A NETWORK ANALYSIS OF TWITTER DISCOURSE ABOUT THE DEATH OF GEORGE FLOYD

By
Jenna Korobova and Matt Zacharski

Advised by Dr. Katie St. Clair

Senior Comprehensive Exercise
Carleton College

March 2021

Abstract

User interaction data from Twitter—including likes, follows, and retweets—provides valuable information about what kinds of discussions are taking place online, and who is participating in them. This data can be represented using graphs/networks, which maintain the inherent structure of the data, and then analyzed using various mathematical and statistical methods. This research study uses Twitter data from the week following the murder of George Floyd on May 25, 2020 to construct a retweet network from user discussion about the implications of his death. We analyze this data using R by constructing a retweet network, investigating influential users, and modeling their attributes using Exponential Random Graph Models (ERGMs). We determine that @YourAnonCentral and @AttorneyCrump were the most influential users, and that larger differences in the average sentiment score of users were more likely to result in a retweet.

Introduction

As the power and presence of social media around the globe continues to grow, data about users and their interactions with one another is becoming increasingly valuable. This data often has an underlying structure in the form of friends, followers, and likes, so it is important that analyses of this data use graph objects that can preserve this structure. As such, social media research has become increasingly popular in the field of network analysis. In particular, the complexity and breadth of Twitter data makes it conducive to creating informative network graphs. Networks constructed using Twitter data help demystify who is guiding discourse, the characteristics of the communities that are forming, and what kinds of information users are interacting with.

We are interested in understanding how Twitter users were participting in and starting discourse about the murder of George Floyd by Minneapolis Police. In the days immediately following his death on May 25, 2020, protests condemning police brutality and systemic racism were widespread in the United States and around the globe. Throughout this time, Twitter was a resource many used for general information (about protests, how to help those protesting, petitions, etc.) as well as discussion (of personal experiences, solidarity, grief, etc.).

With all of this in mind, the goals of this analysis are as follows: **(i)** construct a retweet network using tweets about George Floyd, **(ii)** identify "influential" users in this network, **(iii)** identify what attributes of a user are most important to the network structure using Exponential Random Graph Models (ERGM), and **(iv)** analyze how networks differ by date.

Prior to proceeding to our analysis, we would like to acknowledge the severity of the topics of George Floyd, police brutality, and systemic racism. It is important to note that behind the numbers and the tweets, there are human stories. This paper will serve strictly as a statistical analysis, without making any ethical claims.

Background

Network Terminology

A network graph, *G*, is a structure that can be defined as a set of *N* nodes and a set of *E* edges where *E* consists of pairs of distinct nodes from *N*. A network is directed if *E* contains ordered pairs of distinct nodes from *N*. Nodes in the graph *G* can be associated with numerical or categorical characteristics called node attributes. Similarly, edges between nodes in *G* can be associated with numerical or categorical characteristics called edge attributes. The number of edges that are connected to any given node is called the degree of a node. In a directed graph, nodes have two different degree values: in-degrees and out-degrees. Each value represents how many incoming and outgoing edges are connected to a given node, respectively. Visualizing the distribution of node degrees of a network is one of the simplest ways to explore the network structure.

There are several other network specific terms related to conducting exploratory analysis on network structures. Namely: density, cliques, transitivity, and reciprocity. The density of a network is the proportion of observed edges to the maximum number of possible edges. Consequently this value ranges from 0 to 1 with greater values implying a more interconnected, or dense, network. A clique refers to a subset of nodes from a network that have the maximum number of possible ties between them. Similar to density, the transitivity value of a network is the proportion of closed triangles to the total number of open and closed triangles in the network. A closed triangle in a network is where three nodes are connected by three edges. An open triangle in a network is where three nodes are connected by only two edges. Another structural term related to networks is reciprocity. The reciprocity of a network is the proportion of observed mutual connections, which is the case of an edge in both directions between two nodes, to the total number of edges in a given network.

To conduct analysis beyond exploratory measures we use an Exponential Random Graph Model (ERGM) to predict the probability of a user retweeting another user. ERGMs are similar to logistic regression models. They are capable of predicting the probability of two nodes within a network being connected. One form of the ERGM can be written as:

$$P(y_{ij} = 1 \mid {Y_{ij}}^{C}) = (\tfrac{1}{c})\ exp\{\sum_{k=1}^{K} \theta_k z_k(y)\}$$

where $\frac{1}{c}$ is a normalizing term, θ_k is a vector of length K containing coefficients, and $z_k(y)$ is a vector of length K representing sufficient statistics. The underlying intuition behind the ERGM is that by comparing the observed network to random networks we can determine which included attributes significantly affect the probability of two nodes being connected. A random network is a network with the same number of nodes as the observed network. However, the edges of a random network are determined by Bernoulli trials with a p typically equal to the density of the observed graph.

Twitter Overview

Twitter is a social networking platform that allows users to post tweets, short microblogs consisting of 280 characters or less. Since its founding in 2006, Twitter has become a hub for online discussions ranging from sports and memes to heated debates and discussions around current events. As of February 7th 2019, 22% adults in the US used Twitter (Pew Research, 2020). Additionally, there were 326 million monthly active users during the time frame we collected our data in (BusinessOfApps, 2021).

Similar to other social networking sites, Twitter allows ways for its users to follow others, be followed, and interact with others' tweets in a variety of ways. Specifically users can like, retweet, mention, and reply to other users' tweets among other things.

Users can also optionally include a hashtag to their tweet, a short string of words with no punctuation or spaces following the # character. Hashtags are used to index topics on Twitter and allow users to denote the topic of their tweet, although it is possible to include a hashtag unrelated to the content of the tweet (Twitter, 2021). Consequently, using hashtags to collect samples of tweets may not capture tweets that are exclusively about the topic of interest. However, this seems to be relatively rare and hashtags still seem to be the optimal way to track and analyze topics on Twitter.

