A Dynamical Systems Perspective Reveals Coordination in Russian Twitter Operations

Sarah Rajtmajer, Ashish Simhachalam, and Thomas Zhao

College of Information Science and Technology, The Pennsylvania State University, University Park, PA 16802, USA

Brady Bickel and Christopher Griffin

Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

(Dated: (Preprint - January 28, 2020- Distribution Limited))

We study Twitter data from a dynamical systems perspective. In particular, we focus on the large set of data released by Twitter Inc. and asserted to represent a Russian influence operation. We propose a mathematical model to describe the per-day tweet production that can be extracted using spectral analysis. We show that this mathematical model allows us to construct families (clusters) of users with common harmonics. We define a labeling scheme describing user strategy in an information operation and show that the resulting strategies correspond to the behavioral clusters identified from their harmonics. We then compare these user clusters to the ones derived from text data using a graph-based topic analysis method. We show that spectral properties of the user clusters are related to the number of user-topic groups represented in a spectral cluster. Bulk data analysis also provides new insights into the data set in the context of prior work.

I. INTRODUCTION

In this paper, we analyze a data set purported to represent Russian troll twitter messages, provided by Twitter Inc. [1, 2] using techniques from dynamical systems analysis. There have been several reports on data released by Twitter representing what is believed to be disinformation originating from Russian state-controlled accounts. Of particular focus has been possible targeted interference in the 2016 U.S. Presidential election. However, evidence continues to mount that foreign influence campaigns operationalized through social media remain a persistent threat.

The data discussed in this paper and closely related data have been analyzed by several other authors. Manual analysis using data scraped from Twitter by Clemson University suggests that Russian twitter agents were pro-NRA [3]. Griffin and Bickel [4, 5] initially reported on the use of unsupervised machine learning for extracting information from a subset of the data analyzed in this paper that was provided by NBC News [6]. Kelner et al. [7] analyze this data set to investigate influence in the German Federal election. In et al. [8] build language-based classifiers for twitter bots using this data set. Cable and Hugh [9] as well as Lim et al. [10] study the problem of detecting twitter bots using general machine learning techniques. In addition, several reports have been written on the subject including the well known Mueller Report [11].

The objective of this paper is to investigate the data provided by Twitter Inc. from a dynamical systems perspective. We propose a mathematical model to describe the per-day tweet rate (as the dynamic system) for a subset of the user (bot) population and show how this model leads to a coherent set of user clusters based on the dynamic behavior. In addition, we construct a novel strategy assignment for message data and show that clustered accounts yield consistent strategic behaviors. We then compare these clusters to constructed topic clusters (using the method in [4]) and show that message content diversity is directly linked to diversity in per-day tweet rate. Subsequently, we treat some of the assertions made in [4] as registered hypotheses and use the larger data set to validate or refute them.

The aim of this work is substantively different from the literature on bot and fake news detection [12–17], as it is from the work on general modeling of tweet dynamics [18, 19]. To our knowledge, we are the first to adopt a dynamical systems approach to describe information operations, and in particular, the strategic behavior of coordinated actors in this space. As the detrimental effects of these operations are continually brought to our attention (e.g., [20]), the need for robust, formal approaches to modeling increasingly sophisticated strategic actors has become clear.

The remainder of this paper is organized as follows: In Section II we discuss notation and some prerequisite information. Bulk data analysis is considered in Section III. We then turn to specific dynamical systems inspired analysis of the data set in Section IV. The results from this analysis is compared and expanded using topic analysis in Section V. Conclusions and future directions are presented in Section VI. Two appendices include specific information generated from the unsupervised clustering approaches discussed.
II. NOTATION AND PRELIMINARIES

In this section we establish notation and fundamental definitions needed for the remainder of this paper. All vectors are assumed to be column vectors unless stated otherwise and are denoted by boldface symbols in lower case; matrices are denoted by boldface symbols in upper case. An m dimensional vector \( \mathbf{x} = (x_1, \ldots, x_m) \) has components that are unbolded and indexed. If \( \mathbf{X} \in \mathbb{R}^{m \times n} \) is a matrix, then \( \mathbf{X}^T \in \mathbb{R}^{n \times m} \) is its transpose. If \( \mathbf{x} = (x_t)_t \) is a time series, then its discrete Fourier transform is denoted \( \hat{\mathbf{x}} = \{\hat{x}_\omega\}_\omega = \mathcal{F}(\mathbf{x}) \).

A. Bag of Words Space

We use a bag-of-words (BOW) formalism in this paper because we are interested in a dynamical system analysis of data that is linked to text. This approach has been extensively used. The interested reader may consult \cite{21} as a reference. Assume there is a finite dictionary of words (or symbols) \( \mathcal{W} \) with size \( m \). Assume \( \mathcal{W} = \{w^1, w^2, \ldots, w^m\} \) is ordered, \( i \) is the index of word \( w^i \), and associate with word \( w^i \) the Euclidean unit vector \( \mathbf{e}_i \) via the map \( g(w^i) = \mathbf{e}_i \). In a BOW model, the representation \( g \) is extended to any sequence of words \( (w_1 \cdot w_2 \cdots w_k) \) in the following way:

\[
\mathbf{x} = g(w^1 \cdot w^2 \cdots w^k) = \sum_{i=1}^k g(w^i) \tag{1}
\]

The cosine metric between two vectors \( \mathbf{x} \) and \( \mathbf{y} \) in \( \mathbb{R}^m \) is given by:

\[
d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|},
\]

where \( \|\cdot\| \) denotes the standard Euclidean metric and \( \langle \cdot, \cdot \rangle \) is the standard Euclidean inner product. When \( \mathbf{x}, \mathbf{y} \geq 0 \) (and non-zero), then \( d(\mathbf{x}, \mathbf{y}) \leq 1 \) and \( 1 - d(\mathbf{x}, \mathbf{y}) \) is the cosine similarity.

