a more robust understanding of the eclipse and success on **NASA**'s part.

Day of Tweet Date

Average Happiness 0.03 18.52

Science Word	Grand Total	Aug 14	Aug 15	Aug 16	Aug 17	Aug 18	Aug 19	Aug 20	Aug 21	Aug 22	Aug 23	Aug 24	Aug 25	Aug 26	Aug 27	Aug 28
eclipse	497	81	97	135	179	267	305	504	2,249	2,159	774	210	158	147	114	84
solar	215	41	44	53	71	119	139	186	775	1,168	359	74	49	64	54	30
eclipse2017	100	10	11	19	24	36	35	70	452	580	185	35	17	12	8	7
day	74	16	20	26	36	54	61	74	311	223	71	26	69	79	27	15
solareclipse	70	4	5	6	9	14	12	21	333	461	134	24	11	8	6	5
solareclipse2017	65	3	3	4	5	8	7	13	305	442	129	23	10	8	5	5
total	59	15	16	29	36	36	33	59	222	238	100	33	26	12	10	15
sun	55	10	11	12	15	27	88	92	253	218	57	15	9	6	6	7
glasses	38	16	13	15	20	27	27	56	229	108	40	9	6	4	2	2
moon	28	3	4	5	8	11	19	21	113	154	44	12	11	8	8	6
star	25	10	7	7	11	11	16	34	110	76	32	9	10	30	10	4
space	23	27	17	22	17	14	10	16	34	67	28	12	20	26	21	15
rocket	19	2	1	0	1	4	1	1	1	1	0	1	141	119	16	2
totality	17	2	2	3	6	7	6	10	60	85	45	11	5	3	2	3
plane	14	2	3	3	3	6	7	6	43	51	19	16	14	13	10	10
event	12	2	3	3	4	6	5	8	18	12	50	20	9	5	24	6
science	12	11	6	11	8	11	12	24	44	24	11	4	4	4	2	2
ice	10	5	4	6	6	8	7	12	35	34	13	6	7	3	5	3
planet	10	1	3	3	3	4	3	4	37	41	12	6	6	11	9	9
station	10	15	3	13	6	3	2	1	18	49	12	3	10	5	7	3
totaleclipse	9	0	0	0	1	1	1	1	38	50	13	7	4	3	4	6
sky	7	2	2	2	2	3	3	5	27	38	13	4	4	3	2	1
dark	7	1	2	2	3	4	4	7	31	30	11	4	2	1	1	1
black	6	1	1	1	1	2	2	5	18	22	13	4	3	6	3	1
hole	6	1	1	2	2	4	4	10	28	18	7	2	2	1	1	1
unit	5	1	1	2	2	2	2	3	20	32	7	2	1	1	1	1
partial	5	1	1	1	2	3	3	5	17	26	7	1	1	0	0	0
full	4	1	2	2	2	6	4	9	15	9	4	2	3	1	1	1
corona	4	2	2	2	2	3	3	2	4	4	3	4	4	4	17	3
satellite	4	3	3	1	4	10	1	1	4	8	5	4	5	7	2	3
astronomy	4	2	2	3	3	3	2	3	6	4	4	3	3	8	8	3
free	4	2	3	2	3	3	4	4	11	9	3	2	3	2	2	1
north	4	1	1	1	1	3	2	5	11	17	8	1	1	0	0	1
universe	3	1	1	1	1	1	1	9	10	12	3	2	2	1	1	4
fall	3	1	1	1	1	1	2	7	13	4	2	1	1	1	0	2
body	3	0	1	1	2	2	2	5	12	6	2	1	1	1	1	1
force	2	0	0	0	0	0	0	1	5	21	7	1	1	0	0	0
ray	2	2	1	1	1	2	2	3	10	6	2	1	2	1	2	1
superior	2	0	0	0	0	0	0	1	1	24	9	0	0	0	0	0
safety	2	1	2	3	3	4	2	7	11	1	1	0	0	0	0	0
gas	2	0	1	1	1	1	3	5	9	6	3	1	0	0	1	0
pinhole	2	0	1	1	1	2	2	6	12	1	0	0	0	0	0	0
astronaut	2	1	2	3	1	1	1	1	2	6	3	4	1	1	1	0
orbit	2	2	1	1	1	2	2	3	5	2	1	2	2	2	1	0
mass	2	1	1	1	1	5	4	2	3	3	2	1	1	0	0	0
lunar	2	1	1	1	1	1	1	2	3	3	3	1	1	0	1	3
astronomer	2	1	1	1	2	1	2	2	4	7	2	1	0	0	0	0
rings	1	0	1	1	1	1	1	2	3	2	2	1	1	3	3	1
umbra	1	0	0	0	0	0	0	1	3	8	7	1	0	0	0	0
lens	1	0	0	0	0	1	0	1	3	10	4	0	0	0	0	0