Users can also reply to the tweets of other users through an action that Twitter calls mentions (Twitter, 2021). To mention another user, users will type the @ character followed by another user's Twitter handle. However, users can also reply to another's tweet through the comment section.

Previous Research

Due to Twitter's widespread usage, ability to shape public discourse, and accessibility of data, it has become an area of interest for many researchers from a variety of fields, such as journalism, politics, and sociology. There are several examples of Twitter being used as an organizing tool by protestors across the world such as the Iranian election protests of 2009 - 2010 (Morozov, 2009). However, since we are interested in how Twitter users participated in and started the discourse about the murder of George Floyd by Minneapolis Police our literatrue review focused on analysis of protest organization and discourse on Twitter for US based events. Namely, the Occupy Wall Street and Unite the Right movements that took place in US cities during 2011 and 2017 respectively.

Mark Tremayne (Tremayne 2014) conducted a network analysis of the Occupy Wall Street movement by creating a network of approximately 3,000 Twitter users connected to each other either by one or both of them mentioning the other, or both of them using the same hashtag in a tweet. Tremayne's analysis focused on determining the central hubs on Twitter in the OWS protests as well as considering how the network shifted over time. To answer these questions, Tremayne calculated various measures of centrality and looked at several diffusion periods, timeframes where influential posts about the movement were made. Tremayne's analysis found several influential users, posts, and hashtags that seemed to shape the discourse.

Tien et al. (Tien et al. 2020) conducted a network analysis of the Charlottesville Unite the Right rally by analyzing a retweet network. A retweet network is a network that consists of nodes representing twitter users that are connected if one retweeted the other. Their network contained 238,892 nodes and 365,589 edges and their data consisted of 389,736 retweets that used the hashtag #Charlottesville. The edges were weighted by the number of times one user retweeted another. Additionally, it is worth noting that 94% of the users in the network were not retweeted at all by other users in the dataset. This is somewhat similar to our dataset in which 70% of the users were not retweeted by any other users in the data. In their analysis they considered node degree distributions, several centrality measures, and two community detection algorithms. Additionally, they also did a principal component analysis on the twitter accounts associated with media that users in their network were following to compliment their network analysis. Consequently, Tien et al. found that the polarization as determined by their PCA was evident in the structure of their retweet network and that Left leaning accounts with higher in-degrees were more likely to be retweeted by communities other than their own, among other findings.

The Data

rtweet

Twitter data must be accessed via an account approved through the Twitter Developer API. There are several packages that exist that allow for R users to interact with the Twitter API, all of which have similar functions. However, because of its extensive documentation, our data was pulled using Michael Kearny's rtweet (Kearny 2019). rtweet allows for users to set many different parameters to narrow the Twitter API search.

For the purposes of our research question, we were interested in querying tweets from May 26 - June 1, 2020, the five days immediately following George Floyd's death. Moreover, we restricted our queries by specific hashtags, which are not case-sensitive on Twitter. The goal of our list of hashtags was twofold. First, we were interested in considering several of the most commonly used hashtags in the Twitter community in response to the murder of George Floyd. Second, we chose hashtags that we believed encompassed a variety of users and discourse about George Floyd, systemic racism, police brutality, Black Lives Matter, and protests. With this in mind, we queried tweets containing one or multiple of the following hashtags: #justiceforfloyd, #blacklivesmatter, #GeorgeFloyd, #acab, and #fuck12. We pulled data from arbitrary time frames during the day, which were consistent across each of the 7 days we consider in our analysis. This was intended to diversify the kinds of tweets our dataset would contain, as rtweet only pulls the last tweets from a given day. As such, our dataset contains tweets from mornings, afternoons, evenings, and late at night.

The majority of the tweets in our dataset were retweets, so we restricted our data to only retweets, for a total of 11,300 tweets. By default, rtweet pulls 90 columns of data for each tweet, which includes both user-specific information and tweet-specific information. User-specific data includes information such as username, verified status, date of account creation, number of tweets, and number of followers, while tweet-specific data includes the text of the tweet itself, time of tweet, hashtags used, language, and mentions. To limit the scope of our analysis to only users and retweets, we subsetted our data to only include information relevant to our research question. Prior to any manipulation to accommodate a graph structure, our dataset was composed of 11,300 observations of retweets and 23 columns.

Methodology

In order to analyze and visualize our network data, we relied on several R packages that are specific to network data: igraph, network, ggraph, and ergm (Csardi et al. 2006; Butts 2008; Pedersen 2020; Hunter et al. 2008). For more information about functionality available in these packages, please refer to their respective vignettes. This section provides an overview of the sentiment analysis and ERGM processes, and how we used them.

Sentiment Analysis

Sentiment Analysis is the process of systematically identifying and categorizing opinions expressed in text. There are several ways to go about analyzing the sentiment of a piece of text, each with their own pros and cons. For our project, we decided to use a rule-based approach that would allow us to identify the polarity of each tweet. To do this, we used a lexicon sentiment analysis tool called the Valence Aware Dictionary and sEntiment Reasoner or VADER for short. VADER is an open source rule based model for sentiment analysis developed by C.J. Hutto and Eric Gilbert.

A lexicon in sentiment analysis is a collection of words and phrases that contain pre-labeled scores that allow us to score our tweets based on the scores in the lexicon. One draw-back to using sentiment analysis is that it does not take into account how words are organized in a sequence, each word is scored individually. However, we decided to use the VADER lexicon because it is tailored specifically to sentiments expressed in social media. For example, VADER has the ability to score many emojis, slang such as ‘sux’, acronyms such as ‘lol’, and more. We focused our analysis on VADER’s compound score, which is computed by summing valence scores for each word in the lexicon and then normalized with -1 being the most negative and +1 being the most positive.

ergm

In order to fit exponential random graph models to our networks, we used the “ergm” package, which allows users to specify ergms using syntax similar to that of an ordinary regression. Like network, the ergm package is part of the statnet suite of network analysis packages for R. The choices of covariates that we made were based on information provided by Laurence Brandenberger and Sebastián Martínez’s RPubs post from February 2019 (Brandenberger).