B. Twitter Dataset

We assume a passing familiarity with the microblogging service Twitter. Each Twitter user generates short broadcast messages (tweets) that can refer to any other user by their username prepended with the \( \@ \) symbol. A Twitter user who follows another Twitter user will receive the broadcast tweets made by that user. A tweet is defined as a retweet if it is a repost of another user’s tweet. Users can be directionally connected through the act of retweeting forming a retweet network. See \cite{22}.

For an arbitrary sampling frequency (e.g., once per hour, once per day, etc.) the output of a Twitter user creates two discrete time series of interest. These can be thought of as observations from a dynamic system. Fix a user \( i \). During sampling period \( t \) a word sequence \( M^i_t \) is observed which contains the concatenated tweets sent by user \( i \) during period \( t \).

\[
\mathbf{X} = \{g(M^i_t)\}_t \tag{2}
\]

The second time series of interest is the tweet count series which simply counts the number of tweets sent by user \( i \) during period \( k \). We denote this \( \{v^i_k\}_t \) or in functional form \( v^i(t) \). In Section \[V\] we use two timescales for topic analysis to compensate for the fact that tweets are short texts and therefore, short timescales are not instructive for understanding the behavior of the dynamical system in message space.

In this work, we are interested in the generation and representation of content by the IRA user group within Twitter, as identified by by Twitter Inc. For this reason, we categorize tweets in one of three ways:

Original: A tweet by an IRA user is original if it is not a retweet of any other user’s content.

Spreading: A tweet by an IRA user is spreading if it is a retweet of another user’s content and that user is within the IRA user group.

Amplifying: A tweet by an IRA user is amplifying if it is a retweet of another user’s content and that user is outside the IRA user group.

Assigning these descriptors to a tweet can be done automatically. The data set provides information on whether a tweet is a retweet and if so who is being retweeted. It is then straightforward to check whether the user who is retweeted is part of the IRA data set.

C. User Strategy Space

Define:

\[
\Delta_n = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0 \right\}
\]

This is the (n-1)-dimensional simplex embedded in \( \mathbb{R}^n \) and used extensively in three-strategy evolutionary games (see e.g., \cite{23}).

We represent user \( i \)’s decision to post an original, spreading or amplifying tweet as strategic. Any time \( \nu^i(t) > 0 \), we can divide user \( i \)’s tweets into these three categories and calculate the three dimensional strategy vector:

\[
\mathbf{\pi}^i(t) = \begin{pmatrix} \nu^i_1(t) \\ \nu^i_2(t) \\ \nu^i_3(t) \end{pmatrix}, \tag{3}
\]

where \( \nu^i_j(t) \) is the number of times strategy \( j \) is selected by user \( i \) during time period \( t \). Each strategy vector \( \mathbf{\pi}^i(t) \in \Delta_3 \). Thus, \( \{\mathbf{\pi}^i\}_t \) is a time series of mixed strategies in \( \Delta_3 \). To understand this dynamical system, we will
adopt a symbolization of $\Delta_3$ allowing us to study (at a gross scale) the distribution of user strategies over longer time periods in the data. The symbolization is shown in Fig. 1. Denote by $\{\sigma_i^t\} \subset \Delta_3$ the symbolized strategy sequence of User $i$, where $\sigma_i^t \in A = \{A, B, C, D, E, F, G\}$.

FIG. 1: A symbolization of $\Delta_3$ that will be used in analyzing user behavior.

III. BULK DATA ANALYSIS

Twitter Inc. maintains the Twitter Elections’ Integrity Hub\(^1\). This site contains tweets and accounts alleged to be part of a Russian active measures campaign against the West, led by the Internet Research Agency (IRA). In the datasets shared on the site, all user names for accounts with less than 5000 followers are hashed, that is, user names have been obscured through a string hashing function. Un-hashed versions of this data can be obtained upon request to Twitter Inc. Data analysis in this paper was performed with the un-hashed version. User account information will be protected\(^2\) unless it is already clear that the user is a known figure and has appeared in a publicly available data set; e.g., the user TEN\_GOP is mentioned in the Mueller report \(^11\).

The data set we use is an aggregate of the October 2018, January 2019, and June 2019 batch releases on the Elections’ Integrity Hub. Our combined data set consists of 8,768,633 individual tweets spanning over 9 years. There are 3,116 users identified in the data. Fig. 2 shows the number of tweets per month over the time spanned by the data set. Others authors \(^4, 7, 24–27\) have spent substantial time correlating twitter activity to world events and there is non-trivial correlation to these events.

The internal retweet network of IRA users (see Fig. 3) suggests a somewhat fragmented network with the top retweeted accounts tweeting primarily in Russian. We identify 28 communities in the retweet network using modularity-based community detection. Fragmentation

FIG. 2: A temporal histogram showing the total distribution of IRA tweets over a period from May-2009 to June 2018.

FIG. 3: The retweet network amongst IRA accounts over the complete dataset. Using modularity-based community detection, we identify 28 communities and attain modularity of 0.395.

within the retweet network will be mirrored in the dynamical systems analysis in the sequel, suggesting that a consistent picture emerges at multiple scales of analysis.

A. English Language Tweets

We now focus on English language tweets because this will allow us to test hypotheses posed in \(^3\) using the full data set provided by Twitter Inc. In \(^4\) the Mathematical language detector was used to classify tweet language. In our present dataset, Twitter Inc. has provided a language for each tweet in the data set under investigation. We note that the accuracy of Twitter’s language detector is unknown. A temporal histogram of English language tweets is shown in Fig. 4. For simplicity, we focus our study in English on the time period between January 1,
2015 and December 31, 2017. This includes the 2016 US presidential election as well as the majority (88.86%) of English tweets in the data set. In the sequel, we will focus even more specifically on users who tweet sufficiently often to ensure that derived time series are not subject to small sample problems.

1. Model of tweet frequency

Let \( \{N(t)\}_t \) denote the total number of tweets posted in English each day between January 1, 2015 and December 31, 2017 and let the accumulation function be given by:

\[
S(t) = \sum_{s \leq t} N(s)
\]

This function is shown in Fig. 5. Visual inspection suggests that the function is piecewise linear.