Figure 5: Occurrence of STEM Words

Shown in Figure 5, the most common STEM words were *eclipse* and *solar*, unsurprisingly; each of these saw a dramatic and exponential increase in occurrence between the eclipse date and the week prior. As well, the average happiness measure saw growth in this time. Both the occurrences and average happiness measure dropped off in the days following August 21^{st}, as attention inevitably began to shift elsewhere. Other STEM words, such as *totality*, *plane*, and *planet*, showed a similar behaviour.

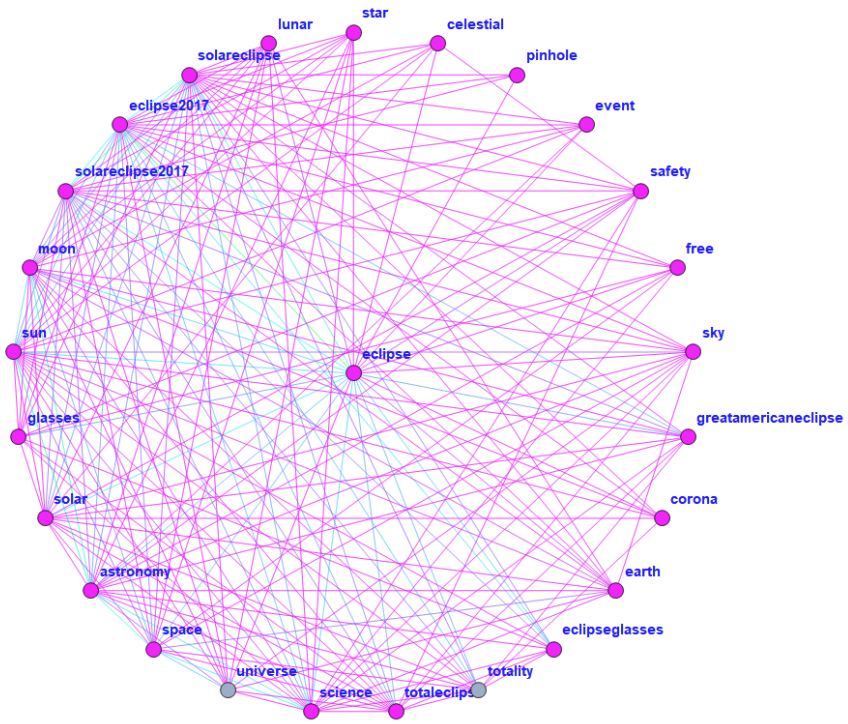

Figure 6: Occurrence of STEM Words

Figure 6 illustrates how the STEM words were found to be connected. The lighter blue lines indicate a stronger relation between the words. The strong pink lines are the most freqent found connections between words. We can clearly see that words like “moon” and “sun” are often paired with “eclipse”, and less often with “totality”. An interesting open problem would be the study of such graphs.

3.3. *Geolocation Analysis*

Although the percentage of tweets containing geolocation data is very much on the low end, being only 0.25% on August 21st, this information may still provides valuable insight into the engagement.