We began constructing our ergm formula with the network equivalent of a basic linear model’s (lm) or generalized linear model’s (glm) intercept term: edges. Like in an lm or glm, an ergm can be fit using just the edges term. However, because we were interested in analyzing the significance of our networks’ node attributes, we proceeded by adding more terms. Note that unlike in regression analysis, when fitting an ergm, it is important to add all of the covariates of interest to the model immediately, rather than adding and removing sequentially based on significance. This is because changing the predictors in an ergm one by one makes the fit less likely to converge (Brandenberger).

Prior to adding covariates based upon our networks’ node attributes, we considered adding variables that accounted for out-degree, in-degree, geometrically weighted in-degree, and geometrically weighted out-degree. Geometrically weighted degree assigns a value to a node based upon how its “tendency” to have incoming or outgoing nodes (Brandenberger). That is, a node with a high geometrically weighted degree would indicate that it is likely to have many ties. Geometrically weighted degree predictors tend to be more robust, and are therefore preferable when possible.

Covariates in an ergm need to be specified differently depending on their type, which can be either a factor variable (i.e. verified status, in our case) or a continuous variable (i.e. number of followers). Both factor and continuous parameters can be set to fit node parameters differently depending on whether nodes have incoming ties or outgoing ties. To allow us to compare our directed network fit to a fit on an undirected version of our network, we chose to use ergm covariates that did not distinguish between incoming and outgoing ties. It is also possible to add an additional layer of analysis by including covariates that allow us to determine whether users with similar or different levels for a factor or continuous attribute are more likely to have a tie.

Users who wish to extend our work or explore other node covariates for networks should refer to Brandenberger and Martínez’s post.

Graph Structure and EDA

Our networks were constructed using edgelists. Each observation in our original dataset represented a tweet, which in turn corresponded to two users: the retweeter and the original poster. We created a new two-column dataset consisting of both of these users for each tweet—an edgelist where each row represented a retweet. It is important to note that because retweets are directional, our retweet network is a directed graph where nodes represent users and ties represent retweets. For example, if user A retweeted user B, our network would show an arrow coming out of node A towards node B. Moreover, our retweet network is a multigraph. That is, multiple edges can come out of and into a single vertex, representing the fact that a Twitter user can retweet and be retweeted multiple times. Thus, our final network contains 15,300 nodes and 11,130 edges, representing the 11,130 retweets we queried.

We also created 7 subsets of the larger network, one for each of the dates that we considered in our analysis of May 26 - June 1. These smaller networks allow us to explore how the date of retweet impacted network EDA, which users were influential, and the ERGM output. While the number of tweets pulled on each of the 7 days varied, each of the 7 smaller networks contains roughly 2300 nodes and 1400 edges. These 8 networks will serve as the basis for subsequent analysis.

Data EDA

Exploratory data analysis of several prominent user characteristics revealed that the distributions of number of followers, number following, and number of statuses (which includes both original tweets and retweets) in our dataset are all highly right-skewed. This is because our dataset included retweets of celebrities with millions of followers on Twitter, such as Katy Perry (over 109 million followers) and Chance the Rapper (over 8 million followers). Similarly, our dataset contained users following several million people as well as users with millions of statuses. To combat these skewed distributions, we first added one to each user's value for the variable, then log-transformed. However, it is important to note that even after transformation, the distribution of number of followers remains right-skewed, and the distribution of number of statuses is left-skewed.

Our sentiment analysis also allowed us to create a variable for the average sentiment of a given user's retweets in the dataset. The distribution of compound average sentiment for the retweets in our dataset was unimodal, with 0 (neutral) being the most common value for users. Histograms of these four variables can be seen in **Figure 1**. It appeared that the number of tweets pulled for each day of our analysis was relatively consistent, at around 1500. Finally, there appeared to be no clear pattern in the distribution of account age, though the number of new users joining Twitter was highest in 2009 and 2019.