2. Model of tweet strategy

We illustrate that in addition to a change in messaging rate, there is also a change point in the strategy used by the IRA users. Here we consider the symbolic dynamic systems \( \{\sigma_i^t\}_t \) for all users \( i \) sampled daily. We compare the distribution of visited partition elements in \( \mathcal{A} \) before and after September 9, 2016, the change point used in the previous analysis. Define the indicator:

\[
I_\alpha(\sigma_i^t) = \begin{cases} \ 1 & \sigma_i^t = \alpha \\ 0 & \text{otherwise} \end{cases}
\]

and define the discrete distributions:

\[
p^1(\alpha) = \frac{\sum_i \sum_{t \in [t_0^1, t_0^2]} I_\alpha(\sigma_i^t)}{\sum_\alpha \sum_i \sum_{t \in [t_0^1, t_0^2]} I_\alpha(\sigma_i^t)}
\]

\[
p^2(\alpha) = \frac{\sum_i \sum_{t \in [t_0^3, t_f]} I_\alpha(\sigma_i^t)}{\sum_\alpha \sum_i \sum_{t \in [t_0^3, t_f]} I_\alpha(\sigma_i^t)}
\]

where \( t_f \) is May 10, 2017. All data up to \( t_f \) are considered. The two distributions are shown in Fig. 4. It is the hypothesis that there is a change point in the total messaging velocity \( N(t) \) at some point near the 2016 presidential election. To see this, we fit two linear models of the form:

\[
\hat{S}^j(t) = \beta_0^j + \beta_1^j (t - t_0^j)
\]

The first model uses data from \( S(t) \) starting on July 20, 2015 and ending on September 8, 2016. The second model uses data from \( S(t) \) starting on September 9, 2016 and ending on May 9, 2017. The fits for these models are both shown in Fig. 5. Both models have \( r^2 - \text{Adj} > 0.99 \) (i.e., more than 99% of the variance in the data is explained by the linear models). Moreover statistically we can show that a change point occurs at some point in the data set near September 8 by observing the model parameters:

**Model 1 Parameters:**

| Estimate | Std. Error | Confidence Interval |
|----------|------------|---------------------|
| \( t - 200 \) | 1973.81 | 5.20356 | (1949.91, 1997.71) |

**Model 2 Parameters:**

| Estimate | Std. Error | Confidence Interval |
|----------|------------|---------------------|
| \( t - 616 \) | 3647.54 | 6.76713 | (3616.05, 3679.03) |

The 5\( \sigma \) confidence intervals on \( \beta_1^1 \) and \( \beta_1^2 \) are non-intersecting, thus suggesting that these two linear models are different. Thus, if \( S(t) \) is modeled by a piecewise function of the form in Eq. (4), then with high confidence we can assert there is a change point in this function near \( t_0^2 = 616 \) (September 9, 2016).
After September 9, 2016

FIG. 6: The discrete distributions of visiting patterns of the symbolic dynamical systems \(\{\sigma_i^t\}_t\) in two temporal regions identified.

clear (visually) that these two distributions are distinct. However, we compute \(\chi^2 = 30,827\) using \(p^\ast(\alpha)\) as the reference distribution. Thus, with extremely high certainty we can assert that a strategy change occurred at some point between 2016 and 2017. In particular, IRA users shifted their most common strategy from posting original tweets to posting amplifying tweets. This pattern is investigated further in the sequel when we focus on the subset of IRA users who post the highest volume of tweets.

IV. SPECTRAL ANALYSIS OF THE DYNAMICAL SYSTEM

Observing samples from \(\nu^i(t)\) for (almost) any arbitrary user illustrates that the underlying dynamics of users is non-stationary. This is illustrated in Fig. 7 using TEN_GOP, however any user would suffice. To analyze the data set as a dynamical system, we now restrict our attention to two shorter time periods and a smaller set of users. We select users in the 90th percentile for total volume of tweets from January 1, 2015 through December 31, 2017. In particular, these users posted at least 1,093 tweets during this period. Restricting to this set of users ensures there is sufficient text information available to make analysis of message content meaningful. We refer to this set of 312 users by the set \(\mathcal{U}^*\). Moreover, we restrict our investigation to the time periods from March 9, 2016 to November 8, 2016 and November 29, 2016 to July 31, 2017. Both of these time periods are 244 days long. This ensures we have a sufficient time span for a meaningful time series analysis. Note we omit the period immediately after the 2016 US presidential election because many users had a dramatic decrease in their output. This is illustrated in Fig. 7. Finally, in order to ensure we have sufficient observations within each of the 244-day time periods, we consider only those users who tweet at least once per day over 60% of the days surveyed. As a result, we analyze 24 user dynamics during the March 9, 2016 - November 8, 2016 time period and 117 user dynamics during the November 29, 2016 - July 31, 2017 time period.

The objective of the remainder of this section is to illustrate that the behavior of some users in \(\mathcal{U}_1^*\) and \(\mathcal{U}_2^*\) can be described by noisy quasi-periodic functions and can be grouped by common operational frequencies. We further show how these groups produce consistent topical content and use consistent strategies.

A. Families of Quasi-Periodic Behavior

Using a small subset of the data used in this paper, [4] observed that some Twitter users exhibit periodic behavior. In this section we test this observation for multiple users in \(\mathcal{U}_i^*\), including TEN_GOP, the user discussed in [4]. We show that in this larger data set the following model describes a subset (88.8%) of per-day user tweet volume in both the pre- and post-election time periods we consider:

\[
\nu^i(t) = \mu^i(u) + \sum_j A_j^i(u) \cos(\omega_j^i t + \varphi_j^i) + \epsilon^i(t). \tag{6}
\]

Here \(\mu^i(t)\) is a drift term affected by a (hidden) control signal \(u\), \(A_j^i(u)\) is an amplitude affected by the same \(u\), \(\varphi_j^i\) is a phase specific to \(i\) and \(\omega_j^i\) is a frequency common to members of a group \([i]\) to which user \(i\) belongs. The term \(\epsilon^i(t) \sim (0, \sigma^j)\) is a mean-zero noise term. More importantly, the parameter \(\omega_j^i\) exhibits some consistency over the two time periods, though some users do switch their fundamental frequencies. Furthermore, the data suggest that the number of terms in the Fourier expansion (indexed by \(j\)) is generally small (6 or fewer); therefore we assume the sum in Eq. (6) is finite. We note that not all users can be modeled this way, with some users exhibiting extremely noisy or bursty behavior.