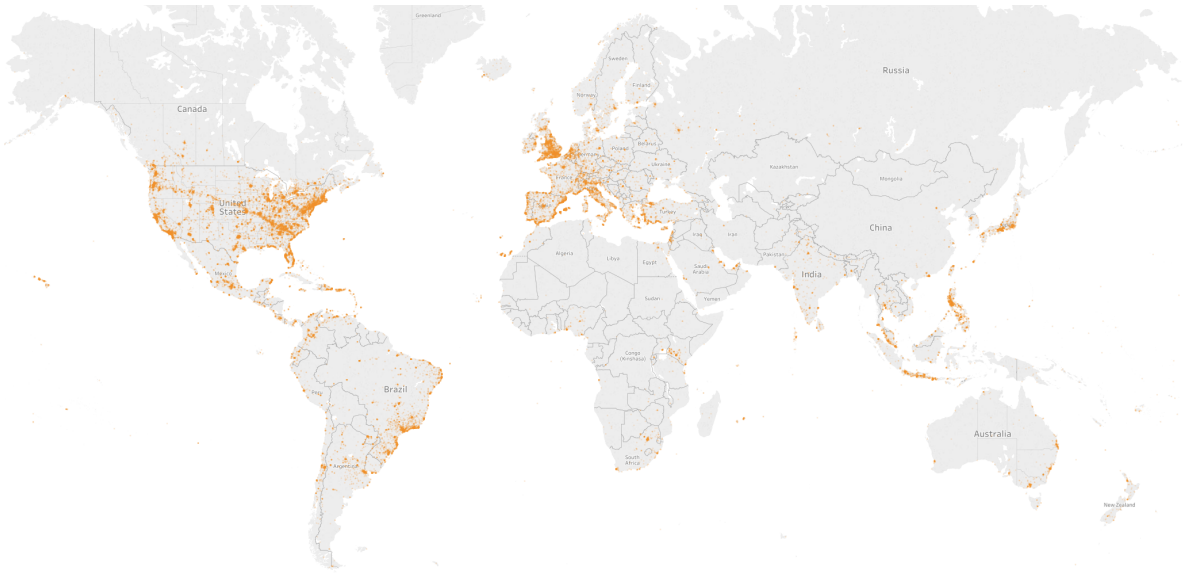

Figure 7: Geolocation of tweets worldwide across entire collection period.

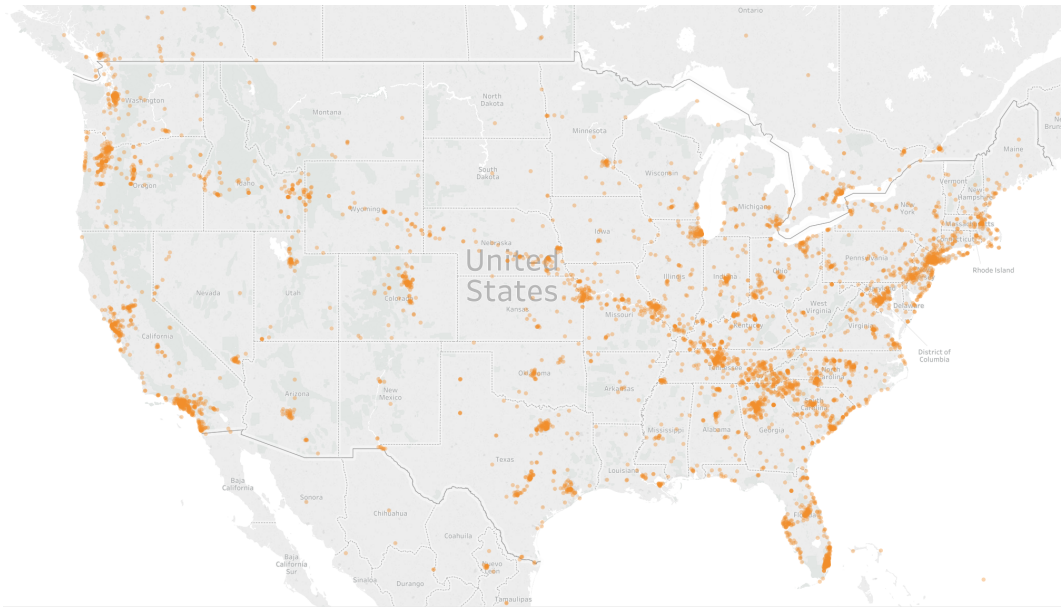

Figure 8: Geolocation of tweets within the United States on August 21st.

In Figures 7 and 8, we plot the points of user-provided geolocations of all collected tweets. In Figure 7 we can clearly see the pockets of the world with the most "chatter" regarding the eclipse. A clear streak of engagement existed along the path of totality during August 21st, as would be expected. There is significantly greater activity in the Eastern half of the continent, although this does follow the population density distribution of the United States.

Europe had experienced a surge in activity, attributable to the area having gotten a glimpse of a partial eclipse late in the evening. Indonesia was also a major talker, having recently experienced a total solar eclipse in March of 2016, and no doubt were still interested in the occurrence of another one.