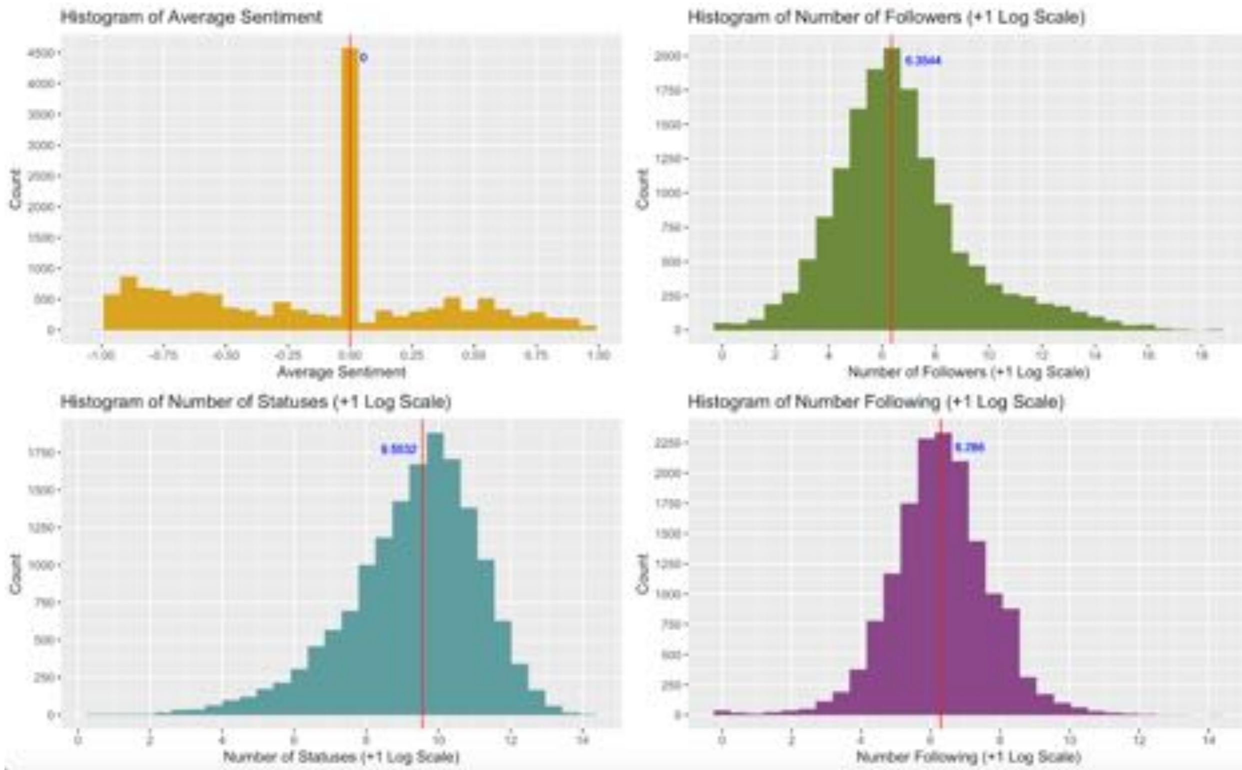

Figure 1: Grid of histograms for user EDA

<u>Network EDA</u>

Table 1 reports some of the network summary statistics we used as part of our exploratory data analysis. From this we observe that although there is variability in the structure of each of the smaller date networks, they all have similar values for density, and identical values for transitivity and reciprocity.

Table 1: Network EDA for each network

Network	Density	Full Network Density % Diff.	Transitivity	Reciprocity
Full Network	4.742×10^{-5}	-	0	0
May 26	5.506×10^{-4}	1061.113	0	0
May 27	3.100×10^{-4}	553.733	0	0
May 28	2.930×10^{-4}	517.883	0	0
May 29	2.770×10^{-4}	484.142	0	0
May 30	2.587×10^{-4}	445.550	0	0
May 31	2.749×10^{-4}	479.713	0	0
June 1	2.728×10^{-4}	475.285	0	0

Graph density was lowest in the full network, and highest in the May 26 network. Given that George Floyd was killed on May 25, it makes sense that the network of retweets on May 26 was the densest. However, across all of our networks, graph density was low. This might be explained by the relatively low volume of tweets we were able to query using rtweet. It is possible that pulling more tweets from each given day would have allowed for more interactions between users, and thus a higher density of realized edges (retweets). With this in mind, it also makes sense that transitivity was 0 for each of our 8 networks. The short time intervals in our queries were not conducive to there being multiple interactions between users in our dataset. Likely for the same reason, all of our networks consist only of cliques of size 1 or 2, with none of the networks containing any instances of triangles. Finally, each of our 8 networks had a reciprocity of 0. As in the case of density and transitivity, this is reasonable given the fact that our dataset often contained short time intervals for a particular time query, where only a few minutes of a query for several hours were represented.

It is also common to analyze measures of centrality when conducting network EDA. These include betweenness centrality, a measure of how often a given node is on the shortest path between any two users in the network, and edge betweenness centrality, a measure of how often a particular edge is part of the shortest path between users. Both of these metrics are used to understand how "important" a node is in a network, which is a question we were interested in exploring in this analysis. However, because our network is constructed exclusively of single nodes and dyads, we cannot use these in our understanding of the influence of a given node. This is again because of the short time intervals of tweets that we obtained as part of our queries to the Twitter API.

Results

Influential Users

Using in-degree, we determined the most influential users for our large network and for each of the smaller date networks. The results of the top five most influential users are presented in **Table 2**, where blue entries denote a verified user and parentheses indicate the in-degree value. Entries with slashes between two users indicates that the given two accounts were tied for the fifth-highest in-degree.

Table 2: Top five most influential (highest in-degree) users by network, blue users verified

Full Network	May 26	May 27	May 28
@YourAnonCentral (598)	@AttorneyCrump (148)	@ava (63)	@narcissariddles (80)
@AttorneyCrump (158)	@flywithkamala (106)	@RealQDaKidd (41)	@storgazing (43)
@narcissariddles (126)	@chancetherapper (44)	@chescaleigh (36)	@jecaylen (42)
@flywithkamala (125)	@AOC (38)	@chloexhalle (24)	@theelovelylaya (39)
@Mediavenir (83)	@QasimRashid (34)	@CoreyPaulMusic / @byersfilms (23)	@ethnickoo (34)

May 29	May 30	May 31	June 1
@lipsredcensored (42)	@thegreatkhalid (59)	@YourAnonCentral (344)	@YourAnonCentral (254)
@asjadnazir (41)	@NedWhat (56)	@tkerssen (43)	@scarlettdevries (33)
@narcissariddles (37)	@karbrulee (34)	@alexbaretv (19)	@JoshuaPotash (30)
@karbrulee (35)	@officialshyasia (33)	@g0ldie_teee (19)	@richardgrant88 (25)
@RandyOrton (31)	@COTimesRecorder (32)	@ajplus / @violadavis (18)	@killmeslowerth (23)

We observed that @YourAnonCentral and @AttorneyCrump were the users with the highest in-degree in our network, with 598 and 158 retweets respectively. @YourAnonCentral (5.9 million followers as of March 2021) is one of the Twitter accounts associated with the political "hacktivist" group Anonymous, which often speaks out against human rights violations, and @AttorneyCrump is a lawyer and political activist (715k followers as of March 2021). Notably, @AttorneyCrump is run by Benjamin Crump, the attorney representing George Floyd's family. These two users also had the highest in-degree on the days that they appeared in the network, where @AttorneyCrump was also the most influential user on May 26, and @YourAnonCentral was the most influential account on May 31 and June 1. It is important to note that the only other user that appears more than once across all of the date networks is @narcissariddles, a teenage activist, who was dominating discourse on May 28 and May 29.