Before discussing the full set of data from \(\mathcal{U}_i^*\), we illustrate the phenomena we discuss using TEN_GOP. This allows us to confirm the hypothesis set out in [4] that this user has a periodicity of 4 days in his/her tweet pattern. Note first from Fig. 6 that there is a trend in the underlying data. To compensate for the trend, we use...
a detrended price oscillator approach [28] replacing \( \nu^i(t) \) (the daily tweet volume) with:

\[
\xi^i(t) = \nu^i(t) - \tilde{\nu}^i_{SM,7}(t).
\]

Here \( \tilde{\nu}^i_{SM,7} \) is the simple 7 day moving average prior to \( t \). The result of detrending is shown in Fig. 8. Notice the de-trended data do not show a linear increase in the run up to the 2016 US presidential election. Removing these trends also allows dominant frequencies in the spectrum to appear more clearly. Thus we can replace Eq. (6) with the more convenient expression:

\[
\xi^i(t) = \sum_j A^i_j(u) \cos (\omega^i_j t + \varphi^i_j) + \epsilon^i(t). \tag{7}
\]

The Fourier transform and a fit using 6 terms for \( \xi(t) \) for \text{TEN\_GOP} is shown in Fig. 9 for the period before and after the 2016 US elections. We note that in both spectra there is a high amplitude frequency at bin 60 indicating a cyclic behavior recurring (approximately) every 4 days. This is consistent with the results found in [21]. We also note that there are a reasonably small number of high amplitude frequencies, particularly before the election. For completeness we illustrate the spectra and fit of \text{KansasDailyNews} in Appendix [A]. This user has much more periodic behavior and a cleaner spectrum.

### B. Behavior Spectrum Clustering

Before proceeding, we layout the general analytic procedure to be followed in determining families of behavior describing \( \xi^i(t) \).

1. Cluster users based on the spectra of \( \xi^i(t) \), thereby constructing the family \([i]\).

2. Compare the derived clusters between the two time periods.

3. Examine the trajectories \( \pi^i(t) \) for those individuals in \([i]\) to determine whether there is a correlation between tweet strategy and tweet volume.

In Section [V], we discuss clustering users in word space and relate these clusters to the behavioral clusters.

Even after detrending the data, the spectra in Fig. 9 are noisy. This is common to all users and is described by the noise term \( \epsilon^i(t) \). To address noise in the spectra, we apply a de-noising procedure and following [29], we use a principal components analysis to project the spectra onto a lower-dimensional space.

The spectral de-noising procedure is simple: we zero out frequency bins whose square amplitude is in the bottom \( q \)-percentile. For this data, we used \( q = 0.33 \). Let \( D(\hat{x},q) \) denote the de-noising procedure on a Fourier transform \( \hat{x} \) using quantile \( q \) as a threshold. For each user \( p_i \in U_p \ (p \in \{1,2\}) \) this leaves a matrix:

\[
\Xi_p = \begin{bmatrix}
\xi_{p1}^T \\
\vdots \\
\xi_{pm}^T
\end{bmatrix},
\]

where:

\[
\hat{\xi}_{pi} = D(\mathcal{F}(\xi_{pi}), q)
\]

is the de-noised Fourier transform of the time series \( \xi_{pi} = \{\xi_{pi}^t\}_t \). We then apply principal components analysis [30] to the data in \( \Xi_p \). The result is a reduced dimensional data set that exhibits good separation. The eigenvalues of the covariance matrix computed during the principal components analysis for both the pre- and post-election data sets are shown in Fig. 10. We note in both cases that there is a clear knee in the curve of the eigenvalues after the third eigenvalue. Therefore, we reconstruct the data in three dimensional space. Visual inspection of the reduced dimensional data suggests that the data may divide into four clusters. We use the \( k \)-medoid method [31] to cluster the data. The results are shown in Fig. 11.
The resulting clusters each have specific spectral properties that define them. These are shown in Fig. 13, where we use box-and-whisker plots to show the ranges of each frequency bin over all elements of the cluster. Cluster 1 is composed of users whose behavior is not modeled by Eq. 7, except to say that their per-day tweet volume is composed of noise; i.e., there is no dominant frequency in their spectra. Cluster 2 is composed of users like TEN\_GOP whose behavior is harmonic with period (approximately) 4 days. This cluster could be subdivided in the post-election period since we can see a substantial amount of variation in the lower-frequency amplitudes. The 4 day periodicity is clearer in the pre-election period. Clusters 3 and 4 both have extremely sharp defining principal frequencies between bins 30 and 35 (corresponding to a 7 day periodicity). Cluster 3 has a second frequency spike between bin 60 and 65, implying a periodicity of 4 days like Cluster 2. Cluster 4 has a second frequency spike between bins 100 and 105, corresponding to a faster 2-3 day cycle, which is difficult to explain from a practical perspective. These results support Eq. 7 as modeling the tweet rate of the users in this data set and also support the idea that these accounts are controlled Twitter bots.

The user names in the themselves have interesting semantic properties. In both the pre-election period and the post-election period, Cluster 3 consists of user names that appear to be regionally selected news sources. For example, OnlineMemphis, KansasDailyNews and ChicagoDailyNew are all in Cluster 3. However, there are some interesting anomalies to this rule. DetroitDailyNew and DailySanJose both appear in Cluster 1 in the post-election time period. This can be explained by anomalies in their spectra. (See Fig. 12) While both users have a spike between bins 30 and 35, DetroitDailyNew has a second frequency spike between bin 60 and 65 like TEN\_GOP, while DailySanJose has some low frequency noise and a frequency spike between bins 100 and 105, consistent with Cluster 4. The confusion results in these two users being placed in Cluster 1.