3.4. Graph Clustering

One of our interests was to discover how Twitter users were interconnected and communicating with each other. For this, we deployed some basic graph analysis to determine the clustering coefficient of each day.

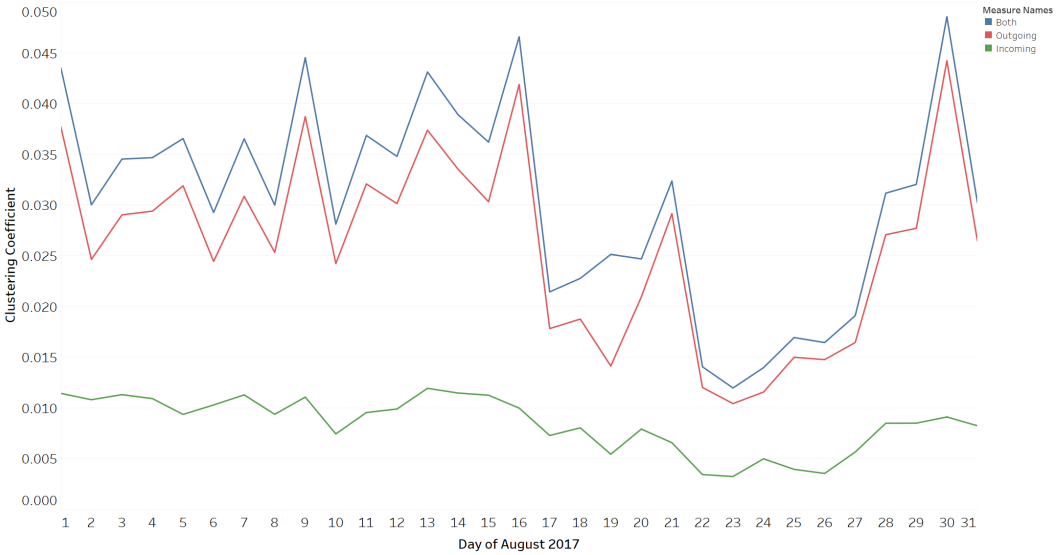

Figure 9: Clustering coefficient analysis of both replies and retweets.

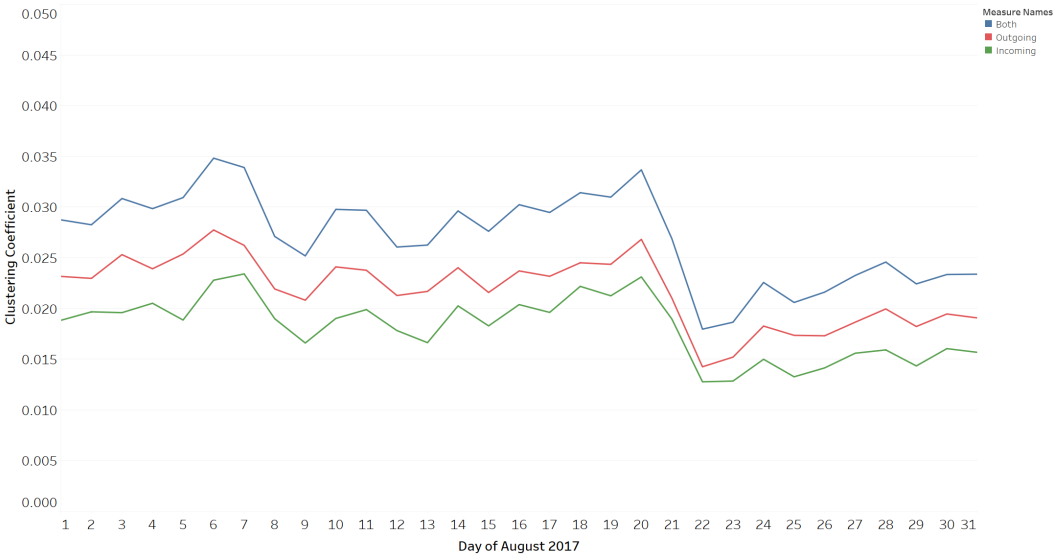

Figure 10: Clustering coefficient analysis of replies.