It is helpful to visualize the above results by plotting our retweet network. We chose to create a visualization for one of the date networks instead of our full network because it is easier to observe the graph's structure. Although we only include the graph of the May 26 retweet network in this paper, the same visualization can be constructed for any of our date networks. **Figure 2** shows the May 26 retweet network, plotted using the "ggraph" package. In **Figure 2**, the color of the node represents the verified status of the user, arrows indicate the direction of the retweet, and the size of the node denotes the square root of the in-degree for a particular user. The square root is used purely for visualization purposes, and does not affect the relative "influence" that we determined for a Twitter account. For readability, only some of the most influential users are labeled in the figure, which was also determined using a threshold for in-degree values. Finally, for the graph layout, we used "graphopt" because of its ability to work well with larger graphs.

By analyzing **Figure 2**, we noticed that the distribution of verified and unverified accounts in the May 26 network is relatively even. This means that users that were active in the conversation about the death of George Floyd were both public figures with large platforms and users with smaller or no platforms. Moreover, reinforcing what we saw in **Table 2**, we observed that @AttorneyCrump and @flywithkamala, a lawyer and political activist (24.9k followers as of March 2021), were the two most influential users on May 26.

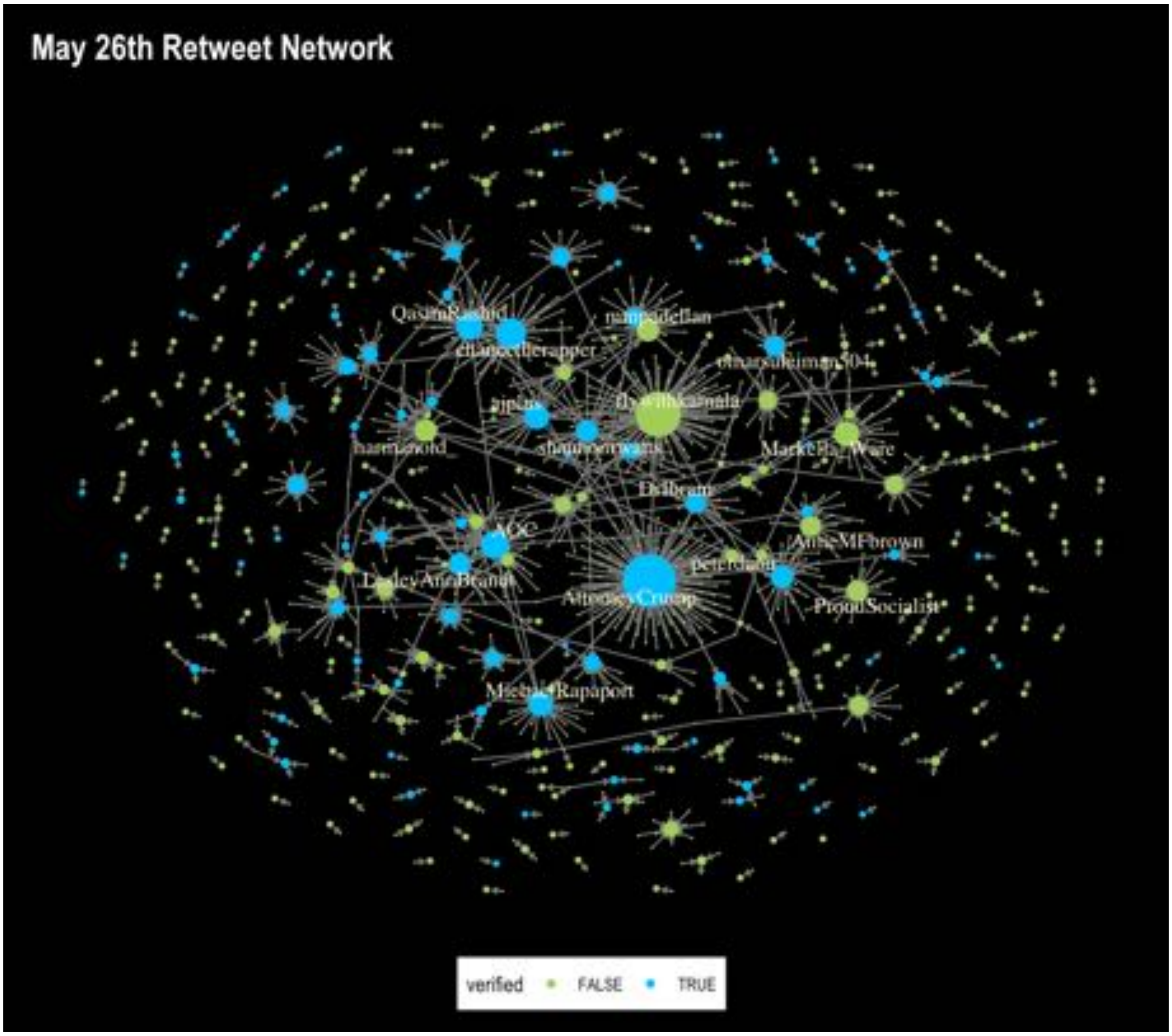

Figure 2: Network of May 26 retweets

<u>ERGM</u>

As mentioned earlier, ERGM allows us to conduct analysis of our data beyond exploratory measures. Using ERGM, we can calculate the probability of a user with specific attributes retweeting another user with their own set of user-level attributes. Through the ERGM fitting process, we were able to identify several user-level attributes that appear to have a significant effect on the probability of a tweet from one user in our data being retweeted by another user from our data.