We analyze the strategic behavior present in the clusters in Fig. 13. The three strategies are shown along with the sample points and a smoothed density histogram of the strategies. The weighted mean point of the strategies is shown as an open circle (in red). Consistent with the results illustrated in Fig. 9 there is a transition from the originate strategy to the amplify strategy for users in Cluster 1, 2 and 4. We note, however, that Cluster 3, which contains users with news-related names, continues to follow the originate strategy both before and after the US presidential election. It is also interesting to note that Cluster 2 (which contains TEN\_GOP) maintains a balance between the originate and amplify strategies in contrast to users in Clusters 1 and 4.

V. TOPICAL EXTRACTION BEHAVIOR CORRELATION

In this section, we construct a second set of clusters during the pre- and post-election periods based on the topics mentioned in posted tweets. We investigate the word-space dynamical system over two long time periods, and then compare the resulting clusters of behaviors (and extracted topics) with the clusters derived in the previous sections. We show evidence to support the hypothesis that the better-modeled a cluster is by Eq. 7, the fewer topical-user clusters are represented, that is, the greater the topical cohesion observed.
Pre-Election Cluster 2
Post-Election Cluster 3
Pre-Election Strategy 3
Pre-Election Cluster 3
Pre-Election Strategy 2
Pre-Election Strategy 4
Post-Election Strategy 3
Post-Election Strategy 4

FIG. 13: Both pre- and post-election spectra and the corresponding strategies are illustrated. The four clusters in the pre-election period can be matched the four clusters in the post-election period showing consistent dynamics across time periods exit. Strategies show a clear transition from originating to amplifying in some clusters but not in others.
A. Algorithm for User Clustering by Text

We discuss the algorithm (Algorithm 1) used to generate clusters of users by their topical signatures. Algorithm 1 is a compilation of several pre-existing methods in natural language processing [21] and manifold learning [33, 34] with a novel variant that makes use of graph analytics developed in the physics community [35–37]. For this reason, we present the algorithm in detail.

We assume some minimal knowledge from statistical language processing. In particular, a stopword is a word with little value to text meaning: examples in English include “the” and “a”. Word stemming removes common endings to ground the text into a more systematic vocabulary. For example, “ending” is replaced with “end”.

We assume the input is a sequence of tweets and users. In Line 1, the tweets from User \( i \) are concatenated to form a synthetic document. For the remainder of this algorithm, if \( w \) is a word, let \( \#(w, i) \) be the number of times User \( i \) uses word \( w \). In Line 2, the words in the document are passed through a stemmer. Since we did not know that multiple languages were in use a priori, we used the standard English (Porter) stemmer available in Mathematica 11.3. However, multiple stemmers and language detectors could be used at this step. In Line 3, we use a custom stopword list consisting of the (stems of) most common words in a language. Again, this was specialized to English, however it is possible to generalize this step. The remaining non-stopwords were used to create term count vectors for each user in Line 4. In Line 5, a binary term-user matrix is constructed. The entries are indicators determining whether a term is used by a user. A dynamic stopword set is initialized in Step 7. In Lines 8-10, we iterate through the term-user matrix and determine whether a term is used by more than \( p \times 100\% \) of the users, with \( p \in (0, 1) \). For our study we set \( p = 0.5 \). The rows corresponding to the words that were used by more than \( p \times 100\% \) are added to \( T \).

After deleting the dynamic stopwords, we identify keywords across all documents in Lines 13 - 18 using frequency. While it is the case that many documents follow the Zipf–Mandelbrot law and their word counts follow a Zipf-Mandelbrot distribution [32, 39], we have found it useful to smooth these distributions and use a percentile cutoff. To do this, we fit a Gamma distribution to the word counts and then remove words that do not fall in the \( q \times 100\% \) percentile of the distribution. For this study, we set \( q = 0.9 \). We chose the Gamma distribution because it has non-negative support, and being a two parameter distribution is widely adaptable. We note, there are alternate way of identifying keywords [21].

After removing non-keywords, the term and user set is fixed and a new term-user matrix \( \tilde{X} \) can be computed where each column is the (reduced term) term counts for the corresponding user. In Line 26, we then replace each column of \( \tilde{X} \), denoted \( \tilde{X}_i \) with its unit vector form. That is, column \( i \) of \( \tilde{X} \) becomes \( \tilde{X}_i/\|\tilde{X}_i\| \).

To understand the remainder of the algorithm, we

**Input:** Raw tweets with users

**Result:** Clusters of Users

1. Concatenate tweets per user to make documents
2. Stem all words in documents
3. Remove stopwords by their stems
4. Generate word counts per user
5. Create binary term-user matrix:

\[
X_{ij} = \begin{cases} 
1 & \text{if User } j \text{ uses term } i \\
0 & \text{otherwise}
\end{cases}
\]

6. Define \( \{n_r, n_c\} = \text{Dimensions}[X] \)
7. Define \( T \leftarrow \emptyset \)
8. for \( i \in \{1, \ldots, n_r\} \) do
9. \( \text{if } \sum_j X_{ij} > p \times n_c \) then
10. \( T \leftarrow T \cup \{i\} \)
11. end if
12. end for
13. Delete all terms corresponding to indexes in \( T \)
14. Set \( T \leftarrow \emptyset \)
15. for \( i \in \{1, \ldots, n_c\} \) do
16. Compute mean word count \( \bar{w} \) and word count variance \( S^2_w \) for User \( i \),
17. Fit a Gamma distribution \( \Gamma(k, \theta) \) for User \( i \)’s word use with:

\[
k = \frac{\bar{w}^2}{S^2_w} \quad \theta = \frac{S^2_w}{\bar{w}}
\]
18. for Each word used by User \( i \) do
19. \( \text{if } \#(w, i) \text{ is in the } q \times 100\% \text{ percentile of } \Gamma(k, \theta) \) then
20. \( T \leftarrow T \cup \{w\} \)
21. end if
22. end for
23. end for
24. Delete all terms not in \( T \).
25. Create new term-user matrix: \( \tilde{X}_i \) is the number of times User \( j \) uses term \( i \).
26. Replace each column of \( \tilde{X}_i \) with \( \tilde{X}_i/\|\tilde{X}_i\| \).
27. Compute modified adjacency matrix: \( \tilde{A} \leftarrow \tilde{X}^T \cdot \tilde{X} \)
28. Compute \( \hat{A} \) by zeroing the diagonal of \( \tilde{A} \)
29. Set \( B_i \leftarrow k^{th} \text{ largest value in } \tilde{A}_i \)
30. for \( i \in \{1, \ldots, n_c\} \) do
31. for \( j \in \{1, \ldots, n_r\} \) do
32. \( \text{if } \tilde{A}_{ij} < \min\{B_i, B_j\} \) then
33. \( \tilde{A}_{ij} \leftarrow 0 \)
34. end if
35. end for
36. end for
37. Create weighted graph \( G \) with weight matrix \( \tilde{A} \)
38. Cluster vertices using maximum modularity clustering.

**ALGORITHM 1:** Construct User Clusters from Linguistic Features

Think of \( \tilde{X} \) as a sub-matrix of the weighted adjacency matrix \( \tilde{B} \) of a bipartite graph between terms and documents where:

\[
\tilde{B} = \begin{bmatrix} 0 & \tilde{X} \\ \tilde{X}^T & 0 \end{bmatrix}
\]

Suppose that \( B \) has the same structure as \( \tilde{B} \), but with
all positive values replaced by 1. It is a classic result in graph theory that $B^2$ has in its $(i, j)$ position the number of walks of length 2 from vertex $i$ to vertex $j$. Thus, if $X$ is again the binarized form of $X$, then:

$$B^2 = \begin{bmatrix} X \cdot X^T & 0 \\ 0 & X^T \cdot X \end{bmatrix}$$

Here $X^T \cdot X$ is a square matrix whose $(i, j)$ position is the number of terms shared between User $i$ and User $j$ when $i \neq j$. By normalizing the columns of $X$ in Line 26, we ensure that the $(i, j)$ element of $X^T \cdot X$ is just the cosine similarity measure of the term vectors used by User $i$ and User $j$ (when $i \neq j$).

In Lines 27-29, we create the adjacency matrix $\hat{A}$ by using the process described above and zero the diagonal. We then compute bounds (define neighbors) for each user, by finding the $k^{th}$ largest value. This is the graph theoretic equivalent of using $k$-nearest neighbors. In our work we set $k = 10$. In Lines 30 - 36, we zero an edge weight (i.e., delete edges) at index $(i, j)$ if it is less than the bound computed for user $i$ or user $j$. In essence, this creates a version of a preferential attachment graph where (i) preference is based on shared topics, (ii) each user prefers to attach to $k$ other users but (iii) user $i$ may attach to more than $k$ users if there are more than $k$ other users who have ranked user $i$ as one of their top $k$ connections. If modeled as a directed graph, this would be an $(n_e, k)$ directed Barabasi-Albert network with a hidden preference based on topic.

Finally, in Lines 37 - 38, we form the graph and use maximum modularity clustering defined by Newman to build user communities. This is a variation on the Isomap and Multidimensional Scaling approaches to manifold learning.

### B. User-Topic Clustering Results

Applying Algorithm 1 yields 5 language-based user clusters in the pre-election period and 7 language-based clusters in the post-election period. The resulting user graphs generated in Line 37 of Algorithm 1 are shown in Fig. 14 along with the resulting maximum modularity clusters generated in Line 38. The user clusters are listed in Appendix C.

We can generate word clouds from the user clusters by merging the stemmed, non-stop words used by the individuals in these clusters. This provides general information about the topical information that caused the users to be clustered. For the word clouds shown in Fig. 14 we do not remove common words.

Topics emergent in the pre-election time period are largely consistent with what is already known about this Twitter data set. There are two groups whose comments may be designed to stoke racial tension (Word Clouds 1 and 4), one group that focuses on news (Word Cloud 5), one group devoted to politics (Word Cloud 3) and in particular Donald Trump, and a final group that seems to re-tweet information from a variety of Twitter handles (Word Cloud 2).

Topics in the post-election time period are similar, but with some apparent changes in composition. In particular, one cluster (Word Cloud 1) seems to be the result of posts about the holiday(s) that occurred during the post-election period. There is a user group that focuses on news (Word Cloud 3). Interestingly, this group contains the majority of the news-themed user names (e.g., KansasDailyNews). There are at least two groups that post on race relations (Word Clouds 6 and 7). The remaining groups seem to focus on music (Word Cloud 2) and general topics related to Donald Trump (Word Clouds 5 and 6) with various subtle language differences separating these groups. As is to be expected based on the Mueller report, TENGOP is a member of post election user group 5 and pre-election user group 3.

### C. Comparison with Previous Clusters

We compare the spectrum-based cluster in the pre- and post-election time periods with the topic-based clusters. This is shown in Fig. 15.

We noted in Section IV that clusters 3 and 4 in both the pre- and post-election periods had the clearest defining spectral characteristics. It is not surprising, therefore, that these groups of users show the least topical diversity. Members of pre-election (spectral) cluster 3 can be found only in pre-election user-topic cluster 5 (the news cluster), while members of pre-election (spectral) cluster 4 can be found only in pre-election user-topic cluster 2. In a similar pattern, members of post-election (spectral) cluster 3 are members of post-election user-topic cluster 3 (a news cluster). Members of post-election (spectral) cluster 4 are members of post-election user-topic clusters 1 (holiday) and 4 (Trump-related). Other spectral clusters have users that are more spread out, corresponding to less well-defined spectral properties. It is possible to create sub-clusterings of the users corresponding to both their topics and their spectra. This helps to decrease noise in the spectral clusters. We illustrate this using post-election spectral cluster 2 and post-election user-topic cluster 4. The smaller cluster contains only the users hashed_26, JennAbrams, PamelaMoore13, TEN GOP, TheFoundingSon, and todayinsyria, but has a clear dominating frequency corresponding to a 4 day period. This is shown in Fig. 16. Exhaustive sub-clustering using both frequency and user-topic information is left for future work.