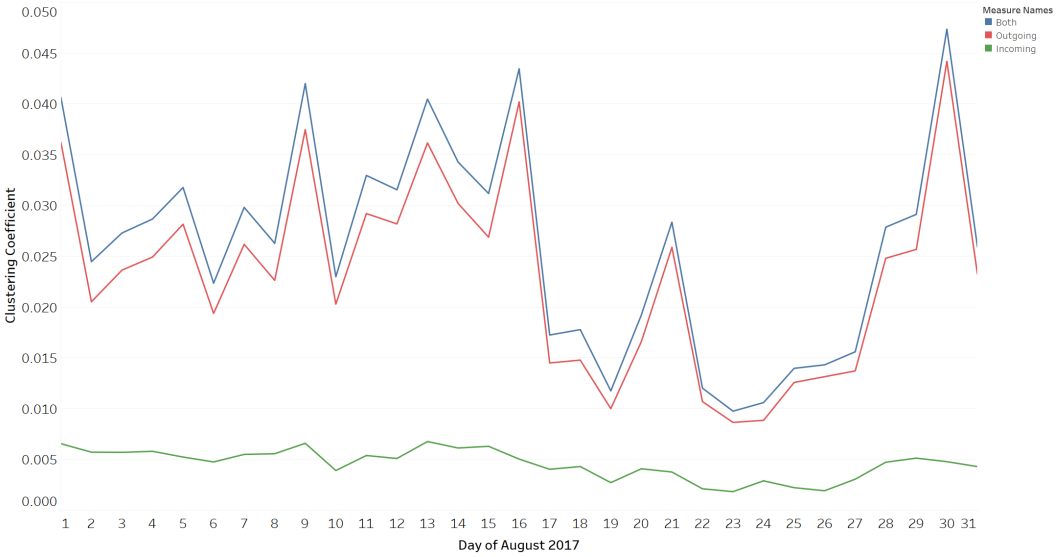

Figure 11: Clustering coefficient analysis of retweets.

In Figure 9, *Outgoing* clustering coefficients were derived from graphs that only included directed edges from userX to userY where userX replied or retweeted to a tweet of userY, while *Incoming* is the opposite. *Both* clustering coefficients were derived from undirected graphs containing both replies and retweets. Figure 10 shows this analysis on only replies, while Figure 11 shows this analysis on only retweets.

Looking at the trends in the clustering coefficient measures, and comparing them with the days, we can clearly see that a more diverse group of tweeters and tweets were noticed around the time of the solar eclipse. This leads to the conclusion that during the days closest to the solar eclipse, a more diverse group of Twitter users were enganging with the eclipse. This can be attributed to NASA's large production of the solar eclipse on twitter, since most of the data we collected was connected to NASA accounts. Under similar circumstances, if we had seen an increase in clustering coefficient, we would have believed the opposite, namely a smaller distinct group were reposnsible for most of the tweet action on Twitter.

4. Future Work

The in depth study into the feasability of analyzing a disctrete event in a social media platform like Twitter in a low-cost manner, has led to some very interesting open directions of research. First, the idea posed earlier regarding the study of graphs of either keywords or hashtags and the connection between them would be interesting to see how a given topic develops and changes over time from a graph theoretic perspective. Secondly, looking at the text provided in the tweets, and seeing if they can help identify where a user is Geographically located, namely for a discrete event whethere they are at the event or just tweeted about it. Lastly, looking at how user sentiment of an event in real time on social media, may help those in a position to actually influence the event itself, fix any issues addressed via social media in real time.

5. Conclusion

Twitter is a very influential social media platform, with many uses in today's digital world. On August 21^{st}, 2017, the Earth experienced a total solar eclipse that stretched geographically across the continental United States. **NASA** rolled out a comprehensive social media campaign to promote the eclipse and gain the public's interest. Their interactions with the public left a trail that could easily be studied and quantified. Moreover, the methods used here to quantify **NASA**'s media influence may be used for future discrete or ongoing events.

NASA clearly had a great impact on the conversation surrounded the Great American Eclipse, with their services being vital to the public's understanding of the event.

In this paper, we summarize some of our efforts in sentiment analysis. Overall, Twitter provides a great opportunity to better understand the general public's views, sentiments, and interpretations of countless issues and events, all in byte-sized pieces.

References

[1] M. Bakich. 25 facts you should know about the august 21, 2017, total solar eclipse, Aug. 2017.
[2] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science*, pages 1–15. Springer, 2010.
[3] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.
[4] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.
[5] K. Makice. *Twitter API: Up and Running Learn How to Build Applications with the Twitter API*. O'Reilly Media, Inc., 1st edition, 2009.
[6] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010.