When estimating coefficients for ERGMs that include structural parameters, the “ergm” package uses Monte Carlo maximum likelihood estimation. Otherwise the “ergm” package uses maximum pseudolikelihood estimation. Unfortunately, adding structural predictors while attempting to generate an ERGM fit increases the likelihood of the Monte Carlo maximum likelihood estimation failing in a timely manner or at all. Consequently, despite our efforts to simplify our network, we were unable to produce an ERGM fit that included a structural predictor, such as a term to account for nodes' in-degrees, and was able to converge. Coefficient estimates for our final fit on the largest component of our large network can be observed in **Table 3.**

The edges term in each ergm can be interpreted similarly to the intercept term in generalized linear models. Each term preceding the '.' in **Table 3** refers to the ergm specific predictor term. Ergm offers a wide range of terms used for modeling. Nodematch is a term used to test the effect of two users sharing the same level of a categorical variable, such as verified status, affects the likelihood of an edge between two users. Nodecov is just a term to test how numerical variables affect the likelihood of a tie between users and is interpreted similarly to numerical terms in logistic regression. Absdiff is a term that tests, as the name suggests, whether the absolute difference between two numerical variables affects the likelihood of a tie between two users.

Table 3: Coefficient estimates from ERGM using component network

Term	Estimate	S.E.	z-value	p-value
edges	-7.4689	0.4122	-18.121	0.0001
nodematch.verified	-0.4457	0.1135	-3.925	0.0001
nodecov.followers_count	0.3005	0.01329	22.614	0.0001
absdiff.followers_count	0.2334	0.01643	14.202	0.0001
nodecov.friends_count	-0.0409	0.0208	-1.962	0.0498
absdiff.friends_count	0.0158	0.0283	0.557	0.5773
nodecov.statuses_count	-0.110	0.0185	-0.543	0.5869
absdiff.statuses_count	-0.0467	0.0255	-1.832	0.0670
absdiff.compound_avg	-4.8147	0.1700	-28.321	0.0001
nodecov.compound_avg	0.3640	0.0345	10.558	0.0001
absdiff.account_age_months	-0.0065	0.0009	-6.948	0.0001
nodecov.account_age_months	-0.0063	0.0006	-10.207	0.0001

We ran the above ERGM on 9 different networks: the 7 date networks, the full directed network, and the undirected component network. **Figure 3** presents the range of estimated coefficient values for each of the 12 ERGM parameters we used. Note that these are separated by magnitude for ease of readability. We found that the coefficient for matching verification status had the largest point range amongst the small magnitude coefficients, ranging from about -0.5 to just over 1.75. That is, depending on the network, the odds of tie for users with matching verification statuses ranged from being 0.6 to 5.755 times that of users with different verification statuses. We also observed that each of the 9 networks estimated that both of the terms for account age do not substantially impact the odds of a tie occurring in any of the networks. Finally, we found that covariates for the difference in compound average sentiment and edges had the highest magnitude estimates, and relatively large point ranges. However, because both ranges are well within negative values, we can determine that users with different compound sentiments are less likely to retweet each other, and that our edge term is significant. This is because the estimate for the edges term for an ERGM model fit with just the edges term as a covariate will be the equivalent of the graph's density (in this case, only edges are accounting for ties). However, as more parameters are added to the ERGM fit, the edges term alone becomes less predictive of a tie, thus resulting in a negative coefficient estimate of relatively high magnitude.