### VI. CONCLUSION

In this paper, we have analyzed the IRA data provided by Twitter Inc. from a dynamical systems perspective. We proposed a mathematical model to describe the per-
FIG. 14: Topical user networks and clusters are generated by Algorithm [1] for pre- and post-election data (a,b). Word clouds represent frequent topics in each cluster (c,d).
day tweet output of a subset of the users and showed that this model could be validated using spectral analysis. This analysis allowed us to cluster user behavior in a way that was distinct from prior language-based work. We also developed a novel strategy assignment that allowed us to label each tweet as being original to the IRA, amplifying of external information or spreading existing IRA information. We showed that a subset of the spectral clusters had interpretable strategic behaviors. We also showed the existence of a substantial behavioral shift within the English data set from originating information to amplifying external information just after the 2016 US presidential election. Finally, we compared the spectrally-derived user clusters to clusters derived from topics discussed and showed that more cohesive spectral clusters (based on the distinctiveness of frequency peaks) also have greater topical cohesion. We used this observation to illustrate the utility of sub-clustering using both dynamic behavior and linguistic features.

In future work, we may investigate joint clustering based on dynamic behavior and linguistic features. Additionally, it would be very instructive to conduct a very large scale analysis on general Twitter data to determine whether there are hidden attractors or fractal properties within the per-day (or per-week) time series. While this data set seems large (9M+ tweets), in reality many of the users exhibited bursty behavior making it practically impossible to reconstruct an attracting set. In the case of this data set, the presence of dominant frequencies strongly suggests that chaotic behavior is not present. However, this fact may be an important feature in differentiating bots from humans, who might exhibit far more organic (and therefore complex) tweeting patterns. In any future investigation with even more data the application of more sophisticated analytical methods using wavelet analysis or (at a minimum) spectrograms might be more instructive in understanding bursty data.

**ACKNOWLEDGEMENT**

Portions of CG’s work were supported by the National Science Foundation under grant IIS-1909255. SR gratefully acknowledges support from the Center for Security Research and Education at The Pennsylvania State University.

**Appendix A: Alternative Spectrum and Fit Example**

In this appendix, we provide an alternate spectrum and fit example from a [KansasDailyNews](https://www.kansascity.com/) (see Figure 17). This user has a 7 day period and is substantially less noisy that many other users.
Appendix B: Spectrum Clustered User IDs for each Time Period

The following user names are associated to the two periods of investigation. We have clustered them according to their primary frequencies of posting.

Period 1 (March 8, 2016 - November 8, 2016):

Cluster 1:
TrayneshaCole, BleepThePolice, WarfareWW, Jenn_Abrams, Blk_Voice, blackmattersus, BlackToLive, hashed_68, todayinsyria, TheFoundingSon

Cluster 2:
USA_Gunslinger, TEN_GOP, gloed_up, Crystal1Johnson,

Cluster 3:
DailySanJose, hashed_37, NewspeakDaily, RoomOfRumor, hashed_57, OaklandOnline, ScreamyMonkey, KansasDailyNews,

Cluster 4:
DanaGeezus, GiselleEvns

Period 2 (November 29, 2016 - July 31, 2017)

Cluster 1:
hashed_19, hashed_39, DetroitDailyNew, DailySanJose, hashed_55, hashed_54, hashed_12, hashed_44, hashed_1, hashed_32, hashed_16, hashed_18, hashed_36, hashed_63, hashed_4, hashed_33, hashed_28, hashed_46, hashed_77, hashed_35, hashed_41, hashed_5, hashed_78, hashed_25, hashed_38, hashed_2, hashed_61, hashed_60, hashed_21, hashed_51, hashed_66, hashed_17, hashed_34, hashed_66, hashed_62, hashed_40, hashed_24, hashed_9, hashed_47, hashed_45, hashed_14, hashed_74, hashed_73, hashed_3, hashed_58, hashed_13, hashed_31, hashed_71, hashed_67, hashed_53, hashed_42, hashed_76, hashed_64, hashed_48, hashed_27, hashed_59, hashed_8, hashed_15, hashed_43, hashed_10, hashed_50, hashed_7, hashed_75, hashed_65, hashed_69

Cluster 2:
TrayneshaCole, BleepThePolice, hashed_11, hashed_52, TEN_GOP, Jenn_Abrams, hashed_29, ElPasoTopNews, hashed_49, gloed_up, blackmattersus, Crystal1Johnson, BlackToLive, Pamela_Moore13, hashed_30, hashed_26, todayinsyria, BlackNewsOutlet, TheFoundingSon, hashed_72

Cluster 3:
hashed_20, Seattle_Post, DallasTopNews, DailyLosAngeles, PhiladelphiaON, NewOrleansON, TodayNYCity, TodayCincinnati, TodayMiami, PhoenixDailyNew, ChicagoDailyNew, OnlineCleveland, BaltimoreOnline, Atlanta_Online, TodayBostonMA, TodayPittsburgh, DailySanFran, Politweecs, DailySanDiego, WashingtonOnline, OnlineMemphis, HoustonTopNews, StLouisOnline, KansasDailyNews, SanAntoTopNews

Cluster 4:
hashed_70, DanaGeezus, GiselleEvns, ChrizMorgan, hashed_22, hashed_23, hashed_6

Appendix C: Topic Clustered Users for each Time Period

Here user names are clustered according to the topic about which they tweet during the two time periods.

Period 1 (March 8, 2016 - November 8, 2016):

Cluster 1
Blk_Voice, Crystal1Johnson, blackmattersus

Cluster 2
DanaGeezus, GiselleEvns

Cluster 3
Jenn_Abrams, TEN_GOP, TheFoundingSon, USA_Gunslinger

Cluster 4
BlackToLive, BleepThePolice, TrayneshaCole, gloed_up

Cluster 5
DailySanJose, hashed_37, KansasDailyNews, hashed_57, NewspeakDaily, OaklandOnline, hashed_68, RoomOfRumor, ScreamyMonkey, WarfareWW, todayinsyria

Period 2 (November 29, 2016 - July 31, 2017)

Cluster 1
1. D. Silva, “Twitter releases massive data trove on russian, iranian foreign influence campaigns,” https://www.nbcnews.com/tech/social-media/twitter-releases-massive-data-trove-russianiranian-foreign-influence-campaigns-n921146 (2017).

2. TwitterInc., “Information operations,” https://about.twitter.com/en_us/values/elections-integrity.html (2017).

3. T. Mak, “Russia’s Divisive Twitter Campaign Took A Rare Consistent Stance: Pro-Gun,” https://www.npr.org/2018/09/21/648803459/russias-2016-twitter-campaign-was-strongly-progun-with-echoes-of-the-nra (2018).

4. C. Griffin and B. Bickel, “Unsupervised machine learning of open source russian twitter data reveals global scope and operational characteristics,” (2018), arXiv:1810.01466 [cs.SI].

5. M. T. Review, “Data mining has revealed previously unknown russian twitter troll campaigns,” https://www.technologyreview.com/s/612252/data-mining-has-revealed-previously-unknown-russian-twitter-troll-campaigns/ (2018).

6. B. Popken, NBC News https://www.nbcnews.com/tech/social-media/no-nov-available-more-200-000-deleted-russian-troll--/tweets-n844731 (2018).

7. A. Kellner, L. Rangosch, C. Wressnegger, and K. Rieck, “Political elections under (social) fire? analysis and detection of propaganda on twitter,” (2019), arXiv:1912.04473 [cs.CY].

8. J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert, “Still out there: Modeling and identifying russian troll accounts on twitter,” (2019), arXiv:1901.11162.

9. J. Cable and G. Hugh, Bots in the Net: Applying Machine Learning to Identify Social Media Trolls, Tech. Rep. (Stanford University, 2019).

10. J. Lim, Z. Liu, and L. Zhou, in 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) (2019) pp. 203–205.

11. R. S. Mueller, Report On The Investigation Into Russian Interference In The 2016 Presidential Election, Tech. Rep. (Special Counsel’s Office, United States Department of Justice, 2019).

12. Z. Chen, R. S. Tanash, R. Stoll, and D. Subramanian, in International Conference on Social Informatics (Springer 2017) pp. 501–510.

13. D. M. Beskow and K. M. Carley, in SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (2018).

14. N. Chavoshi, H. Hamooni, and A. Mueen, in 2016 IEEE 16th International Conference on Data Mining (ICDM) (2016) pp. 817–822.

15. N. M. Radziwill and M. C. Benton, ArXiv abs/1605.06555 (2016).

16. Y. Khan and S. Thakur, in 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC) (2018) pp. 1–5.

17. F. A. Ozbay and B. Alatas, Physica A: Statistical Mechanics and its Applications 540, 123174 (2020).

18. M. C. Watson, in 2015 IEEE International Conference on Big Data (Big Data) (2015) pp. 793–800.

19. J. Gao, H. Shen, S. Liu, and X. Cheng, in Proceedings of the 25th International Conference Companion on World Wide Web, WWW ’16 Companion (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016) pp. 33–34.

20. Stillgherrian, “Twitter bots and trolls promote conspiracy theories about Australian bushfires,” https://www.zdnet.com/article/twitter-bots-and-trolls-promote-conspiracy-theories-about-australian-bushfires/ (2020).

21. C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing (MIT Press, 1999).

22. T. Inc., “Getting Started: Twitter Help Center,” https://help.twitter.com/en/twitter-guide (2017).

23. C. Griffin and A. Belmonte, Physical Review E 95, 052309 (2017).

24. A. Badawy, E. Ferrara, and K. Lerman, CoRR abs/1802.04291 (2018) arXiv:1802.04291.

25. G. Cleary, “Twitterbots: Anatomy of a Propaganda Campaign,” https://www.symantec.com/blogs/threat-intelligence/twitterbots-propaganda-disinformation (2019).

26. A. C. D. F. R. Lab, “TrollTracker: Twitter Troll Farm Archives,” https://medium.com/dfrlab/trolltracker-twitter-troll-farm-archives--/8d5dd61c486b (2018).

27. A. Thompson, “This is How We Troll,” https://www.ceros.com/originals/russian-tweet-bot-data/ (2018).

28. R. Di Lorenzo, in Basic Technical Analysis of Financial Markets (Springer, 2013) pp. 189–220.

29. D. Friedlander, C. Griffin, N. Jacobson, S. Phoha, and R. R. Brooks, EURASIP J. Applied Signal Processing 4, 371 (2003).

30. I. Jolliffe, Principal component analysis (Springer, 2011).

31. L. Kaufman, P. Rousseau, and Y. Dodge, “Clustering by means of medoids in statistical data analysis based on the,” (1987).

32. N. Meves and K. Allgemein, “Der Bot Boost #merkelmussbleiben,” https://www.wahl.de/aktuell/2016/08/05/social-bot-bundesregierung/ (2016).
[33] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, Science 290, 2319 (2000), http://science.sciencemag.org/content/290/5500/2319.full.pdf.
[34] M. Belkin and P. Niyogi, in Advances in neural information processing systems (2002) pp. 929–936.
[35] M. E. J. Newman, Proceedings of the National Academy of Sciences 103, 8577 (2006), http://www.pnas.org/content/103/23/8577.full.pdf.
[36] M. E. J. Newman, Phys. Rev. E 94, 052315 (2016).
[37] M. E. J. Newman, Phys. Rev. E 88, 042822 (2013).
[38] G. K. Zipf, Selected Studies of the Principle of Relative Frequency in Language (Harvard University Press, 1932).
[39] B. Mandelbrot, in Language (Penguin Books, 1965).
[40] P. E. Hart, D. G. Stork, and R. O. Duda, Pattern Classification (Wiley-Interscience, 2000).
[41] R. Albert and A.-L. Barabási, Phys. Rev. Lett. 85, 5234 (2000).
[42] I. Borg and P. Groenen, Modern Multidimensional Scaling: theory and applications, (Springer-Verlag, New York, NY, 1997).