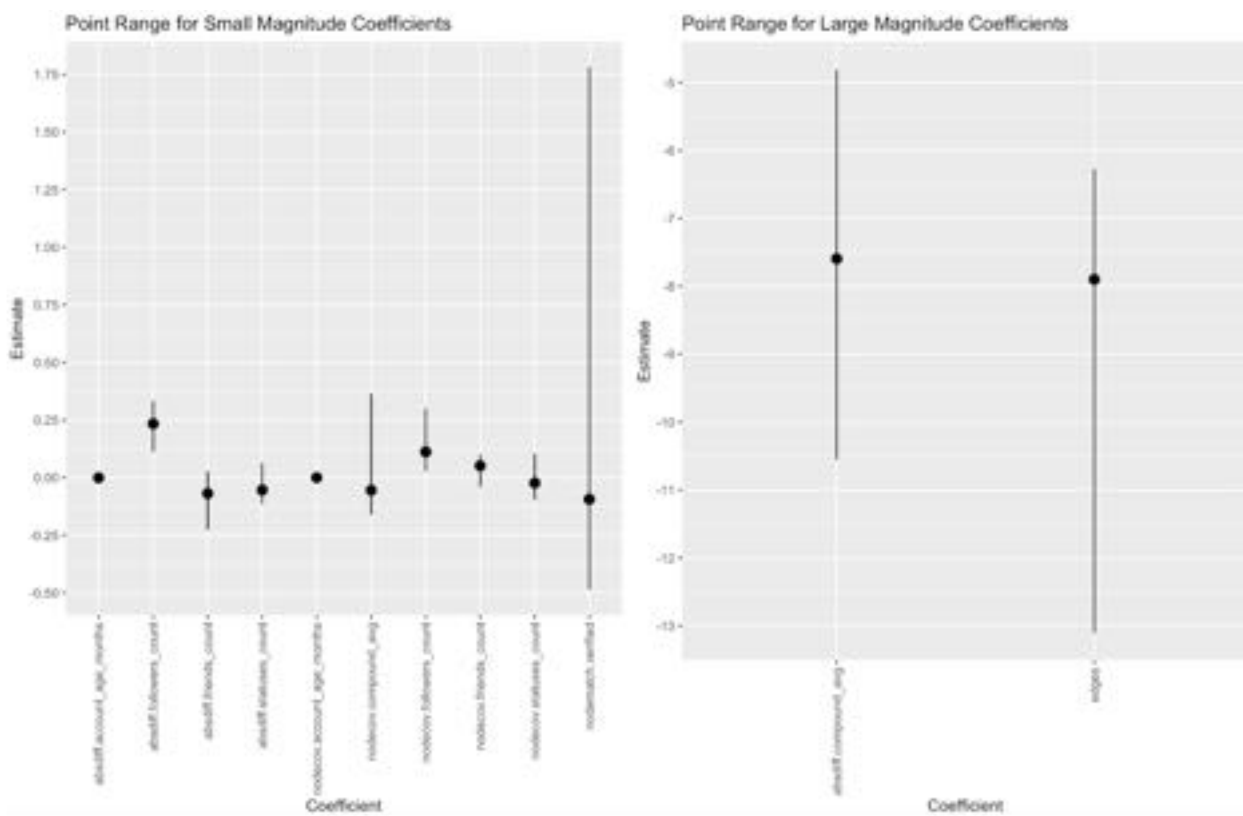

Figure 3: Point ranges of ERGM coefficients for all 9 networks

Discussion and Conclusions

In the aftermath of George Floyd's murder, there was an eruption of discourse around police brutality, systemic racsim, and inequality that seemed ubiquitous accross all social media. Our analysis has attempted to take an objective look at these conversations in order to identify influential contributors to the conversation on Twitter. Additionally, we were interested in identifying which user-level attributes increased the probability of a user being retweeted and observing how the conversations evolved over time.

When a Twitter user retweets a tweet, the tweet will appear on the user's profile, and then will be shown in the feeds of all the users that follow the user who retweeted. Consequently, we are assuming that when someone retweets someone else's tweet, the 'retweeter' finds the original tweet worthwhile to share with their followers. It follows that someone whose tweets have high retweet counts generally produces tweets that Twitter-sphere finds worth sharing. With these assumptions in mind, we return to our discussion of influential users.

It is unsurprising that users with a large platform in the civil rights world such as @YourAnonCentral and @AttorneyCrump drove the discussion in the aftermath of Geoge Floyd's death. As described above, @YourAnonCentral purports to be associated with political "hacktivist" group Anonymous and Benjamin Crump has represented the families of Trayvon Martin, Breonna Taylor, and other high profile civil rights and personal injury cases. While it is almost expected that @AttorneyCrump would have shaped the discourse on Twitter in the days following George Floyd's death, it is notable that @YourAnonCentral is unverified and allegedly gained several million new followers over the same time period while tweeting, in some cases, half-truths (Thalen, 2020).

After noting the significance levels and coefficients of sizes for several ERGM fits on our large network, a smaller simplified undirected component of our large network, and for each network for each day of the time period we collected data over, it seems that the term testing for homophily of average compound sentiment scores of retweets is the most substantial variable in determining the probability of a user retweeting another user. All of our fits found that the greater the difference between two users' average compound sentiment score, the less likely one of them would retweet the other. Notably, the other largest term by absolute magnitude in our network fits besides a term for edges was a term testing for the homophily of user verification status. Interestingly, this term was insignificant for some fits but not for others. Additionally, it ranged from positive to negative, in our date networks implying that on some days it increased the likelihood of a retweet between two users and decreased the likelihood on other days. We observed this behavior, flipping signs, changing significance over different days, or close to zero, in several of our other terms. Consequently, it seems that the other most useful term across all of our fits was the term testing for homophily among the numbers of followers each user had. This seems to make intuitive sense, the larger the difference between number of followers between two users being positively associated with a multiplicative change the odds of one of them retweeting another (ranging from approximately a 12% increase to about 44%), seems to suggest that users with a small number of followers retweeting those who also have small number of followers would not be likely to produce a highly retweeted tweet.

Assumptions, Limitations, Motivations

As mentioned earlier, Tremayne, Tien et al., and others have contributed significantly to a growing body of literature around Twitter and protests. However, there seems to be little network research that utilizes ERGM with data from Twitter. Consequently, our analysis used ERGM to better understand who and how Twitter users used the platform during the aftermath of George Floyd's murder.

During our modeling process we encountered several notable difficulties, perhaps explaining why there are no other instances of ERGMs being used on network data from Twitter. Namely, using structural parameters with ERGM frequently caused the Monte Carlo Markov Chain simulation algorithm used by the ERGM package in R to fail to converge for several of our proposed models. This may have been caused by the nature of our network, which has more nodes than edges as well as many unique node pairings. However, it may have also been caused by the difficult nature of structural parameters within ERGM models (Jones, Ready, Hazel, 2018).

Despite these difficulties, we were still able to develop models with structural parameters that did converge by simplifying our original graph and using a subgraph of it. Consequently, we are assuming that the actors within our subgraph behave in a similar fashion to the rest of our network and Twitter when making our interpretations.

Future studies that plan to use ERGM and data from Twitter may benefit from focusing their collection of data on several pre-identified communities. That is, focusing on a number of highly connected users from an observable community, or a subset of a community, instead of a large number of loosely connected individuals from many communities would most likely be beneficial when using ERGM.

As seen in **Figure 2**, our networks contained many unique dyads, instances of one user retweeting another and no one else within our network. As such, we believe this may have hindered the ability of ERGM to make stable predictions that did not fluctuate in significance by network and to converge.

Future Work

In the future, we hope to see more analysis of social networks found on Twitter that use ERGMs to predict connections between actors. However, we believe that being more conscientious about data collection would be beneficial, including a focus on constructing networks that are highly interconnected. As mentioned earlier, the density of the original graph is used as the p-value for each Bernoulli trial that determines the presence of an edge in a random graph used in the ERGM modeling process. With this in mind, we believe that extremely low density graphs may cause ERGMs to run into problems with convergence and estimation.

Another area for future research would be to explore using Principal Component Analysis (PCA) on hashtags. This analysis would allow researchers to add another layer of understanding when it comes to what effects the likelihood of two users on Twitter interacting with each other in some way.

Lastly, we would be interested in seeing analysis of other forms of networks from Twitter, such as networks formed by tweets that share hashtags, or networks formed by individuals who follow the same users, etc. Studying the structure of interactions between Twitter users in the form of retweets is one of many possible directions for a study using similar data. Exploring a different group of interactions between users on Twitter or all of the interactions together would allow us to better understand how and why discourse on Twitter behaves the way it does.

Acknowledgements

This research would not have been possible without the guidance and support of Dr. Katie St. Clair. Thank you for being patient with us, and for leading such an interesting project. We would also like to thank Carleton's mathematics and statistics department for molding us into statisticians over the last four years, and Sue Jandro for fostering an organized and supportive math and stats community.

References

Brandenberger, Laurence, and Sebastián Martínez. “Lab 2 and 3 - Intro to ERGMs, Marginal Effects and Goodness-of-Fit.” RPubs, February 2019. https://rstudio-pubs-static.s3.amazonaws.com/471073_d45a4acd780b4987932dc8fc47c46dd5.html.

Butts C (2008). “network: a Package for Managing Relational Data in R.” *Journal of Statistical Software*, **24**(2). https://www.jstatsoft.org/v24/i02/paper.

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. https://igraph.org

Demographics of social media users and adoption in the United States. (2020, June 05). Retrieved March 16, 2021, from https://www.pewresearch.org/internet/fact-sheet/social-media/

Hunter D, Handcock M, Butts C, Goodreau S, Morris M. “ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks.” *Journal of Statistical Software* 24, no. 3 (2008): 1–29.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Jones, J. H., Ready, E., & Hazel, A. (2018). Introduction to ergms. Retrieved March 16, 2021, from https://eehh-stanford.github.io/SNA-workshop/ergm-intro.html#why-not-use-standard-regression-methods

Kearney, M. W. (2019). rtweet: Collecting Twitter Data. R package version 0.6.9 Retrieved from https://cran.r-project.org/package=rtweet

Kolaczyk, Eric D., and Gábor Csárdi. *Statistical Analysis of Network Data with R*. New York, NY: Springer, 2014.

Pedersen, Thomas Lin (2020). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.0.4. https://CRAN.R-project.org/package=ggraph

Luke, Douglas A. *A User's Guide to Network Analysis in R*. Springer International Publishing, 2015.

Morozov, E. (2009). Iran: Downside to the “Twitter Revolution”. Dissent, 56, 10–14

Thalen, M. (2020, June 04). How @YOURANONCENTRAL hijacked the news with fake stories. Retrieved March 16, 2021, from https://www.dailydot.com/debug/youranoncentral-anonymous-hacking-twitter/

Tien, J.H., Eisenberg, M.C., Cherng, S.T. *et al.* , “Online reactions to the 2017 ‘Unite the right’ rally in Charlottesville: measuring polarization in Twitter networks using media followership.” *Appl Netw Sci* 5, no. 10 (2020). https://doi.org/10.1007/s41109-019-0223-3

Tremayne, Mark, “Anatomy of Protest in the Digital Era: A Network Analysis of *Twitter* and Occupy Wall Street.” *Social Movement Studies* 13, no.1 (2014): 110-126. https://doi.org/10.1080/14742837.2013.830969

Twitter. (2021). How to post Twitter replies and mentions | Twitter Help Retrieved from https://help.twitter.com/en/using-twitter/mentions-and-replies

Twitter. (2021). How to use hashtags Retrieved from https://help.twitter.com/en/using-twitter/how-to-use-hashtags

Twitter revenue and usage Statistics (2020). (2021, March 08). Retrieved March 16, 2021, from https://www.businessofapps.com/data/twitter-statistics/

Appendix

```
> summary(main_fit_directed)
Call:
ergm(formula = bigfit_form_directed, control = control.ergm(parallel = 2,
    parallel.type = "PSOCK"))

Iterations:  11 out of 20

Monte Carlo MLE Results:
                            Estimate Std. Error MCMC %  z value Pr(>|z|)
edges                     -1.095e+01  1.010e-01      0 -108.471   <1e-04 ***
nodematch.verified         3.326e-01  3.049e-02      0   10.909   <1e-04 ***
nodecov.followers_count    1.164e-01  3.515e-03      0   33.104   <1e-04 ***
absdiff.followers_count    2.382e-01  4.422e-03      0   53.872   <1e-04 ***
nodecov.friends_count      5.093e-02  5.229e-03      0    9.741   <1e-04 ***
absdiff.friends_count     -1.061e-01  7.501e-03      0  -14.142   <1e-04 ***
nodecov.statuses_count    -5.522e-03  4.872e-03      0   -1.134    0.257
absdiff.statuses_count    -3.466e-02  6.694e-03      0   -5.178   <1e-04 ***
absdiff.compound_avg      -7.776e+00  8.448e-02      0  -92.045   <1e-04 ***
nodecov.compound_avg      -1.130e-01  1.275e-02      0   -8.864   <1e-04 ***
absdiff.account_age_months -5.694e-03  2.750e-04     0  -20.707   <1e-04 ***
nodecov.account_age_months -1.676e-03  1.711e-04     0   -9.795   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 324517647  on 234090000  degrees of freedom
 Residual Deviance:    209502  on 234089988  degrees of freedom

AIC: 209526    BIC: 209733    (Smaller is better.)
```